

Ю.В. Быченков, Е.В. Чижонков

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СЕДЛОВЫХ ЗАДАЧ



ИЗДАТЕЛЬСТВО

БИНОМ



МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ



Ю.В. Быченков, Е.В. Чижонков

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СЕДЛОВЫХ ЗАДАЧ



Москва
БИНОМ. Лаборатория знаний
2010

УДК 519.6
ББК 22.19
Б95

Быченков Ю. В.

Б95 Итерационные методы решения седловых задач / Ю. В. Быченков, Е. В. Чижонков. — М. : БИНОМ. Лаборатория знаний, 2010. — 349 с. : ил. — (Математическое моделирование).

ISBN 978-5-9963-0118-8

Впервые в одной книге рассматриваются все известные итерационные методы для больших систем линейных алгебраических уравнений блочной структуры, которые имеют в качестве решения седловую точку: подробно анализируются идеи построения, условия сходимости и вопросы оптимизации. Результаты анализа представлены в виде удобных для использования формул. Имеющееся приложение ориентировано на применение теории для численного моделирования в гидродинамике и смежных областях.

Для научных работников в области вычислительной математики, аспирантов и студентов, а также для инженеров и исследователей в прикладных областях.

УДК 519.6
ББК 22.19



Издание осуществлено при финансовой поддержке
Российского фонда фундаментальных исследований
по проекту № 09-01-07025

Научное издание

Серия: «Математическое моделирование»

Быченков Юрий Владимирович

Чижонков Евгений Владимирович

**ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ
СЕДЛОВЫХ ЗАДАЧ**

Ведущий редактор *М. С. Стригунова*

Художественный редактор *Н. А. Новак*

Технический редактор *Е. В. Денюкова*

Корректор *Д. И. Мурадян*

Оригинал-макет подготовлен *М. Ю. Копаницкой* в пакете \LaTeX 2_ε

Подписано в печать 30.03.10. Формат 60×90/16.

Усл. печ. л. 22. Тираж 400 экз. Заказ 1170.

Издательство «БИНОМ. Лаборатория знаний»

125167, Москва, проезд Аэропорта, д. 3

Телефон: (499) 157-5272, e-mail: binom@Lbz.ru, <http://www.Lbz.ru>

Отпечатано в ООО ПФ «Полиграфист»,

160001, г. Вологда, ул. Челюскинцев, 3.

Тел.: 8(817-2) 72-61-75; 8(817-2) 72-60-63.

ISBN 978-5-9963-0118-8

© БИНОМ. Лаборатория знаний,
2010

ПРЕДИСЛОВИЕ

Не будет преувеличением сказать, что за последние 15–20 лет одной из самых значительных и популярных тем в численном анализе является исследование сеточных седловых задач, имеющих выраженную блочную структуру. Это объясняется двумя основными факторами: широтой приложений и новизной идей, что необходимо порождает разнообразие конкретных формулировок проблем и методов их решения.

Термины «седловая задача» или «оператор с седловой точкой» имеют происхождение из теории математического программирования ([1], с. 602). Применительно к системам линейных алгебраических уравнений это, как правило, означает симметрию матрицы и наличие у нее собственных значений разных знаков, хотя допустима и более широкая трактовка. Понятие «сеточная задача» относится к происхождению постановки, часто следующей из дискретизации дифференциальных уравнений в частных производных. Блочная структура по форме характеризует специфику задачи, а по сути делает постановку доступной для анализа.

Важным сигналом, что настало время провести определенную систематизацию результатов в этой области, стала публикация обзора Benzi M., Golub G., Liesen J. Numerical solution of saddle point problems (2005) [117], содержащего более пятисот ссылок на первоисточники. Конечно, эту работу следует воспринимать только в качестве путеводителя по тематике, так как изложение материала с необходимой полнотой потребовало бы не полторы сотни страниц, а во много раз больше.

Настоящая книга имеет целью осветить круг проблем вычислительной математики, традиционно считающихся одними из самых сложных не только при решении седловых задач. Здесь имеется в виду оптимизация итерационных методов для решения линейных систем, т. е. процесс наилучшего в некотором смысле выбора ограниченного множества скалярных параметров.

Теория итерационных методов для линейных алгебраических уравнений сама по себе достаточно обширна. Видимо, в настоящее время библиография работ в этой области уже не поддается учету. Поэтому отметим ее приоритетное направление — решение сеточных задач, отличительными особенностями матриц которых являются: большая размерность, плохая обусловленность и разреженность (ленточная структура). Достаточно полное представление о развитии и современном состоянии теории итерационных методов для решения таких задач можно получить из следующих монографий:

- Young D.M. *Iterative Solution of Large Linear Systems* (1971) [200];
- Самарский А. А., Николаев Е. С. *Методы решения сеточных уравнений* (1978) [76];
- Марчук Г. И., Лебедев В. И. *Численные методы в теории переноса нейтронов* (1981) [64];
- Hackbusch W. *Multi-Grid Methods and Applications* (1985) [159];
- Дьяконов Е. Г. *Минимизация вычислительной работы. Асимптотически оптимальные алгоритмы для эллиптических задач* (1989) [41];
- Saad Y. *Iterative methods for sparse linear systems* (1996) [183].

Решение седловых систем в них практически не затронуто. Поэтому предлагаемая книга является связующим звеном между ставшей уже классической теорией итерационных методов и необходимостью решать актуальные седловые задачи. Она написана на основе многочисленных журнальных публикаций авторов, а также с учетом сжатия и качественной переработки материала из монографии [96].

При решении линейных систем алгебраических уравнений ключевым понятием является предобусловливание, что связано с подбором некоторых матриц, удобных с вычислительной точки зрения. Поэтому к настоящему времени для изложения итерационной теории сложилась следующая структура. Сначала рассматриваются простейшие методы: простой итерации, Якоби, Зейделя, верхней релаксации (последние три традиционно называют релаксационными) и т. д., при этом объектом анализа являются условия сходимости и задача асимптотической оптимизации, т. е. выбор постоянных итерационных параметров, обеспечивающих наивысшую скорость сходимости. Кроме того, представляет интерес получение оценок погрешности метода для этого оптимального случая. Отличительной особенностью релаксационных методов является применение элементов матрицы исходной системы для построения предобусловливающих операторов (простейшее предобусловливание). Это приводит

к выявлению предельных, или асимптотически наилучших, характеристик методов. На следующем этапе изучаются возможности ускорения сходимости алгоритмов за счет использования переменных итерационных параметров, причем их выбор может осуществляться как по явным формулам (когда известными считаются постоянные из матричных неравенств), так и из вариационных принципов. И наконец, делается завершающий шаг, связанный с введением спектрально-эквивалентных операторов (нетривиальным предобуславливанием), т. е. производится обобщение наиболее важных полученных результатов на случай, когда матрицы в алгоритмах обращаются неточно.

Книга состоит из двух частей и приложения. Первая часть посвящена изложению теории релаксационных методов для решения седловых задач. Используя элементарные средства анализа, здесь удастся показать базовые идеи и основные результаты, а также проследить преемственность с классической теорией. Этот материал вполне доступен для понимания молодыми исследователями, уверенно владеющими основами линейной алгебры, математического анализа и численных методов. Во второй части книги речь идет об обобщенных (спектрально-эквивалентных) методах, что требует более основательной подготовки от читателя. Здесь для цельности изложения исходная постановка переформулирована с необходимой общностью и с достаточной полнотой приведены вспомогательные утверждения. В этой части подвергнуты анализу итерационные методы, основанные на всех основополагающих идеях блочного седлового случая. В отличие от симметричной знакоопределенной сеточной задачи (типа дискретного аналога уравнения Пуассона) анализ обобщенных методов решения седловых задач не является формальной процедурой. Более того, введение спектрально-эквивалентных операторов приводит к такому расширению множества постановок оптимизационных задач, что влечет за собой необходимость разработки новой методологии исследования, включающей нестандартные технические элементы. Важно отметить, что главными качествами найденных решений задач по оптимизации методов являются неулучшаемость оценок и наличие явных формул для итерационных параметров.

Интерес к проблемам оптимизации алгоритмов для седловых задач возник у авторов при моделировании гидродинамических эффектов, в первую очередь, из необходимости численного решения уравнений Навье—Стокса. Именно поэтому приложение книги ориентировано на вычислительную гидродинамику. В рассматриваемом случае оценка некоторых величин в формулах для оптимальных

параметров тесно связана с определением констант в дискретном условии Ладыженской—Бабушки—Брецци и непрерывном $\inf\text{-sup}$ -неравенстве. Проблематика наилучших констант в функциональных неравенствах очень глубока и интересна сама по себе; нашей целью было зафиксировать содержательное родство между далекими на первый взгляд областями исследований.

В книге принята сквозная нумерация глав во всех частях, аналогично нумеруются формулы внутри главы. Имеется дополнительное разбиение на параграфы, в соответствии с ним организована нумерация утверждений. Для обозначения методов используются унифицированные англоязычные аббревиатуры, что способствует удобству в восприятии и обозначает преемственность в развитии теории.

Авторы глубоко признательны коллегам и учителям, общение с которыми позволило сформировать точку зрения на предмет; выделим среди них Е. Г. Дьяконова, В. И. Лебедева и Г. М. Кобелькова. Исключительная благодарность адресуется Н. С. Бахвалову, чью роль в обсуждении замысла и советах по написанию книги невозможно переоценить.

ЧАСТЬ I

РЕЛАКСАЦИОННЫЕ МЕТОДЫ

ВВОДНЫЕ СВЕДЕНИЯ

В главе приводятся основные обозначения и постановка задачи, рассматривается базовый для решения блочных седловых задач метод Узава и доказываются вспомогательные результаты, используемые в дальнейшем.

1.1. ОСНОВНЫЕ ОБОЗНАЧЕНИЯ И ПОСТАНОВКА ЗАДАЧИ

Обозначим через U и P евклидовы пространства векторов размерностей N_u и N_p соответственно, $Z = U \times P$. Запись вектора z в виде $z = \{u, p\} \in Z$ означает, что он состоит из двух компонент: $u \in U$, $p \in P$. Для системы линейных алгебраических уравнений $Lz = F$ это соответствует разбиению квадратной матрицы L на блоки L_{ij} ($1 \leq i, j \leq 2$):

$$L = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix},$$

размерности которых определяются размерностями компонент вектора z : L_{11} является $N_u \times N_u$ матрицей, $L_{12} - N_u \times N_p$, $L_{21} - N_p \times N_u$, $L_{22} - N_p \times N_p$. Правая часть системы F имеет представление, аналогичное z : $F = \{f, \varphi\} \in Z$.

Если L_{11} невырождена, то можно определить матрицу

$$S = -L/L_{11} \equiv -(L_{22} - L_{21}L_{11}^{-1}L_{12}),$$

часто называемую дополнением Шура для матрицы L относительно L_{11} (для удобства взятым со знаком минус). Значимость S определяется следующим факторизованным представлением L :

$$L = \begin{pmatrix} I & 0 \\ L_{21}L_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} L_{11} & 0 \\ 0 & -S \end{pmatrix} \begin{pmatrix} I & L_{11}^{-1}L_{12} \\ 0 & I \end{pmatrix}, \quad (1.1)$$

где I имеет смысл единичной матрицы соответствующего размера. Выражение (1.1) формально сводит решение системы $Lz = F$ к обращению двух подматриц L_{11} и S , а фактически — только S , так как в определении последней уже входит L_{11}^{-1} . Таким образом, этот прием

позволяет понизить размерность решаемой задачи путем сведения ее к равносильной системе с матрицей специальной структуры.

Далее мы будем иметь дело с вещественной системой линейных алгебраических уравнений $L_\varepsilon z = F$ с параметром $\varepsilon \geq 0$ наиболее распространенного вида:

$$L_\varepsilon z \equiv \begin{pmatrix} A & B \\ B^T & -\varepsilon D \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ \varphi \end{pmatrix} \equiv F, \quad (1.2)$$

где $A = A^T > 0, D = D^T > 0$ — квадратные матрицы размеров $N_u \times N_u$ и $N_p \times N_p$, а B — прямоугольная, в общем случае, матрица размера $N_u \times N_p$. С другими постановками блочных седловых задач можно ознакомиться в [117].

Будем предполагать, что матрица L_ε невырождена при любом $\varepsilon \geq 0$. Это условие, в силу факторизации вида (1.1), означает невырожденность дополнения Шура

$$S_\varepsilon = B^T A^{-1} B + \varepsilon D.$$

По построению матрица S_0 симметрична и положительно полуопределена, т. е. $(S_0 p, p) \geq 0$ для произвольного $p \in P$. Поэтому при $\varepsilon > 0$ матрица S_ε (и следовательно, L_ε) невырождена всегда, а при $\varepsilon = 0$ условия невырожденности матриц L_0 и S_0 носят эквивалентный характер. Это означает исключительность ситуации с $\varepsilon = 0$, поэтому основное изложение будет посвящено именно этому случаю, а обобщение результатов для $\varepsilon > 0$ будет проводиться по мере необходимости.

Обратим внимание на соотношение между размерностями пространств U и P , следующее из условия невырожденности матрицы L_0 . Это — неравенство $N_u \geq N_p$. Действительно, рассмотрим в противном случае решение однородной системы $L_0 z = 0$ вида $z = \{0, p\}$, которое порождает для компоненты p условие $Bp = 0$, т. е. систему из N_u уравнений с $N_p > N_u$ неизвестными. Такая однородная система всегда имеет нетривиальное решение. Поэтому вместо неконструктивного предположения о невырожденности L_0 часто требуют, чтобы матрица B имела полный ранг при указанном выше условии на размерности.

Задачу (1.2) часто называют задачей с седловой точкой (или задачей с седловым оператором). В этом случае имеют в виду, что функция Лагранжа

$$\begin{aligned} \Phi(u, p) &\equiv (L_\varepsilon z, z) - 2(F, z) = \\ &= (Au, u) + 2(B^T u, p) - \varepsilon(Dp, p) - 2(f, u) - 2(\varphi, p) \end{aligned}$$

имеет седловую точку $z^* = (u^*, p^*)$, совпадающую с решением (1.2), т. е. справедливы равенства

$$\Phi(u^*, p^*) = \min_{u \in U} \Phi(u, p^*) = \max_{p \in P} \Phi(u^*, p).$$

Содержательными по смыслу следствиями этого факта являются следующие свойства симметричной матрицы L_ε : фиксированная блочная структура и принципиальное отсутствие знакоопределенности.

1.2. МЕТОД УЗАВЫ – СОПРЯЖЕННЫХ ГРАДИЕНТОВ

Рассмотрим для задачи $L_0 z = F$ в покомпонентной форме

$$\begin{cases} Au + Bp = f, \\ B^T u = \varphi \end{cases}$$

процедуру исключения Гаусса. Выделим компоненту решения u из первого уравнения

$$u = A^{-1}(f - Bp) \quad (1.3)$$

с последующей подстановкой во второе. В результате получим систему уравнений $S_0 p = b$ следующего вида:

$$S_0 p \equiv B^T A^{-1} B p = B^T A^{-1} f - \varphi \equiv b. \quad (1.4)$$

Отсюда имеем, что если матрица L_0 невырождена, то для нахождения $z = \{u, p\}$ достаточно сначала решить систему (1.4) с симметричной положительно определенной матрицей S_0 , а затем определить недостающую компоненту u по формуле (1.3). Под методом Узавы в самом широком смысле, как правило, понимают реализацию этого подхода.

Для решения (1.4) в качестве предобусловливателя матрицы S_0 обычно вводится матрица $C = C^T > 0$ размерности $N_p \times N_p$ (в простейшем случае $C = I$ — единичная матрица) и затем применяется обобщенный метод сопряженных градиентов. Приведем одну из его наиболее распространенных версий.

Зададим начальный вектор p^0 и последовательно вычислим векторы:

$$r^0 = S_0 p^0 - b, \quad w^0 = C^{-1} r^0, \quad s^0 = w^0,$$

затем с помощью величины

$$a_1 = \frac{(w^0, r^0)}{(S_0 s^0, s^0)}$$

определим следующее приближение

$$p^1 = p^0 - a_1 s^0.$$

Далее для $k = 1, 2, \dots$ формулы имеют вид:

$$\begin{aligned} r^k &= r^{k-1} - a_k S_0 s^{k-1}, & w^k &= C^{-1} r^k, \\ d_k &= \frac{(w^k, r^k)}{(w^{k-1}, r^{k-1})}, & s^k &= w^k + d_k s^{k-1}, \\ a_{k+1} &= \frac{(w^k, r^k)}{(S_0 s^k, s^k)}, & p^{k+1} &= p^k - a_{k+1} s^k. \end{aligned}$$

Хорошо известно (см., например, [15], с. 296), что при отсутствии ошибок округлений метод сопряженных градиентов приводит к точному решению за число итераций, не превышающее размерность системы. Однако на практике критерием останова часто служит абсолютная или относительная малость невязки

$$r^k = S_0 p^k - b$$

в какой-либо норме.

Для наших целей важной является оценка погрешности (ошибки) метода сопряженных градиентов. Обозначим через γ и Γ постоянные в матричном неравенстве (здесь и далее будем считать их точными)

$$\gamma C \leq S_0 \leq \Gamma C, \quad 0 < \gamma, \quad (1.5)$$

что эквивалентно принадлежности спектра матрицы

$$C^{-\frac{1}{2}} S_0 C^{-\frac{1}{2}}$$

отрезку $[\gamma, \Gamma]$. Тогда для любой итерации с номером k будет справедливо неравенство

$$\|p^k - p\|_{S_0} \leq \frac{2q_0^k}{1 + q_0^{2k}} \|p^0 - p\|_{S_0}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}. \quad (1.6)$$

Здесь и далее выражение $\|r\|_A$ будет использоваться для обозначения нормы вектора r , порожденной симметричной положительной матрицей A , в данном случае

$$\|p\|_{S_0} = (S_0 p, p)^{1/2}.$$

Приведенная оценка означает, что норма ошибки $p^k - p$ асимптотически убывает как геометрическая прогрессия со знаменателем q_0 . Отметим также, что в процессе реализации алгоритма Узавы в качестве промежуточных величин получаются приближения к u по формуле

$$u^{k+1} = A^{-1}(f - Bp^k),$$

для которых также справедливо неравенство вида (1.6)

$$\|u^{k+1} - u\|_A \leq \frac{2q_0^k}{1 + q_0^{2k}} \|u^1 - u\|_A.$$

При фиксированных γ и Γ величина

$$\frac{2q_0^k}{1 + q_0^{2k}}$$

является неувлучшаемой, поскольку определяется нормой так называемого оптимального полинома (в данном случае — чебышевского). Этот факт лежит в основе распространенной точки зрения:

Если матрица A исходной задачи (1.2) легко обратима, то метод Узавы — сопряженных градиентов является наилучшим для ее решения.

Конечно, так дело обстоит не всегда (в комментариях к главе приведен пример задачи, «неудобной» для этого алгоритма), однако отрицать приоритетность метода Узавы — сопряженных градиентов для решения очень широкого класса седловых задач было бы неправильно. Вследствие этого актуальной является разработка предобусловливателей S для дополнения Шура S_0 (или S_ε) в конкретных задачах, сводящихся к системам с седловой точкой (1.2). Использование специфики постановок, следующей, например, из исходных дифференциальных уравнений или геометрии областей, делает эту тематику практически неисчерпаемой.

Имеется еще одно важное следствие оценки (1.6). Поскольку в ней явно присутствует зависимость от γ и Γ (констант спектральной эквивалентности матриц в (1.5)), то для представления об эффективности любого другого итерационного метода решения системы (1.2) желательно иметь оценку его погрешности, выраженную в тех же величинах. Это приводит к принципиально новым постановкам задач для оптимизации итерационных методов в отличие от классической теории.

Закончим раздел указанием на границу применимости алгоритма Узавы — сопряженных градиентов. Метод существенно теряет свою привлекательность, если требуются большие вычислительные затраты для обращения матрицы A (или, что эквивалентно, отсутствует хороший предобусловливатель для нее). В такой ситуации даже использование внутренних итераций для приближенного вычисления величины $A^{-1}x$ делает алгоритм малоэффективным (см. также раздел 6.2).

1.3. ВСПОМОГАТЕЛЬНЫЕ УТВЕРЖДЕНИЯ

Обозначим через $\Phi_T(\lambda) = \det(\lambda I - T)$ характеристический многочлен квадратной матрицы T и рассмотрим связь между спектрами произведений матриц, взятых в прямом и обратном порядках.

Лемма 1.3.1. Пусть T_1 и T_2 — матрицы размерностей $N_p \times N_u$ и $N_u \times N_p$ соответственно ($N_p \leq N_u$). Тогда справедливо равенство

$$\Phi_{T_2 T_1}(\lambda) = \lambda^{N_u - N_p} \Phi_{T_1 T_2}(\lambda),$$

т. е. матрица $T_2 T_1$ имеет те же, с учетом кратностей, собственные значения, что и $T_1 T_2$, и, кроме того, еще $N_u - N_p$ собственных значений, равных нулю.

Доказательство. Рассмотрим следующие тождества для блочных матриц размерности $(N_u + N_p) \times (N_u + N_p)$:

$$\begin{pmatrix} T_1 T_2 & 0 \\ T_2 & 0 \end{pmatrix} \begin{pmatrix} I & T_1 \\ 0 & I \end{pmatrix} = \begin{pmatrix} T_1 T_2 & T_1 T_2 T_1 \\ T_2 & T_2 T_1 \end{pmatrix},$$

$$\begin{pmatrix} I & T_1 \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ T_2 & T_2 T_1 \end{pmatrix} = \begin{pmatrix} T_1 T_2 & T_1 T_2 T_1 \\ T_2 & T_2 T_1 \end{pmatrix}.$$

Поскольку блочная матрица

$$K = \begin{pmatrix} I & T_1 \\ 0 & I \end{pmatrix}$$

размерности $(N_u + N_p) \times (N_u + N_p)$ невырождена, имеем

$$\begin{pmatrix} I & T_1 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} T_1 T_2 & 0 \\ T_2 & 0 \end{pmatrix} \begin{pmatrix} I & T_1 \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ T_2 & T_2 T_1 \end{pmatrix}.$$

Таким образом, две матрицы размерности $(N_u + N_p) \times (N_u + N_p)$

$$M_1 = \begin{pmatrix} T_1 T_2 & 0 \\ T_2 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & 0 \\ T_2 & T_2 T_1 \end{pmatrix}$$

подобны: $K^{-1} M_1 K = M_2$. Собственные значения матрицы M_1 — это собственные значения матрицы $T_1 T_2$ вместе с N_u нулями, а собственные значения матрицы M_2 — это собственные значения матрицы $T_2 T_1$ вместе с N_p нулями. Поскольку характеристические многочлены подобных матриц совпадают

$$\begin{aligned} \Phi_{M_2}(\lambda) &= \det(\lambda I - M_2) = \det(\lambda K^{-1} K - K^{-1} M_1 K) = \\ &= \det K^{-1} \det(\lambda I - M_1) \det K = \Phi_{M_1}(\lambda), \end{aligned}$$

то отсюда следует утверждение леммы. ■

1.3.1. Две задачи на собственные значения

Рассмотрим две обобщенные задачи на собственные значения:

$$S_0 p \equiv B^T A^{-1} B p = t C p, \quad (1.7)$$

$$B_0 u \equiv B C^{-1} B^T u = \omega A u, \quad (1.8)$$

где t, ω — спектральные параметры, а матрицы A, B и C имеют прежний смысл: $A = A^T > 0$, $C = C^T > 0$ — квадратные матрицы размеров $N_u \times N_u$ и $N_p \times N_p$, а B — прямоугольная, в общем случае, матрица размера $N_u \times N_p$.

Теорема 1.3.1. Пусть $\det(L_\epsilon) \neq 0$ для любого $\epsilon \geq 0$. Тогда:

- 1) все собственные значения задачи (1.7) положительны;
- 2) ненулевые собственные значения задачи (1.8) положительны и совпадают с учетом кратностей с собственными значениями задачи (1.7);
- 3) задача (1.8) имеет ровно $N_u - N_p \geq 0$ нулевых собственных значений;
- 4) каждому решению (t_i, p_i) задачи (1.7) можно поставить в соответствие единственное (с точностью до постоянного множителя) решение (ω_i, u_i) задачи (1.8) по следующему правилу:

$$t_i = \omega_i, \quad p_i = C^{-1} B^T u_i, \quad i = 1, \dots, N_p.$$

Доказательство. Из невырожденности матрицы L_ϵ при $\epsilon = 0$ следует положительная определенность дополнения Шура S_0 в силу факторизации (1.1). Поэтому задача (1.7) имеет N_p положительных собственных значений $0 < t_1 \leq \dots \leq t_{N_p}$ вместе с соответствующими C -ортогональными собственными векторами p_1, \dots, p_{N_p} .

Если ввести обозначения

$$R = C^{-1} B^T, \quad Q = A^{-1} B,$$

то матрицы $C^{-1} S_0$ и $A^{-1} B_0$ представимы в виде RQ и QR соответственно. На основании леммы 1.3.1 это приводит к равенству

$$\Phi_{A^{-1} B_0}(\lambda) = \lambda^{N_u - N_p} \Phi_{C^{-1} S_0}(\lambda),$$

откуда следуют второе и третье утверждения теоремы.

Отметим теперь свойства решений задачи (1.8). Так как

$$B_0 = B_0^T \geq 0,$$

задача (1.8) имеет N_u действительных собственных значений

$$0 = \omega_1 = \dots = \omega_{N_u - N_p} < \omega_{N_u - N_p + 1} \leq \dots \leq \omega_{N_u}$$

вместе с соответствующими A -ортогональными собственными векторами u_1, \dots, u_{N_u} . Кроме того, векторы

$$p_i = C^{-1}B^T u_i, \quad i = 1, \dots, N_u,$$

являются C -ортогональными. Действительно, в силу A -ортогональности системы векторов $\{u_i\}_{i=1}^{N_u}$ имеем при $i \neq j$:

$$0 = (u_j, BC^{-1}B^T u_i) = (B^T u_j, C^{-1}B^T u_i) = (Cp_j, p_i),$$

где $p_i = C^{-1}B^T u_i$, причем некоторые из них могут быть нулевыми.

Рассмотрим теперь некоторый вектор u_i , соответствующий ненулевому собственному значению ω_i :

$$BC^{-1}B^T u_i = \omega_i A u_i.$$

Применив к этому равенству последовательно операторы A^{-1} и B^T , будем иметь

$$B^T A^{-1} B C^{-1} B^T u_i = \omega_i B^T u_i.$$

Вводя обозначения $t_i = \omega_i$, $p_i = C^{-1}B^T u_i$, можно переписать последнее соотношение в виде

$$S_0 p_i = t_i C p_i.$$

Учитывая, что ненулевые собственные значения задачи (1.8) совпадают с учетом кратностей с собственными значениями задачи (1.7) и преобразование $C^{-1}B^T$ переводит A -ортогональные векторы u в C -ортогональные векторы p , можем положить в последнем равенстве $i = N_u - N_p + 1, \dots, N_u$, что и завершает доказательство последнего утверждения. ■

1.3.2. Базис специального вида из собственных векторов

Введем $H = \ker(A^{-1}B_0)$ — множество векторов размерности $N_u - N_p$. Допустимо и эквивалентное определение:

$$H = \{u \in U : B^T u = 0\}.$$

Далее будем использовать разложение пространства U в прямую сумму: $U = H \oplus G$, где G есть обозначение ортогонального дополнения к H (в смысле скалярного произведения, порожденного матрицей A).

Разложим базис пространства U , состоящий из собственных векторов задачи (1.8), на два подмножества: базис пространства H и базис его ортогонального дополнения G так, что

$$\{u_i\}_{i=1}^{N_u} = \{h_i\}_{i=1}^{N_u - N_p} \cup \{g_i\}_{i=1}^{N_p}.$$

Отметим, что каждую из подсистем векторов можно считать ортонормированной в метрике, порождаемой матрицей A , в силу

$$B_0 = B_0^T, \quad A = A^T > 0.$$

Аналогично будем считать C -ортонормированными собственные векторы p_i задачи (1.7) в силу

$$S_0 = S_0^T > 0, \quad C = C^T > 0.$$

Введем в пространстве $Z = U \times P$ скалярное произведение

$$(z_1, z_2)_Z = \chi_1(Au_1, u_2) + \chi_2(Cp_1, p_2),$$

$$z_i = (u_i, p_i) \in Z, \quad \chi_i > 0, \quad i = 1, 2,$$

и построим в нем базис специального вида, используя собственные векторы задач (1.7) и (1.8).

Теорема 1.3.2. Система векторов $\{z_i\}_{i=1}^{N_u+N_p}$ вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

$$z_j^{(2,3)} = \{g_j, \kappa_j^{(2,3)} p_j\}, \quad j = 1, \dots, N_p,$$

где $p_j = C^{-1}B^T g_j$, образует базис в пространстве Z , если при фиксированном значении j конечные коэффициенты $\kappa_j^{(2)}$ и $\kappa_j^{(3)}$ различны.

Доказательство. Проверим сначала ортогональность в метрике пространства Z таких векторов из множества

$$\{z_i\}_{i=1}^{N_u+N_p},$$

у которых первые компоненты различны. Действительно, при любых k и j векторы h_k и g_j принадлежат множеству $\{u_i\}_{i=1}^{N_u}$ собственных векторов задачи (1.8), для которых, если $n \neq l$, справедливо равенство

$$(Au_n, u_l) = 0.$$

При этом для вторых компонент векторов z_j — решений задачи (1.7) — справедливо равенство

$$(Cp_n, p_l) = 0, \quad n \neq l.$$

Пусть теперь некоторый ненулевой вектор

$$z = \{u, p\} \in Z$$

ортogonalен произвольному вектору из множества

$$\{z_i\}_{i=1}^{N_u+N_p}.$$

Тогда, поскольку системы

$$\{h_i\}_{i=1}^{N_u-N_p}, \quad \{g_i\}_{i=1}^{N_p} \quad \text{и} \quad \{p_i\}_{i=1}^{N_p}$$

являются базисами в пространствах H , G и P соответственно, получаем, что компоненты z могут иметь только следующий вид:

$$u = \sum_{i=1}^{N_p} c_i g_i, \quad p = \sum_{i=1}^{N_p} d_i p_i.$$

Но для каждого g_i имеется пара различных векторов $z_i^{(2)}$ и $z_i^{(3)}$ вида

$$\{g_i, \varkappa_i^{(2)} p_i\} \quad \text{и} \quad \{g_i, \varkappa_i^{(3)} p_i\},$$

откуда, предполагая наличие некоторого c_k (или d_k , порознь или одновременно), не равного нулю, после скалярного умножения z на $z_k^{(2)}$ и $z_k^{(3)}$ получим равенства

$$c_k \chi_1(Ag_k, g_k) + d_k \chi_2(Cp_k, p_k) \varkappa_k^{(2)} = 0,$$

$$c_k \chi_1(Ag_k, g_k) + d_k \chi_2(Cp_k, p_k) \varkappa_k^{(3)} = 0.$$

Отличие от нуля определителя линейной системы (так как $\varkappa_k^{(2)} \neq \varkappa_k^{(3)}$, $g_k \neq 0$, $p_k \neq 0$) дает только тривиальное решение $c_k = d_k = 0$, откуда и следует искомое утверждение. ■

1.3.3. Полезное начальное приближение

Так как все рассматриваемые далее алгоритмы являются итерационными, то важным является выбор стартового (начального) приближения. Пусть в процессе итераций мы получаем приближения к решению $\{u^k, p^k\}$; тогда обозначим погрешность решения на k -й итерации через

$$y^k = \{v^k, r^k\} = \{u^k - u, p^k - p\},$$

где $\{u, p\}$ — точное решение задачи $L_0 z = F$, и выберем начальное приближение $z^0 = \{u^0, p^0\}$ из условия

$$Au^0 + Bp^0 = f \tag{1.9}$$

(например, если возьмем произвольный вектор p^0 и положим $u^0 = A^{-1}(f - Bp^0)$). Для такого начального приближения имеем

$$Av^0 + Br^0 = 0.$$

Отсюда следует, что v^0 является элементом подпространства $G \subset U$. Действительно, для произвольного элемента $h \in H$ справедливо $B^T h = 0$, поэтому

$$(Av^0, h) = -(Br^0, h) = -(r^0, B^T h) = 0.$$

Выясним, когда свойство принадлежности к G сохраняется в процессе итераций. Справедлива

Лемма 1.3.2. *Предположим, что одно из соотношений итерационного метода определяется формулой*

$$A \frac{u^{k+1} - u^k}{\tau_{k+1}} + (A + \beta BC^{-1} B^T) u^k + Bp^k = f + \beta BC^{-1} \varphi$$

с произвольными $\beta \in \mathbb{R}$, $\tau_{k+1} > 0$. Тогда для любой итерации k первая компонента v^k погрешности y^k итерационного метода, стартового с начального приближения вида (1.9), является элементом подпространства G , т. е. $(Av^k, h) = 0$ для $\forall h \in H$.

Доказательство. Из приведенной в условии леммы формулы следует, что компонента v^k удовлетворяет соотношению

$$Av^{k+1} = [(1 - \tau_{k+1})A - \tau_{k+1}\beta BC^{-1} B^T]v^k + \tau_{k+1}Br^k.$$

Отсюда для $\forall h \in H$ имеем

$$(Av^{k+1}, h) = (1 - \tau_{k+1})(Av^k, h).$$

Это означает, что если $v^k \in G$, то и $v^{k+1} \in G$. Таким образом, выбор начального приближения вида (1.9) гарантирует принадлежность для любой итерации k вектора v^k к подпространству G . ■

Из леммы 1.3.2 и теоремы 1.3.2 вытекает покомпонентное разложение погрешности следующего вида:

$$v^k = \sum_{i=1}^{N_p} c_i^{(k)} g_i, \quad r^k = \sum_{i=1}^{N_p} d_i^{(k)} p_i,$$

где $\{g_i\}$ и $\{p_i\}$ — базисы пространств G и P , порожденные задачами (1.7), (1.8). Основной задачей любого итерационного алгоритма является уменьшение нормы ошибки как элемента некоторого пространства. В нашем случае основным является пространство $Z = U \times P$. Использование рассматриваемого начального приближения (1.9) позволяет его сузить до подпространства $Z' = G \times P$. Часто это позволяет ускорить сходимость алгоритма. Так как Z' является подпространством Z , то итерации с таким свойством обычно называют *итерациями в подпространстве*.

1.4. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Первая попытка систематического изложения вопросов, связанных с постановкой задач с седловым оператором, их разрешимостью, корректностью, а также с итерационными методами их решения, была предпринята в монографии [41] (см. также ее переработанный перевод [151]). Особенностью изложения является ориентация на сеточные системы, возникающие при дискретизации уравнений типа Стокса и Навье–Стокса. Анализ эффективности итерационных методов связан в первую очередь с их асимптотической оптимальностью относительно параметра дискретизации. Позже появились изложения такого рода и в зарубежной литературе (см. [127]). Наиболее полным обзором в настоящее время является работа [117].

Первоначальный вариант алгоритма Узавы приведен в §5 главы 10 часто цитируемой книги [108] (имеется перевод — [103]). В используемых здесь обозначениях это — метод простой итерации для решения уравнения (1.4):

$$\frac{p^{k+1} - p^k}{\tau} + S_0 p^k = b.$$

Применительно к задаче Стокса в дифференциальной форме алгоритм Узавы рассматривался в ряде работ. Для случая фиксированного параметра τ в [79] получено достаточное условие сходимости: $0 < \tau < 2$. Наличие здесь абсолютной постоянной, равной двум, связано с тем, что спектр дифференциального аналога оператора $B^T A^{-1} B$ принадлежит отрезку $[\delta, 1]$, $\delta > 0$.

Этот результат был обобщен в работе [144] на случай использования переменного итерационного параметра:

$$0 < \inf_k (\tau_k) \leq \sup_k (\tau_k) < 2.$$

Кроме того, там же приведены формулы для параметров и оценки погрешностей для оптимального одношагового метода, циклического k -шагового метода и полуитерационного метода Чебышева.

Позднее в книге [156] была доказана сходимость методов наискорейшего спуска и сопряженных градиентов для решения системы (1.4), а в работе [174] получены оценки их скоростей сходимости для сеточных систем, возникающих при аппроксимации задачи Стокса.

Дальнейшее развитие метода Узавы заключается в построении различными способами для конкретных сеточных систем операторов

предобусловливания C таких, что отношение констант эквивалентности Γ/γ в матричном неравенстве

$$\gamma C \leq S_0 \leq \Gamma C$$

минимально. Следует отметить, что здесь в полной мере используется специфика конкретных задач, и поэтому эта тематика чрезвычайно обширна и разнообразна. Например, для задачи Стокса с параметром $\alpha \geq 0$

$$\begin{aligned} -\Delta \mathbf{u} + \operatorname{grad} p + \alpha \mathbf{u} &= \mathbf{f} && \text{в } \Omega, \\ \operatorname{div} \mathbf{u} &= 0 && \text{в } \Omega, \\ \mathbf{u} &= 0 && \text{на } \partial\Omega, \end{aligned} \quad (1.10)$$

возникающей при неявной дискретизации по времени нестационарной первой краевой задачи Стокса, в [131] предложено, а в работе [70] обосновано использование оператора предобусловливания с константами эквивалентности γ и Γ , не зависящими как от шагов сетки, так и от параметра α . Для классической задачи Стокса ($\alpha = 0$) в области типа вытянутого прямоугольника в работе [177] построен оператор предобусловливания с константами, не зависящими как от шагов сетки, так и от отношения сторон прямоугольника. Хорошая библиография на эту тему представлена в [171].

Рассмотрим также пример задачи, для которой метод Уза-вы — сопряженных градиентов является не очень привлекательным. Пусть в постановке (1.2) зафиксировано: $N_u = N_p = n - 1$, $\varepsilon = 1$, $A = D = I$ (единичный оператор), а матрица $B = B^T$ является дискретизацией на равномерной сетке с шагом $h = 1/n$ дифференциального оператора из задачи

$$-\alpha y'' = f(x), \quad x \in (0, 1), \quad y(0) = y(1) = 0,$$

$\alpha \geq 1$ — параметр, т. е.

$$b_{kj} = \begin{cases} 2\alpha n^2 & \text{при } k = j, \\ -\alpha n^2 & \text{при } |k - j| = 1, \\ 0 & \text{при } |k - j| > 1. \end{cases}$$

Легко заметить, что в этом случае спектральное число обусловленности дополнения Шура $S_1 = (\alpha B)^2 + I$ имеет порядок $O(\alpha^2 n^4)$, что требует для решения (1.2) разработки и анализа алгоритмов, основанных на принципиально других идеях [65]. Этот пример, конечно, является одним из простейших, поскольку эллиптический оператор в нем можно усложнить как увеличением размерности

пространства, так и за счет введения переменных (возможно, разрывных) коэффициентов.

Метод сопряженных градиентов предложен в работе [161], его поведение для задач большой размерности, в том числе возможность неустойчивости, проанализировано в [59]. Ускорение скорости сходимости итерационных алгоритмов за счет использования операторов, эквивалентных по спектру, предложено в [38].

Лемма 1.3.1 является изложением в подходящих обозначениях теоремы 1.3.20 из [82]. Предложенные в [88] теоремы 1.3.1 и 1.3.2 самостоятельной ценности, по-видимому, не имеют, но полезны для решения вспомогательных спектральных задач при исследовании сходимости различных алгоритмов релаксационного типа.

Методы, использующие итерации в подпространстве, появились в вычислительной практике в начале 60-х годов прошлого века. Их привлекательными свойствами являются экономия памяти и ускорение сходимости [58], что эффективно используется при решении сложных с вычислительной точки зрения задач, например при моделировании переноса нейтронов или лазерно-плазменных взаимодействий.

МОДИФИЦИРОВАННЫЕ МЕТОДЫ РЕЛАКСАЦИИ. ОБЩИЙ АНАЛИЗ

В главе приводятся краткие сведения из теории методов релаксации и для системы

$$L_0 z \equiv \begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ \varphi \end{pmatrix} \equiv F \quad (2.1)$$

рассматриваются вопросы, связанные со сходимостью и асимптотической оптимизацией модифицированных релаксационных алгоритмов: параметризация спектра оператора перехода, решение соответствующих неравенств и минимаксных задач, итерации в подпространстве и т. п.

2.1. СВЕДЕНИЯ О МЕТОДАХ РЕЛАКСАЦИИ

2.1.1. Общие понятия

Методы релаксации решения невырожденной системы $Au = f$ относятся к классу линейных стационарных одношаговых итерационных методов, для записи которых используется формула

$$B \frac{u^{k+1} - u^k}{\tau} + Au^k = f,$$

где B — невырожденная матрица, а τ — числовой параметр.

Пусть I — единичная матрица, тогда матрица

$$T = I - \tau B^{-1} A$$

называется *матрицей (оператором) перехода* итерационного метода, а величина $\rho(T)$, равная максимальному по модулю собственному значению матрицы T , — ее *спектральным радиусом*.

Условие

$$\rho(T) < 1$$

необходимо и достаточно для сходимости с произвольного начального приближения линейного стационарного одношагового итерационного метода.

Асимптотическая скорость сходимости линейного стационарного одношагового итерационного метода с матрицей перехода T , для которой выполнено условие $\rho(T) < 1$, определяется формулой

$$r_{\infty} = -\ln \rho(T).$$

Пусть оператор перехода T непрерывно зависит (как от параметров) от компонент вектора $\gamma = (\gamma_1, \dots, \gamma_p) \in \mathbb{R}^p$ (при некотором $p \geq 1$), т. е. $T = T(\gamma)$, и $G_{\gamma} \subseteq \mathbb{R}^p$ — множество векторов $\gamma \in \mathbb{R}^p$, для которых $\rho(T) < 1$. Тогда задача асимптотической оптимизации метода эквивалентна максимизации его асимптотической скорости сходимости и заключается в нахождении такого вектора $\gamma_{opt} \in G_{\gamma}$, чтобы

$$\rho(T(\gamma_{opt})) = \min_{\gamma \in G_{\gamma}} \rho(T(\gamma)),$$

и является минимаксной задачей вида

$$\min_{\gamma \in G_{\gamma}} \max_{\lambda \in \sigma(T(\gamma))} |\lambda|.$$

Здесь было использовано обозначение *спектра матрицы* $\sigma(T)$ — множества из n собственных значений матрицы T порядка n (каждое отличное от других собственное значение берется столько раз, какова его кратность). Отметим, что решение задачи асимптотической оптимизации метода тесно связано с условием его сходимости, и поэтому в дальнейшем эти две проблемы будут рассматриваться одновременно.

2.1.2. Метод Якоби

Пусть $A = (a_{ij})$ — некоторая невырожденная $(n \times n)$ -матрица с ненулевыми диагональными элементами и определена матрица $\Lambda = \text{diag}(a_{11}, \dots, a_{nn})$. Тогда итерационный метод

$$\Lambda(u^{k+1} - u^k) + Au^k = f \quad (2.2)$$

называется *точечным методом Якоби* решения системы $Au = f$. Пусть

$$A = \begin{pmatrix} A_{11} & \dots & A_{1s} \\ \dots & \dots & \dots \\ A_{s1} & \dots & A_{ss} \end{pmatrix}$$

блочная матрица с квадратными невырожденными блоками A_{ii} , $i = 1, \dots, s$, и

$$\Lambda = A_{11} \oplus A_{22} \oplus \dots \oplus A_{ss}.$$

В этом случае итерационный метод (2.2) называется *блочным методом Якоби*.

Пусть Λ — матрица какого-либо из вариантов (точечного или блочного) метода Якоби и ω — некоторый числовой параметр. Тогда итерационный метод

$$\Lambda \frac{u^{k+1} - u^k}{\omega} + Au^k = f \quad (2.3)$$

называется *релаксированным методом Якоби* (Jacobi Over Relaxation — JOR).

Пусть $s = 2$ и ω_1, ω_2 — некоторые числовые параметры. Тогда блочный вариант метода (2.3) — *модифицированный релаксированный метод Якоби* (MJOR) — имеет вид:

$$\begin{cases} A_{11} \frac{u_1^{k+1} - u_1^k}{\omega_1} + A_{11}u_1^k + A_{12}u_2^k = f_1, \\ A_{22} \frac{u_2^{k+1} - u_2^k}{\omega_2} + A_{21}u_1^k + A_{22}u_2^k = f_2. \end{cases} \quad (2.4)$$

Пусть $A = A^T > 0$. Релаксированный метод Якоби (2.3) сходится с произвольного начального приближения тогда и только тогда, когда $\omega \in (0, 2/\rho(\Lambda^{-1}A))$. При этом оптимальное значение параметра (в смысле максимума асимптотической скорости сходимости) вычисляется по формуле

$$\omega_0 = 2/(\gamma + \Gamma),$$

где $\gamma = 1/\rho(A^{-1}\Lambda)$ — минимальное собственное значение матрицы $\Lambda^{-1}A$ и $\Gamma = \rho(\Lambda^{-1}A)$.

Пусть $s = 2$ и $A = A^T > 0$. Тогда если величина μ является собственным значением оператора перехода T_ω в блочном релаксированном методе Якоби (2.3) при $\omega = 1$, то и величина $(-\mu)$ также является собственным значением T_1 . Отсюда, в частности, следует, что оптимальное значение параметра ω (в смысле максимума асимптотической скорости сходимости) в (2.3) определяется формулой $\omega_0 = 1$. Для модифицированного метода (2.4) это влечет за собой неравенство

$$\rho(T_{\omega_1, \omega_2}) \geq \rho(T_{\omega_0, \omega_0}).$$

Таким образом, в случае (2×2) -блочных линейных алгебраических систем с симметричной положительно определенной матрицей A оптимальным вариантом модифицированного релаксированного метода Якоби является выбор

$$\omega_1 = \omega_2 = 1,$$

и следовательно, не имеет смысла вводить в классический вариант блочного метода Якоби итерационные параметры.

2.1.3. Метод SOR

Пусть $A = (a_{ij})$ — некоторая невырожденная $(n \times n)$ -матрица с ненулевыми (невырожденными) диагональными элементами (блоками), $\Lambda = \text{diag}(a_{11}, \dots, a_{nn})$ (или $\Lambda = A_{11} \oplus A_{22} \oplus \dots \oplus A_{ss}$), L — строго нижняя треугольная матрица, U — строго верхняя треугольная матрица, так что $A = \Lambda + L + U$. Тогда итерационный метод

$$\left(\frac{1}{\omega}\Lambda + L\right)(u^{k+1} - u^k) + Au^k = f \quad (2.5)$$

называется *методом последовательной верхней релаксации* решения системы $Au = f$, а ω — *релаксационным параметром*. В зарубежной литературе для метода последовательной верхней релаксации принято обозначение SOR (Successive Over Relaxation). Если все блоки матрицы имеют порядок единица, метод называется *точечным*, в противном случае — *блочным*.

Метод Гаусса–Зейделя является частным случаем метода последовательной верхней релаксации, когда $\omega = 1$.

Пусть $s = 2$ и ω_1, ω_2 — некоторые числовые параметры. Тогда блочный вариант метода (2.5) — *модифицированный метод последовательной верхней релаксации* (MSOR) — имеет вид:

$$\begin{cases} A_{11} \frac{u_1^{k+1} - u_1^k}{\omega_1} + A_{11}u_1^k + A_{12}u_2^k = f_1, \\ A_{22} \frac{u_2^{k+1} - u_2^k}{\omega_2} + A_{21}u_1^{k+1} + A_{22}u_2^k = f_2. \end{cases} \quad (2.6)$$

Пусть $A = A^T > 0$. Метод последовательной верхней релаксации (2.5) сходится с произвольного начального приближения тогда и только тогда, когда $\omega \in (0, 2)$.

Пусть $s = 2$ и $A = A^T > 0$. Тогда собственные значения λ оператора перехода T_{ω_1, ω_2} в методе (2.6) связаны с собственными значениями μ оператора перехода T_1 в блочном методе (2.3) соотношением

$$(\lambda + \omega_1 - 1)(\lambda + \omega_2 - 1) = \omega_1 \omega_2 \mu^2 \lambda.$$

Более того, если $\mu \neq 0$ является собственным значением матрицы T_1 , то оба корня этого уравнения являются собственными значениями матрицы T_{ω_1, ω_2} , в противном случае либо одна из величин $\omega_1 - 1$ и $\omega_2 - 1$ является собственным значением матрицы T_{ω_1, ω_2} , либо обе эти величины.

Пусть $s = 2$ и $A = A^T > 0$. Модифицированный метод последовательной верхней релаксации (2.6) сходится с произвольного начального приближения тогда и только тогда, когда $\omega_1, \omega_2 \in (0, 2)$.

Пусть $s = 2$, $A = A^T > 0$ и $\omega_1 = \omega_2 = \omega$. Тогда оптимальное значение параметра ω (в смысле максимума асимптотической скорости сходимости) в (2.5) вычисляется по формуле

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \rho^2(T_1)}}.$$

При этом $\rho(T_{\omega_0}) = \omega_0 - 1$ и соответственно

$$r_\infty = -\ln(\omega_0 - 1).$$

Для модифицированного метода (2.6) это влечет за собой неравенство

$$\rho(T_{\omega_1, \omega_2}) \geq \rho(T_{\omega_0, \omega_0}).$$

Таким образом, в случае (2×2) -блочных линейных алгебраических систем с симметричной положительно определенной матрицей A оптимальным вариантом модифицированного метода верхней релаксации является выбор $\omega_1 = \omega_2 = \omega_0$, и следовательно, не имеет смысла вводить в блочный метод верхней релаксации дополнительные итерационные параметры.

2.1.4. Метод SSOR

Пусть $A = (a_{ij})$ — некоторая невырожденная $(n \times n)$ -матрица с ненулевыми (невырожденными) диагональными элементами (блоками), $\Lambda = \text{diag}(a_{11}, \dots, a_{nn})$ (или $\Lambda = A_{11} \oplus A_{22} \oplus \dots \oplus A_{ss}$), L — строго нижняя треугольная матрица, U — строго верхняя треугольная матрица, так что $A = \Lambda + L + U$. Тогда итерационный метод

$$\begin{cases} \left(\frac{1}{\omega} \Lambda + L \right) (u^{k+1/2} - u^k) + Au^k = f, \\ \left(\frac{1}{\omega} \Lambda + U \right) (u^{k+1} - u^{k+1/2}) + Au^{k+1/2} = f \end{cases} \quad (2.7)$$

называется *методом симметричной последовательной верхней релаксации* решения системы $Au = f$. В зарубежной литературе для него принято обозначение SSOR. Если все блоки матрицы имеют порядок единица, метод называется *точечным*, в противном случае — *блочным*.

Пусть $A = A^T > 0$. Метод симметричной последовательной верхней релаксации (2.7) сходится с произвольного начального приближения тогда и только тогда, когда $\omega \in (0, 2)$.

Асимптотическая оптимизация метода симметричной последовательной верхней релаксации эквивалентна минимизации A -нормы

его матрицы перехода $T(\omega)$:

$$T(\omega) = I - \left(\frac{2}{\omega} - 1 \right) \left(\frac{1}{\omega} \Lambda + U \right)^{-1} \Lambda \left(\frac{1}{\omega} \Lambda + L \right)^{-1} A.$$

Пусть $s = 2$ и ω_1, ω_2 — некоторые числовые параметры. Тогда блочный вариант метода (2.7) — *модифицированный метод симметричной последовательной верхней релаксации* (MSSOR) — имеет вид:

$$\begin{cases} A_{11} \frac{u_1^{k+1/2} - u_1^k}{\omega_1} + A_{11} u_1^k + A_{12} u_2^k = f_1, \\ A_{22} \frac{u_2^{k+1/2} - u_2^k}{\omega_2} + A_{21} u_1^{k+1/2} + A_{22} u_2^k = f_2, \\ A_{22} \frac{u_2^{k+1} - u_2^{k+1/2}}{\omega_2} + A_{21} u_1^{k+1/2} + A_{22} u_2^{k+1/2} = f_2, \\ A_{11} \frac{u_1^{k+1} - u_1^{k+1/2}}{\omega_1} + A_{11} u_1^{k+1/2} + A_{12} u_2^{k+1} = f_1. \end{cases} \quad (2.8)$$

Пусть

$$s = 2 \quad \text{и} \quad A = A^T > 0.$$

Модифицированный метод симметричной последовательной верхней релаксации (2.8) сходится с произвольного начального приближения тогда и только тогда, когда $\omega_1, \omega_2 \in (0, 2)$.

Пусть

$$s = 2, \quad A = A^T > 0 \quad \text{и} \quad \omega_1 = \omega_2 = \omega.$$

Тогда оптимальное значение параметра ω метода симметричной последовательной верхней релаксации (2.7) (в смысле максимума асимптотической скорости сходимости)

$$\min_{\omega \in (0, 2)} \|T(\omega)\|_A = \|T(1)\|_A$$

равно единице, т. е. $\omega_0 = 1$.

Для модифицированного метода (2.8) оптимальным является выбор

$$\omega_1 + \omega_2 - \omega_1 \omega_2 = \frac{2}{1 + \sqrt{1 - \rho^2(T_1)}},$$

при этом

$$\rho(T(\omega_1, \omega_2)) = \frac{1 - \sqrt{1 - \rho^2(T_1)}}{1 + \sqrt{1 - \rho^2(T_1)}},$$

где T_1 — оператор перехода в блочном методе (2.3). Таким образом, в случае (2×2) -блочных линейных алгебраических систем с симметричной положительно определенной матрицей A не имеет

смысла вводить в блочный метод симметричной верхней релаксации дополнительные итерационные параметры, поскольку он имеет скорость сходимости, совпадающую с асимптотической скоростью метода SOR, но в то же время требует на каждой итерации в два раза больше арифметических действий. Все это делает рассматриваемый в данных условиях алгоритм малопривлекательным с практической точки зрения.

2.2. МОДИФИЦИРОВАННЫЙ МЕТОД ЯКОБИ (MJOR)

2.2.1. Построение метода

Классические релаксационные алгоритмы строятся в предположении отличия от нуля диагональных элементов (невырожденности диагональных блоков) матрицы исходной системы уравнений. Однако построение модифицированных методов не требует выполнения этого условия. Приведем пример.

Пусть вектор $z = \{u, p\}$ является решением системы $L_0 z = F$:

$$\begin{cases} Au + Bp = f, \\ B^T u = \varphi. \end{cases}$$

Рассмотрим систему $L_\varepsilon z = \tilde{F}$, равносильную исходной и полученную с помощью параметра $\varepsilon \geq 0$ и матрицы $C = C^T > 0$:

$$\begin{cases} Au + Bp = f, \\ B^T u - \varepsilon Cp = \tilde{\varphi}, \end{cases}$$

где $\tilde{\varphi} = \varphi - \varepsilon Cp$. Запишем для нее модифицированный, т. е. с двумя итерационными параметрами ω_1, ω_2 , релаксированный метод Якоби:

$$\begin{cases} A \frac{u^{k+1} - u^k}{\omega_1} + Au^k + Bp^k = f, \\ -\varepsilon C \frac{p^{k+1} - p^k}{\omega_2} + B^T u^k - \varepsilon Cp^k = \tilde{\varphi}. \end{cases}$$

Положим далее $\omega_2 = \varepsilon \tilde{\omega}_2$ и перейдем к пределу при $\varepsilon \rightarrow 0$. В результате будем иметь

$$\begin{cases} A \frac{u^{k+1} - u^k}{\omega_1} + Au^k + Bp^k = f, \\ -C \frac{p^{k+1} - p^k}{\tilde{\omega}_2} + B^T u^k = \varphi. \end{cases}$$

Отсюда после формального переобозначения параметров

$$\omega_1 \longrightarrow \tau, \quad \tilde{\omega}_2 \longrightarrow \tau/\alpha$$

следуют искомые формулы метода MJOR:

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau} + Au^k + Bp^k = f, \\ -\alpha C \frac{p^{k+1} - p^k}{\tau} + B^T u^k = \varphi. \end{cases} \quad (2.9)$$

2.2.2. Спектр оператора перехода

Обозначим через T оператор перехода в алгоритме (2.9) и рассмотрим спектральную задачу $Tz = \lambda z$:

$$Tz \equiv \begin{pmatrix} (1-\tau)I & -\tau A^{-1}B \\ (\tau/\alpha)C^{-1}B^T & I \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} u \\ p \end{pmatrix}. \quad (2.10)$$

Теорема 2.2.1. Спектр $\sigma(T)$ оператора перехода T в методе (2.9) принадлежит множеству

$$\Lambda = \{1 - \tau\} \cup \left\{ 1 - \frac{\tau}{2} \pm \frac{\tau}{2} \sqrt{1 - \frac{4t}{\alpha}} \right\}, \quad t \in [\gamma, \Gamma].$$

Доказательство. Для вывода формул собственных значений оператора T используем базис пространства Z , построенный в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (2.10) с $\lambda_k^{(1)} = 1 - \tau$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^T g_j\}, \quad j = 1, \dots, N_p.$$

После применения к первому уравнению (2.10) матрицы $C^{-1}B^T$ и замены $C^{-1}B^T g_j = p_j$ будем иметь

$$\begin{cases} (1 - \tau)p_j + \tau \kappa^{-1}C^{-1}S_0 p_j = \lambda p_j, \\ -\frac{\tau}{\alpha}p_j + \kappa^{-1}p_j = \lambda \kappa^{-1}p_j. \end{cases}$$

Перепишем эту систему в виде

$$\begin{cases} S_0 p_j = \frac{\kappa(\lambda - 1 + \tau)}{\tau} C p_j, \\ p_j \left(\lambda - 1 + \kappa \frac{\tau}{\alpha} \right) = 0. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0 p = t C p$ соответствует собственное значение t_j , $j = 1, \dots, N_p$ по теореме 1.3.1. Зафиксировав его, из полученной системы для λ и κ выведем соотношения

$$t_j = \frac{\kappa(\lambda - 1 + \tau)}{\tau}, \quad \lambda - 1 + \kappa \frac{\tau}{\alpha} = 0.$$

Исключая из этих уравнений κ , приходим к выражению

$$\lambda_j^{(2,3)} = 1 - \frac{\tau}{2} \pm \frac{\tau}{2} \sqrt{1 - 4t_j/\alpha}$$

и, соответственно,

$$\kappa_j^{(2,3)} = \frac{\alpha}{2} \left(1 \mp \sqrt{1 - 4t_j/\alpha} \right).$$

Перейдем к обоснованию того факта, что указанным способом найдены все собственные значения задачи (2.10).

Рассмотрим вначале более простой случай различных собственных значений $\lambda_j^{(2,3)}$ при фиксированном j . Пусть параметр α таков, что выражение $1 - 4t_j/\alpha$ не обращается в нуль ни при каком значении t_j . Тогда величины $\kappa_j^{(2)}$ и $\kappa_j^{(3)}$ для любого j также будут различны, что следует из их явного представления. Это означает, что полученная система собственных векторов

$$\{z_i^{(1,2,3)}\}_{i=1}^{N_u+N_p}$$

задачи (2.10) удовлетворяет теореме 1.3.2, следовательно, все искомые собственные значения найдены.

Обобщим доказательство на случай кратных собственных значений $\lambda_j^{(2,3)}$, т. е. обращения в нуль выражения $1 - 4t_j/\alpha$ при некотором t_j . Сразу необходимо отметить, что в силу линейности рассматриваемого выражения при любом фиксированном α такое значение t_j может быть только одно. Его кратность при этом значения не имеет, так как по теореме 1.3.1 даже одинаковым значениям t_j соответствуют различные C -ортогональные векторы p_j . Итак, пусть в рассматриваемом случае $\lambda_j^{(2)} = \lambda_j^{(3)} = \lambda_j$ имеется только один собственный вектор $z_j^{(2)}$ оператора T с первой компонентой g_j . Тогда для построения полной в Z системы векторов достаточно добавить к нему в пару корневой вектор высоты два

$$z_j^{(3)} = \{g_j, -p_j(\lambda_j + \tau)/t_j\tau\},$$

удовлетворяющий уравнению $(T - \lambda_j I)^2 z_j^{(3)} = 0$ и линейно независимый с соответствующим собственным вектором

$$z_j^{(2)} = \{g_j, -2/\alpha p_j\}.$$

Теперь уточненная система собственных векторов $\{z_i^{(1,2,3)}\}_{i=1}^{N_u+N_p}$ удовлетворяет теореме 1.3.2.

Таким образом, полнота найденной системы векторов в пространстве Z , состоящей либо только из собственных векторов оператора T , либо с добавлением к ним корневых по указанной схеме, дает основание утверждать, что других собственных значений в задаче (2.10), отличных от указанных, не существует. Вспомним, что $t_j \in [\gamma, \Gamma]$ (см. (1.5)). Это порождает принадлежность спектра оператора перехода $\sigma(T)$ множеству

$$\Lambda = \{1 - \tau\} \cup \left\{1 - \frac{\tau}{2} \pm \frac{\tau}{2} \sqrt{1 - 4t/\alpha}, t \in [\gamma, \Gamma]\right\}. \quad \blacksquare$$

2.2.3. Условие сходимости

Полученное представление спектра оператора перехода дает возможность выяснить условия сходимости метода (2.9). Имеет место

Теорема 2.2.2. При любом $\alpha > 0$ и произвольном начальном приближении $z^0 \in Z$ необходимым и достаточным условием сходимости метода (2.9) является выполнение неравенства

$$0 < \tau < \min\left(2, \frac{\alpha}{\Gamma}\right). \quad (2.11)$$

Доказательство. Введем следующие обозначения для элементов множества Λ :

$$\begin{aligned} \lambda_1 &= 1 - \tau, \\ \lambda_{2,3} &= 1 - \frac{\tau}{2} \pm \frac{\tau}{2} \sqrt{1 - 4t/\alpha}. \end{aligned}$$

Знак «+» относится к λ_2 .

Достаточность. Отметим, что ограничение $|\lambda_1| < 1$ дает $0 < \tau < 2$. Рассмотрим теперь некоторую фиксированную точку $t \in [\gamma, \Gamma]$ и выясним соотношение между параметрами τ и α , при котором $|\lambda_{2,3}| < 1$.

Изучим сначала случай различных вещественных значений $\lambda_{2,3}$, т. е. когда $1 - 4t/\alpha > 0$. Так как $\lambda_2 > \lambda_3$, то достаточно исследовать неравенства

$$-1 < \lambda_3 < \lambda_2 < 1.$$

Условие $\lambda_2 < 1$ выполняется всегда, так как $\sqrt{1 - 4t/\alpha} < 1$. Несложные преобразования выражения $\lambda_3 > -1$ приводят к неравенству

$$-\tau^2 \frac{t}{\alpha} < 2(2 - \tau),$$

которое также справедливо при любом $0 < \tau < 2$. Таким образом, условие (2.11) гарантирует выполнение неравенства

$$\max_t |\lambda_{1,2,3}| < 1$$

в случае различных вещественных $\lambda_{2,3}$.

Рассмотрим теперь случай комплексных (или кратных) значений $\lambda_{2,3}$ при некотором $t \in [\gamma, \Gamma]$, т. е. когда выполнено неравенство $1 - 4t/\alpha \leq 0$. При этом

$$|\lambda_{2,3}|^2 = 1 - \tau + \tau^2 t / \alpha.$$

Неравенство $|\lambda_{2,3}|^2 < 1$ в силу положительности параметра τ равносильно следующему: $\tau < \alpha/t$, откуда получаем, что для нахождения комплексных и кратных значений $\lambda \in \Lambda$ внутри единичного круга достаточно выполнения неравенства $0 < \tau < \alpha/\Gamma$.

Завершение доказательства достаточности следует из объединения двух рассмотренных выше случаев.

Необходимость. Доказательство будем проводить от противного. Пусть условие (2.11) не выполнено. Возможны два варианта. В первом, при $\alpha/\Gamma \geq 2$, положим $\tau = 2$, тогда $|\lambda_1| = 1$ и метод не будет сходиться при любом начальном приближении вида $z^0 = (u^0, p^0)$, $u^0 \in H$.

Во втором варианте, при $\alpha/\Gamma < 2$, покажем, что даже для единственной точки отрезка $[\gamma, \Gamma]$, а именно $t = \Gamma$, невыполнение (2.11) приводит к неравенству $|\lambda_{2,3}| \geq 1$. Пусть

$$\tau = \frac{\alpha}{\Gamma} + \delta < 2, \quad 0 \leq \delta < 2 - \frac{\alpha}{\Gamma}.$$

При этом λ_2 и λ_3 будут комплексно-сопряженными, так как дискриминант, равный $1 - 4\Gamma/\alpha$, отрицателен (он строго меньше, чем -1).

Рассмотрим изменение определяющей величины

$$|\lambda_{2,3}|^2 = 1 - \tau + \tau^2 \frac{\Gamma}{\alpha}$$

в зависимости от параметра δ . Так как $\alpha/\Gamma < 2$, то справедлива оценка снизу

$$|\lambda_{2,3}|^2 = 1 + \delta + \delta^2 \frac{\Gamma}{\alpha} \geq \left(1 + \frac{\delta}{2}\right)^2 + \frac{\delta^2}{4}.$$

Последнее неравенство приводит к условию $1 \leq |\lambda_{2,3}|$ при

$$0 \leq \delta < 2 - \frac{\alpha}{\Gamma}.$$

Напомним, что $t = \Gamma$ является точным собственным значением задачи $S_0 p = t C p$ (см. раздел 1.3) и, следовательно, $\lambda_{2,3}$ при $t = \Gamma$ являются собственными значениями оператора T в (2.10). Отсюда следует, что при невыполнении условия (2.11) существует собственный вектор оператора T такой, что отвечающее ему одно из собственных значений $\lambda_{2,3}$ по модулю не меньше единицы. С учетом этого замечания и теоремы о необходимом и достаточном условии сходимости метода простой итерации (см., например, [15], с. 269) получаем, что выполнение неравенства (2.11) является необходимым и достаточным для сходимости метода (2.9). ■

2.2.4. Задача асимптотической оптимизации

Знание аналитического представления спектра оператора перехода $\sigma(T)$ позволяет сформулировать и решить задачу асимптотической оптимизации метода: *найти положительные значения τ_0 и α_0 , минимизирующие спектральный радиус оператора перехода T*

$$q = \max_{t \in [\gamma, \Gamma]} \left\{ |1 - \tau|, \left| 1 - \frac{\tau}{2} \pm \frac{\tau}{2} \sqrt{1 - \frac{4t}{\alpha}} \right| \right\}, \quad 0 < \gamma \leq \Gamma. \quad (2.12)$$

Отметим, что случай $\gamma = \Gamma$ не представляет интереса (это означает явную обратимость матрицы S_0 в методе Узава из раздела 1.2), поэтому далее всюду будем предполагать, что $\gamma < \Gamma$.

Докажем несколько вспомогательных утверждений. Введем обозначения:

$$D(t, \alpha) = 1 - \frac{4t}{\alpha}, \quad \lambda_1(\tau) = 1 - \tau, \quad \lambda_{2,3} = 1 - \frac{\tau}{2} \pm \frac{\tau}{2} \sqrt{1 - \frac{4t}{\alpha}},$$

$$\lambda(t, \tau, \alpha) = \begin{cases} \lambda_2 & \text{при } D(t, \alpha) > 0, \\ \sqrt{\tau^2 t / \alpha - \tau + 1} & \text{при } D(t, \alpha) \leq 0. \end{cases}$$

Знак «+» в $\lambda_{2,3}$ относится к λ_2 . Переписав задачу (2.12) в новых обозначениях

$$q = \max_{t \in [\gamma, \Gamma]} \{ |\lambda_1(\tau)|, |\lambda_2|, |\lambda_3| \},$$

упростим целевую функцию. Справедлива

Лемма 2.2.1. *Функция q представима в следующем виде:*

$$q = \max \{ |\lambda_1(\tau)|, \lambda(\gamma, \tau, \alpha), \lambda(\Gamma, \tau, \alpha) \}.$$

Доказательство. Сначала покажем, что

$$q = \max_{t \in [\gamma, \Gamma]} \{ |\lambda_1(\tau)|, \lambda(t, \tau, \alpha) \}.$$

Действительно, при любом фиксированном $t \in [\gamma, \Gamma]$ в случае $D(t, \alpha) > 0$ справедливо

$$\lambda(t, \tau, \alpha) = \lambda_2 = |\lambda_2| > |\lambda_3|.$$

Если $D(t, \alpha) \leq 0$, то имеем равенства

$$|\lambda_2| = |\lambda_3| = \sqrt{\tau^2 t / \alpha - \tau + 1} = \lambda(t, \tau, \alpha).$$

Предположим теперь, что

$$q = \max_{t \in [\gamma, \Gamma]} \{|\lambda_1(\tau)|, \lambda(t, \tau, \alpha)\} = \max_{t \in (\gamma, \Gamma)} \{|\lambda_1(\tau)|, \lambda(t, \tau, \alpha)\},$$

т. е. максимум достигается во внутренней точке отрезка $[\gamma, \Gamma]$. В этом случае справедливо $q = |\lambda_1(\tau)|$. Показав это, придем к утверждению леммы.

Пусть t_0 — некоторая внутренняя точка отрезка $[\gamma, \Gamma]$ ($\gamma < t_0 < \Gamma$) такая, что

$$q = \lambda(t_0, \tau, \alpha),$$

тогда для произвольного достаточно малого $\varepsilon > 0$ найдутся две точки $t_\varepsilon = t_0 - \varepsilon$ и $\tilde{t}_\varepsilon = t_0 + \varepsilon$ из интервала (γ, Γ) такие, что

$$1 - \frac{\tau}{2} + \frac{\tau}{2} \sqrt{1 - \frac{4t_\varepsilon}{\alpha}} > 1 - \frac{\tau}{2} + \frac{\tau}{2} \sqrt{1 - \frac{4t_0}{\alpha}} \quad \text{при } D(t_0, \alpha) > 0,$$

$$\sqrt{\tau^2 \frac{\tilde{t}_\varepsilon}{\alpha} - \tau + 1} > \sqrt{\tau^2 \frac{t_0}{\alpha} - \tau + 1} \quad \text{при } D(t_0, \alpha) \leq 0.$$

Эти неравенства противоречат определению максимума непрерывной функции, поэтому такой точки t_0 не существует. ■

Отметим некоторые свойства целевой функции q .

Лемма 2.2.2. Решение задачи (2.12) обладает следующим свойством:

- 1) если $q = \lambda(\gamma, \tau, \alpha)$, то $D(\gamma, \alpha) > 0$;
- 2) если $q = \lambda(\Gamma, \tau, \alpha)$, то $D(\Gamma, \alpha) \leq 0$.

Доказательство. Рассмотрим первый вариант. Пусть справедливы соотношения $q = \lambda(\gamma, \tau, \alpha)$ и $D(\gamma, \alpha) \leq 0$, тогда

$$\lambda(\Gamma, \tau, \alpha) = \sqrt{\tau^2 \Gamma / \alpha - \tau + 1} > \sqrt{\tau^2 \gamma / \alpha - \tau + 1} = \lambda(\gamma, \tau, \alpha) = q.$$

Это противоречие завершает доказательство. Аналогично доказывается и второе утверждение леммы. ■

Обозначим через τ_0, α_0 экстремальные значения параметров, через q_0 — соответствующее значение целевой функции.

Лемма 2.2.3. Для решения задачи (2.12) выполняется равенство

$$q_0 = |\lambda_1(\tau_0)|.$$

Доказательство. Покажем, что если утверждение леммы не выполнено, то значение целевой функции q можно уменьшить. А это, в свою очередь, противоречит определению решения задачи (2.12), так как минимальное значение q есть q_0 . Пусть $q_0 \neq |\lambda_1(\tau_0)|$, тогда

$$q_0 = \max\{\lambda(\gamma, \tau_0, \alpha_0), \lambda(\Gamma, \tau_0, \alpha_0)\}.$$

Теперь на основании леммы 2.2.2 можно записать

$$q_0 = \max\left\{1 - \frac{\tau_0}{2} + \frac{\tau_0}{2} \sqrt{1 - \frac{4\gamma}{\alpha_0}}, \sqrt{\tau_0^2 \frac{\Gamma}{\alpha_0} - \tau_0 + 1}\right\}.$$

Пусть максимум достигается на первом выражении, тогда положим $\tilde{\tau}_0 = \tau_0 + \varepsilon$, $\tilde{\alpha}_0 = \alpha_0$, где ε — достаточно малое положительное число. В силу непрерывности первого аргумента по τ_0 выполняется неравенство

$$1 - \frac{\tilde{\tau}_0}{2} + \frac{\tilde{\tau}_0}{2} \sqrt{1 - \frac{4\gamma}{\tilde{\alpha}_0}} < 1 - \frac{\tau_0}{2} + \frac{\tau_0}{2} \sqrt{1 - \frac{4\gamma}{\alpha_0}}.$$

Пусть теперь максимум достигается на втором выражении. Положим $\tilde{\tau}_0 = \tau_0$, $\tilde{\alpha}_0 = \alpha_0 + \varepsilon$, где ε — достаточно малое положительное число. В силу непрерывности второго аргумента по α_0 выполняется неравенство

$$\sqrt{\tilde{\tau}_0^2 \frac{\Gamma}{\tilde{\alpha}_0} - \tilde{\tau}_0 + 1} < \sqrt{\tau_0^2 \frac{\Gamma}{\alpha_0} - \tau_0 + 1}.$$

Осталось рассмотреть случай, когда максимум достигается на двух аргументах одновременно. Покажем, что в этом случае также возможно уменьшить целевую функцию. Действительно, из равенств

$$q_0 = 1 - \frac{\tau_0}{2} + \frac{\tau_0}{2} \sqrt{1 - \frac{4\gamma}{\alpha_0}} = \sqrt{\tau_0^2 \frac{\Gamma}{\alpha_0} - \tau_0 + 1}$$

следует

$$\tau_0 = \frac{2(1 - q_0)}{1 - \sqrt{1 - 4\gamma/\alpha_0}}, \quad q_0 = \frac{2(\Gamma - \gamma)}{2(\Gamma + \gamma) - \alpha_0(1 - \sqrt{1 - 4\gamma/\alpha_0})}.$$

Рассмотрим функцию

$$\varphi(\alpha) = \frac{2(\Gamma - \gamma)}{2(\Gamma + \gamma) - \alpha(1 - \sqrt{1 - 4\gamma/\alpha})},$$

обладающую свойством $\varphi(\alpha_0) = q_0$, и вычислим ее производную

$$\varphi'(\alpha) = \frac{-2(\Gamma - \gamma)(-1 + \sqrt{1 - 4\gamma/\alpha} + 2\gamma/\alpha\sqrt{1 - 4\gamma/\alpha})}{2(\Gamma + \gamma) - \alpha(1 - \sqrt{1 - 4\gamma/\alpha})}.$$

В условиях рассматриваемого случая $\varphi'(\alpha)$ не имеет нулей в окрестности точки α_0 , поэтому значение $\varphi(\alpha_0) = q_0$ может быть уменьшено за счет выбора α . ■

Продолжим анализ необходимых условий оптимальности.

Лемма 2.2.4. Для решения задачи (2.12) необходимо, чтобы выполнялось равенство

$$q_0 = \lambda(\gamma, \tau_0, \alpha_0).$$

Доказательство. Покажем, как и в предыдущем случае, что если утверждение леммы не выполнено, то значение целевой функции можно уменьшить. Пусть $q_0 \neq \lambda(\gamma, \tau_0, \alpha_0)$, тогда

$$q_0 = \max\{|\lambda_1(\tau_0)|, \lambda(\Gamma, \tau_0, \alpha_0)\}.$$

Используя лемму 2.2.2, это можно записать в явном виде

$$q_0 = \max\left\{|1 - \tau_0|, \sqrt{\tau_0^2 \frac{\Gamma}{\alpha_0} - \tau_0 + 1}\right\}.$$

Если максимум достигается на одном из аргументов, то, как в лемме 2.2.3, используя непрерывность аргументов по соответствующим параметрам в некоторых окрестностях α_0 и τ_0 , показываем, что можно уменьшить значение целевой функции. Пусть теперь максимум реализуется на равных аргументах. Тогда справедливо равенство

$$q_0 = |1 - \tau_0| = \sqrt{\tau_0^2 \frac{\Gamma}{\alpha_0} - \tau_0 + 1}.$$

В случае $\tau_0 \leq 1$ имеем

$$\tau_0 = 1 - q_0, \quad q_0 = \frac{\Gamma}{\Gamma - \alpha_0}.$$

Рассмотрим функцию $\varphi(\alpha) = \Gamma/(\Gamma - \alpha)$ такую, что $\varphi(\alpha_0) = q_0$. Она монотонно возрастает.

В случае $\tau_0 > 1$ имеем

$$\tau_0 = q_0 + 1, \quad q_0 = -\frac{\Gamma}{\Gamma - \alpha_0}.$$

Соответствующая функция $\varphi(\alpha) = -\Gamma/(\Gamma - \alpha)$ монотонно убывает. Таким образом, в обоих случаях за счет выбора α можно уменьшить значение целевой функции. ■

Докажем еще одно утверждение. Справедлива

Лемма 2.2.5. Для решения задачи (2.12) должно выполняться равенство

$$q_0 = \lambda(\Gamma, \tau_0, \alpha_0).$$

Доказательство. Будем проводить доказательство аналогично доказательству лемм 2.2.3 и 2.2.4. Пусть

$$q_0 \neq \lambda(\Gamma, \tau_0, \alpha_0).$$

Тогда

$$q_0 = \max\{|\lambda_1(\tau_0)|, \lambda(\gamma, \tau_0, \alpha_0)\}.$$

На основании лемм 2.2.3 и 2.2.4 это эквивалентно равенству

$$q_0 = |\lambda_1(\tau_0)| = \lambda(\gamma, \tau_0, \alpha_0).$$

Теперь из леммы 2.2.2 следует

$$q_0 = |1 - \tau_0| = 1 - \frac{\tau_0}{2} + \frac{\tau_0}{2} \sqrt{1 - \frac{4\gamma}{\alpha_0}}.$$

Откуда имеем

$$\tau_0 = \frac{2(1 - q_0)}{1 - \sqrt{1 - 4\gamma/\alpha_0}}, \quad q_0 = \frac{1 + \sqrt{1 - 4\gamma/\alpha_0}}{3 - \sqrt{1 - 4\gamma/\alpha_0}}.$$

Рассмотрим функцию

$$\varphi(\alpha) = \frac{1 + \sqrt{1 - 4\gamma/\alpha}}{3 - \sqrt{1 - 4\gamma/\alpha}},$$

обладающую свойством $\varphi(\alpha_0) = q_0$, и вычислим ее производную

$$\varphi'(\alpha) = \frac{4\gamma}{(3 - \sqrt{1 - 4\gamma/\alpha})^2 \alpha^2 \sqrt{1 - 4\gamma/\alpha}}.$$

Монотонность $\varphi'(\alpha)$ в окрестности точки α_0 позволяет за счет выбора α уменьшить значение целевой функции. Получаем противоречие. ■

Теперь перейдем непосредственно к решению задачи (2.12). Имеет место

Теорема 2.2.3. Решение задачи (2.12) имеет следующий вид:

$$q_0 = \frac{2 - \xi}{2 + \xi}, \quad \tau_0 = \frac{4}{2 + \xi}, \quad \alpha_0 = \frac{4\Gamma}{2 - \xi},$$

Доказательство. Леммы 2.2.3, 2.2.4 и 2.2.5 порождают необходимые условия оптимальности:

$$q_0 = |\lambda_1(\tau_0)| = \lambda(\gamma, \tau_0, \alpha_0) = \lambda(\Gamma, \tau_0, \alpha_0),$$

что эквивалентно системе трех нелинейных уравнений с тремя неизвестными.

Из уравнения $\lambda(\Gamma, \tau_0, \alpha_0) = |\lambda_1(\tau_0)|$ следует

$$(1 - \tau_0)^2 = \frac{\tau_0^2 \Gamma}{\alpha_0} - \tau_0 + 1,$$

откуда, учитывая положительность τ_0 , находим

$$\tau_0 = \frac{\alpha_0}{\alpha_0 - \Gamma}.$$

Из уравнения $\lambda(\gamma, \tau_0, \alpha_0) = |\lambda(\Gamma, \tau_0, \alpha_0)|$ и предыдущей формулы получаем

$$\alpha_0 = \frac{4\Gamma}{2 - \xi}, \quad \tau_0 = \frac{4}{2 + \xi},$$

где $\xi = \gamma/\Gamma$. Используя явное выражение для τ_0 , завершаем доказательство:

$$|\lambda_1(\tau_0)| = q_0 = \frac{2 - \xi}{2 + \xi}. \quad \blacksquare$$

2.2.5. Оптимизация в подпространстве

Обозначим погрешность решения на k -й итерации через

$$y^k = \{v^k, r^k\} = \{u^k - u, p^k - p\},$$

где $\{u, p\}$ — точное решение задачи $L_0 z = F$, и выберем начальное приближение $z^0 = \{u^0, p^0\}$ из условия (1.9)

$$Au^0 + Bp^0 = f.$$

Для такого начального приближения, в силу леммы 1.3.2, первая компонента v^k погрешности y^k итерационного метода (2.9) на любой итерации k является элементом подпространства G (т. е. $(Av^k, h) = 0$ для $\forall h \in H$). Это приводит к тому, что для определения асимптотически оптимальных параметров в методе (2.9), стартующего с начального приближения (1.9), достаточно рассмотреть следующую задачу: найти положительные значения τ_0 и α_0 , доставляющие минимум функции \tilde{q} :

$$\tilde{q} = \max_{t \in [\gamma, \Gamma]} \left\{ \left| 1 - \frac{\tau}{2} \pm \tau/2 \sqrt{1 - 4 \frac{t}{\alpha}} \right| \right\}. \quad (2.13)$$

Имеет место

Теорема 2.2.4. При выборе начального приближения, удовлетворяющего условию (1.9), спектральный радиус \tilde{q}_0 оператора перехода и асимптотически оптимальные параметры τ_0, α_0 в итерационном методе (2.9) определяются по формулам

$$\tilde{q}_0 = \sqrt{\frac{1-\xi}{1+\xi}}, \quad \tau_0 = 2, \quad \alpha_0 = 2(\gamma + \Gamma),$$

где $\xi = \gamma/\Gamma$.

Доказательство. Рассмотрим вспомогательную задачу для дискриминанта в (2.13)

$$\bar{q} = \min_{\alpha} \max_{t \in [\gamma, \Gamma]} |1 - 4t/\alpha|.$$

Ее решение ([15], с. 278) имеет вид

$$\bar{q} = \tilde{q}_0^2 = \frac{1-\xi}{1+\xi}, \quad \alpha_0 = 2(\gamma + \Gamma), \quad \xi = \gamma/\Gamma.$$

Теперь на отрезке $\gamma \leq t \leq (\gamma + \Gamma)/2$ дискриминант при $\alpha = \alpha_0$ неотрицателен, и поэтому на нем можно рассмотреть «половинку» задачи (2.13) в виде

$$q_1 = \min_{\tau} \max_{s \in [-\tilde{q}_0, \tilde{q}_0]} \left| 1 - \frac{\tau}{2} + \frac{\tau}{2}s \right|.$$

Используя линейность по τ , несложно получить решение этой подзадачи: $q_1 = \tilde{q}_0$, $\tau = 2$. Для завершения доказательства достаточно заметить, что на отрезке $(\gamma + \Gamma)/2 \leq t \leq \Gamma$ величина

$$\max_{s \in [-\tilde{q}_0, \tilde{q}_0]} \left| 1 - \frac{\tau}{2} + i \frac{\tau}{2}s \right|$$

при $\tau = 2$ не превосходит \tilde{q}_0 . ■

Выясним, насколько улучшается сходимость метода (2.9) при специальном выборе начального приближения. Для этого разложим функции q_0 и \tilde{q}_0 в ряд Тейлора в нуле:

$$q_0 = \frac{1-\xi/2}{1+\xi/2} = 1 - \xi + \frac{1}{2}\xi^2 - \frac{1}{4}\xi^3 + O(\xi^4),$$

$$\tilde{q}_0 = \sqrt{\frac{1-\xi}{1+\xi}} = 1 - \xi + \frac{1}{2}\xi^2 - \frac{5}{9}\xi^3 + O(\xi^4).$$

Полученные выражения означают их совпадение с точностью до величин третьего порядка малости, т. е.

$$q_0 = \tilde{q}_0 + O(\xi^3).$$

Отсюда следует, что влияние специального начального приближения, т. е. проектирования начальной ошибки на некоторое подпространство, оказывает слабое влияние на асимптотическую скорость сходимости алгоритма при правильном выборе параметров.

2.3. МОДИФИЦИРОВАННЫЙ МЕТОД SOR (MSOR)

Рассмотрим модифицированный метод SOR (метод MSOR) для системы $L_0 z = F$

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau} + Au^k + Bp^k = f, \\ -\alpha C \frac{p^{k+1} - p^k}{\tau} + B^T u^{k+1} = \varphi. \end{cases} \quad (2.14)$$

Его построение может быть осуществлено, как в предыдущем разделе, поэтому сразу перейдем к анализу алгоритма.

2.3.1. Спектр оператора перехода

Обозначим через T оператор перехода в алгоритме (2.14) и рассмотрим спектральную задачу $Tz = \lambda z$:

$$Tz \equiv \begin{pmatrix} (1-\tau)I & -\tau A^{-1}B \\ \frac{\tau(1-\tau)}{\alpha} C^{-1}B^T & I - \frac{\tau^2}{\alpha} C^{-1}S_0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} u \\ p \end{pmatrix}. \quad (2.15)$$

Имеет место

Теорема 2.3.1. *Спектр $\sigma(T)$ оператора перехода T в методе (2.14) принадлежит множеству*

$$\Lambda = \{1 - \tau\} \cup \left\{ 1 - \tau\theta \pm \tau\sqrt{\theta^2 - \frac{t}{\alpha}}, \theta = \frac{1 + \tau t/\alpha}{2}, t \in [\gamma, \Gamma] \right\}.$$

Доказательство. Проведем доказательство аналогично теореме 2.2.1. Для этого используем базис пространства Z , построенный в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (2.15) с $\lambda_k^{(1)} = 1 - \tau$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^T g_j\}.$$

После применения к первому уравнению (2.15) матрицы $C^{-1}B^T$ и замены $C^{-1}B^T g_j = p_j$ будем иметь

$$\begin{cases} (1 - \tau)p_j + \tau\kappa^{-1}C^{-1}S_0p_j = \lambda p_j, \\ \frac{\tau(\tau - 1)}{\alpha}p_j + \kappa^{-1}\left(I - \frac{\tau^2}{\alpha}C^{-1}S_0\right)p_j = \lambda\kappa^{-1}p_j. \end{cases}$$

Перепишем эту систему в виде

$$\begin{cases} S_0p_j = \frac{\kappa(\lambda - 1 + \tau)}{\tau}Cp_j, \\ S_0p_j = \frac{\alpha(1 - \lambda) - \kappa\tau(1 - \tau)}{\tau^2}Cp_j. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0p = tCp$ соответствует собственное значение t_j , $j = 1, \dots, N_p$ по теореме 1.3.1. Зафиксировав его, из полученной системы для λ и κ выведем соотношения

$$t_j = \frac{\kappa(\lambda - 1 + \tau)}{\tau} = \frac{\alpha(1 - \lambda) - \kappa\tau(1 - \tau)}{\tau^2}.$$

Исключая из этих уравнений κ , приходим к выражению

$$\lambda_j^{(2,3)} = 1 - \tau\theta_j \pm \tau\sqrt{\theta_j^2 - \frac{t_j}{\alpha}},$$

где $\theta_j = (1 + t_j\tau/\alpha)/2$ и, соответственно,

$$\kappa_j^{(2,3)} = \frac{\alpha}{\lambda_j^{(2,3)}} \left(\theta_j \mp \sqrt{\theta_j^2 - \frac{t_j}{\alpha}} \right).$$

Завершение доказательства проводится аналогично теореме 2.2.1. Отличие состоит только в другой формуле для корневого вектора $z_j^{(3)}$ при $\lambda_j^{(2)} = \lambda_j^{(3)} = \lambda_j$:

$$z_j^{(3)} = \left\{ g_j, -p_j \frac{\lambda_j + \tau}{t_j\tau} \right\}. \quad \blacksquare$$

2.3.2. Условие сходимости

Рассмотрим условия сходимости метода (2.14). Справедлива

Теорема 2.3.2. При любом $\alpha > 0$ и произвольном начальном приближении $z^0 \in Z$ необходимым и достаточным условием сходимости метода (2.14) является выполнение неравенства

$$0 < \tau < \sqrt{\frac{\alpha^2}{\Gamma^2} + 4\frac{\alpha}{\Gamma}} - \frac{\alpha}{\Gamma}. \quad (2.16)$$

Доказательство. Введем следующие обозначения для элементов множества Λ :

$$\lambda_1 = 1 - \tau,$$

$$\lambda_{2,3} = 1 - \tau\theta \pm \tau\sqrt{\theta^2 - \frac{t}{\alpha}}.$$

Знак «+» относится к λ_2 .

Достаточность. Рассмотрим некоторую фиксированную точку $t \in [\gamma, \Gamma]$ и выясним соотношение между параметрами τ и α , при котором $|\lambda_{2,3}| < 1$.

Изучим сначала случай различных вещественных значений $\lambda_{2,3}$, т. е. когда $\theta^2 - t/\alpha > 0$. Так как $\lambda_2 > \lambda_3$, достаточно исследовать неравенства

$$-1 < \lambda_3 < \lambda_2 < 1.$$

Условие $\lambda_2 < 1$ выполнено всегда, так как $\sqrt{\theta^2 - t/\alpha} < \theta$. Несложные преобразования неравенства $\lambda_3 > -1$ приводят к выражению

$$0 < \tau < \sqrt{\frac{\alpha^2}{t^2} + 4\frac{\alpha}{t}} - \frac{\alpha}{t},$$

из монотонности по t правой части которого следует неравенство (2.16).

Ограничение $|\lambda_1| < 1$ дает $0 < \tau < 2$. Поскольку правая часть (2.16) монотонно возрастает по α и ограничена величиной 2, получим, что условие (2.16) гарантирует выполнение неравенства

$$\max_t |\lambda_{1,2,3}| < 1$$

в случае различных вещественных $\lambda_{2,3}$.

Рассмотрим далее случай комплексных (или кратных) значений $\lambda_{2,3}$ при некотором $t \in [\gamma, \Gamma]$, т. е. когда $\theta^2 - t/\alpha \leq 0$. При этом $|\lambda_{2,3}|^2 = 1 - \tau$, откуда в силу положительности параметра τ следует, что комплексные и кратные значения $\lambda \in \Lambda$ всегда лежат внутри единичного круга.

Завершение доказательства достаточности следует из объединения двух рассмотренных выше случаев.

Необходимость. Доказательство проведем от противного. Покажем, что даже для единственной точки отрезка $[\gamma, \Gamma]$, а именно при $t = \Gamma$, невыполнение (2.16) приводит к неравенству $|\lambda_3| \geq 1$. Пусть

$$\tau = \sqrt{\frac{\alpha^2}{\Gamma^2} + 4\frac{\alpha}{\Gamma}} - \frac{\alpha}{\Gamma} + \delta\frac{\alpha}{\Gamma}, \quad \delta \geq 0.$$

При этом λ_2 и λ_3 будут вещественны, так как дискриминант, равный $\theta^2 - \Gamma/\alpha$, положителен для любого $\delta \geq 0$ (он строго больше $1/4$). Далее рассмотрим изменение определяющей величины

$$\lambda_3 = 1 - \tau\theta - \tau\sqrt{\theta^2 - \frac{\Gamma}{\alpha}}$$

в зависимости от параметра θ

$$\frac{\partial \lambda_3}{\partial \theta} = -\tau \left(1 + \frac{\theta}{\sqrt{\theta^2 - \Gamma/\alpha}} \right) < 0.$$

В свою очередь, $\partial\theta/\partial\delta > 0$, и при $\delta = 0$ имеем $\lambda_3 = -1$. Последние два неравенства для производных приводят к условию $\lambda_3 \leq -1$ при $\delta \geq 0$.

Напомним, что $t = \Gamma$ является точным собственным значением задачи $S_0p = tCp$ (см. раздел 1.3) и, следовательно, λ_3 при $t = \Gamma$ — собственным значением оператора T в (2.15). Отсюда следует, что при невыполнении условия (2.16) существует собственный вектор оператора T такой, что отвечающее ему собственное значение λ_3 по модулю не меньше единицы. Теперь на основании теоремы о необходимом и достаточном условии сходимости метода простой итерации (см., например, [15], с. 269) получаем, что выполнение неравенства (2.16) является необходимым и достаточным для сходимости метода (2.14). ■

2.3.3. Задача асимптотической оптимизации

Для определения асимптотически оптимальных параметров в методе (2.14) рассмотрим следующую задачу: *найти положительные значения τ_0 и α_0 , доставляющие минимум функции*

$$q = \max_{t \in [\gamma, \Gamma]} \left\{ \left| 1 - \tau \right|, \left| 1 - \tau\theta \pm \tau\sqrt{\theta^2 - \frac{t}{\alpha}} \right| \right\}, \quad (2.17)$$

где, как и ранее, $\theta = (1 + \tau t/\alpha)/2$. Задачу асимптотической оптимизации метода (2.14) решает

Теорема 2.3.3. *Спектральный радиус q_0 оператора перехода и асимптотически оптимальные параметры τ_0, α_0 в итерационном методе (2.14) определяются по формулам*

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \tau_0 = \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}, \quad \alpha_0 = \frac{4\gamma}{(1 + \sqrt{\xi})^2},$$

где $\xi = \gamma/\Gamma$.

Доказательство. Рассмотрим выражение (2.17). Вторым аргументом в процедуре \max — это модуль корней уравнения

$$\lambda^2 - \lambda \left(2 - \tau - \tau^2 \frac{t}{\alpha} \right) + 1 - \tau = 0. \quad (2.18)$$

Пусть $0 < \tau < 1$. Тогда максимальный по модулю корень уравнения (2.18) удовлетворяет оценке

$$\max |\lambda| \geq \sqrt{1 - \tau},$$

причем равенство достигается в случае комплексных или кратных корней. Изучим подробнее этот предельный случай неположительного дискриминанта (2.18) для всех $t \in [\gamma, \Gamma]$, т. е. когда

$$\theta^2 - \frac{t}{\alpha} \leq 0.$$

Преобразования последнего неравенства приводят к выражению

$$\tau \leq \min \left\{ 2\sqrt{\frac{\alpha}{\Gamma}} - \frac{\alpha}{\Gamma}, 2\sqrt{\frac{\alpha}{\gamma}} - \frac{\alpha}{\gamma} \right\},$$

откуда максимально возможное значение τ достигается при

$$\alpha_0 = \frac{4\gamma}{(1 + \sqrt{\xi})^2}$$

и равно

$$\tau_0 = \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}.$$

При этом все корни уравнения (2.18) лежат на окружности

$$|\lambda| = q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Неулучшаемость полученной оценки при $0 < \tau < 1$ следует из наличия пары двукратных вещественных корней и непрерывности корней многочлена в зависимости от коэффициентов. Действительно, при $\tau > \tau_0$ немедленно получаем $\max |\lambda| > q_0$, причем максимум достигается на одном из вещественных корней.

Рассмотрим далее значение $\tau = 1$. Максимальный по модулю корень (2.18) имеет вид $\lambda = 1 - t/\alpha$, и решение задачи (см. [15], с. 278)

$$q_1 = \min_{\alpha} \max_{t \in [\gamma, \Gamma]} \left| 1 - \frac{t}{\alpha} \right| = \frac{1 - \xi}{1 + \xi}$$

для любых $\gamma < \Gamma$ больше, чем q_0 .

Осталось рассмотреть интервал $\tau \in (1, 2)$, поскольку ограничение $\tau < 2$ следует из явного вида первого аргумента в процедуре \max

в (2.17). В этом случае при любых положительных α и τ корни (2.18) вещественны и имеют различные знаки, что дает возможность записать (2.17) в следующем виде:

$$\begin{aligned} q_2 &= \max_t \left\{ \tau - 1, \tau \sqrt{\theta^2 - \frac{t}{\alpha}} + 1 - \tau\theta, \tau \sqrt{\theta^2 - \frac{t}{\alpha}} + \tau\theta - 1 \right\} = \\ &= \max_t \left\{ \tau - 1, \tau \sqrt{\theta^2 - \frac{t}{\alpha}} + |1 - \tau\theta| \right\}. \end{aligned}$$

Теперь приведем цепочку неравенств:

$$\min_{\alpha} q_2 = \min_{\alpha} \max_t \left\{ \tau - 1, \tau \sqrt{\theta^2 - \frac{t}{\alpha}} + |1 - \tau\theta| \right\} >$$

(в первом слагаемом под знаком радикала учтем, что $\tau > 1$)

$$> \min_{\alpha} \max_t \left\{ \frac{\tau}{2} \left| 1 - \frac{t}{\alpha} \right| + \left| 1 - \frac{\tau}{2} - \frac{\tau^2}{2\alpha} t \right| \right\} =$$

(во втором слагаемом вынесем положительный множитель за знак модуля)

$$= \min_{\alpha} \max_t \left\{ \frac{\tau}{2} \left| 1 - \frac{t}{\alpha} \right| + \left(1 - \frac{\tau}{2} \right) \left| 1 - \frac{t}{\alpha} \frac{\tau^2}{2(1 - \tau/2)} \right| \right\} \geq$$

(расширим область минимизации, введя во втором слагаемом вместо α параметр κ)

$$\begin{aligned} &\geq \min_{\alpha, \kappa} \max_t \left\{ \frac{\tau}{2} \left| 1 - \frac{t}{\alpha} \right| + \left(1 - \frac{\tau}{2} \right) \left| 1 - \frac{t}{\kappa} \frac{\tau^2}{2(1 - \tau/2)} \right| \right\} = \\ &= \min_{\alpha} \max_t \left\{ \frac{\tau}{2} \left| 1 - \frac{t}{\alpha} \right| \right\} + \min_{\kappa} \max_t \left\{ \left(1 - \frac{\tau}{2} \right) \left| 1 - \frac{t}{\kappa} \frac{\tau^2}{2(1 - \tau/2)} \right| \right\} = \\ &= \frac{\tau}{2} \frac{1 - \xi}{1 + \xi} + \left(1 - \frac{\tau}{2} \right) \frac{1 - \xi}{1 + \xi} = \frac{1 - \xi}{1 + \xi}. \end{aligned}$$

Таким образом, при $\tau \in (1, 2)$ имеем $q_2 > q_1$. ■

2.4. МОДИФИЦИРОВАННЫЙ МЕТОД SSOR (MSSOR)

Рассмотрим модифицированный метод SSOR (метод MSSOR) для системы $L_0 z = F$:

$$\begin{cases} A \frac{u^{k+1/2} - u^k}{\tau} + Au^k + Bp^k = f, \\ -\alpha C \frac{p^{k+1} - p^k}{2\tau} + B^T u^{k+1/2} = \varphi, \\ A \frac{u^{k+1} - u^{k+1/2}}{\tau} + Au^{k+1/2} + Bp^{k+1} = f. \end{cases} \quad (2.19)$$

Его построение может быть осуществлено, как и выше. Отметим лишь, что здесь требуется обращаться матрицу C один раз, а не два, как в (2.8) (в силу особенности структуры матрицы исходной системы).

2.4.1. Спектр оператора перехода

Обозначим через T оператор перехода в алгоритме (2.19), который имеет вид

$$\begin{pmatrix} ((1-\tau)^2 I - \frac{2\tau^2(1-\tau)}{\alpha} A^{-1} B_0 & \tau A^{-1} B \left(\frac{2\tau^2}{\alpha} C^{-1} S_0 - (2-\tau) I \right) \\ \frac{2\tau(1-\tau)}{\alpha} C^{-1} B^T & I - \frac{2\tau^2}{\alpha} C^{-1} S_0 \end{pmatrix},$$

и рассмотрим спектральную задачу

$$Tz = \lambda z. \quad (2.20)$$

Имеет место

Теорема 2.4.1. Спектр $\sigma(T)$ оператора перехода T в методе (2.19) принадлежит множеству

$$\Lambda = \{(1-\tau)^2\} \cup \{1 - \tau(2-\tau)\theta \pm \tau(2-\tau)\sqrt{\theta^2 - \frac{2t}{\alpha(2-\tau)}}\},$$

$$\theta = \frac{\alpha + 2\tau t}{2\alpha}, \quad t \in [\gamma, \Gamma].$$

Доказательство. Получим формулы собственных значений оператора T с помощью базиса пространства Z , построенного в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (2.20) с $\lambda_k^{(1)} = (1-\tau)^2$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\kappa^{-1} C^{-1} B^T g_j\}.$$

После применения к первому уравнению (2.20) матрицы $C^{-1} B^T$ и замены $C^{-1} B^T g_j = p_j$ будем иметь

$$\begin{cases} \lambda p_j = (1-\tau)^2 p_j - \frac{2\tau^2(1-\tau)}{\alpha} C^{-1} S_0 p_j + \\ \quad + \tau \kappa^{-1} \left[(2-\tau) I - \frac{2\tau^2}{\alpha} C^{-1} S_0 \right] C^{-1} S_0 p_j, \\ \lambda p_j = -\kappa \frac{2\tau^2(1-\tau)}{\alpha} p_j + \left(I - \frac{2\tau}{\alpha} C^{-1} S_0 \right) p_j. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0 p = t C p$ соответствует собственное значение t_j , $j = 1, \dots, N_p$ по теореме 1.3.1. Поэтому при фиксированном t_j из полученной системы для λ и κ следуют соотношения

$$\begin{cases} \lambda = (1 - \tau)^2 - \frac{2\tau^2(1 - \tau)}{\alpha} t_j + \tau \kappa^{-1} \left[(2 - \tau) - \frac{2\tau^2}{\alpha} t_j \right] t_j, \\ \lambda = -\kappa \frac{2\tau(1 - \tau)}{\alpha} + 1 - \frac{2\tau^2}{\alpha} t_j. \end{cases}$$

Исключая из этих уравнений κ , приходим к выражению

$$\lambda_j^{(2,3)} = 1 - \tau(2 - \tau)\theta_j \pm \tau(2 - \tau) \sqrt{\theta_j^2 - \frac{2t_j}{\alpha(2 - \tau)}},$$

где

$$\theta_j = \frac{\alpha + 2\tau t_j}{2\alpha}, \quad t_j \in [\gamma, \Gamma],$$

и, соответственно,

$$\kappa_j^{(2,3)} = \frac{1 - 2\tau^2 t_j / \alpha - \lambda_j^{(2,3)}}{2\tau^2(1 - \tau) / \alpha}.$$

Покажем, что указанным способом найдены все собственные значения задачи (2.20).

Пусть $\tau \neq 1$. Рассмотрим вначале простой случай — случай различных собственных значений $\lambda_j^{(2,3)}$ при фиксированном j . Пусть параметры α и τ таковы, что выражение

$$\theta_j^2 - \frac{2t_j}{\alpha(2 - \tau)}$$

не обращается в нуль ни при каком значении t_j . Тогда величины $\kappa_j^{(2)}$ и $\kappa_j^{(3)}$ для любого j также будут различны, что следует из их явного представления. Это означает, что найденная система собственных векторов $\{z_i^{(1,2,3)}\}_{i=1}^{N_u + N_p}$ задачи (2.20) удовлетворяет теореме 1.3.2, следовательно, все искомые собственные значения найдены.

Обобщим доказательство на случай кратных собственных значений, т. е. обращения в нуль выражения

$$\theta_j^2 - \frac{2t_j}{\alpha(2 - \tau)}$$

при некотором t_j . Сразу необходимо отметить, что при любых фиксированных α и τ таких различных значений t_j — не более двух, причем даже одинаковым t_j по теореме 1.3.1 соответствуют

различные p_j . Пусть в рассматриваемом случае $\lambda_j^{(2)} = \lambda_j^{(3)} = \lambda_j$ имеется только один собственный вектор $z_j^{(2)}$ оператора T с первой компонентой g_j . Тогда для построения полной в Z системы функций достаточно добавить к нему в пару корневой вектор $z_j^{(3)}$ высоты два следующего вида:

$$z_j^{(3)} = \left\{ g_j, -p_j \frac{1 - (1 - \tau)^2 + \lambda + \frac{2\tau^2(1-\tau)}{\alpha} t_j}{t_j (2 - \tau - \frac{2\tau^2}{\alpha} t_j)} \right\},$$

удовлетворяющий уравнению

$$(T - \lambda_j I)^2 z_j^{(3)} = 0$$

и линейно независимый с соответствующим собственным. Теперь уточненная система собственных векторов

$$\{z_i^{(1,2,3)}\}_{i=1}^{N_u + N_p}$$

удовлетворяет теореме 1.3.2.

Проведенные рассуждения, как это легко видеть из явной формулы для $\kappa_j^{(2,3)}$, некорректны при $\tau = 1$, поэтому данный случай рассмотрим отдельно. Здесь соотношения для λ и κ имеют вид

$$\begin{cases} \lambda = \kappa^{-1} t_j \left(1 - \frac{2}{\alpha} t_j \right), \\ \lambda = 1 - \frac{2}{\alpha} t_j. \end{cases}$$

Отсюда следует, что каждому значению λ_j в нормальной жордановой форме оператора T соответствует клетка второго порядка. Поэтому для построения канонического базиса достаточно для каждого λ_j построить в пару к собственному $z_j^{(2)}$ корневой вектор $z_j^{(3)}$ высоты два, удовлетворяющий уравнению

$$(T - \lambda_j I)^2 z_j^{(3)} = 0.$$

Теперь, аналогично предыдущему случаю, дополненная система собственных векторов

$$\{z_i^{(1,2,3)}\}_{i=1}^{N_u + N_p}$$

будет удовлетворять теореме 1.3.2.

Таким образом, полнота найденной системы векторов в пространстве Z , состоящей либо только из собственных векторов оператора T , либо с добавлением к ним корневых по указанной схеме, дает основание утверждать, что других собственных значений в задаче (2.20) (отличных от найденных) не существует. ■

2.4.2. Условие сходимости

Полученное представление спектра оператора перехода T в методе (2.19) дает возможность определить условия сходимости метода. Вначале нам потребуется

Лемма 2.4.6. *Нелинейная замена итерационных параметров*

$$\nu = \tau(2 - \tau), \quad \delta = \alpha \frac{\nu}{2\tau}$$

порождает следующую параметризацию спектра $\sigma(T)$ оператора перехода T в методе (2.19):

$$\Lambda = \{1 - \nu\} \cup \left\{ 1 - \nu\theta \pm \nu \sqrt{\theta^2 - \frac{t}{\delta}}, \theta = \frac{1 + \nu t/\delta}{2}, t \in [\gamma, \Gamma] \right\}.$$

Доказательство. К искомому результату приводят две последовательные подстановки. Сначала выражение $\tau(2 - \tau)$ в утверждении теоремы 2.4.1 меняем на ν и, соответственно, $(2 - \tau)$ — на ν/τ , а затем используем подстановку

$$\alpha \frac{\nu}{2\tau} = \delta. \quad \blacksquare$$

Смысл этих формальных преобразований заключается в сведении интересующих задач к уже решенным. Справедлива

Теорема 2.4.2. *При любом $\delta > 0$ и произвольном начальном приближении $z^0 \in Z$ необходимым и достаточным условием сходимости метода (2.19) является выполнение неравенства*

$$0 < \nu < \sqrt{\frac{\delta^2}{\Gamma^2} + 4\frac{\delta}{\Gamma}} - \frac{\delta}{\Gamma},$$

где новые параметры ν, δ определяются по формулам:

$$\nu = \tau(2 - \tau), \quad \delta = \alpha \frac{\nu}{2\tau}.$$

Доказательство. Построенная в лемме 2.4.6 параметризация спектра оператора перехода и теорема 2.3.2 о принадлежности единичному кругу произвольной точки множества Λ при условии (2.16) гарантируют справедливость сформулированного утверждения. \blacksquare

2.4.3. Задача асимптотической оптимизации

Знание аналитического представления спектра оператора перехода позволяет сформулировать и решить задачу асимптотической оптимизации метода: *найти положительные значения τ_0 и α_0 ,*

минимизирующие спектральный радиус оператора перехода. Имеет место

Теорема 2.4.3. Спектральный радиус q_0 оператора перехода и асимптотически оптимальные параметры τ_0, α_0 в итерационном методе (2.19) определяются по формулам

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \tau_0^\pm = 1 \pm q_0, \quad \alpha_0^\pm = 2\tau_0^\pm \frac{\delta_0}{\nu_0},$$

где

$$\nu_0 = \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}, \quad \delta_0 = \frac{4\gamma}{(1 + \sqrt{\xi})^2}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Доказательство. Как и в предыдущем случае, ключевую роль играют утверждения леммы 2.4.6 и теоремы 2.3.3, решающей поставленную оптимизационную задачу в терминах « $\nu - \delta$ ». Поэтому для нахождения выражений для оптимальных параметров достаточно выполнить преобразование перехода к исходным параметрам τ, α . ■

Отсюда следует, что существует ровно два набора исходных итерационных параметров, которые обеспечивают наивысшую асимптотическую скорость сходимости рассматриваемого алгоритма, совпадающую с асимптотической скоростью сходимости метода MSOR.

2.5. ТРЕХПАРАМЕТРИЧЕСКИЙ МЕТОД (3MSOR)

Рассмотрим трехпараметрический метод типа MSOR для системы уравнений, равносильной системе $L_0 z = F$:

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau} + (A + \beta B_0)u^k + Bp^k = f + \beta BC^{-1}\varphi, \\ -\alpha C \frac{p^{k+1} - p^k}{\tau} + B^T u^{k+1} = \varphi. \end{cases} \quad (2.21)$$

Эта система получена прибавлением к обеим частям первого уравнения следствия второго: $\beta B_0 u = \beta BC^{-1}\varphi$, и содержит дополнительный свободный параметр β .

2.5.1. Спектр оператора перехода

Обозначим через T оператор перехода в алгоритме (2.21), который имеет следующий вид:

$$T = \begin{pmatrix} (1 - \tau)I - \tau\beta A^{-1}B_0 & -\tau A^{-1}B \\ \frac{\tau(1-\tau)}{\alpha} C^{-1}B^T - \frac{\tau^2\beta}{\alpha} C^{-1}S_0 C^{-1}B^T & I - \frac{\tau^2}{\alpha} C^{-1}S_0 \end{pmatrix},$$

и рассмотрим спектральную задачу

$$Tz = \lambda z. \quad (2.22)$$

Имеет место

Теорема 2.5.1. *Спектр $\sigma(T)$ оператора перехода T в методе (2.21) принадлежит множеству*

$$\Lambda = \{1 - \tau\} \cup \left\{ 1 - \tau\theta \pm \tau\sqrt{\theta^2 - \frac{t}{\alpha}} \right\},$$

где

$$\theta = \frac{1 + \beta t + \tau t/\alpha}{2}, \quad t \in [\gamma, \Gamma].$$

Доказательство этого утверждения может быть получено таким же способом, как были доказаны теоремы 2.2.1, 2.3.1 и 2.4.1 (см. [129]). Однако этот результат является частным случаем более общего утверждения из соответствующего раздела главы 7 второй части настоящей книги и потому здесь его обоснование опущено.

2.5.2. Задача асимптотической оптимизации

Знание аналитического представления спектра оператора перехода позволяет сформулировать и решить задачу асимптотической оптимизации метода: *найти положительные значения τ_0, α_0 и $\beta_0 \in \mathbb{R}$, минимизирующие спектральный радиус оператора перехода*

$$q = \max_{t \in [\gamma, \Gamma]} \left\{ |1 - \tau|, \left| 1 - \tau\theta \pm \tau\sqrt{\theta^2 - \frac{t}{\alpha}} \right| \right\}, \quad (2.23)$$

где

$$\theta = \frac{1}{2} \left(1 + \beta t + \frac{\tau t}{\alpha} \right).$$

Задачу асимптотической оптимизации метода (2.21) решает

Теорема 2.5.2. *Спектральный радиус q_0 оператора перехода и асимптотически оптимальные параметры $\tau_0, \alpha_0, \beta_0$ в итерационном методе (2.21) определяются по формулам*

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \tau_0 = \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}, \quad \alpha_0 = \frac{4\gamma}{(1 + \sqrt{\xi})^2}, \quad \beta_0 = 0,$$

где $\xi = \gamma/\Gamma$.

Доказательство этого утверждения требует привлечения технического аппарата из второй части настоящей книги (см. [129]). Кроме

того, оно является следствием решения более общей задачи оптимизации, рассмотренной в разделе 7.2. В силу указанных причин, вывод формул для оптимальных параметров рассматриваемого алгоритма здесь не приводится.

2.5.3. Частный случай: (β, τ) -метод

Рассмотрим исторически важный частный случай трехпараметрического алгоритма: (β, τ) — метод [48], следующий из формул (2.21) при формальной замене

$$\beta \longrightarrow \frac{1}{\beta}, \quad \alpha \longrightarrow \beta\tau, \quad \beta > 0.$$

Это дает

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau} + (A + \frac{1}{\beta} B_0) u^k + B p^k = f + \frac{1}{\beta} B C^{-1} B^T \varphi, \\ -\beta \tau C \frac{p^{k+1} - p^k}{\tau} + B^T u^{k+1} = \varphi. \end{cases} \quad (2.24)$$

В данном случае спектр $\sigma(T)$ оператора перехода T в методе (2.24) принадлежит множеству Λ :

$$\Lambda = \{1 - \tau\} \cup \left\{ 1 - \tau\theta \pm \tau \sqrt{\theta^2 - \frac{t}{\beta}} \tau, \quad \theta = \frac{1}{2} + \frac{t}{\beta}, \quad t \in [\gamma, \Gamma] \right\},$$

а основной результат о сходимости метода формулируется следующим образом:

Теорема 2.5.3. При любом $\beta > 0$ и произвольном начальном приближении $z^0 \in Z$ необходимым и достаточным условием сходимости метода (2.24) является выполнение неравенства

$$0 < \tau < \frac{4\beta}{2\beta + 3\Gamma}. \quad (2.25)$$

Доказательство. Введем следующие обозначения для элементов множества Λ :

$$\lambda_1 = 1 - \tau, \quad \lambda_{2,3} = 1 - \tau\theta \pm \tau \sqrt{\theta^2 - \frac{t}{\beta}} \tau.$$

Знак «+» относится к λ_2 .

Достаточность. Рассмотрим некоторую фиксированную точку $t \in [\gamma, \Gamma]$ и выясним соотношение между параметрами τ и β , при котором $|\lambda_{2,3}| < 1$.

Проанализируем сначала случай различных вещественных $\lambda_{2,3}$, когда дискриминант $\theta^2 - t/\beta\tau > 0$. Так как $\lambda_2 > \lambda_3$, достаточно исследовать неравенства $-1 < \lambda_3 < \lambda_2 < 1$. Условие $\lambda_2 < 1$ выполнено всегда, в силу

$$\sqrt{\theta^2 - \frac{t}{\beta}\tau} < \theta.$$

Несложные преобразования неравенства $\lambda_3 > -1$ приводят к выражению

$$0 < \tau < \frac{4\beta}{2\beta + 3t},$$

из монотонности по t правой части которого следует неравенство (2.25). Ограничение $|\lambda_1| < 1$ дает $\tau < 2$. Поскольку правая часть (2.25) монотонно возрастает по β и ограничена величиной 2, получим, что условие (2.25) гарантирует выполнение неравенства

$$\max_t |\lambda_{1,2,3}| < 1$$

в случае различных вещественных λ .

Рассмотрим далее случай комплексных (или кратных) значений λ при некотором $t \in [\gamma, \Gamma]$, когда справедливо

$$\theta^2 - \frac{t}{\beta}\tau \leq 0.$$

При этом

$$|\lambda_{2,3}|^2 = 1 - \tau - \tau \frac{t}{\beta},$$

откуда в силу положительности параметров τ, β и t следует, что комплексные и кратные значения $\lambda \in \Lambda$ всегда лежат внутри единичного круга.

Завершение доказательства достаточности следует из объединения двух рассмотренных выше случаев.

Необходимость. Доказательство проведем от противного. Покажем, что даже для единственной точки отрезка $[\gamma, \Gamma]$, а именно, $t = \Gamma$, невыполнение (2.25) приводит к неравенству $|\lambda_3| \geq 1$. Пусть

$$\tau = \delta \frac{4\beta}{2\beta + 3\Gamma}, \quad \delta \geq 1.$$

Отметим, что при этом λ_2 и λ_3 будут вещественны, так как дискриминант

$$\theta^2 - \frac{\Gamma}{\beta}\tau > \frac{1}{4}$$

для любого $\delta \geq 1$. Далее рассмотрим изменение определяющей величины

$$\lambda_3 = 1 - \tau\theta - \tau\sqrt{\theta^2 - \frac{t}{\beta}\tau}$$

в зависимости от параметра τ :

$$\frac{\partial \lambda_3}{\partial \tau} = - \left(\theta + \frac{1}{2} \frac{\theta^2 - t/2\beta\tau}{\sqrt{\theta^2 - t/\beta\tau}} \right) < 0.$$

В свою очередь, $\partial\tau/\partial\delta > 0$, и при $\delta = 1$ получим $\lambda_3 = -1$. Последние два неравенства для производных приводят к условию $\lambda_3 \leq -1$ при $\delta \geq 1$.

Напомним, что $t = \Gamma$ является точным собственным значением оператора задачи $S_0 p = t C p$ (см. раздел 1.3) и, следовательно, λ_3 при $t = \Gamma$ — собственным значением оператора T в (2.22). Отсюда следует, что при невыполнении условия (2.25) существует собственный вектор оператора T такой, что отвечающее ему собственное значение λ_3 по модулю не меньше единицы. Теперь на основании теоремы о необходимом и достаточном условии сходимости метода простой итерации (см., например, [15], с. 269) получаем, что выполнение неравенства (2.25) является необходимым и достаточным для сходимости метода (2.24). ■

2.5.4. Задача асимптотической оптимизации (β, τ) -метода

Знание аналитического представления спектра оператора перехода позволяет сформулировать и решить задачу асимптотической оптимизации метода: *найти положительные значения τ_0 и β_0 , минимизирующие спектральный радиус оператора перехода*

$$q = \max_{t \in [\gamma, \Gamma]} \left\{ |1 - \tau|, \left| 1 - \tau\theta \pm \tau\sqrt{\theta^2 - \frac{t}{\beta}\tau} \right| \right\}, \quad (2.26)$$

где $\theta = 1/2 + t/\beta$.

Задачу асимптотической оптимизации (β, τ) — метода решает

Теорема 2.5.4. *Спектральный радиус q_0 оператора перехода и асимптотически оптимальные параметры τ_0, β_0 в итерационном методе (2.24) определяются по формулам*

$$q_0 = \frac{\beta_0}{\beta_0 + 2\gamma}, \quad \tau_0 = \frac{4\gamma\beta_0}{(2\gamma + \beta_0)^2},$$

$$\beta_0 = \begin{cases} \frac{2\Gamma}{3} \sqrt{\xi(9 - 5\xi)} \cos \frac{\delta_0}{3} - \frac{2}{3}\gamma & \text{при } \xi < 1/3, \\ \frac{\Gamma + \gamma}{2} & \text{при } \xi \geq 1/3, \end{cases}$$

где

$$\xi = \frac{\gamma}{\Gamma}, \quad \delta_0 = \arccos \left(- \left[2 \sqrt{\frac{\xi}{9 - 5\xi}} \right]^3 \right).$$

Кроме того, имеет место асимптотическое представление

$$q_0 = 1 - \sqrt{\frac{4}{3}}\xi + O(\xi), \quad \xi \rightarrow 0.$$

Вывод этих формул требует привлечения технического аппарата из второй части настоящей книги и поэтому здесь не приводится, тем более что значение этих формул теоретически невелико.

2.6. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Краткие сведения о методах релаксации цитируются по справочнику [30], доказательства большинства утверждений на эту тему можно найти в [200]. Принято считать, что в методе верхней релаксации параметр удовлетворяет неравенствам $1 < \omega < 2$. При $0 < \omega < 1$ тот же самый алгоритм называют методом нижней релаксации. В случае модифицированных (с двумя параметрами) релаксационных методов подобной детализации в терминологии не существует. Симметричный метод верхней релаксации предложен и исследован в [186].

Важным результатом теории методов релаксации является тот факт, что для решения систем с симметричными положительно определенными (2×2) -блочными матрицами не требуется вводить дополнительные параметры, т. е. вполне достаточно варианта $\omega_1 = \omega_2 = \omega$. В случае симметричных седловых задач одного итерационного параметра заведомо не достаточно, что приводит к необходимости анализировать более сложные модифицированные релаксационные методы.

Основные идеи общего анализа релаксационных методов для решения седловых задач предложены на дифференциальном уровне в [83], где впервые были выведены формулы для комплексных собственных значений оператора перехода.

Результаты по сходимости и оптимизации модифицированного метода Якоби для седловых задач получены в [93].

За (2×2) -блочными методами решения седловых задач, использующими на втором полушаге ранее вычисленное значение u^{k+1} , закрепилось название «алгоритмы типа Эрроу-Гурвица». Отметим,

что первоначальный вариант метода имел вид [108]:

$$\begin{cases} \frac{u^{k+1} - u^k}{\tau} + Au^k + Bp^k = f, \\ -\frac{p^{k+1} - p^k}{\nu} + B^T u^{k+1} = \varphi. \end{cases}$$

В дальнейшем, в основном под влиянием французских математиков [79, 144], он постепенно трансформировался к виду метода MSOR, и в настоящее время эти термины стали практически синонимами.

Применительно к дифференциальной задаче Стокса один из первых результатов для метода MSOR при $C = I$

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau} + Au^k + Bp^k = f, \\ -\frac{p^{k+1} - p^k}{\tau/\alpha} + B^T u^{k+1} = \varphi, \end{cases}$$

был получен в [144]: достаточное условие сходимости вида

$$0 < \tau \leq 1, \quad 0 < \frac{\tau}{\alpha} < 2.$$

В [79] приведено другое достаточное условие:

$$0 < \tau < \frac{2\alpha}{\alpha + 1} \quad \forall \alpha > 0.$$

Окончательный результат в виде необходимого и достаточного условия сходимости метода MSOR для дифференциальной задачи Стокса получен в [85].

Результаты по сходимости и оптимизации модифицированного метода SSOR для седловых задач получены в [89].

Возможность использования слагаемого вида $B^T A^{-1} B u$ (или в дифференциальной форме для задачи Стокса — $\text{grad div } u$) была впервые предложена в [144]. Его введение в алгоритм типа Эрроу–Гурвица привело к построению семейства (β, τ) -методов для решения систем типа Стокса и Навье–Стокса (см. обзор [48] и цитируемую там литературу). В 80-е годы прошлого века велась активная научная дискуссия о преимуществах и недостатках двух подходов при численном решении гидродинамических задач: сравнивались алгоритмы в переменных скорость — давление и в переменных функция тока — вихрь (см., например, [75]). Математически обоснованные (β, τ) -методы сыграли важную роль в выборе стратегического направления для дальнейших исследований в вычислительной гидродинамике. Однако вопрос об актуальности слагаемого $B^T A^{-1} B u$ оказался достаточно сложным. Теорема 2.5.2 утверждает, что для

ускорения сходимости метода MSOR при решении линейных задач оно не требуется. Численные эксперименты по решению нелинейных задач ([16, 21], глава 15 настоящей книги), наоборот, свидетельствуют о его крайней полезности. Кроме того, оно оказывается необходимым при решении седловых линейных задач с вырожденной матрицей A (см. [26], [27], раздел 6.3 о нерегулярных задачах), а также для улучшения аппроксимативных свойств (в виде $\text{grad div } u$) при переходе от непрерывных задач к дискретным [178].

Отметим любопытный факт, что если в исходные уравнения задачи $L_0 z = F$ добавить слагаемое вида $BC^{-1}B^T u$ с параметром β и в новой постановке решить задачу асимптотической оптимизации для метода MJOR, то полученный показатель скорости сходимости будет совпадать с показателем для метода MSOR [94]. При этом с вычислительной точки зрения новый алгоритм будет более эффективен, чем MSOR, так как его реализация требует только одного обращения матрицы C и естественным образом допускает параллелизм вычислений.

Трехпараметрический метод 3MSOR был предложен и исследован в [129].

Результаты по сходимости и оптимизации (β, τ) -метода приведены в работе [83]. До ее публикации расчетные параметры алгоритма брались при $\Gamma = 1$ из полуэмпирических соображений [16]: $\beta = \tau/(1 - \tau)$, а τ определялся экспериментально.

Теорема 2.5.4 является более точной, чем соответствующий результат из [83].

ОЦЕНКИ ПОГРЕШНОСТИ МЕТОДОВ MJOR И MSOR

Глава посвящена получению оценок погрешности модифицированных методов Якоби и SOR (методов MJOR и MSOR) при наилучшем выборе постоянных итерационных параметров. Рассматриваются также возможности использования переменных параметров и их выбор из вариационных принципов.

Обозначим погрешность решения на k -й итерации через

$$y^k = \{v^k, r^k\} = \{u^k - u, p^k - p\},$$

где $\{u, p\}$ — точное решение задачи $L_0 z = F$, введем операторы

$$D_u = D_u^T > 0 \quad \text{и} \quad D_p = D_p^T > 0,$$

такие что

$$D_u A^{-1} B_0 = (D_u A^{-1} B_0)^T, \quad D_p C^{-1} S_0 = (D_p C^{-1} S_0)^T,$$

и определим норму погрешности в пространстве Z как

$$\|y\|_D^2 = (D_u v, v) + (D_p r, r), \quad y = \{v, r\} \in Z.$$

Напомним, что для стационарных методов всегда существует формальная оценка следующего вида:

$$\|y^k\| \leq \|T^k\| \|y^0\|,$$

и лемма 5.6.10 [82, с. 359] утверждает, что для любого $\delta > 0$ существует по крайней мере одна матричная норма такая, что

$$q \leq \|T\| \leq q + \delta,$$

где q — величина спектрального радиуса оператора перехода T . Однако если матрица T недиагонализуема, то построение таких норм весьма обременительно для практического использования. Поэтому представляет большой интерес получение оценок в естественных нормах, например при $D_u = A$, $D_p = C$, хотя величина $\|T^k\|$ в соответствующей подчиненной норме может ухудшиться.

3.1. ОЦЕНКИ ИЗ ОБЩЕЙ ТЕОРИИ

В разделе приводятся формулы и оценки погрешностей из общей теории итерационных методов решения линейных систем с симметричными положительно определенными матрицами

$$Au = f, \quad A = A^T > 0. \quad (3.1)$$

Рассматриваются следующие обобщенные алгоритмы: оптимальный одношаговый метод, циклический метод с чебышевскими параметрами, трехслойный метод с постоянными параметрами, полуитерационный метод Чебышева, методы сопряженных направлений (градиентов, невязок, поправок, погрешностей).

3.1.1. Оптимальный одношаговый метод

Рассмотрим для решения задачи (3.1) обобщенный метод простой итерации

$$B \frac{u^{k+1} - u^k}{\tau} + Au^k = f \quad (3.2)$$

с симметричным положительно определенным (здесь и далее в настоящем разделе) оператором B : $B = B^T > 0$. Пусть оператор $D = D^T > 0$ таков, что $DB^{-1}A = (DB^{-1}A)^T$, и имеет место матричное неравенство

$$\gamma D \leq DB^{-1}A \leq \Gamma D$$

с постоянными $0 < \gamma \leq \Gamma$. Тогда при $\tau = 2/(\gamma + \Gamma)$ метод (3.2) называется *оптимальным одношаговым методом*, который сходится к решению u с оценкой погрешности

$$\|u^k - u\|_D \leq q_1^k \|u^0 - u\|_D, \quad q_1 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma}{\Gamma}. \quad (3.3)$$

3.1.2. Циклический метод с чебышевскими параметрами

Рассмотрим следующую оптимизационную задачу: для фиксированного N указать набор итерационных параметров τ_1, \dots, τ_N в алгоритме

$$B \frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k = f, \quad (3.4)$$

минимизирующий D -норму погрешности за N итераций. При использовании введенных выше обозначений решение поставленной задачи имеет вид

$$\tau_k = \frac{2}{(\gamma + \Gamma)(1 + q_1 \mu_k)}, \quad k = 1, 2, \dots, N, \quad (3.5)$$

где

$$\mu_k \in \aleph_N = \left\{ -\cos \frac{2i-1}{2N} \pi, i = 1, 2, \dots, N \right\}.$$

Метод (3.4) – (3.5) называется *циклическим методом с чебышевскими параметрами*, который сходится к решению u с оценкой погрешности

$$\|u^k - u\|_D \leq \frac{2q_0^k}{1 + q_0^{2k}} \|u^0 - u\|_D, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}. \quad (3.6)$$

Отметим здесь необходимость построения множества \aleph_N для заранее определенного количества итераций N и, кроме того, требование упорядоченности его элементов для вычислительной устойчивости [61, 62, 76].

3.1.3. Полуитерационный метод Чебышева

Рассмотрим следующую оптимизационную задачу: указать набор итерационных параметров $\tau_1, \alpha_2, \tau_2, \dots, \alpha_l, \tau_l, \dots$ в алгоритме

$$\begin{aligned} Bu^{k+1} &= \alpha_{k+1}(B - \tau_{k+1}A)u^k + (1 - \alpha_{k+1})Bu^{k-1} + \alpha_{k+1}\tau_{k+1}f, \\ Bu^1 &= (B - \tau_1A)u^0 + \tau_1f, \end{aligned} \quad (3.7)$$

приводящий для произвольного начального приближения u^0 к наилучшей для любого $k = 1, 2, \dots$ оценке погрешности

$$\|u^k - u\|_D \leq \frac{2q_0^k}{1 + q_0^{2k}} \|u^0 - u\|_D, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}. \quad (3.8)$$

При использовании введенных выше обозначений решение поставленной задачи имеет вид

$$\tau_k = \frac{2}{\gamma + \Gamma}, \quad \alpha_{k+1} = \frac{4}{4 - q_1^2 \alpha_k}, \quad k = 1, 2, \dots, \quad (3.9)$$

где для замыкания последнего рекуррентного соотношения необходимо положить $\alpha_1 = 2$. Метод (3.7), (3.9) называется *полуитерационным методом Чебышева*.

Далее нам потребуются следующие свойства формул итерационных параметров (3.9):

Лемма 3.1.1. Итерационные параметры τ_k, α_k в полуитерационном методе Чебышева (3.7), (3.9) удовлетворяют для любого $k = 1, 2, \dots$ неравенствам:

$$\tau_k > 0, \quad 2 \geq \alpha_k > 1,$$

причем только $\alpha_1 = 2$.

Доказательство. Первое неравенство вытекает из явной формулы $\tau_k = 2/(\gamma + \Gamma)$, $k = 1, 2, \dots$, и свойств спектра

$$\sigma(B^{-1/2}AB^{-1/2}) \in [\gamma, \Gamma], \quad \gamma > 0,$$

а второе следует из представления ([76], с.322)

$$\alpha_{k+1} = \frac{2q_0(1 + q_0^{2k})}{q_1(1 + q_0^{2k+2})}, \quad k = 1, 2, \dots, \quad \text{и} \quad \alpha_1 = 2.$$

Действительно, так как справедливо

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma} < 1 \quad \text{и} \quad q_1 = \frac{1 - \xi}{1 + \xi} = \frac{2q_0}{1 + q_0^2},$$

то имеем при $k \geq 1$

$$2 > \alpha_{k+1} = 1 + \frac{q_0^2 + q_0^{2k}}{1 + q_0^{2k+2}} > 1. \quad \blacksquare$$

3.1.4. Стационарный трехслойный метод

Рассмотрим для решения задачи (3.1) трехслойный итерационный метод с постоянными параметрами α , τ

$$\begin{aligned} Bu^{k+1} &= \alpha(B - \tau A)u^k + (1 - \alpha)Bu^{k-1} + \alpha\tau f, \\ Bu^1 &= (B - \tau A)u^0 + \tau f, \end{aligned} \quad (3.10)$$

где

$$\tau = \frac{2}{\gamma + \Gamma}, \quad \alpha = \lim_{k \rightarrow \infty} \alpha_k = 1 + q_0^2, \quad (3.11)$$

а последовательность α_k определена в (3.9). Такой алгоритм сходится к решению u с оценкой погрешности

$$\begin{aligned} \|u^k - u\|_D &\leq q_0^k \left(1 + k \frac{1 - q_0^2}{1 + q_0^2} \right) \|u^0 - u\|_D, \\ q_0 &= \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}, \end{aligned} \quad (3.12)$$

и называется *стационарным трехслойным методом*.

3.1.5. Методы сопряженных направлений

Самостоятельный интерес представляет группа методов (3.7)

$$\begin{aligned} Bu^{k+1} &= \alpha_{k+1}(B - \tau_{k+1}A)u^k + (1 - \alpha_{k+1})Bu^{k-1} + \alpha_{k+1}\tau_{k+1}f, \\ Bu^1 &= (B - \tau_1A)u^0 + \tau_1f, \end{aligned}$$

также приводящих с произвольного начального приближения u^0 к наилучшей для любого $k = 1, 2, \dots$ оценке погрешности (3.8):

$$\|u^k - u\|_D \leq \frac{2q_0^k}{1 + q_0^{2k}} \|u^0 - u\|_D, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Однако здесь итерационные параметры выбираются из вариационных принципов и имеют следующий вид:

$$\begin{aligned} \tau_{k+1} &= \frac{(r^k, Dw^k)}{(w^k, Dw^k)}, \quad k = 0, 1, 2, \dots, \\ \alpha_{k+1} &= \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r^k, Dw^k)}{(r^{k-1}, Dw^{k-1})} \frac{1}{\alpha_k} \right)^{-1}, \quad k = 1, 2, \dots, \end{aligned} \quad (3.13)$$

где $\alpha_1 = 1$, $w^k = B^{-1}x^k$ — поправка, $r^k = A^{-1}x^k = (u^k - u)$ — погрешность, $x^k = (Au^k - f)$ — невязка. Эти алгоритмы называются *методами сопряженных направлений*, а конкретный вариант метода определяется выбором оператора D :

- $D = A$ — метод сопряженных градиентов,
- $D = AB^{-1}A$ — метод сопряженных поправок,
- $D = A^2$ — метод сопряженных невязок,
- $D = AB$ — метод сопряженных погрешностей.

Отметим, что в двух последних методах дополнительно требуется выполнение условия $AB = BA$.

Далее нам потребуются следующие свойства формул итерационных параметров (3.13).

Лемма 3.1.2. *Итерационные параметры τ_k , α_k в методе сопряженных направлений (3.7), (3.13) удовлетворяют для любого $k \geq 1$ неравенствам*

$$\tau_k > 0, \quad \alpha_k \geq 1,$$

причем только $\alpha_1 = 1$.

Доказательство. Справедливость первого неравенства следует из преобразования первой формулы (3.13). Поскольку

$$w^k = B^{-1}x^k = B^{-1}Ar^k,$$

то

$$\tau_{k+1} = \frac{(r^k, Dw^k)}{(w^k, Dw^k)} = \frac{(r^k, DB^{-1}Ar^k)}{(w^k, Dw^k)}.$$

Далее, в силу

$$D = D^T > 0 \quad \text{и} \quad DB^{-1}A = (DB^{-1}A)^T > 0,$$

получаем $\tau_k > 0$ для всех $k = 1, 2, \dots$

Для доказательства неравенства $\alpha_{k+1} > 1$ при $k = 1, 2, \dots$ перепишем вторую формулу (3.13) в виде

$$\alpha_{k+1} = 1 + \frac{\alpha_{k+1}\tau_{k+1}}{\alpha_k\tau_k} \frac{(r^k, DB^{-1}Ar^k)}{(r^{k-1}, DB^{-1}Ar^{k-1})}.$$

Теперь искомое неравенство легко доказывается по индукции от противного, так как все сомножители во втором слагаемом положительны. Напоминание о том, что $\alpha_1 = 1$, завершает доказательство леммы. ■

Чтобы не загромождать изложение, далее будем обсуждать применение только метода сопряженных градиентов, формулы для итерационных параметров которого имеют вид:

$$\begin{aligned} \tau_{k+1} &= \frac{(x^k, w^k)}{(w^k, Aw^k)}, \quad k = 0, 1, 2, \dots, \\ \alpha_{k+1} &= \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(x^k, w^k)}{(x^{k-1}, w^{k-1})} \frac{1}{\alpha_k} \right)^{-1}, \quad k = 1, 2, \dots \end{aligned}$$

3.2. ПОГРЕШНОСТЬ МЕТОДА MJOR В СЛУЧАЕ ПОСТОЯННЫХ ПАРАМЕТРОВ

3.2.1. Преобразование формул

Сначала из формул (2.9) получим отдельные трехслойные соотношения для компонент погрешности v^k, r^k . Имеет место

Лемма 3.2.3. Компоненты погрешности v^{k+1}, r^{k+1} при $k \geq 1$ удовлетворяют соотношениям

$$\begin{aligned} v^{k+1} &= (2 - \tau)v^k - \left[(1 - \tau)I + \frac{\tau^2}{\alpha} A^{-1}B_0 \right] v^{k-1}, \\ r^{k+1} &= (2 - \tau)r^k - \left[(1 - \tau)I + \frac{\tau^2}{\alpha} C^{-1}S_0 \right] r^{k-1}. \end{aligned}$$

Доказательство. Перепишем формулы итерационного метода (2.9) для погрешности $y^k = \{v^k, r^k\}$:

$$\begin{cases} A \frac{v^{k+1} - v^k}{\tau} + Av^k + Br^k = 0, \\ -\alpha C \frac{r^{k+1} - r^k}{\tau} + B^T v^k = 0, \end{cases}$$

откуда имеем

$$v^{k+1} = (1 - \tau)v^k - \tau A^{-1}Br^k, \quad (3.14)$$

$$r^{k+1} = r^k + \frac{\tau}{\alpha} C^{-1} B^T v^k. \quad (3.15)$$

Увеличим в (3.15) индекс k на единицу и заменим в полученном выражении v^{k+1} с помощью соотношения (3.14). В результате получим

$$r^{k+2} = r^{k+1} - \frac{\tau^2}{\alpha} C^{-1} S_0 r^k + \frac{\tau}{\alpha} (1 - \tau) C^{-1} B^T v^k. \quad (3.16)$$

Теперь выразим из (3.15) величину

$$\frac{\tau}{\alpha} C^{-1} B^T v^k = r^{k+1} - r^k$$

и подставим ее в (3.16):

$$r^{k+2} = (2 - \tau) r^{k+1} - \left[(1 - \tau) I + \frac{\tau^2}{\alpha} C^{-1} S_0 \right] r^k.$$

Искомое соотношение для второй компоненты погрешности получено. Аналогичным образом получается трехсложное соотношение и для первой. Увеличим в (3.14) индекс k на единицу и заменим в полученном выражении r^{k+1} с помощью соотношения (3.15). В результате будем иметь

$$v^{k+2} = (1 - \tau) v^{k+1} - \frac{\tau^2}{\alpha} A^{-1} B_0 v^k - \tau A^{-1} B r^k. \quad (3.17)$$

Теперь выразим из (3.14) величину

$$-\tau A^{-1} B r^k = v^{k+1} - (1 - \tau) v^k$$

и подставим ее в (3.17):

$$v^{k+2} = (2 - \tau) v^{k+1} - \left[(1 - \tau) I + \frac{\tau^2}{\alpha} A^{-1} B_0 \right] v^k. \quad \blacksquare$$

3.2.2. Начальное приближение

Выберем начальное приближение $\{u^0, p^0\}$ из условия (1.9):

$$A u^0 + B p^0 = f.$$

Для такого начального приближения имеем:

$$v^1 = v^0, \quad r^1 = \left(I - \frac{\tau}{\alpha} C^{-1} S_0 \right) r^0, \quad (3.18)$$

и, кроме того, в силу леммы 1.3.2 первая компонента v^k погрешности y^k итерационного метода (2.9) для любой итерации k является элементом подпространства G (т. е. $(A v^k, h) = 0$ для $\forall h \in H$).

Поэтому справедливо покомпонентное разложение погрешности следующего вида:

$$v^k = \sum_{i=1}^{N_p} c_i^{(k)} g_i, \quad r^k = \sum_{i=1}^{N_p} d_i^{(k)} p_i,$$

где $\{g_i\}$ и $\{p_i\}$ — базисы пространств G и P , порожденные задачами (1.7) и (1.8).

3.2.3. Оценка погрешности с постоянными параметрами

Для используемого начального приближения в пункте 2.2.5 была решена задача асимптотической оптимизации метода в подпространстве с величиной спектрального радиуса \tilde{q}_0 . Получим оценку погрешности, соответствующую этому значению. Справедлива

Теорема 3.2.1. *Итерационный метод (2.9) с асимптотически оптимальными параметрами τ_0, α_0 и спектральным радиусом оператора перехода*

$$\tilde{q}_0 = \sqrt{\frac{1-\xi}{1+\xi}}, \quad \xi = \frac{\gamma}{\Gamma},$$

стартующий с начального приближения вида (1.9), сходится с оценкой погрешности при четных k

$$\|y^k\|_D \leq \tilde{q}_0^k \|y^0\|_D.$$

Доказательство. Пусть $\tau_0 = 2$, $\alpha_0 = 2(\gamma + \Gamma)$ — асимптотически оптимальные итерационные параметры, а \tilde{q}_0 — соответствующий им спектральный радиус оператора перехода из теоремы 2.2.4. Если ввести обозначение $\tilde{\tau} = 2/(\gamma + \Gamma)$, то полученные трехслойные соотношения для погрешности метода (2.9) можно переписать в более удобном виде:

$$v^{k+1} = (I - \tilde{\tau} A^{-1} B_0) v^{k-1}, \quad r^{k+1} = (I - \tilde{\tau} C^{-1} S_0) r^{k-1}. \quad (3.19)$$

С помощью этих выражений можно оценить компоненты погрешности v^k и r^k порознь. Поскольку из теоремы 1.3.1 следует, что область изменения собственных значений матриц $A^{-1/2} B_0 A^{-1/2}$ и $C^{-1/2} S_0 C^{-1/2}$ в соотношениях (3.19) одинакова и является отрезком $[\gamma, \Gamma]$, то в силу определения величины $\tilde{\tau}$ имеем

$$\|y^{k+2}\|_D \leq \tilde{q}_0^2 \|y^k\|_D,$$

откуда и следует искомая оценка. ■

При нечетных k оценка погрешности немного ухудшается за счет выражений (3.18): показатель степени уменьшается на единицу.

3.3. ПОГРЕШНОСТЬ МЕТОДА MJOR В СЛУЧАЕ ПЕРЕМЕННЫХ ПАРАМЕТРОВ

Рассмотрим алгоритм из предыдущего раздела, но величины α и τ будем считать переменными

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k + Bp^k = f, \\ -\alpha_{k+1} C \frac{p^{k+1} - p^k}{\tau_{k+1}} + B^T u^k = \varphi. \end{cases} \quad (3.20)$$

Здесь τ_{k+1}, α_{k+1} — вещественные положительные итерационные параметры. Обозначим через $R(\alpha)$ участвующий в методе (3.20) оператор

$$R(\alpha) = \begin{pmatrix} I & A^{-1}B \\ -\frac{1}{\alpha}C^{-1}B^T & 0 \end{pmatrix}. \quad (3.21)$$

Тогда для погрешности (ошибки) метода y^k будет справедливо соотношение

$$y^{k+1} = T(\alpha_{k+1})y^k, \quad T(\alpha_k) = I - \tau_k R(\alpha_k). \quad (3.22)$$

После N итераций по формуле (3.20) для ошибки y^N отсюда следует

$$y^N = (I - \tau_N R(\alpha_N)) \dots (I - \tau_k R(\alpha_k)) \dots (I - \tau_1 R(\alpha_1)) y^0. \quad (3.23)$$

В этой формуле оператор уменьшения ошибки за N шагов представляет собой обобщенный многочлен от матрицы R , в котором множители не обязаны коммутировать при различных α_k . Проверим это.

Рассмотрим спектральную задачу

$$R(\alpha)z = \lambda z, \quad z \in Z. \quad (3.24)$$

Имеет место

Теорема 3.3.1. *Собственные значения $\lambda_1, \dots, \lambda_{N_u+N_p}$ в задаче (3.24) представимы в виде*

$$\lambda_k^{(1)} = 1, \quad k = 1, \dots, N_u - N_p, \quad (3.25)$$

$$\lambda_j^{(2,3)} = \frac{1}{2} \left(1 \pm \sqrt{1 - 4t_j/\alpha} \right), \quad j = 1, \dots, N_p, \quad (3.26)$$

где t_1, \dots, t_{N_p} — собственные значения задачи $S_0 p = t C p$.

Доказательство. Так как $R(\alpha) = \tau^{-1}(I - T(\alpha))$, то искомые формулы для собственных значений следуют из теоремы 2.2.1. При этом собственные векторы, соответствующие значениям (3.26), представимы в виде

$$z_j = \{g_j, -\alpha^{-1}C^{-1}B^T g_j\}$$

с коэффициентами $\kappa_j^{(2,3)} = \lambda_j^{(2,3)}\alpha$ при $\lambda_j^{(2)} \neq \lambda_j^{(3)}$. В случае кратных собственных значений коэффициенты определяются формулами:

$$\kappa_j^{(2)} = \frac{\alpha}{2} = 2t_j, \quad \kappa_j^{(3)} = -2t_j. \quad \blacksquare$$

Из полученного результата следует, что собственные векторы $z_j^{(2,3)}$ в (3.24) зависят от α . Поэтому известные стандартные методы оптимизации алгоритмов, основанные на минимизации спектрального радиуса оператора перехода в (3.20), в рассматриваемом случае переменных α_k неприменимы и требуется более детальное изучение.

Задав особым образом начальное приближение, будем проводить итерации, для которых вектор ошибки y^k при произвольном k будет принадлежать подпространству размерности $2N_p$ и который может быть представлен в виде специального разложения по линейно-независимой системе векторов.

Найдем соответствующие системы двухмерных соотношений. Выберем начальное приближение $\{u^0, p^0\}$ из условия (1.9)

$$Au^0 + Bp^0 = f.$$

Тогда из леммы 1.3.2 и теоремы 1.3.2 вытекает покомпонентное разложение погрешности $y^k = \{v^k, r^k\}$ следующего вида:

$$v^k = \sum_{i=1}^{N_p} c_i^{(k)} g_i, \quad r^k = \sum_{i=1}^{N_p} d_i^{(k)} p_i, \quad (3.27)$$

где $\{g_i\}$ и $\{p_i\}$ — базисы пространств G и P , порожденные задачами (1.7) и (1.8). Подставляя (3.27) в (3.22), получим для каждого i

$$y_i^{k+1} = (I_2 - \tau_{k+1} R_i(\alpha_{k+1})) y_i^k. \quad (3.28)$$

Здесь использованы обозначения

$$y_i^k = (c_i^{(k)}, d_i^{(k)}), \quad R_i(\alpha) = \begin{pmatrix} 1 & t_i \\ -\alpha^{-1} & 0 \end{pmatrix},$$

а I_2 имеет смысл единичной матрицы второго порядка. Следовательно, формулу (3.23) можно уточнить:

$$y_i^N = (I_2 - \tau_N R_i(\alpha_N)) \dots (I_2 - \tau_k R_i(\alpha_k)) \dots (I_2 - \tau_1 R_i(\alpha_1)) y_i^0. \quad (3.29)$$

Собственными значениями матрицы $R_j(\alpha)$ будут величины из (3.26), а собственными векторами — векторы $(\alpha \lambda_j^{(2,3)}, -1)$. При различных α_k множители в (3.29), как в (3.23), не коммутируют.

Запишем равенства (3.28) в координатной форме:

$$\begin{aligned} c_i^{(k+1)} &= (1 - \tau_{k+1})c_i^{(k)} - \tau_{k+1}t_i d_i^{(k)}, \\ d_i^{(k+1)} &= \frac{\tau_{k+1}}{\alpha_{k+1}}c_i^{(k)} + d_i^{(k)}. \end{aligned} \quad (3.30)$$

Отсюда можно получить трехслойные соотношения, связывающие либо только величины $c_i^{(k)}$, либо только $-d_i^{(k)}$. Приведем выкладки для первого случая. Из второго уравнения (3.30) получаем

$$d_i^{(k+1)} - d_i^{(k)} = \frac{\tau_{k+1}}{\alpha_{k+1}}c_i^{(k)}$$

и, учитывая первое уравнение (3.30) относительно $c_i^{(k+2)}$ и $c_i^{(k+1)}$, имеем

$$c_i^{(k+2)} = c_i^{(k+1)} - \frac{\tau_{k+2}\tau_{k+1}}{\alpha_{k+1}}t_i c_i^{(k)} + \tau_{k+2} \left(\frac{1}{\tau_{k+1}} - 1 \right) (c_i^{(k+1)} - c_i^{(k)}). \quad (3.31)$$

Таким образом, для всех $m \geq 0$ справедливо

$$\begin{aligned} c_i^{(2m)} &= Q_m^{(1)}(t_i)c_i^{(0)}, & c_i^{(2m+1)} &= Q_m^{(2)}(t_i)c_i^{(0)}, \\ d_i^{(2m)} &= Q_m^{(3)}(t_i)d_i^{(0)}, & d_i^{(2m+1)} &= Q_{m+1}^{(4)}(t_i)d_i^{(0)}, \\ Q_0^{(1)}(t) &= Q_0^{(2)}(t) \equiv 1, & Q_0^{(3)}(t) &\equiv 1, & Q_1^{(4)}(t) &= 1 - \frac{\tau_1}{\alpha_1}t, \end{aligned}$$

где через $Q_m^{(l)}(t)$, $l = 1, 2, 3, 4$, обозначены многочлены m -й степени, удовлетворяющие нестандартным трехчленным рекуррентным соотношениям вида:

- при $k = 2m$

$$\begin{aligned} Q_{m+1}^{(1)}(t) &= Q_m^{(2)}(t) - \frac{\tau_{2m+2}\tau_{2m+1}}{\alpha_{2m+1}}tQ_m^{(1)}(t) + \\ &+ \tau_{2m+2} \left(\frac{1}{\tau_{2m+1}} - 1 \right) (Q_m^{(2)}(t) - Q_m^{(1)}(t)); \end{aligned}$$

- при $k = 2m + 1$

$$\begin{aligned} Q_{m+2}^{(2)}(t) &= Q_{m+1}^{(1)}(t) - \frac{\tau_{2m+3}\tau_{2m+2}}{\alpha_{2m+2}}tQ_{m+1}^{(2)}(t) + \\ &+ \tau_{2m+3} \left(\frac{1}{\tau_{2m+2}} - 1 \right) (Q_{m+1}^{(1)}(t) - Q_{m+1}^{(2)}(t)). \end{aligned}$$

Это означает, что ошибки v^k выражаются через операторные многочлены от $A^{-1}B_0$:

$$v^{2m} = Q_m^{(1)}(A^{-1}B_0)v^0, \quad v^{2m+1} = Q_m^{(2)}(A^{-1}B_0)v^0,$$

а, соответственно, ошибки r^k выражаются через операторные многочлены от $C^{-1}S_0$:

$$r^{2m} = Q_m^{(3)}(C^{-1}S_0)r^0, \quad r^{2m+1} = Q_{m+1}^{(4)}(C^{-1}S_0)r^0.$$

Рассмотрим сначала случай, когда $\alpha_k = \alpha$, $k = 1, \dots, N$. Тогда операторные множители в формулах (3.23), (3.29) коммутируют друг с другом и задачу оптимизации метода удастся исследовать до конца. В самом деле, каждый из операторов $R_i(\alpha)$ будет иметь два собственных значения

$$\frac{1}{2} \left(1 \pm \sqrt{1 - 4t_i/\alpha} \right),$$

а спектр оператора $R(\alpha)$ будет состоять из объединения этих значений, причем $t_i \in [\gamma, \Gamma]$.

Иследуем вид множества $\sigma(\alpha) = \sigma(R(\alpha))$ в плоскости комплексного переменного при различных значениях α .

При $\alpha^{-1} < 1/(4\Gamma)$ множество $\sigma(\alpha)$ представляет собой два отрезка одинаковой длины, лежащих на действительной оси симметрично относительно точки $(1/2, 0)$ с крайними точками

$$\frac{1}{2} \left(1 \pm \sqrt{1 - \frac{4\gamma}{\alpha}}, 0 \right), \quad \frac{1}{2} \left(1 \pm \sqrt{1 - \frac{4\Gamma}{\alpha}}, 0 \right).$$

Причем при $\alpha > 4\Gamma$ эти отрезки расположены по одну сторону от начала координат, а при $\alpha < 0$ — по разные.

При $0 < \alpha < 4\gamma$ множество $\sigma(\alpha)$ — два отрезка одинаковой длины, расположенных перпендикулярно действительной оси, с концами в точках

$$\frac{1}{2} \left(1 \pm i\sqrt{\frac{4\gamma}{\alpha} - 1} \right), \quad \frac{1}{2} \left(1 \pm i\sqrt{4\frac{\Gamma}{\alpha} - 1} \right).$$

При $4\gamma < \alpha < 4\Gamma$ множество $\sigma(\alpha)$ — крест с центром в точке $(1/2, 0)$ и с концами в точках

$$\left(\frac{1}{2} \left(1 \pm \sqrt{1 - \frac{4\gamma}{\alpha}} \right), 0 \right), \quad \frac{1}{2} \left(1 \pm i\sqrt{\frac{4\Gamma}{\alpha} - 1} \right).$$

Единым преобразованием

$$t = \alpha\lambda(1 - \lambda) \tag{3.32}$$

множество $\sigma(\alpha)$ для всех случаев отображается на отрезок $[\gamma, \Gamma]$.

Решим теперь задачу о наилучшем выборе параметров в методе (3.20). В случае $\alpha_k = \alpha$, $k = 1, \dots, N$ формулы (3.23), (3.29)

принимают вид

$$y^N = P_N(R(\alpha))y^0, \quad P_N(\lambda) = \prod_{k=1}^N (1 - \tau_k \lambda). \quad (3.33)$$

Будем считать выбор параметров наилучшим, для которого $P_N^0(\lambda)$ в рассматриваемом алгоритме есть решение задачи

$$P_N^0(\lambda) = \arg \inf_{P_N(\lambda)} \max_{\lambda \in \sigma(\alpha)} |P_N(\lambda)|, \quad (3.34)$$

где \inf берется по всем многочленам степени N вида (3.33). Решение этой задачи для $N = 2M$ определяется явно [61, с. 235]:

$$P_{2M}^0(\lambda) = \frac{1}{T_M(\theta)} T_M \left(\frac{\gamma + \Gamma - 2t}{\Gamma - \gamma} \right), \quad (3.35)$$

где $T_M(x)$ — многочлен Чебышева первого рода степени M , $\theta = (\Gamma + \gamma)/(\Gamma - \gamma) > 1$, а t определяется через λ по формуле (3.32). При этом

$$E_{2M} = \max_{\lambda \in \sigma(\alpha)} |P_{2M}^0(\lambda)| = \frac{2q_0^M}{1 + q_0^{2M}}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}. \quad (3.36)$$

Отметим, что величина E_{2M} не зависит от конкретного значения α . Осталось определить формулы для параметров τ_k , $k = 1, \dots, N$: они будут величинами, обратными к корням (3.35). Пусть

$$\omega_k = \frac{2j_k - 1}{2M}, \quad 1 \leq j_k \leq M, \\ \bar{t}_k = \frac{1}{2} (\Gamma + \gamma - (\Gamma - \gamma) \cos \pi \omega_k),$$

тогда τ_k , $k = 1, \dots, N = 2M$, есть множество, состоящее из чисел

$$\left\{ \frac{2}{1 \pm \sqrt{1 - 4\bar{t}_k/\alpha}} \right\}, \quad k = 1, \dots, M, \quad (3.37)$$

порядок следования которых определяется перестановкой (j_1, j_2, \dots, j_M) при условии, что два параметра τ_k с одинаковыми \bar{t}_k употребляются подряд.

Удобно проводить вычисления (3.20) в действительной арифметике. Для этого α должно удовлетворять одному из неравенств: $\alpha < 0$ или $\alpha > 4\Gamma$. После проведения нескольких циклов из $2M$ итераций вследствие ошибок округлений итерационные приближения могут выйти из рабочего подпространства, связанного с представлением погрешности (3.27). Тогда следует повторить операцию проектирования подобно (1.9).

Окончательный результат можно сформулировать так:

Теорема 3.3.2. Пусть $\alpha_k = \alpha \neq 0$, а τ_k определено формулой (3.37). Тогда после $2M$ итераций для погрешности метода (3.20) будет справедлива оценка

$$\|y^{2M}\|_D \leq \frac{2q_0^M}{1 + q_0^{2M}} \|y^0\|_D, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Заметим, что полученная оценка носит неупрощаемый характер, т. е. использование переменных параметров α_k (некоммутативный случай) не может уменьшить норму погрешности. Действительно, из формул (3.30), (3.31) следует, что степень многочлена, определяющего норму погрешности, не зависит от того, параметры α_k переменные или постоянные, а именно: за $2M$ итераций степень многочлена не превышает M . Поскольку среди всех таких рассматриваемых многочленов на отрезке $[\gamma, \Gamma]$ построенный нами многочлен Чебышева за счет только переменных τ_k при произвольном $\alpha \neq 0$ реализует теоретически возможный минимум нормы, то использование различных α_k может привести только к неединственности наилучшего многочлена ошибки, но не изменит величину его нормы.

3.4. ПОГРЕШНОСТЬ МЕТОДА MSOR В СЛУЧАЕ ПОСТОЯННЫХ ПАРАМЕТРОВ

3.4.1. Преобразование формул

Получим из формул (2.14) отдельные трехслойные соотношения для компонент погрешности v^k, r^k . Имеет место

Лемма 3.4.4. Компоненты погрешности v^{k+1}, r^{k+1} при $k \geq 1$ в области сходимости метода ($0 < \tau < 2$) удовлетворяют соотношениям

$$v^{k+1} = \tilde{\alpha}(I - \tilde{\tau}A^{-1}B_0)v^k + (1 - \tilde{\alpha})v^{k-1}, \quad (3.38)$$

$$r^{k+1} = \tilde{\alpha}(I - \tilde{\tau}C^{-1}S_0)r^k + (1 - \tilde{\alpha})r^{k-1}, \quad (3.39)$$

где новые параметры определяются формулами

$$\tilde{\alpha} = 2 - \tau, \quad \tilde{\tau} = \frac{\tau^2}{\alpha(2 - \tau)}.$$

Доказательство. Перепишем формулы итерационного метода (2.14) для погрешности $y^k = \{v^k, r^k\}$

$$\begin{cases} A \frac{v^{k+1} - v^k}{\tau} + Av^k + Br^k = 0, \\ -\alpha C \frac{r^{k+1} - r^k}{\tau} + B^T v^{k+1} = 0, \end{cases}$$

откуда имеем

$$v^{k+1} = (1 - \tau)v^k - \tau A^{-1} B r^k, \quad (3.40)$$

$$\begin{aligned} r^{k+1} &= r^k + \frac{\tau}{\alpha} C^{-1} B^T v^{k+1} = \\ &= \left(I - \frac{\tau^2}{\alpha} C^{-1} S_0 \right) r^k + \frac{\tau}{\alpha} (1 - \tau) C^{-1} B^T v^k. \end{aligned} \quad (3.41)$$

Увеличим в выражении (3.41) индекс k на единицу

$$r^{k+2} = \left(I - \frac{\tau^2}{\alpha} C^{-1} S_0 \right) r^{k+1} + \frac{\tau}{\alpha} (1 - \tau) C^{-1} B^T v^{k+1}, \quad (3.42)$$

выразим из первого равенства (3.41) величину

$$\frac{\tau}{\alpha} C^{-1} B^T v^{k+1} = r^{k+1} - r^k$$

и подставим ее в (3.42)

$$r^{k+2} = (2 - \tau) \left(I - \frac{\tau^2}{\alpha(2 - \tau)} C^{-1} S_0 \right) r^{k+1} + (1 - (2 - \tau)) r^k.$$

Теперь после замены

$$\tilde{\alpha} = 2 - \tau, \quad \tilde{\tau} = \frac{\tau^2}{\alpha(2 - \tau)}$$

при $0 < \tau < 2$ имеем искомое соотношение для второй компоненты погрешности. Аналогичным образом получается трехслойное соотношение и для первой. Увеличим в (3.40) индекс k на единицу и заменим в полученном выражении r^{k+1} с помощью левой части соотношения (3.41). В результате получим

$$v^{k+2} = \left((1 - \tau)I - \frac{\tau^2}{\alpha} A^{-1} B_0 \right) v^{k+1} - \tau A^{-1} B r^k. \quad (3.43)$$

Теперь выразим из (3.40) величину

$$-\tau A^{-1} B r^k = v^{k+1} - (1 - \tau)v^k$$

и подставим ее в (3.43):

$$v^{k+2} = (2 - \tau) \left(I - \frac{\tau^2}{\alpha(2 - \tau)} A^{-1} B_0 \right) v^{k+1} + (1 - (2 - \tau))v^k.$$

Теперь после замены

$$\tilde{\alpha} = 2 - \tau, \quad \tilde{\tau} = \frac{\tau^2}{\alpha(2 - \tau)}$$

приходим к искомому соотношению для первой компоненты погрешности. ■

3.4.2. Начальное приближение

Выберем начальное приближение $\{u^0, p^0\}$ из условия (1.9):

$$Au^0 + Bp^0 = f.$$

Для такого начального приближения имеем:

$$v^1 = v^0, \quad r^1 = \left(I - \tilde{\tau} \frac{2 - \tau}{\tau} C^{-1} S_0 \right) r^0, \quad (3.44)$$

и, кроме того, в силу леммы 1.3.2, первая компонента v^k погрешности y^k итерационного метода (2.9) для любой итерации k является элементом подпространства G (т. е. $(Av^k, h) = 0$ для $\forall h \in H$).

3.4.3. Полином ошибки

Положим в формулах для $\tilde{\alpha}, \tilde{\tau}$ асимптотически оптимальные значения α_0, τ_0 . Теперь для получения оценок каждой из компонент погрешности достаточно найти алгебраический полином $P_k(t)$, определяемый соотношениями

$$\begin{aligned} P_{k+1}(t) &= \tilde{\alpha}(1 - \tilde{\tau}t)P_k(t) + (1 - \tilde{\alpha})P_{k-1}(t), \\ P_1(t) &= 1 - \kappa\tilde{\tau}t, \quad P_0(t) = 1 \end{aligned} \quad (3.45)$$

(параметр κ может принимать значения 0 или $(2 - \tau_0)/\tau_0$), и определить его норму

$$\|P_k(t)\| = \max_{t \in [\gamma, \Gamma]} |P_k(t)|.$$

Для этого нам потребуются многочлены Чебышева степени k первого рода $T_k(x)$ и второго рода $U_k(x)$ (см. [76], с. 57). Имеет место

Лемма 3.4.5. Многочлен $P_k(t)$, определяемый соотношениями (3.45), имеет вид

$$P_k(t) = q_0^k \left\{ T_k(x) + \left[\left(\frac{\kappa q_1}{q_0} - 1 \right) x + \frac{1 - \kappa}{q_0} \right] U_{k-1}(x) \right\}, \quad (3.46)$$

где

$$x = \frac{1 - \tilde{\tau}t}{q_1} \in [-1, 1], \quad q_1 = \frac{1 - \xi}{1 + \xi}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Доказательство. Полагая $t = (1 - q_1x)/\tilde{\tau}$, отображим отрезок $[\gamma, \Gamma]$ на $[-1, 1]$. Тогда $P_k(t) = Q_k(x)$, $x \in [-1, 1]$. Учитывая явные формулы для параметров $\tilde{\alpha}, \tilde{\tau}$, получим рекуррентные соотношения для многочленов $Q_k(x)$:

$$\begin{aligned} Q_{k+1}(x) &= 2q_0xQ_k(x) - q_0^2Q_{k-1}(x), \\ Q_1(x) &= 1 - \kappa + \kappa q_1x, \quad Q_0(x) = 1. \end{aligned}$$

Отсюда, при помощи замены $Q_k(x) = q_0^k R_k(x)$, имеем

$$\begin{aligned} R_{k+1}(x) &= 2xR_k(x) - R_{k-1}(x), \\ R_1(x) &= \frac{1-\kappa}{q_0} + \frac{\kappa q_1}{q_0}x, \quad R_0(x) = 1. \end{aligned}$$

Полученному рекуррентному соотношению удовлетворяют многочлены Чебышева первого и второго рода, поэтому методом неопределенных коэффициентов $R_k(x) = C_1(x)T_k(x) + C_2(x)U_{k-1}(x)$ несложно определить, что

$$R_k(x) = T_k(x) + \left[\left(\frac{\kappa q_1}{q_0} - 1 \right) x + \frac{1-\kappa}{q_0} \right] U_{k-1}(x),$$

откуда и следует формула (3.46). ■

3.4.4. Оценка погрешности

Теорема 3.4.1. Итерационный метод (2.14) с асимптотически оптимальными параметрами t_0, α_0 и спектральным радиусом оператора перехода

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma},$$

стартующий с начального приближения вида (1.9), сходится с оценкой погрешности

$$\|y^k\|_D \leq q_0^k (1 + ck) \|y^0\|_D,$$

где постоянная $c > 0$ не зависит от номера итерации.

Доказательство. Обозначим многочлен $P_k(t)$ при $\kappa = 0$ за $P_k^v(t)$, тогда из (3.38), (3.44) следует $v^k = P_k^v(A^{-1}B_0)v^0$, $v^k \in G$ и

$$(D_u v^k, v^k) \leq \|P_k^v(t)\|^2 (D_u v^0, v^0).$$

Оценим величину $\|P_k^v(t)\|$. В данном случае из (3.46) имеем

$$P_k^v(t) = q_0^k \left\{ T_k(x) + \left(\frac{1}{q_0} - x \right) U_{k-1}(x) \right\}, \quad x = \frac{1 - \tilde{r}t}{q_1} \in [-1, 1].$$

Учитывая свойства многочленов Чебышева

$$\begin{aligned} \max_{x \in [-1, 1]} |T_k(x)| &= 1, \quad U_k(-x) = (-1)^k U_k(x), \\ \max_{x \in [-1, 1]} |U_k(x)| &= U(1) = k + 1, \end{aligned}$$

из явной формулы для $P_k^v(t)$ получим

$$\|P_k^v(t)\| \leq q_0^k \left\{ 1 + \left(1 + \frac{1}{q_0} \right) k \right\} = q_0^k \left\{ 1 + \frac{2k}{1 - \sqrt{\xi}} \right\}.$$

Аналогичным образом введем обозначение $P_k^r(t)$ для многочлена $P_k(t)$ при $\kappa = (2 - \tau_0)/\tau_0 > 1$. Тогда из (3.39), (3.44) следует

$$r^k = P_k^r(C^{-1}S_0)r^0$$

и

$$(D_p r^k, r^k) \leq \|P_k^r(t)\|^2 (D_p r^0, r^0).$$

Оценим величину $\|P_k^r(t)\|$. Проводя те же рассуждения, что и выше, получим

$$\|P_k^r(t)\| \leq q_0^k \left\{ 1 + \left(\frac{\kappa q_1}{q_0} - 1 + \frac{\kappa - 1}{q_0} \right) k \right\} = q_0^k \left\{ 1 + \frac{k}{\sqrt{\xi}} \right\}.$$

Теперь, определяя постоянную c формулой

$$c = \max \left\{ \frac{2}{1 - \sqrt{\xi}}, \frac{1}{\sqrt{\xi}} \right\},$$

получим искомую оценку погрешности:

$$\|y^k\|_D^2 = (D_u v^k, v^k) + (D_p r^k, r^k) \leq q_0^{2k} (1 + ck)^2 \|y^0\|_D^2. \quad \blacksquare$$

Прокомментируем полученный результат. Из доказательства теоремы 2.3.3 следует, что при оптимальных значениях итерационных параметров τ_0, α_0 жорданова форма оператора перехода в методе (2.14) содержит клетки второго порядка, поэтому множитель $q_0^k(1 + ck)$ в полученной оценке является асимптотически правильным. Кроме того, в доказательстве теоремы 3.4.1 при получении оценок для норм многочленов $P_k^v(t), P_k^r(t)$ множители при k определялись точно, поэтому постоянная c , вообще говоря, неупрощается.

Заметим также, что при специальном выборе нормы для погрешности ($D_u = A$) от выбора специального начального приближения можно отказаться. Действительно, в силу равенства $(Ag, h) = 0$ для произвольных $g \in G$ и $h \in H$, имеем разложение для первой компоненты ошибки $v^k = h^k + g^k$ и соответствующее равенство

$$(D_u v^k, v^k) = (D_u h^k, h^k) + (D_u g^k, g^k)$$

на произвольной итерации с номером k . Учитывая справедливость неравенства (см. теорему 2.3.3)

$$(D_u h^k, h^k) \leq (1 - \tau_0)^{2k} (D_u h^0, h^0) = q_0^{4k} (D_u h^0, h^0),$$

имеем оценку как в теореме 3.4.1:

$$\|y^k\|_D^2 = (D_u h^k, h^k) + (D_u g^k, g^k) + (D_p r^k, r^k) \leq q_0^{2k} (1 + ck)^2 \|y^0\|_D^2,$$

поскольку

$$q_0^k \leq 1 + ck$$

для любого $k \geq 0$.

3.5. ПОГРЕШНОСТЬ МЕТОДА MSOR В СЛУЧАЕ ПЕРЕМЕННЫХ ПАРАМЕТРОВ

Рассмотрим алгоритм из предыдущего раздела, но с переменными итерационными параметрами:

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k + Bp^k = f, \\ -C \frac{p^{k+1} - p^k}{\nu_{k+1}} + B^T u^{k+1} = \varphi. \end{cases} \quad (3.47)$$

Здесь для удобства используется обозначение $\nu_k = \tau_k / \alpha_k$.

В данном разделе определяются последовательности итерационных параметров, приводящие либо к наилучшей оценке погрешности относительно p :

$$\|p^k - p\|_{D_p} \leq \epsilon_k \|p^0 - p\|_{D_p}, \quad (3.48)$$

либо к наилучшей оценке погрешности относительно u :

$$\|u^{k+1} - u\|_{D_u} \leq \epsilon_k \|u^1 - u\|_{D_u}, \quad (3.49)$$

где

$$\epsilon_k = \frac{2q_0^k}{1 + q_0^{2k}}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Термин «наилучшая оценка» здесь и далее означает, что при заданной информации величину ϵ_k в (3.48) и (3.49) уменьшить теоретически невозможно как для наперед заданного, так и для произвольного номера итерации k .

3.5.1. Преобразование формул

Из формул (3.47) получим отдельные трехслойные соотношения для компонент погрешности $y^k = \{v^k, r^k\}$. Имеет место

Лемма 3.5.6. Пусть $0 < \tau_k \leq 1$. Тогда компоненты погрешности v^{k+1}, r^{k+1} при $k \geq 1$ удовлетворяют соотношениям

$$v^{k+1} = \tilde{\mu}_{k+1}(I - \tilde{\rho}_{k+1}A^{-1}B_0)v^k + (1 - \tilde{\mu}_{k+1})v^{k-1}, \quad (3.50)$$

$$r^{k+1} = \mu_{k+1}(I - \rho_{k+1}C^{-1}S_0)r^k + (1 - \mu_{k+1})r^{k-1}, \quad (3.51)$$

где новые параметры определяются по формулам

$$\tilde{\mu}_{k+1} = 1 + \frac{\tau_{k+1}}{\tau_k}(1 - \tau_k), \quad \tilde{\rho}_{k+1} = \tilde{\mu}_{k+1}^{-1}\tau_{k+1}\nu_k, \quad (3.52)$$

$$\mu_{k+1} = 1 + \frac{\nu_{k+1}}{\nu_k}(1 - \tau_{k+1}), \quad \rho_{k+1} = \mu_{k+1}^{-1}\tau_{k+1}\nu_{k+1}. \quad (3.53)$$

Доказательство. Перепишем формулы итерационного метода (3.47) для погрешности $y^k = \{v^k, r^k\}$

$$\begin{cases} A \frac{v^{k+1} - v^k}{\tau_{k+1}} + Av^k + Br^k = 0, \\ -C \frac{r^{k+1} - r^k}{\nu_{k+1}} + B^T v^{k+1} = 0, \end{cases}$$

откуда имеем

$$v^{k+1} = (1 - \tau_{k+1})v^k - \tau_{k+1}A^{-1}Br^k, \quad (3.54)$$

$$\begin{aligned} r^{k+1} &= r^k + \nu_{k+1}C^{-1}B^T v^{k+1} = \\ &= (I - \tau_{k+1}\nu_{k+1}C^{-1}S_0)r^k + \nu_{k+1}(1 - \tau_{k+1})C^{-1}B^T v^k. \end{aligned} \quad (3.55)$$

Увеличим в выражении (3.55) индекс k на единицу

$$\begin{aligned} r^{k+2} &= (I - \tau_{k+2}\nu_{k+2}C^{-1}S_0)r^{k+1} + \\ &+ \nu_{k+2}(1 - \tau_{k+2})C^{-1}B^T v^{k+1}, \end{aligned} \quad (3.56)$$

выразим из левого равенства (3.55) величину

$$C^{-1}B^T v^{k+1} = \frac{r^{k+1} - r^k}{\nu_{k+1}}$$

и подставим ее в (3.56). Теперь после замены

$$\mu_{k+1} = 1 + \frac{\nu_{k+1}}{\nu_k}(1 - \tau_{k+1}), \quad \rho_{k+1} = \mu_{k+1}^{-1}\tau_{k+1}\nu_{k+1}$$

имеем искомое соотношение для второй компоненты погрешности. Аналогичным образом получается трехслойное соотношение и для первой. Увеличим в (3.54) индекс k на единицу и заменим в полученном выражении r^{k+1} с помощью левой части соотношения (3.55). В результате будем иметь

$$v^{k+2} = [(1 - \tau_{k+2})I - \tau_{k+2}\nu_{k+1}A^{-1}B_0]v^{k+1} - \tau_{k+2}A^{-1}Br^k. \quad (3.57)$$

Выразим из (3.54) величину

$$-A^{-1}Br^k = \frac{v^{k+1} - (1 - \tau_{k+1})v^k}{\tau_{k+1}}$$

и подставим ее в (3.57). Теперь после замены

$$\tilde{\mu}_{k+1} = 1 + \frac{\tau_{k+1}}{\tau_k}(1 - \tau_k), \quad \tilde{\rho}_{k+1} = \tilde{\mu}_{k+1}^{-1}\tau_{k+1}\nu_k$$

искоем соотношение и для первой компоненты погрешности получено. ■

3.5.2. Выбор параметров для p , как в циклическом методе

Выпишем для произвольного фиксированного $N = 1, 2, \dots$ чебышевский набор параметров:

$$\begin{aligned} \tilde{\tau}_k &= \frac{2}{(\gamma + \Gamma)(1 + q_1 \mu_k)}, \quad k = 1, 2, \dots, N, \\ q_1 &= \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma}{\Gamma}, \quad \mu_k \in \mathbb{N}_N = \left\{ -\cos \frac{2i-1}{2N} \pi, \quad i = 1, 2, \dots, N \right\}. \end{aligned} \quad (3.58)$$

Имеет место

Теорема 3.5.1. Пусть в методе (3.47) итерационные параметры для произвольного фиксированного $N = 1, 2, \dots$ выбираются следующим образом: $\tau_k = 1$, $\nu_k = \tilde{\tau}_k$, $k = 1, 2, \dots, N$, из (3.58). Тогда для приближения p^N метода справедлива оценка (3.48)

$$\|p^N - p\|_{D_p} \leq \epsilon_N \|p^0 - p\|_{D_p}.$$

Доказательство. Положим в формулах (3.53) леммы 3.5.6 для итерационных параметров μ_{k+1}, ρ_{k+1} значения τ_k, ν_k , указанные в условии теоремы. Формулы примут вид

$$\mu_k = 1, \quad \rho_k = \tilde{\tau}_k,$$

что, в свою очередь, приводит к соотношению для погрешности r^k :

$$r^{k+1} = (I - \tilde{\tau}_k C^{-1} S_0) r^k.$$

Учитывая расположение спектра $\sigma(C^{-1/2} S_0 C^{-1/2}) \in [\gamma, \Gamma]$, $\gamma > 0$, и результаты пункта 3.1.2, имеем искомую оценку погрешности:

$$\|p^N - p\|_{D_p} \leq \epsilon_N \|p^0 - p\|_{D_p}, \quad \epsilon_N = \frac{2q_0^N}{1 + q_0^{2N}}$$

с

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma}{\Gamma}. \quad \blacksquare$$

Этот результат служит иллюстрацией того факта, что двухслойные алгоритмы типа Узавы (случай $\tau_k = 1$) могут быть записаны в форме (3.47).

3.5.3. Выбор параметров для p , как в трехслойных методах

Определим для получения наилучшей оценки погрешности относительно компоненты p последовательности итерационных параметров в полуитерационном методе Чебышева (3.7):

$$\tilde{\tau}_k = \frac{2}{\gamma + \Gamma}, \quad \tilde{\alpha}_{k+1} = \frac{4}{4 - q_1^2 \tilde{\alpha}_k}, \quad k = 1, 2, \dots, \quad (3.59)$$

где

$$\tilde{\alpha}_1 = 2, \quad q_1 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma}{\Gamma},$$

и в методе сопряженных градиентов:

$$\begin{aligned} \tilde{\tau}_{k+1} &= \frac{(x^k, w^k)}{(w^k, S_0 w^k)}, \quad k = 0, 1, 2, \dots, \\ \tilde{\alpha}_{k+1} &= \left(1 - \frac{\tilde{\tau}_{k+1}}{\tilde{\tau}_k} \frac{(x^k, w^k)}{(x^{k-1}, w^{k-1})} \frac{1}{\tilde{\alpha}_k} \right)^{-1}, \quad k = 1, 2, \dots, \end{aligned} \quad (3.60)$$

где $\alpha_1 = 1$, $w^k = C^{-1}x^k$ — поправка, $r^k = S_0^{-1}x^k = (p^k - p)$ — погрешность, $x^k = (S_0 p^k - B^T A^{-1} f - \varphi)$ — невязка. Напомним, что

$$S_0 = B^T A^{-1} B.$$

Докажем следующее вспомогательное предложение.

Лемма 3.5.7. Пусть для итерационных параметров μ_k, ρ_k , определяемых формулами (3.53) леммы 3.5.6, справедливы неравенства при $k = 1, 2, \dots$

$$\mu_{k+1} > 1, \quad \rho_{k+1} > 0.$$

Тогда обратное преобразование к параметрам τ_k, ν_k исходного алгоритма (3.47) однозначно определяется формулами

$$\begin{aligned} \nu_{k+1} &= \nu_k(\mu_{k+1} - 1) + \mu_{k+1}\rho_{k+1} > 0, \\ 1 > \tau_{k+1} &= \mu_{k+1} \frac{\rho_{k+1}}{\nu_{k+1}} > 0 \end{aligned} \quad (3.61)$$

для $k = 1, 2, \dots$, если $\nu_1 > 0$.

Доказательство. Преобразуем формулу для ρ_{k+1} :

$$\tau_{k+1}\nu_{k+1} = \mu_{k+1}\rho_{k+1}.$$

Теперь для некоторого $\nu_k > 0$ (напомним, что по предположению $\nu_1 > 0$) из формулы для μ_{k+1} имеем

$$\nu_{k+1} = \nu_k(\mu_{k+1} - 1) + \mu_{k+1}\rho_{k+1}.$$

Отсюда, в силу условия леммы, получаем $\nu_{k+1} > 0$, что дает

$$\tau_{k+1} = \mu_{k+1} \frac{\rho_{k+1}}{\nu_{k+1}} > 0.$$

Подстановка в полученную формулу для τ_{k+1} явного вида ν_{k+1} приводит к неравенству $\tau_{k+1} < 1$. ■

На основании имеющихся свойств итерационных параметров получим наилучшую для любого k оценку погрешности относительно компоненты p .

Теорема 3.5.2. Пусть в методе (3.47) итерационные параметры выбираются следующим образом:

$$\begin{aligned}\tau_1 &= 1, \quad \nu_1 = \tilde{\tau}_1, \quad \nu_{k+1} = \nu_k(\tilde{\alpha}_{k+1} - 1) + \tilde{\alpha}_{k+1}\tilde{\tau}_{k+1}, \\ \tau_{k+1} &= \tilde{\alpha}_{k+1}\tilde{\tau}_{k+1}/\nu_{k+1}, \quad k = 1, 2, \dots,\end{aligned}$$

где $\tilde{\tau}_k, \tilde{\alpha}_k$ определяются формулами из полуитерационного метода Чебышева (3.59) или формулами из метода сопряженных градиентов (3.60). Тогда для любого k приближения p^k метода удовлетворяют оценке (3.48)

$$\|p^k - p\|_{D_p} \leq \epsilon_k \|p^0 - p\|_{D_p}.$$

Доказательство. В силу справедливости лемм 3.1.1 и 3.1.2, параметры полуитерационного метода Чебышева (3.59) и методов сопряженных направлений (3.60) удовлетворяют неравенствам

$$\tilde{\tau}_k > 0, \quad \tilde{\alpha}_k > 1, \quad k = 2, 3, \dots,$$

и следовательно, допускают обратное преобразование (3.61). Поэтому выбор параметров по формулам (3.61) в методе (3.47), в силу леммы 3.5.7, порождает для погрешности $r^k = p^k - p$ соотношение (3.51) леммы 3.5.6 с $\mu_{k+1} = \tilde{\alpha}_{k+1}, \rho_{k+1} = \tilde{\tau}_{k+1}$ для $k = 1, 2, \dots$ и $r^1 = (I - \nu_1 C^{-1} S_0) r^0$, что и приводит к искомой оценке погрешности (см. [76], с. 347). ■

Следует отметить, что если дополнительно имеется ограничение $\tilde{\alpha}_{k+1} < 2$ (как в полуитерационном методе Чебышева), то формулы (3.61) устойчивы к ошибкам округлений.

Далее, если необходимо определить вектор u^{k+1} , удовлетворяющий оценке (3.49), достаточно взять $\tau_{k+1} = 1$ и после этого шага завершить вычисления. Таким образом, формулы (3.47), (3.61) обобщают алгоритмы типа Узава на случай использования переменных итерационных параметров.

3.5.4. Выбор параметров для u , как в трехслойных методах

Прежде чем приступить к построению искомого набора итерационных параметров, проанализируем следствие первого шага в методе (3.47) при $\tau_1 = 1$ с произвольного начального приближения $\{u^0, p^0\}$. Из леммы 3.5.6 имеем

$$v^1 = -A^{-1} B r^0, \quad v^2 = (I - \tau_2 \nu_1 A^{-1} B_0) v^1, \quad (3.62)$$

откуда следует, что начальное приближение u^0 фактически не участвует в вычислениях, и следовательно, v^0 — в оценках погрешности. Другими словами, можно считать, что в методе (3.47) величина u^1 выбирается специальным образом, а именно:

$$Au^1 + Bp^0 = f. \quad (3.63)$$

Теперь покажем, что метод (3.47) при начальном приближении вида (3.63) эквивалентен итерированию ошибки v^k в подпространстве G . Имеет место

Лемма 3.5.8. Для любой итерации k первая компонента v^k погрешности итерационного метода (3.47), стартующего с начального приближения вида (3.63), является элементом подпространства G (т. е. $(Av^k, h) = 0$ для $\forall h \in H$).

Доказательство. Проведем рассуждения, как при выборе полезного начального приближения в разделе 1.3.3.

Из соотношения (3.63) следует, что начальная погрешность $\{v^1, r^0\}$ удовлетворяет равенству

$$Av^1 + Br^0 = 0$$

и, следовательно, v^1 является элементом G . Действительно, для произвольного элемента $h \in H$ справедливо $B^T h = 0$, поэтому

$$(Av^1, h) = -(Br^0, h) = -(r^0, B^T h) = 0.$$

Далее покажем, что если $v^k \in G$, то и $v^{k+1} \in G$. Компонента v^k удовлетворяет соотношению

$$v^{k+1} = (1 - \tau_{k+1})v^k - \tau_{k+1}A^{-1}Br^k,$$

поэтому для $\forall h \in H$ имеем

$$(Av^{k+1}, h) = ((1 - \tau_{k+1})Av^k - \tau_{k+1}Br^k, h) = (1 - \tau_{k+1})(Av^k, h).$$

Таким образом, индуктивный переход и начальное условие гарантируют, что для любой итерации k справедливо $v^k \in G$. ■

Отсюда на основании теоремы 1.3.1 и леммы 3.5.6 можно сделать вывод, что метод (3.47) для решения задачи (1.2) $L_0 z = F$, стартующий с начального приближения (3.63), эквивалентен, с точки зрения соотношений для погрешности v^k , методу простой итерации для отыскания решения

$$A^{-1}B_0 u = A^{-1}BC^{-1}\varphi, \quad u \in G. \quad (3.64)$$

При этом оператор $A^{-1}B_0$ симметризуем и на подпространстве G справедливо неравенство

$$\gamma(y, y) \leq (A^{-1/2}B_0A^{-1/2}y, y) \leq \Gamma(y, y) \quad \forall y \in G.$$

Если положить $D_u = B_0$, то аналогично (3.60) можно определить итерационные параметры метода сопряженных градиентов для нахождения u из (3.64):

$$\begin{aligned} \tilde{\tau}_{k+1} &= \frac{(r^k, w^k)}{(w^k, B_0 w^k)}, \quad k = 0, 1, 2, \dots, \\ \tilde{\alpha}_{k+1} &= \left(1 - \frac{\tilde{\tau}_{k+1}}{\tilde{\tau}_k} \frac{(r^k, w^k)}{(r^{k-1}, w^{k-1})} \frac{1}{\tilde{\alpha}_k} \right)^{-1}, \quad k = 1, 2, \dots, \end{aligned} \quad (3.65)$$

где

$$\begin{aligned} \tilde{\alpha}_1 &= 1, \quad w^k = A^{-1}x^k, \\ B_0 v^k &= x^k \quad (\text{причем } v^k \in G), \\ x^k &= B_0 u^k - BC^{-1}\varphi. \end{aligned}$$

Определим теперь алгоритм выбора параметров. Используя свойства полушага с $\tau_1 = 1$, для произвольного p^0 вычислим u^1 . Это дает возможность определить $\tilde{\tau}_2$ по формулам (3.59) или (3.65). В соответствии с (3.62) получаем $\nu_1 \tau_2 = \tilde{\tau}_2$, что порождает некоторый произвол в выборе ν_1 (необходимо, только чтобы $\tau_2 \neq 1$). Зафиксировав каким-либо образом ν_1 , сразу же по формулам (3.47) имеем возможность вычислить p^1 и u^2 . Это, в свою очередь, порождает значения $\tilde{\alpha}_{k+1}, \tilde{\tau}_{k+1}$ при $k = 2$ по формулам (3.59) или (3.65). Теперь, используя обратное преобразование формул леммы 3.5.6, получим

$$\tau_{k+1} = \frac{(\tilde{\mu}_{k+1} - 1)\tau_k}{1 - \tau_k}, \quad \nu_k = \frac{\tilde{\mu}_{k+1}\tilde{\rho}_{k+1}}{\tau_{k+1}}, \quad (3.66)$$

где

$$\tilde{\mu}_{k+1} = \tilde{\alpha}_{k+1}, \quad \tilde{\rho}_{k+1} = \tilde{\tau}_{k+1}$$

из (3.59) или (3.65). Это приводит к нахождению p^k и u^{k+1} для $k = 2$, и далее этот процесс может быть продолжен до достижения искомого результата.

Из формул (3.66) следует, что их применимость определяется условием $0 < \tau_k < 1$, $k = 2, 3, \dots$, что, в свою очередь, зависит от выбора ν_1 . Имеет место

Лемма 3.5.9. Для произвольного $k = 2, 3, \dots, k_0$ существует $0 < \tau_2 = \tau_2(k_0) < 1$ такое, что если $1 < \tilde{\mu}_{k+1} < 2$, то все τ_{k+1} из (3.66) удовлетворяют неравенству $0 < \tau_{k+1} < 1$.

Доказательство. Введем обозначения

$$\Delta_k = \tau_k^{-1}, \quad \delta_{k+1} = (\tilde{\mu}^{k+1} - 1)^{-1}.$$

Обратим внимание, что все $\delta_{k+1} > 1$, и перепишем первую формулу (3.66) в виде

$$\Delta_{k+1} = \delta_{k+1}(\Delta_k - 1), \quad k = 2, 3, \dots, k_0.$$

Это дает

$$\Delta_{k_0+1} = \left(\prod_{j=3}^{k_0+1} \delta_j \right) \Delta_2 - \sum_{j=3}^{k_0+1} \prod_{s=j}^{k_0+1} \delta_s = \left(\prod_{j=3}^{k_0+1} \delta_j \right) \left(\Delta_2 - 1 - \sum_{j=3}^{k_0} \prod_{s=j}^{k_0} \delta_s^{-1} \right).$$

Пусть

$$\Delta_2 \geq 2 + \sum_{j=3}^{k_0} \prod_{s=j}^{k_0} \delta_s^{-1},$$

тогда для любого $k = 2, 3, \dots, k_0$ справедливо

$$\Delta_{k+1} \geq \prod_{j=3}^{k+1} \delta_j > 1,$$

или $0 < \tau_{k+1} < 1$. Добиться же выполнения искомого неравенства для Δ_2 несложно, если взять, например,

$$\tau_2 = \tau_2(k_0) = \Delta_2^{-1} = k_0^{-1}.$$

■

С помощью полученных результатов покажем, что справедлива

Теорема 3.5.3. Для любого фиксированного $k = 1, 2, \dots$ существует набор итерационных параметров: $\tau_1 = 1$, $\tau_2 \nu_1 = \tilde{\tau}_2$, далее

$$\tau_{k+1} = \frac{(\tilde{\mu}_{k+1} - 1)\tau_k}{1 - \tau_k}, \quad \nu_k = \frac{\tilde{\mu}_{k+1}\tilde{\rho}_{k+1}}{\tau_{k+1}},$$

где

$$\tilde{\mu}_{k+1} = \tilde{\alpha}_{k+1}, \quad \tilde{\rho}_{k+1} = \tilde{\tau}_{k+1}$$

из (3.59) или (3.65), такой что приближения u^k метода (3.47) удовлетворяют оценке погрешности (3.49).

Доказательство. Зафиксируем некоторое k_0 и определим

$$\nu_1 = \tilde{\tau}_2 k_0.$$

Тогда, в силу леммы 3.5.9, формулы (3.66) корректны и порождают набор параметров τ_{k+1}, ν_k при $k = 2, 3, \dots, k_0$ для метода (3.47).

В свою очередь, это приводит к соотношению (3.50) леммы 3.5.6 для ошибки $v^k = u^k - u$ с параметрами $\tilde{\mu}_{k+1}, \tilde{\rho}_{k+1}$, гарантирующему для любого $k = 1, 2, \dots, k_0$ оценку

$$\|u^{k+1} - u\|_{D_u} \leq \epsilon_k \|u^1 - u\|_{D_u}. \quad \blacksquare$$

3.6. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Формулы и оценки погрешностей из общей теории итерационных методов решения линейных систем с симметричными положительно определенными матрицами цитируются по книге [76], возможность реализации трехслойных методов в экономичной двухслойной форме указана там же на с. 354.

Для релаксационных методов решения систем с симметричными положительно определенными матрицами оценки погрешностей впервые систематизированы в [200], в более общей форме они приведены в [76].

В случае седловых систем способ получения неухудшаемых оценок погрешностей для релаксационных методов является новым, хотя вывод отдельных трехслойных соотношений для компонент решения из двухслойного алгоритма встречался ранее в [60].

Оценка погрешности метода MJOR с постоянными параметрами получена в [93], а с переменными — в [137].

Оценка погрешности метода MSOR с постоянными параметрами получена в [90], а с переменными — в [91].

Поясним причину невозможности одновременного достижения наилучших оценок для произвольной итерации относительно u и относительно p в рамках алгоритма (3.47). Пусть, например, параметры выбираются для достижения наилучшей оценки относительно p . Тогда, чтобы из нее следовала наилучшая оценка относительно u , необходимо для любого k выполнение соотношения

$$Au^{k+1} + Bp^k = f.$$

Это достигается лишь при $\tau_k = 1$. Но этот случай вырожденный (раздел 3.5.2), так как приводит не к трехслойным соотношениям для r^k , а только к двухслойным (см. лемму 3.5.6). В рамках же двухслойных соотношений для погрешности хорошо известно, что невозможно достигнуть наилучшей оценки для произвольной итерации. Это противоречие и проясняет ситуацию.

Подчеркнем, что невозможность одновременного достижения наилучших оценок относительно u и относительно p касается лишь формул (3.47), т. е. метода с однократным обращением матриц A

и C . В частности, из предыдущих рассуждений следует, что дополнительное обращение матрицы A позволяет получить на любой итерации приближение к u , удовлетворяющее наилучшей оценке. Более того, в процессе реализации метода (3.47) для достижения наилучшей оценки относительно p можно получить приближения \tilde{u}^k , удовлетворяющие неравенству

$$\|\tilde{u}^{k+1} - u\|_{D_u} \leq \epsilon_k \|\tilde{u}^1 - u\|_{D_u},$$

без дополнительных вычислительных затрат, но они в общем случае не будут совпадать с истинными u^k :

$$\tilde{u}^{k+1} = A^{-1}(f - Bp^k) \neq u^{k+1} = (1 - \tau_{k+1})u^k + \tau_{k+1}\tilde{u}^{k+1}.$$

РЕЛАКСАЦИОННЫЕ МЕТОДЫ ДЛЯ СИСТЕМЫ С ПАРАМЕТРОМ

Глава посвящена анализу релаксационных алгоритмов, обобщающих метод MSOR на случай системы уравнений $L_\varepsilon z = F$ с параметром $\varepsilon \geq 0$. Использование значений $\varepsilon > 0$ часто называют *регуляризацией* или *стабилизацией* системы $L_0 z = F$, так как это приводит к улучшению ее спектральных свойств. Естественно выяснить, как это сказывается на сходимости наиболее эффективных из ранее приведенных методов. Для них рассмотрены необходимые и достаточные условия сходимости, сформулированы и решены задачи асимптотической оптимизации. Приведены оценки погрешности явного алгоритма при наилучшем выборе итерационных параметров.

4.1. ЯВНЫЙ МЕТОД ТИПА MSOR (MSOR ε)

Строится и анализируется явный модифицированный метод SOR (метод MSOR ε) для системы уравнений $L_\varepsilon z = F$ с параметром $\varepsilon \geq 0$:

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau} + Au^k + Bp^k = f, \\ -\alpha C \frac{p^{k+1} - p^k}{\tau} + B^T u^{k+1} - \varepsilon C p^k = \varphi. \end{cases} \quad (4.1)$$

4.1.1. Построение метода

Пусть вектор $z = \{u, p\}$ является решением невырожденной алгебраической системы уравнений $L_\varepsilon z = F$ с параметром $\varepsilon \geq 0$:

$$\begin{cases} Au + Bp = f, \\ B^T u - \varepsilon Cp = \varphi. \end{cases}$$

Запишем для ее решения модифицированный, т. е. с двумя итерационными параметрами ω_1, ω_2 , метод SOR

$$\begin{cases} A \frac{u^{k+1} - u^k}{\omega_1} + Au^k + Bp^k = f, \\ -\varepsilon C \frac{p^{k+1} - p^k}{\omega_2} + B^T u^{k+1} - \varepsilon C p^k = \varphi. \end{cases}$$

Положим далее $\omega_2 = \varepsilon \tilde{\omega}_2$ и после формального переобозначения параметров

$$\omega_1 \longrightarrow \tau, \quad \tilde{\omega}_2 \longrightarrow \frac{\tau}{\alpha}$$

будем иметь соотношения (4.1).

4.1.2. Спектр оператора перехода

Обозначим через T оператор перехода в алгоритме (4.1) и рассмотрим спектральную задачу $Tz = \lambda z$:

$$\begin{pmatrix} (1-\tau)I & -\tau A^{-1}B \\ \frac{\tau(1-\tau)}{\alpha} C^{-1}B^T & (1-\frac{\tau\varepsilon}{\alpha})I - \frac{\tau^2}{\alpha} C^{-1}S_0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} u \\ p \end{pmatrix}. \quad (4.2)$$

Имеет место

Теорема 4.1.1. Спектр $\sigma(T)$ оператора перехода T в методе (4.1) принадлежит множеству

$$\Lambda = \{1-\tau\} \cup \left\{ 1-\tau\theta \pm \tau\sqrt{\theta^2 - \frac{t+\varepsilon}{\alpha}}, \theta = \frac{\alpha + \tau t + \varepsilon}{2\alpha} \right\},$$

где $t \in [\gamma, \Gamma]$.

Доказательство. Как и ранее, для вывода формул собственных значений оператора T используем базис пространства Z , построенный в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (4.2) с $\lambda_k^{(1)} = 1 - \tau$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\varkappa^{-1} C^{-1} B^T g_j\}.$$

После применения к первому уравнению (4.2) оператора $C^{-1}B^T$ и замены $C^{-1}B^T g_j = p_j$ будем иметь

$$\begin{cases} (1-\tau)p_j + \tau\varkappa^{-1} C^{-1} S_0 p_j = \lambda p_j, \\ \frac{\tau(1-\tau)}{\alpha} p_j + \varkappa^{-1} \left[\left(1 - \frac{\tau\varepsilon}{\alpha}\right) I - \frac{\tau^2}{\alpha} C^{-1} S_0 \right] p_j = \lambda \varkappa^{-1} p_j. \end{cases}$$

Перепишем эту систему в виде

$$\begin{cases} S_0 p_j = \frac{\kappa(\lambda - 1 + \tau)}{\tau} C p_j, \\ S_0 p_j = \frac{\alpha(1 - \lambda) - \kappa\tau(1 - \tau) - \tau\varepsilon}{\tau^2} C p_j. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0 p = t C p$ соответствует собственное значение $t_j, j = 1, \dots, N_p$ по теореме 1.3.1. Зафиксировав его, из полученной системы для λ и κ выведем соотношения

$$t_j = \frac{\kappa(\lambda - 1 + \tau)}{\tau} = \frac{\alpha(1 - \lambda) - \kappa\tau(1 - \tau) - \tau\varepsilon}{\tau^2}.$$

Исключая из этих уравнений κ , приходим к выражению

$$\lambda_j^{(2,3)} = 1 - \tau\theta_j \pm \tau\sqrt{\theta_j^2 - \frac{t_j + \varepsilon}{\alpha}},$$

где

$$\theta_j = \frac{1 + t_j\tau/\alpha + \varepsilon/\alpha}{2}$$

и, соответственно,

$$\kappa_j^{(2,3)} = \frac{\alpha}{\lambda_j^{(2,3)}} \left(\theta_j \mp \sqrt{\theta_j^2 - \frac{t_j + \varepsilon}{\alpha}} \right).$$

Завершение доказательства, как в теореме 2.2.1. Единственное отличие состоит только в виде корневого вектора $z_j^{(3)}$ для случая $\lambda_j^{(2)} = \lambda_j^{(3)} = \lambda_j$:

$$z_j^{(3)} = \left\{ g_j, -p_j \frac{\lambda_j + \tau}{t_j \tau} \right\}. \quad \blacksquare$$

4.1.3. Условие сходимости

Выберем начальное приближение $\{u^0, p^0\}$ из условия (1.9):

$$A u^0 + B p^0 = f.$$

Для такого начального приближения, в силу леммы 1.3.2, первая компонента v^k погрешности y^k итерационного метода (2.9) на любой итерации k является элементом подпространства G (т. е. $(A v^k, h) = 0$ для $\forall h \in H$). Это дает возможность определить условия сходимости метода (4.1). Вначале нам потребуется

Лемма 4.1.1. *Нелинейная замена итерационных параметров*

$$\tilde{\tau} = 1 - (1 - \tau) \left(1 - \tau \frac{\varepsilon}{\alpha} \right), \quad \tilde{\alpha} = \alpha \frac{\tilde{\tau}^2}{\tau^2}$$

порождает следующую параметризацию спектра $\sigma(T)$ оператора перехода T в методе (4.1) при использовании начального приближения вида (1.9):

$$\Lambda = \{0\} \cup \left\{ 1 - \tilde{\tau}\theta \pm \tilde{\tau} \sqrt{\theta^2 - \frac{\tilde{t}}{\tilde{\alpha}}}, \theta = \frac{\tilde{\alpha} + \tilde{\tau}\tilde{t}}{2\tilde{\alpha}}, \tilde{t} \in [\gamma + \varepsilon, \Gamma + \varepsilon] \right\}.$$

Доказательство. Если применяется начальное приближение вида (1.9), то на основании леммы 1.3.2 первая часть спектра оператора перехода T в методе (4.1) при анализе сходимости не используется, т. е. величину $(1 - \tau)$ можно заменить на нуль. Вторая часть спектра описывается уравнением

$$\lambda^2 - \lambda \left(2 - \tau - \frac{\tau^2 t}{\alpha} - \frac{\tau \varepsilon}{\alpha} \right) + (1 - \tau) \left(1 - \frac{\tau \varepsilon}{\alpha} \right) = 0.$$

К искомому результату приводят две последовательные подстановки. Сначала выражение $(1 - \tau)(1 - (\tau \varepsilon / \alpha))$ меняем на $1 - \tilde{\tau}$. Это дает

$$\lambda^2 - \lambda \left(2 - \tilde{\tau} - \tau^2 \frac{t + \varepsilon}{\alpha} \right) + 1 - \tilde{\tau} = 0.$$

Теперь используем подстановку

$$\frac{\tau^2}{\alpha} = \frac{\tilde{\tau}^2}{\tilde{\alpha}}$$

и обозначение $\tilde{t} = t + \varepsilon$. ■

Смысл этих формальных преобразований заключается в сведении интересующих задач к уже решенным.

Теорема 4.1.2. При любом $\tilde{\alpha} > 0$ и произвольном начальном приближении вида (1.9) необходимым и достаточным условием сходимости метода (4.1) является выполнение неравенства

$$0 < \tilde{\tau} < \sqrt{\frac{\tilde{\alpha}^2}{(\Gamma + \varepsilon)^2} + \frac{4\tilde{\alpha}}{\Gamma + \varepsilon}} - \frac{\tilde{\alpha}}{\Gamma + \varepsilon},$$

где новые параметры $\tilde{\tau}$, $\tilde{\alpha}$ определяются по формулам:

$$\tilde{\tau} = 1 - (1 - \tau) \left(1 - \frac{\tau \varepsilon}{\alpha} \right), \quad \tilde{\alpha} = \alpha \frac{\tilde{\tau}^2}{\tau^2}.$$

Доказательство. Построенная в лемме 4.1.1 параметризация спектра оператора перехода и теорема 2.3.2 о принадлежности единичному кругу произвольной точки множества Λ при условии (2.16), примененная на отрезке $[\gamma + \varepsilon, \Gamma + \varepsilon]$ в терминах $\tilde{\tau}$, $\tilde{\alpha}$, гарантируют справедливость сформулированного утверждения. ■

4.1.4. Задача асимптотической оптимизации

Знание аналитического представления спектра оператора перехода позволяет сформулировать и решить задачу асимптотической оптимизации метода: *найти положительные значения τ_0 и α_0 , минимизирующие спектральный радиус оператора перехода*. Имеет место

Теорема 4.1.3. *Спектральный радиус q_0 оператора перехода и асимптотически оптимальные параметры τ_0, α_0 в итерационном методе (4.1) при произвольном начальном приближении вида (1.9) определяются по формулам*

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \alpha_0 = \frac{\tilde{\alpha}_0 \tau_0^2}{\tilde{\tau}_0^2},$$

$$\tau_0 = \tilde{\tau}_0 \left[\frac{\tilde{\alpha}_0 + \tilde{\tau}_0 \varepsilon}{2\tilde{\alpha}_0} + \sqrt{\left(\frac{\tilde{\alpha}_0 + \tilde{\tau}_0 \varepsilon}{2\tilde{\alpha}_0} \right)^2 - \frac{\varepsilon}{\tilde{\alpha}_0}} \right],$$

где

$$\tilde{\tau}_0 = \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}, \quad \tilde{\alpha}_0 = \frac{4(\gamma + \varepsilon)}{(1 + \sqrt{\xi})^2}, \quad \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon}.$$

Доказательство. Аналогично предыдущему случаю, ключевую роль играют утверждения леммы 4.1.1 и теоремы 2.3.3, решающей поставленную оптимизационную задачу в терминах « $\tilde{\tau} - \tilde{\alpha}$ » на отрезке $[\gamma + \varepsilon, \Gamma + \varepsilon]$. Поэтому при нахождении выражений для оптимальных параметров достаточно выполнить преобразование перехода к исходным параметрам τ, α и проверить при этом неотрицательность подкоренного выражения в формуле для τ_0 . С этой целью первое соотношение в лемме 4.1.1 возьмем в виде

$$\frac{\tau^2 \varepsilon}{\alpha} - \tau \left(1 + \frac{\varepsilon}{\alpha} \right) + \tilde{\tau} = 0.$$

Поскольку $\tau^2/\alpha = \tilde{\tau}^2/\tilde{\alpha}$, то из предыдущего равенства имеем

$$\tau^2 - \tau \left(\tilde{\tau} + \frac{\tilde{\tau}^2 \varepsilon}{\tilde{\alpha}} \right) + \frac{\tilde{\tau}^2 \varepsilon}{\tilde{\alpha}} = 0,$$

а знак «+» в выражении для τ выбирается по непрерывности со случаем $\varepsilon = 0$.

Теперь запишем подкоренное выражение в виде

$$\begin{aligned} \frac{1}{4} \left(1 + \frac{\sqrt{\xi} \varepsilon}{\gamma + \varepsilon} \right)^2 - \frac{(1 + \sqrt{\xi})^2 \varepsilon}{4(\gamma + \varepsilon)} &= \frac{1}{4} \left(1 - \varepsilon \frac{1 + \xi}{\gamma + \varepsilon} + \varepsilon^2 \frac{\xi}{(\gamma + \varepsilon)^2} \right) = \\ &= \frac{1}{4} \left(1 - \frac{\varepsilon}{\gamma + \varepsilon} \right) \left(1 - \frac{\varepsilon \xi}{\gamma + \varepsilon} \right) \geq \frac{1}{4} \left(1 - \frac{\varepsilon}{\gamma + \varepsilon} \right)^2 \geq 0. \end{aligned}$$

Напомним, что в силу выбора начального приближения, часть собственных значений оператора перехода вида $(1 - \tau)$ в решении задачи асимптотической оптимизации не используется. ■

4.2. НЕЯВНЫЙ МЕТОД ТИПА MSOR (IMSORe)

Изучается неявный модифицированный метод SOR (метод IMSORe) для системы уравнений $L_\varepsilon z = F$ с параметром $\varepsilon \geq 0$:

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau} + Au^{k+1} + Bp^k = f, \\ -\alpha C \frac{p^{k+1} - p^k}{\tau} + B^T u^{k+1} - \varepsilon C p^{k+1} = \varphi. \end{cases} \quad (4.3)$$

4.2.1. Построение метода

Пусть вектор $z = \{u, p\}$ является решением невырожденной алгебраической системы уравнений $L_\varepsilon z = F$ с параметром $\varepsilon \geq 0$:

$$\begin{cases} Au + Bp = f, \\ B^T u - \varepsilon Cp = \varphi. \end{cases}$$

Запишем для ее решения неявный модифицированный, т. е. с двумя итерационными параметрами ω_1, ω_2 , метод SOR

$$\begin{cases} A \frac{u^{k+1} - u^k}{\omega_1} + Au^{k+1} + Bp^k = f, \\ -\varepsilon C \frac{p^{k+1} - p^k}{\omega_2} + B^T u^{k+1} - \varepsilon Cp^{k+1} = \varphi. \end{cases}$$

Отметим, что в отличие от метода MSORe эта процедура не является простым обобщением метода MSOR на случай $\varepsilon \neq 0$. Положим далее $\omega_2 = \varepsilon \tilde{\omega}_2$ и после формального переобозначения параметров

$$\omega_1 \rightarrow \tau, \quad \tilde{\omega}_2 \rightarrow \frac{\tau}{\alpha}$$

будем иметь соотношения (4.3).

4.2.2. Спектр оператора перехода

Обозначим через T оператор перехода в алгоритме (4.3) и рассмотрим спектральную задачу $Tz = \lambda z$:

$$\begin{aligned} \frac{1}{1 + \tau} \left(\begin{array}{cc} I & -\tau A^{-1} B \\ \frac{\tau}{\alpha + \varepsilon \tau} C^{-1} B^T & \frac{1}{\alpha + \varepsilon \tau} (\alpha(1 + \tau)I - \tau^2 C^{-1} S_0) \end{array} \right) \begin{pmatrix} u \\ p \end{pmatrix} = \\ = \lambda \begin{pmatrix} u \\ p \end{pmatrix}. \end{aligned} \quad (4.4)$$

Имеет место

Теорема 4.2.1. Спектр $\sigma(T)$ оператора перехода T в методе (4.3) принадлежит множеству

$$\Lambda = \left\{ \frac{1}{1+\tau} \right\} \cup \left\{ \frac{1+\tau\theta \pm \tau\sqrt{\theta^2 - \frac{t+\varepsilon}{\alpha}}}{(1+\tau)(1+\tau\varepsilon/\alpha)}, \theta = \frac{\alpha - \tau t + \varepsilon}{2\alpha} \right\},$$

где $t \in [\gamma, \Gamma]$.

Доказательство. Получим формулы собственных значений оператора T на основе базиса пространства Z , построенного в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (4.4) с $\lambda_k^{(1)} = (1+\tau)^{-1}$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^T g_j\}.$$

После применения к первому уравнению (4.2) оператора $C^{-1}B^T$ и замены $C^{-1}B^T g_j = p_j$ будем иметь

$$\begin{cases} p_j + \tau\kappa^{-1}C^{-1}S_0 p_j = \lambda(1+\tau)p_j, \\ -\tau p_j + \kappa^{-1}[\alpha(1+\tau)I - \tau^2 C^{-1}S_0] p_j = \lambda\kappa^{-1}(1+\tau)(\alpha + \varepsilon\tau)p_j. \end{cases}$$

Перепишем эту систему в виде

$$\begin{cases} S_0 p_j = \frac{\kappa[\lambda(1+\tau) - 1]}{\tau} C p_j, \\ S_0 p_j = \frac{\alpha(1+\tau) - \lambda(1+\tau)(\alpha + \varepsilon\tau) - \kappa\tau}{\tau^2} C p_j. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0 p = t C p$ соответствует собственное значение t_j , $j = 1, \dots, N_p$ по теореме 1.3.1. Зафиксировав его, из полученной системы для λ и κ выведем соотношения

$$t_j = \frac{\kappa[\lambda(1+\tau) - 1]}{\tau} = \frac{\alpha(1+\tau) - \lambda(1+\tau)(\alpha + \varepsilon\tau) - \kappa\tau}{\tau^2}.$$

Исключая из этих уравнений κ , приходим к выражению

$$\lambda_j^{(2,3)} = \frac{1+\tau\theta_j \pm \tau\sqrt{\theta_j^2 - \frac{t_j+\varepsilon}{\alpha}}}{(1+\tau)(1+\varepsilon\tau/\alpha)},$$

где

$$\theta_j = \frac{1 - t_j\tau/\alpha + \varepsilon/\alpha}{2}$$

и, соответственно,

$$x_j^{(2,3)} = -\frac{\alpha}{\lambda_j^{(2,3)}} \left(\theta_j \pm \sqrt{\theta_j^2 - \frac{t_j + \varepsilon}{\alpha}} \right) - \varepsilon.$$

Завершение доказательства такое же, как в теореме 2.2.1. Единственное отличие состоит только в виде корневого вектора $z_j^{(3)}$ в случае $\lambda_j^{(2)} = \lambda_j^{(3)} = \lambda_j$:

$$z_j^{(3)} = \left\{ g_j, -p_j \frac{\lambda_j \tau + \lambda_j + \tau}{t_j \tau} \right\}. \quad \blacksquare$$

4.2.3. Условие сходимости

Выберем начальное приближение $\{u^0, p^0\}$ из условия (1.9):

$$Au^0 + Bp^0 = f.$$

Для такого начального приближения справедлива

Лемма 4.2.2. Для любой итерации k первая компонента v^k погрешности y^k итерационного метода (4.3), стартующего с начального приближения вида (1.9), является элементом подпространства G (т. е. $(Av^k, h) = 0$ для $\forall h \in H$).

Доказательство. Из соотношения (1.9) следует, что начальная погрешность $y^0 = \{v^0, r^0\}$ удовлетворяет равенству

$$Av^0 + Br^0 = 0,$$

и следовательно, v^0 является элементом G . Действительно, для произвольного элемента $h \in H$ справедливо $B^T h = 0$, поэтому

$$(Av^0, h) = -(Br^0, h) = -(r^0, B^T h) = 0.$$

Далее покажем, что если $v^k \in G$, то и $v^{k+1} \in G$. Компонента v^k удовлетворяет соотношению

$$v^{k+1} = (1 + \tau)^{-1} (v^k - \tau A^{-1} B r^k),$$

поэтому для $\forall h \in H$ имеем

$$(Av^{k+1}, h) = (1 + \tau)^{-1} (Av^k - \tau B r^k, h) = (1 + \tau)^{-1} (Av^k, h).$$

Таким образом, индуктивный переход и начальное условие гарантируют, что для любой итерации k справедливо $v^k \in G$. ■

Из леммы 4.2.2 и теоремы 1.3.2 вытекает покомпонентное разложение погрешности следующего вида:

$$v^k = \sum_{i=1}^{N_p} c_i^{(k)} g_i, \quad r^k = \sum_{i=1}^{N_p} d_i^{(k)} p_i,$$

где $\{g_i\}$ и $\{p_i\}$ — базисы пространств G и P , порожденные задачами (1.7) и (1.8). Найденное представление дает возможность получить условия сходимости метода (4.3). Вначале нам потребуется

Лемма 4.2.3. *Нелинейная замена итерационных параметров*

$$\tilde{\tau} = 1 - \frac{1}{(1 + \tau)(1 + \tau\varepsilon/\alpha)}, \quad \tilde{\alpha} = \frac{\alpha\tilde{\tau}^2}{\tau^2}(1 + \tau)\left(1 + \frac{\tau\varepsilon}{\alpha}\right)$$

порождает следующую параметризацию спектра $\sigma(T)$ оператора перехода T в методе (4.3) при использовании начального приближения вида (1.9):

$$\Lambda = \{0\} \cup \left\{ 1 - \tilde{\tau}\theta \pm \tilde{\tau}\sqrt{\theta^2 - \frac{\tilde{t}}{\tilde{\alpha}}}, \theta = \frac{\tilde{\alpha} + \tilde{\tau}\tilde{t}}{2\tilde{\alpha}}, \tilde{t} \in [\gamma + \varepsilon, \Gamma + \varepsilon] \right\}.$$

Доказательство. Если применяется начальное приближение вида (1.9), то на основании леммы 1.3.2 первая часть спектра оператора перехода T в методе (4.3) при анализе сходимости не используется, т. е. величину $(1 + \tau)^{-1}$ можно заменить на нуль. Вторая часть спектра описывается уравнением

$$\lambda^2 - \lambda\left(2 + \tau - \frac{\tau^2 t}{\alpha} + \frac{\tau\varepsilon}{\alpha}\right) + (1 + \tau)^{-1}\left(1 + \frac{\tau\varepsilon}{\alpha}\right)^{-1} = 0.$$

К искомому результату приводят две последовательные подстановки. Сначала выражение

$$(1 + \tau)^{-1}\left(1 + \frac{\tau\varepsilon}{\alpha}\right)^{-1}$$

меняем на $1 - \tilde{\tau}$. Это дает

$$\lambda^2 - \lambda\left(2 - \tilde{\tau} - \frac{\tau^2(t + \varepsilon)}{\alpha(1 + \tau)(1 + \tau\varepsilon/\alpha)}\right) + 1 - \tilde{\tau} = 0.$$

Теперь используем подстановку

$$\frac{\tau^2}{\alpha}(1 + \tau)^{-1}\left(1 + \frac{\tau\varepsilon}{\alpha}\right)^{-1} = \frac{\tilde{\tau}^2}{\tilde{\alpha}}$$

и обозначение $\tilde{t} = t + \varepsilon$. ■

Основной результат о сходимости метода формулируется следующим образом:

Теорема 4.2.2. При любом $\tilde{\alpha} > 0$ и произвольном начальном приближении вида (1.9) необходимым и достаточным условием сходимости метода (4.3) является выполнение неравенства

$$0 < \tilde{\tau} < \sqrt{\frac{\tilde{\alpha}^2}{(\Gamma + \varepsilon)^2} + \frac{4\tilde{\alpha}}{\Gamma + \varepsilon}} - \frac{\tilde{\alpha}}{\Gamma + \varepsilon},$$

где новые параметры $\tilde{\tau}$, $\tilde{\alpha}$ определяются по формулам

$$\tilde{\tau} = 1 - \frac{1}{(1 + \tau)(1 + \tau\varepsilon/\alpha)}, \quad \tilde{\alpha} = \frac{\alpha\tilde{\tau}^2}{\tau^2}(1 + \tau)\left(1 + \frac{\tau\varepsilon}{\alpha}\right).$$

Доказательство. Построенная в лемме 4.2.3 параметризация спектра оператора перехода и теорема 2.3.2 о принадлежности единичному кругу произвольной точки множества Λ при условии (2.16), примененная на отрезке $[\gamma + \varepsilon, \Gamma + \varepsilon]$ в терминах $\tilde{\tau}$, $\tilde{\alpha}$, гарантируют справедливость сформулированного утверждения. ■

4.2.4. Задача асимптотической оптимизации

Знание аналитического представления спектра оператора перехода позволяет сформулировать и решить задачу асимптотической оптимизации метода: найти положительные значения τ_0 и α_0 , минимизирующие спектральный радиус оператора перехода. Имеет место

Теорема 4.2.3. Спектральный радиус q_0 оператора перехода и асимптотически оптимальные параметры τ_0, α_0 в итерационном методе (4.3) при произвольном начальном приближении вида (1.9) определяются по формулам

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \alpha_0 = \frac{\tilde{\alpha}_0\tau_0^2}{\tilde{\tau}_0^2(1 + \tau_0)} - \tau_0\varepsilon,$$

$$\tau_0 = \frac{\tilde{\tau}_0}{1 - \tilde{\tau}_0} \left[\frac{\tilde{\alpha}_0 - \tilde{\tau}_0\varepsilon}{2\tilde{\alpha}_0} + \sqrt{\left(\frac{\tilde{\alpha}_0 - \tilde{\tau}_0\varepsilon}{2\tilde{\alpha}_0}\right)^2 - \frac{\varepsilon(1 - \tilde{\tau}_0)}{\tilde{\alpha}_0}} \right],$$

где

$$\tilde{\tau}_0 = \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}, \quad \tilde{\alpha}_0 = \frac{4(\gamma + \varepsilon)}{(1 + \sqrt{\xi})^2}, \quad \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon}.$$

Доказательство. Аналогично доказательству теоремы 4.1.3, здесь ключевую роль играют утверждения леммы 4.2.3 и теоремы 2.3.3,

решающей поставленную оптимизационную задачу в терминах « $\tilde{\tau} - \tilde{\alpha}$ » на отрезке $[\gamma + \varepsilon, \Gamma + \varepsilon]$. Поэтому при нахождении выражений для оптимальных параметров достаточно выполнить преобразование перехода к исходным параметрам τ, α и проверить при этом неотрицательность подкоренного выражения в формуле для τ_0 . Для этого первое соотношение в лемме 4.2.3 возьмем в виде

$$\frac{\tau^2 \varepsilon}{\alpha} + \tau \left(1 + \frac{\varepsilon}{\alpha}\right) - \frac{\tilde{\tau}}{1 - \tilde{\tau}} = 0.$$

Поскольку

$$\frac{\tau^2}{\alpha} = \frac{\tilde{\tau}^2 / \tilde{\alpha}}{1 - \tilde{\tau}},$$

то из предыдущего равенства имеем

$$\tau^2 - \tau \left(1 - \frac{\tilde{\tau} \varepsilon}{\tilde{\alpha}}\right) \frac{\tilde{\tau}}{1 - \tilde{\tau}} + \frac{\tilde{\tau}^2}{(1 - \tilde{\tau})^2} \frac{(1 - \tilde{\tau}) \varepsilon}{\tilde{\alpha}} = 0,$$

а знак «+» в выражении для τ выбирается по непрерывности со случаем $\varepsilon = 0$. Теперь запишем подкоренное выражение в виде

$$\begin{aligned} \frac{1}{4} \left(1 - \frac{\sqrt{\xi} \varepsilon}{\gamma + \varepsilon}\right)^2 - \frac{(1 - \sqrt{\xi})^2 \varepsilon}{4(\gamma + \varepsilon)} &= \frac{1}{4} \left(1 - \varepsilon \frac{1 + \xi}{\gamma + \varepsilon} + \varepsilon^2 \frac{\xi}{(\gamma + \varepsilon)^2}\right) = \\ &= \frac{1}{4} \left(1 - \frac{\varepsilon}{\gamma + \varepsilon}\right) \left(1 - \frac{\varepsilon \xi}{\gamma + \varepsilon}\right) \geq \frac{1}{4} \left(1 - \frac{\varepsilon}{\gamma + \varepsilon}\right)^2 \geq 0. \quad \blacksquare \end{aligned}$$

Полученные результаты свидетельствуют о том, что усложнение реализации (в методе IMSORe по сравнению с методом MSORe) в данном случае не приводит к ускорению сходимости.

4.3. ПОГРЕШНОСТЬ МЕТОДА MSORe В СЛУЧАЕ ПОСТОЯННЫХ ПАРАМЕТРОВ

Воспользуемся в этом разделе обозначениями для погрешности решения задачи $L_\varepsilon z = F$ и ее нормы из главы 3:

$$\begin{aligned} y^k &= \{v^k, r^k\} = \{u^k - u, p^k - p\}, \\ \|y\|_D^2 &= (D_u v, v) + (D_p r, r), \quad y = \{v, r\} \in Z. \end{aligned}$$

4.3.1. Преобразование формул

Из формул (4.1) получим отдельные трехслойные соотношения для компонент погрешности v^k, r^k . Имеет место

Лемма 4.3.4. Компоненты погрешности v^{k+1}, r^{k+1} при $k \geq 1$ в области сходимости метода удовлетворяют соотношениям

$$v^{k+1} = \tilde{\alpha}(I - \tilde{\tau}(A^{-1}B_0 + \varepsilon I))v^k + (1 - \tilde{\alpha})v^{k-1}, \quad (4.5)$$

$$r^{k+1} = \tilde{\alpha}(I - \tilde{\tau}(C^{-1}S_0 + \varepsilon I))r^k + (1 - \tilde{\alpha})r^{k-1}, \quad (4.6)$$

где новые параметры определяются формулами

$$\tilde{\alpha} = 1 + (1 - \tau) \left(1 - \frac{\tau\varepsilon}{\alpha} \right), \quad \tilde{\tau} = \frac{\tau^2/\alpha}{\tilde{\alpha}}.$$

Доказательство. Перепишем формулы итерационного метода (4.1) для погрешности $y^k = \{v^k, r^k\}$

$$\begin{cases} A \frac{v^{k+1} - v^k}{\tau} + Av^k + Br^k = 0, \\ -\alpha C \frac{r^{k+1} - r^k}{\tau} + B^T v^{k+1} - \varepsilon C r^k = 0, \end{cases}$$

откуда имеем

$$v^{k+1} = (1 - \tau)v^k - \tau A^{-1} B r^k, \quad (4.7)$$

$$\begin{aligned} r^{k+1} &= \left(1 - \frac{\tau\varepsilon}{\alpha} \right) r^k + \frac{\tau}{\alpha} C^{-1} B^T v^{k+1} = \\ &= \left[\left(1 - \frac{\tau\varepsilon}{\alpha} \right) I - \frac{\tau^2}{\alpha} C^{-1} S_0 \right] r^k + \frac{\tau(1 - \tau)}{\alpha} C^{-1} B^T v^k. \end{aligned} \quad (4.8)$$

Увеличим в выражении (4.8) индекс k на единицу

$$r^{k+2} = \left[\left(1 - \frac{\tau\varepsilon}{\alpha} \right) I - \frac{\tau^2}{\alpha} C^{-1} S_0 \right] r^{k+1} + \frac{\tau(1 - \tau)}{\alpha} C^{-1} B^T v^{k+1}, \quad (4.9)$$

выразим из левого равенства (4.8) величину

$$\frac{\tau}{\alpha} C^{-1} B^T v^{k+1} = r^{k+1} - \left(1 - \frac{\tau\varepsilon}{\alpha} \right) r^k$$

и подставим ее в (4.9):

$$r^{k+2} = \left[\left(2 - \tau - \frac{\tau\varepsilon}{\alpha} \right) I - \frac{\tau^2}{\alpha} C^{-1} S_0 \right] r^{k+1} - (1 - \tau) \left(1 - \frac{\tau\varepsilon}{\alpha} \right) r^k.$$

Теперь после замены

$$\tilde{\alpha} = 1 + (1 - \tau) \left(1 - \frac{\tau\varepsilon}{\alpha} \right), \quad \tilde{\tau} = \frac{\tau^2/\alpha}{\tilde{\alpha}},$$

корректной в области сходимости метода (см. теорему 4.1.2), искомое соотношение для второй компоненты погрешности получено. Аналогичным образом получается трехслойное соотношение и для первой.

Увеличим в (4.7) индекс k на единицу и заменим в полученном выражении r^{k+1} с помощью левой части соотношения (4.8). В результате получим

$$v^{k+2} = \left(1 - \tau - \frac{\tau^2}{\alpha} A^{-1} B_0\right) v^{k+1} - \left(1 - \frac{\tau\varepsilon}{\alpha}\right) \tau A^{-1} B r^k. \quad (4.10)$$

Выразим из (4.7) величину

$$-\tau A^{-1} B r^k = v^{k+1} - (1 - \tau) v^k$$

и подставим ее в (4.10):

$$v^{k+2} = \left[\left(2 - \tau - \frac{\tau\varepsilon}{\alpha}\right) I - \frac{\tau^2}{\alpha} A^{-1} B_0\right] v^{k+1} - (1 - \tau) \left(1 - \frac{\tau\varepsilon}{\alpha}\right) v^k.$$

Теперь после замены

$$\tilde{\alpha} = 1 + (1 - \tau) \left(1 - \frac{\tau\varepsilon}{\alpha}\right), \quad \tilde{\tau} = \frac{\tau^2/\alpha}{\tilde{\alpha}},$$

корректной в области сходимости метода (см. теорему 4.1.2), искомое соотношение и для первой компоненты погрешности получено. ■

4.3.2. Начальное приближение

Выберем начальное приближение $\{u^0, p^0\}$ из условия (1.9):

$$A u^0 + B p^0 = f.$$

Для такого начального приближения имеем:

$$v^1 = v^0, \quad (4.11)$$

$$\begin{aligned} r^1 &= \left(I - \frac{\tau}{\alpha} [C^{-1} S_0 + \varepsilon I]\right) r^0 = \\ &= \left(I - \tilde{\tau} \frac{1 + (1 - \tau)(1 - \tau\varepsilon/\alpha)}{\tau} [C^{-1} S_0 + \varepsilon I]\right) r^0, \end{aligned} \quad (4.12)$$

и, кроме того, в силу леммы 1.3.2 первая компонента v^k погрешности y^k итерационного метода (2.9) для любой итерации k является элементом подпространства G (т. е. $(A v^k, h) = 0$ для $\forall h \in H$).

4.3.3. Полином ошибки

Положим в формулах для $\tilde{\alpha}, \tilde{\tau}$ асимптотически оптимальные значения α_0, τ_0 . Теперь для получения оценок каждой из компонент

погрешности достаточно на отрезке $[\gamma + \varepsilon, \Gamma + \varepsilon]$ найти алгебраический полином $P_k(t)$, определяемый соотношениями (3.45)

$$\begin{aligned} P_{k+1}(t) &= \tilde{\alpha}(1 - \tilde{\tau}t)P_k(t) + (1 - \tilde{\alpha})P_{k-1}(t), \\ P_1(t) &= 1 - \kappa\tilde{\tau}t, \quad P_0(t) = 1 \end{aligned}$$

(параметр κ может принимать значения 0 или $(1 + (1 - \tau_0)(1 - \tau_0\varepsilon/\alpha_0))/\tau_0$), и определить его норму

$$\|P_k(t)\| = \max_{t \in [\gamma + \varepsilon, \Gamma + \varepsilon]} |P_k(t)|.$$

Имеет место

Лемма 4.3.5. Многочлен $P_k(t)$, определяемый соотношениями (3.45), имеет вид

$$P_k(t) = q_0^k \left\{ T_k(x) + \left[\left(\frac{\kappa q_1}{q_0} - 1 \right) x + \frac{1 - \kappa}{q_0} \right] U_{k-1}(x) \right\},$$

где

$$x = \frac{1 - \tilde{\tau}t}{q_1}, \quad q_1 = \frac{1 - \xi}{1 + \xi}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon}.$$

Доказательство. Полагая

$$t = \frac{(1 - q_1 x)}{\tilde{\tau}},$$

отобразим отрезок $[\gamma + \varepsilon, \Gamma + \varepsilon]$ на $[-1, 1]$. Продолжение дословно совпадает с доказательством леммы (3.4.5). ■

4.3.4. Оценка погрешности

Теорема 4.3.1. Итерационный метод (4.1) с асимптотически оптимальными параметрами τ_0, α_0 и спектральным радиусом оператора перехода

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon},$$

стартующий с начального приближения вида (1.9), сходится с оценкой погрешности

$$\|y^k\|_D \leq q_0^k (1 + ck) \|y^0\|_D,$$

где постоянная $c > 0$ не зависит от номера итерации.

Доказательство дословно совпадает с доказательством теоремы 3.4.1, отличие состоит только в определении ξ .

4.4. ПОГРЕШНОСТЬ МЕТОДА MSORe В СЛУЧАЕ ПЕРЕМЕННЫХ ПАРАМЕТРОВ

Ниже выводятся оценки погрешности для метода MSORe с переменными итерационными параметрами:

$$\begin{cases} A \frac{u^{k+1} - u^k}{\tau_{k+1}} + Au^k + Bp^k = f, \\ -C \frac{p^{k+1} - p^k}{\nu_{k+1}} + B^T u^{k+1} - \varepsilon C p^k = \varphi. \end{cases} \quad (4.13)$$

Здесь для удобства используется обозначение $\nu_k = \tau_k / \alpha_k$.

В данном разделе определяются последовательности итерационных параметров, приводящие либо к наилучшей оценке погрешности относительно p :

$$\|p^k - p\|_{D_p} \leq \epsilon_k \|p^0 - p\|_{D_p}, \quad (4.14)$$

либо к наилучшей оценке погрешности относительно u :

$$\|u^{k+1} - u\|_{D_u} \leq \epsilon_k \|u^1 - u\|_{D_u}, \quad (4.15)$$

где

$$\epsilon_k = \frac{2q_0^k}{1 + q_0^{2k}}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon}.$$

4.4.1. Преобразование формул

Из формул (4.13) получим отдельные трехслойные соотношения для компонент погрешности v^k, r^k . Имеет место

Лемма 4.4.6. Пусть $0 < \tau_k \leq 1$, $0 < \nu_k < \varepsilon^{-1}$. Тогда компоненты погрешности v^{k+1}, r^{k+1} при $k \geq 1$ удовлетворяют соотношениям

$$v^{k+1} = \tilde{\mu}_{k+1}(I - \tilde{\rho}_{k+1}[A^{-1}B_0 + \varepsilon I])v^k + (1 - \tilde{\mu}_{k+1})v^{k-1}, \quad (4.16)$$

$$r^{k+1} = \mu_{k+1}(I - \rho_{k+1}[C^{-1}S_0 + \varepsilon I])r^k + (1 - \mu_{k+1})r^{k-1}, \quad (4.17)$$

где новые параметры определяются по формулам

$$\tilde{\mu}_{k+1} = 1 + \frac{\tau_{k+1}}{\tau_k}(1 - \tau_k)(1 - \nu_k \varepsilon), \quad \tilde{\rho}_{k+1} = \tilde{\mu}_{k+1}^{-1} \tau_{k+1} \nu_k, \quad (4.18)$$

$$\mu_{k+1} = 1 + \frac{\nu_{k+1}}{\nu_k}(1 - \tau_{k+1})(1 - \nu_k \varepsilon), \quad \rho_{k+1} = \mu_{k+1}^{-1} \tau_{k+1} \nu_{k+1}. \quad (4.19)$$

Доказательство. Перепишем формулы итерационного метода (4.13) для погрешности $y^k = \{v^k, r^k\}$

$$\begin{cases} A \frac{v^{k+1} - v^k}{\tau_{k+1}} + Av^k + Br^k = 0, \\ -C \frac{r^{k+1} - r^k}{\nu_{k+1}} + B^T v^{k+1} - \varepsilon C r^k = 0, \end{cases}$$

откуда имеем

$$v^{k+1} = (1 - \tau_{k+1})v^k - \tau_{k+1}A^{-1}Br^k, \quad (4.20)$$

$$\begin{aligned} r^{k+1} &= (1 - \nu_{k+1}\varepsilon)r^k + \nu_{k+1}C^{-1}B^Tv^{k+1} = \\ &= \left[(1 - \nu_{k+1}\varepsilon)I - \tau_{k+1}\nu_{k+1}C^{-1}S_0 \right] r^k + \\ &\quad + \nu_{k+1}(1 - \tau_{k+1})C^{-1}B^Tv^k. \end{aligned} \quad (4.21)$$

Увеличим в выражении (4.21) индекс k на единицу

$$\begin{aligned} r^{k+2} &= \left[(1 - \nu_{k+2}\varepsilon)I - \tau_{k+2}\nu_{k+2}C^{-1}S_0 \right] r^{k+1} + \\ &\quad + \nu_{k+2}(1 - \tau_{k+2})C^{-1}B^Tv^{k+1}, \end{aligned} \quad (4.22)$$

выразим из левого равенства (4.21) величину

$$C^{-1}B^Tv^{k+1} = \frac{r^{k+1} - (1 - \nu_{k+1}\varepsilon)r^k}{\nu_{k+1}}$$

и подставим ее в (4.22). Теперь после замены

$$\begin{aligned} \mu_{k+1} &= 1 + \frac{\nu_{k+1}}{\nu_k}(1 - \tau_{k+1})(1 - \nu_k\varepsilon), \\ \rho_{k+1} &= \mu_{k+1}^{-1}\tau_{k+1}\nu_{k+1} \end{aligned}$$

имеем искомое соотношение для второй компоненты погрешности. Аналогичным образом получается трехслойное соотношение и для первой. Увеличим в (4.20) индекс k на единицу и заменим в полученном выражении r^{k+1} с помощью левой части соотношения (4.21). В результате получим

$$\begin{aligned} v^{k+2} &= \left[(1 - \tau_{k+2})I - \tau_{k+2}\nu_{k+1}A^{-1}B_0 \right] v^{k+1} - \\ &\quad - \tau_{k+2}(1 - \nu_{k+1}\varepsilon)A^{-1}Br^k. \end{aligned} \quad (4.23)$$

Выразим из (4.20) величину

$$-A^{-1}Br^k = \frac{v^{k+1} - (1 - \tau_{k+1})v^k}{\tau_{k+1}}$$

и подставим ее в (4.23). Теперь после замены

$$\begin{aligned} \tilde{\mu}_{k+1} &= 1 + \frac{\tau_{k+1}}{\tau_k}(1 - \tau_k)(1 - \nu_k\varepsilon), \\ \tilde{\rho}_{k+1} &= \tilde{\mu}_{k+1}^{-1}\tau_{k+1}\nu_k \end{aligned}$$

искмое соотношение для первой компоненты погрешности получено. ■

4.4.2. Выбор параметров для p , как в циклическом методе

Приведем оценку погрешности для p , как в циклическом методе. Выпишем для произвольного фиксированного $N = 1, 2, \dots$ чебышевский набор параметров:

$$\begin{aligned} \tilde{\tau}_k &= \frac{2}{(\gamma + \Gamma + 2\varepsilon)(1 + q_1\mu_k)}, \quad k = 1, 2, \dots, N, \\ q_1 &= \frac{1 - \xi}{1 + \xi}, \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon}, \mu_k \in \aleph_N = \left\{ -\cos \frac{2i-1}{2N}\pi, i = 1, 2, \dots, N \right\}. \end{aligned} \quad (4.24)$$

Справедлива

Теорема 4.4.1. Пусть в методе (4.13) итерационные параметры для произвольного фиксированного $N = 1, 2, \dots$ выбираются следующим образом: $\tau_k = 1$, $\nu_k = \tilde{\tau}_k$, $k = 1, 2, \dots, N$ из (4.24). Тогда для приближений p^N метода справедлива оценка (4.14)

$$\|p^N - p\|_{D_p} \leq \epsilon_N \|p^0 - p\|_{D_p}.$$

Доказательство. Положим в формулах (4.19) леммы 4.4.6 для итерационных параметров μ_{k+1}, ρ_{k+1} значения τ_k, ν_k , указанные в условии теоремы. Тогда эти формулы примут вид

$$\mu_k = 1, \quad \rho_k = \tilde{\tau}_k,$$

что, в свою очередь, приводит к соотношению для погрешности r^k :

$$r^{k+1} = (I - \tilde{\tau}_k[C^{-1}S_0 + \varepsilon I])r^k.$$

Учитывая расположение спектра

$$\sigma(C^{-1/2}S_0C^{-1/2} + \varepsilon I) \subseteq [\gamma + \varepsilon, \Gamma + \varepsilon], \quad \gamma > 0,$$

и результаты пункта 3.1.2, имеем искомую оценку погрешности

$$\|p^N - p\|_{D_p} \leq \epsilon_N \|p^0 - p\|_{D_p}, \quad \epsilon_N = \frac{2q_0^N}{1 + q_0^{2N}}$$

с

$$q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon}. \quad \blacksquare$$

Этот результат служит иллюстрацией того факта, что двухслойные алгоритмы типа Узавы (случай $\tau_k = 1$) для системы с параметром $L_\varepsilon z = F$ могут быть записаны в форме (4.13).

4.4.3. Выбор параметров для p , как в трехслойных методах

Определим для получения наилучшей оценки погрешности относительно компоненты p последовательности итерационных параметров в полуитерационном методе Чебышева (3.7) на отрезке $[\gamma + \varepsilon, \Gamma + \varepsilon]$:

$$\tilde{\tau}_k = \frac{2}{\gamma + \Gamma + 2\varepsilon}, \quad \tilde{\alpha}_{k+1} = \frac{4}{4 - q_1^2 \tilde{\alpha}_k}, \quad k = 1, 2, \dots, \quad (4.25)$$

где

$$\tilde{\alpha}_1 = 2, \quad q_1 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma + \varepsilon}{\Gamma + \varepsilon};$$

и в методе сопряженных градиентов:

$$\begin{aligned} \tilde{\tau}_{k+1} &= \frac{(x^k, w^k)}{(w^k, (S_0 + \varepsilon C)w^k)}, \quad k = 0, 1, 2, \dots, \\ \tilde{\alpha}_{k+1} &= \left(1 - \frac{\tilde{\tau}_{k+1}}{\tilde{\tau}_k} \frac{(x^k, w^k)}{(x^{k-1}, w^{k-1})} \frac{1}{\tilde{\alpha}_k} \right)^{-1}, \quad k = 1, 2, \dots, \end{aligned} \quad (4.26)$$

где $\alpha_1 = 1$, $w^k = C^{-1}x^k - \text{поправка}$, $r^k = (S_0 + \varepsilon C)^{-1}x^k = (p^k - p) - \text{погрешность}$, $x^k = ((S_0 + \varepsilon C)p^k - B^T A^{-1}f - \varphi) - \text{невязка}$. Напомним, что $S_0 = B^T A^{-1}B$.

Справедлива

Лемма 4.4.7. Пусть для итерационных параметров μ_k , ρ_k , определяемых формулами (4.19) леммы 4.4.6, справедливы неравенства при $k = 1, 2, \dots$

$$\mu_{k+1} > 1, \quad \rho_{k+1} > 0.$$

Тогда обратное преобразование к параметрам τ_k, ν_k исходного алгоритма (4.13) однозначно определяется формулами

$$\begin{aligned} \varepsilon^{-1} > \nu_{k+1} &= \nu_k \frac{\mu_{k+1} - 1}{1 - \nu_k \varepsilon} + \mu_{k+1} \rho_{k+1} > 0, \\ 1 > \tau_{k+1} &= \mu_{k+1} \frac{\rho_{k+1}}{\nu_{k+1}} > 0 \end{aligned} \quad (4.27)$$

для $k = 1, 2, \dots$, если $\varepsilon^{-1} > \nu_1 > 0$.

Доказательство. Преобразуем формулу для ρ_{k+1} :

$$\tau_{k+1} \nu_{k+1} = \mu_{k+1} \rho_{k+1}.$$

Теперь для некоторого $\varepsilon^{-1} > \nu_k > 0$ (напомним, что по предположению $\varepsilon^{-1} > \nu_1 > 0$) из формулы для μ_{k+1} имеем

$$\nu_{k+1} = \nu_k \frac{\mu_{k+1} - 1}{1 - \nu_k \varepsilon} + \mu_{k+1} \rho_{k+1}.$$

Отсюда, в силу условия леммы, получаем $\varepsilon^{-1} > \nu_{k+1} > 0$ и, соответственно,

$$\tau_{k+1} = \mu_{k+1} \frac{\rho_{k+1}}{\nu_{k+1}} > 0.$$

Подстановка в полученную формулу явного выражения для ν_{k+1} приводит к неравенству $\tau_{k+1} < 1$.

Теперь для завершения доказательства достаточно показать, что из неравенства $\varepsilon^{-1} > \nu_k > 0$ следует неравенство $\varepsilon^{-1} > \nu_{k+1} > 0$. Предположим, что последнее не выполнено, т. е. $\nu_{k+1} \geq \varepsilon^{-1}$ (так как положительность легко проверяема). Тогда после умножения на ε из формулы для ν_{k+1} получим

$$\varepsilon \nu_k \frac{\mu_{k+1} - 1}{1 - \nu_k \varepsilon} + \varepsilon \mu_{k+1} \rho_{k+1} \geq 1.$$

Отсюда следует

$$\varepsilon \nu_k (\mu_{k+1} - \varepsilon \mu_{k+1} \rho_{k+1}) \geq 1 - \varepsilon \mu_{k+1} \rho_{k+1} > 0,$$

что, в свою очередь, по условию леммы приводит к неравенству $\nu_k > \varepsilon^{-1}$. Полученное противоречие завершает доказательство. ■

На основании имеющихся свойств итерационных параметров получим наилучшую для любого k оценку погрешности для компоненты p .

Теорема 4.4.2. Пусть в методе (4.13) итерационные параметры выбираются следующим образом:

$$\begin{aligned} \tau_1 &= 1, \quad \nu_1 = \tilde{\tau}_1, \\ \nu_{k+1} &= \frac{\nu_k (\tilde{\alpha}_{k+1} - 1)}{1 - \nu_k \varepsilon} + \tilde{\alpha}_{k+1} \tilde{\tau}_{k+1}, \\ \tau_{k+1} &= \tilde{\alpha}_{k+1} \frac{\tilde{\tau}_{k+1}}{\nu_{k+1}}, \quad k = 1, 2, \dots, \end{aligned}$$

где $\tilde{\tau}_k, \tilde{\alpha}_k$ определяются формулами (4.25) или (4.26). Тогда для приближений p^k метода справедлива оценка (4.14).

Доказательство. В силу справедливости лемм 3.1.1 и 3.1.2 параметры полуитерационного метода Чебышева (4.25) и методов сопряженных направлений (4.26) удовлетворяют неравенствам

$$\tilde{\tau}_k > 0, \quad \tilde{\alpha}_k > 1, \quad k = 2, 3, \dots,$$

и следовательно, допускают обратное преобразование (4.27). Поэтому выбор параметров по формулам (4.27) в методе (4.13) в силу леммы 4.4.7 порождает для погрешности $r^k = p^k - p$ соотношение (4.17) леммы 4.4.6 с $\mu_{k+1} = \tilde{\alpha}_{k+1}, \rho_{k+1} = \tilde{\tau}_{k+1}$ для $k = 1, 2, \dots$

и $r^1 = (I - \nu_1[C^{-1}S_0 + \varepsilon I])r^0$, что и приводит к искомой оценке погрешности (см. [76], с. 347). ■

Далее, если необходимо определить вектор u^{k+1} , удовлетворяющий оценке (4.15), достаточно взять $\tau_{k+1} = 1$ и после этого шага завершить вычисления. Таким образом, формулы (4.13), (4.27) обобщают алгоритмы типа Узавы на случай использования переменных итерационных параметров.

4.4.4. Выбор параметров для u , как в трехслойных методах

Прежде чем приступить к построению искомого набора итерационных параметров, проанализируем следствие первого шага в методе (4.13) при $\tau_1 = 1$ с произвольного начального приближения $\{u^0, p^0\}$. Из результатов леммы 4.4.6 имеем

$$v^1 = -A^{-1}Br^0, \quad v^2 = (I - \tau_2\nu_1[A^{-1}B_0 + \varepsilon I])v^1, \quad (4.28)$$

откуда следует, что начальное приближение u^0 фактически не участвует в вычислениях и, следовательно, v^0 — в оценках погрешности. Другими словами, можно считать, что в методе (4.13) величина u^1 выбирается специальным образом, а именно

$$Au^1 + Bp^0 = f. \quad (4.29)$$

Покажем, что метод (4.13) при выборе начального приближения вида (4.29) эквивалентен итерированию ошибки v^k в подпространстве G . Имеет место

Лемма 4.4.8. Для любой итерации k первая компонента v^k погрешности итерационного метода (4.13), стартующего с начального приближения вида (4.29), является элементом подпространства G , т. е. $(Av^k, h) = 0$ для $\forall h \in H$.

Доказательство. Первая компонента погрешности v^k в методе (4.13) удовлетворяет точно такому же соотношению, что и в методе (3.47). Поэтому сформулированный результат следует из леммы 3.4.5. ■

Отсюда на основании теоремы 1.3.1 и леммы 4.4.8 можно сделать вывод, что метод (4.13) для решения задачи (1.2) $L_\varepsilon z = F$, стартующий с начального приближения (4.29), эквивалентен, с точки зрения соотношений для погрешности v^k , методу простой итерации для отыскания решения

$$(A^{-1}B_0 + \varepsilon I)u = A^{-1}BC^{-1}\varphi - \varepsilon A^{-1}f, \quad u \in G. \quad (4.30)$$

При этом оператор $A^{-1}B_0 + \varepsilon I$ симметризуем на подпространстве G и справедливо неравенство

$$(\gamma + \varepsilon)(y, y) \leq ([A^{-1/2}B_0A^{-1/2} + \varepsilon I]y, y) \leq (\Gamma + \varepsilon)(y, y) \quad \forall y \in G.$$

Если ввести оператор $D_u = D_u^T > 0$ такой, что

$$D_u[A^{-1}B_0 + \varepsilon I] = (D_u[A^{-1}B_0 + \varepsilon I])^T,$$

то аналогично (4.26) можно определить итерационные параметры метода сопряженных градиентов для нахождения u из (4.30):

$$\begin{aligned} \tilde{\tau}_{k+1} &= \frac{(x^k, w^k)}{(w^k, (B_0 + \varepsilon A)w^k)}, \quad k = 0, 1, 2, \dots, \\ \tilde{\alpha}_{k+1} &= \left(1 - \frac{\tilde{\tau}_{k+1}}{\tilde{\tau}_k} \frac{(x^k, w^k)}{(x^{k-1}, w^{k-1})} \frac{1}{\tilde{\alpha}_k} \right)^{-1}, \quad k = 1, 2, \dots, \end{aligned} \quad (4.31)$$

где $\tilde{\alpha}_1 = 1$, $w^k = A^{-1}x^k$, $(B_0 + \varepsilon A)v^k = x^k$ (причем $v^k \in G$), $x^k = (B_0 + \varepsilon A)u^k - BC^{-1}\varphi + \varepsilon f$.

Теперь определим алгоритм выбора параметров. Используя свойства полушага с $\tau_1 = 1$, для произвольного p^0 вычислим u^1 . Это дает возможность определить $\tilde{\tau}_2$ по формулам (4.25) или (4.31). В соответствии с (4.28) получаем $\nu_1\tau_2 = \tilde{\tau}_2$, что порождает некоторый произвол в выборе ν_1 (необходимо только, чтобы $\tau_2 \neq 1$). Зафиксировав каким-либо образом ν_1 , сразу же по формулам (4.13) имеем возможность вычислить p^1 и u^2 . Это, в свою очередь, порождает значения $\tilde{\alpha}_{k+1}, \tilde{\tau}_{k+1}$ при $k = 2$ по формулам (4.25) или (4.31). Теперь, используя обратное преобразование формул леммы 4.4.6, получим

$$\tau_{k+1} = \frac{(\tilde{\mu}_{k+1} - 1)\tau_k}{1 - \tau_k} + \varepsilon\tilde{\mu}_{k+1}\tilde{\rho}_{k+1}, \quad \nu_k = \frac{\tilde{\mu}_{k+1}\tilde{\rho}_{k+1}}{\tau_{k+1}}, \quad (4.32)$$

где

$$\tilde{\mu}_{k+1} = \tilde{\alpha}_{k+1}, \quad \tilde{\rho}_{k+1} = \tilde{\tau}_{k+1}$$

из (4.25) или (4.31). Это приводит к нахождению p^k и u^{k+1} для $k = 2$, и далее этот процесс может быть продолжен до достижения искомого результата.

Из формул (4.32) следует, что их применимость определяется условиями $0 < \tau_k < 1$, $0 < \nu_k < \varepsilon^{-1}$, $k = 2, 3, \dots$, что, в свою очередь, зависит от выбора ν_1 . Имеет место

Лемма 4.4.9. Для произвольного $k = 2, 3, \dots, k_0$ существует

$$0 < \tau_2 = \tau_2(k_0) < 1$$

такое, что если

то все τ_{k+1}, ν_k из (4.32) удовлетворяют неравенствам

$$0 < \tau_{k+1} < 1, \quad 0 < \nu_k < \varepsilon^{-1}.$$

Доказательство. Из формул (4.32) следует, что из неравенства $0 < \tau_{k+1} < 1$ сразу следует $0 < \nu_k < \varepsilon^{-1}$, поэтому будем доказывать только первое. Положим

$$\tau_2 = \tau_2(k_0) = (2^{k_0} k_0)^{-1}.$$

Отметим, что из условий леммы

$$1 < \tilde{\mu}_{k+1} < 2, \quad \tilde{\rho}_{k+1} < (2^{k_0+1} k_0 \varepsilon)^{-1}$$

следует

$$\varepsilon \tilde{\mu}_{k+1} \tilde{\rho}_{k+1} < (2^{k_0} k_0)^{-1}.$$

Это дает возможность получить из формулы (4.32) для τ_{k+1} оценку для τ_3 . Действительно,

$$\tau_3 < (2^{k_0} k_0 - 1)^{-1} + (2^{k_0} k_0)^{-1} < (2^{k_0-1} (k_0 - 1))^{-1}.$$

Рекуррентное применение указанной процедуры дает

$$\tau_{k_0+1} < (2^1 2 - 1)^{-1} + (2^{k_0} k_0)^{-1} < 2^{-1}. \quad \blacksquare$$

С помощью полученных результатов покажем, что справедлива

Теорема 4.4.3. Для произвольного фиксированного $k \geq 1$ существует набор итерационных параметров:

$$\tau_1 = 1, \quad \tau_2 \nu_1 = \tilde{\tau}_2,$$

далее

$$\tau_{k+1} = \frac{(\tilde{\mu}_{k+1} - 1)\tau_k}{1 - \tau_k} + \varepsilon \tilde{\rho}_{k+1} \tilde{\mu}_{k+1}, \quad \nu_k = \frac{\tilde{\mu}_{k+1} \tilde{\rho}_{k+1}}{\tau_{k+1}},$$

где

$$\tilde{\mu}_{k+1} = \tilde{\alpha}_{k+1}, \quad \tilde{\rho}_{k+1} = \tilde{\tau}_{k+1}$$

из (3.9) или (4.31), такой что приближения u^k метода (4.13) удовлетворяют оценке погрешности (4.15).

Доказательство. Зафиксируем k_0 и положим $\nu_1 = \tilde{\tau}_2 k_0$. Тогда, в силу леммы 4.4.9, формулы (4.32) корректны и порождают набор параметров τ_{k+1}, ν_k при $k = 2, 3, \dots, k_0$ для метода (4.13). В свою очередь, это приводит к соотношению (4.16) леммы 4.4.6 для ошибки $v^k = u^k - u$ с параметрами $\tilde{\mu}_{k+1}, \tilde{\rho}_{k+1}$, гарантирующему для любого $k = 1, 2, \dots, k_0$ оценку

$$\|u^{k+1} - u\|_{D_u} \leq \epsilon_k \|u^1 - u\|_{D_u}. \quad \blacksquare$$

4.5. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Седловые задачи с параметром $\varepsilon > 0$ возникают как непосредственно из приложений [77, 79], так и в качестве регуляризации (и/или стабилизации) системы для $\varepsilon = 0$ (см., например, [117, 168, 169] и цитированную там литературу). Общая точка зрения: значение $\varepsilon > 0$ приводит к улучшению свойств системы со всех сторон, в том числе к ускорению сходимости итерационных алгоритмов.

Для дифференциальной задачи типа Стокса с параметром исследование явного и неявного алгоритмов проведено в [133].

Выбор переменных итерационных параметров для явного метода MSORe предложен в [92].

С практической точки зрения совпадение матриц C в решаемой системе $L_\varepsilon z = F$ и спектральной задаче $S_0 p = \lambda C p$ является достаточно обременительным ограничением. Но в данном случае нас интересует изменение характеристик скорости сходимости релаксационных алгоритмов только за счет величины $\varepsilon > 0$. Поэтому возмущение исходной системы с $\varepsilon = 0$ с помощью модельной (совпадающей) матрицы C представляется оправданным.

МЕТОДЫ ДЛЯ НОРМАЛЬНЫХ УРАВНЕНИЙ

Глава посвящена анализу алгоритмов, основанных на сочетании идей симметризации и предобуславливания. Такой подход приводит к системе уравнений (называемой *нормальной*) с симметричной положительно определенной матрицей, для решения которой применяется, как правило, метод сопряженных градиентов.

Для задач вида $L_\varepsilon z = F$ с $\varepsilon \geq 0$ построены равносильные нормальные системы уравнений, т. е. уравнения имеющие те же решения, что и исходные. Далее рассмотрены параметризованные представления собственных значений оператора равносильного уравнения, сформулированы и решены задачи оптимизации спектрального числа обусловленности.

Уточним процедуру, лежащую в основе подхода. Систему линейных уравнений $Gu = b$ с невырожденной матрицей G можно привести к равносильной системе $Eu = f$ с матрицей $E = E^T > 0$, где $E = G^T G$, $f = G^T b$. Этот прием называется симметризацией. Отметим, что здесь происходит квадратичное увеличение спектрального числа обусловленности матрицы. Действительно, по определению, имеем

$$\text{cond}_2(G) = \|G\|_2 \|G^{-1}\|_2, \quad \|G\|_2 = \sqrt{\lambda_{\max}(G^T G)},$$

что дает $\text{cond}_2(E) = \text{cond}_2(G^T G) = \text{cond}_2^2(G)$. В свою очередь, предобуславливание $D^{-1}Gu = D^{-1}b$ имеет целью понизить величину $\text{cond}_2(D^{-1}G)$ по сравнению с $\text{cond}_2(G)$. Поэтому имеет смысл нахождение каким-либо эффективным итерационным методом решения задачи

$$(D^{-1}G)^T (D^{-1}G)u = (D^{-1}G)^T D^{-1}b,$$

учитывая, что при $G = G^T$ более устойчивым является обращение матрицы $D^{-1}GD^{-1}G$, а спектр произведения квадратных матриц не зависит от порядка сомножителей (см. лемму 1.3.1).

Далее в главе в качестве предобуславливающего используется оператор D с параметром $\alpha > 0$:

$$D = \begin{pmatrix} A & 0 \\ 0 & \alpha^{-1}C \end{pmatrix},$$

в котором матрицы A и C имеют прежний смысл.

5.1. ОПТИМИЗАЦИЯ МЕТОДА ДЛЯ БАЗОВОЙ СИСТЕМЫ

В разделе анализируется процедура симметризации и предобуславливания базовой системы уравнений $L_0 z = F$.

5.1.1. Спектр оператора равносильной задачи

Рассмотрим зависимость собственных значений от параметра α оператора R равносильной задачи

$$Rz \equiv D^{-1}L_0D^{-1}L_0z = D^{-1}L_0D^{-1}F. \quad (5.1)$$

Имеет место

Теорема 5.1.1. Спектр $\sigma(R)$ оператора R в задаче (5.1) принадлежит множеству

$$\Lambda = \{1\} \cup \left\{ \alpha t + \frac{1}{2} \pm \sqrt{\alpha t + \frac{1}{4}}, t \in [\gamma, \Gamma] \right\}.$$

Доказательство. Рассмотрим спектральную задачу $Rz = \lambda z$:

$$Rz \equiv \begin{pmatrix} I + \alpha A^{-1}B_0 & A^{-1}B \\ \alpha C^{-1}B^T & \alpha C^{-1}S_0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} u \\ p \end{pmatrix}. \quad (5.2)$$

Для нахождения ее решения используем базис пространства Z , построенный в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (5.2) с $\lambda_k^{(1)} = 1$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^Tg_j\}.$$

После применения к первому уравнению (5.2) оператора $C^{-1}B^T$ и замены $C^{-1}B^Tg_j = p_j$ будем иметь

$$\begin{cases} (I + \alpha C^{-1}S_0)p_j - \kappa^{-1}C^{-1}S_0p_j = \lambda p_j, \\ -\alpha p_j + \kappa^{-1}\alpha C^{-1}S_0p_j = \lambda \kappa^{-1}p_j. \end{cases}$$

Перепишем эту систему в виде

$$\begin{cases} S_0p_j = \frac{\kappa(\lambda - 1)}{\kappa\alpha - 1}Cp_j, \\ S_0p_j = \frac{\lambda + \kappa\alpha}{\alpha}Cp_j. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0 p = t C p$ соответствует собственное значение t_j , $j = 1, \dots, N_p$ по теореме 1.3.1. Зафиксировав его, из полученной системы для λ и κ выведем соотношения

$$t_j = \frac{\kappa(\lambda - 1)}{\kappa\alpha - 1} = \frac{\lambda + \kappa\alpha}{\alpha}.$$

Исключая из этих уравнений κ , приходим к выражению

$$\lambda_j^{(2,3)} = \alpha t_j + \frac{1}{2} \pm \sqrt{\alpha t_j + \frac{1}{4}}$$

и, соответственно,

$$\kappa_j^{(2,3)} = -\frac{1}{2\alpha} \mp \sqrt{\frac{t_j}{\alpha} + \frac{1}{4\alpha^2}} \neq 0.$$

Из явного представления величин $\kappa_j^{(2)}$ и $\kappa_j^{(3)}$ следует, что для любого $\alpha > 0$ они будут различны. Это означает, что найденная система собственных векторов

$$\{z_i^{(1,2,3)}\}_{i=1}^{N_u+N_p}$$

задачи (5.2) удовлетворяет теореме 1.3.2, следовательно, все искомые собственные значения найдены.

Поэтому можно сделать вывод о том, что $\sigma(R)$ — спектр оператора R в задаче (5.1) — может быть параметризован с помощью спектра оператора $C^{-1/2} S_0 C^{-1/2}$. Действительно, пусть

$$\sigma(C^{-1/2} S_0 C^{-1/2}) \subseteq [\gamma, \Gamma],$$

тогда $\sigma(R)$ принадлежит множеству Λ :

$$\Lambda = \{1\} \cup \left\{ \alpha t + \frac{1}{2} \pm \sqrt{\alpha t + \frac{1}{4}}, t \in [\gamma, \Gamma] \right\}. \quad \blacksquare$$

5.1.2. Минимизация числа обусловленности

Поскольку оператор R является симметризуемым и положительно определенным по построению, важным является вопрос о минимизации его спектрального числа обусловленности.

Теорема 5.1.2. При $\alpha_0 = 2/\gamma$ спектральное число обусловленности оператора $R = D^{-1} L_0 D^{-1} L_0$ в задаче (5.1) принимает наименьшее значение. При этом спектр R принадлежит отрезку:

$$\sigma(R) \in \left[1, \frac{2\Gamma}{\gamma} + \frac{1}{2} + \sqrt{\frac{2\Gamma}{\gamma} + \frac{1}{4}} \right].$$

Доказательство. Обозначим через λ_{\max} и λ_{\min} выражения $\sup \sigma(R)$ и $\inf \sigma(R)$ соответственно. Кроме того, будем использовать обозначения $\lambda_{2,3}$ для точек спектра оператора R , зависящих от параметра α :

$$\lambda_{2,3} = \alpha t + \frac{1}{2} \pm \sqrt{\alpha t + \frac{1}{4}}.$$

Знак «+» относится к λ_2 .

В силу непрерывности функций $\lambda_{2,3}$ по параметру t и ограниченности $\sigma(R)$, достаточно вместо $\sup(\inf)$ использовать процедуры $\max(\min)$.

Поскольку $\alpha > 0$, то

$$\lambda_{\max} = \max_t \lambda_2 > 1.$$

С учетом монотонности λ_2 по t имеем

$$\lambda_{\max}(\alpha) = \alpha \Gamma + \frac{1}{2} + \sqrt{\alpha \Gamma + \frac{1}{4}},$$

так как $t \in [\gamma, \Gamma]$.

Далее, рассмотрим $\lambda_{\min} = \min \{1, \min_t \lambda_3\}$, откуда получаем

$$\lambda_{\min} = \min \left\{ 1, \alpha \gamma + \frac{1}{2} - \sqrt{\alpha \gamma + \frac{1}{4}} \right\},$$

т. е. спектральное число обусловленности оператора R определяется формулой

$$\text{cond}_2(R) = \max \left\{ \lambda_{\max}(\alpha), \frac{\lambda_{\max}(\alpha)}{\alpha \gamma + 1/2 - \sqrt{\alpha \gamma + 1/4}} \right\}.$$

При этом его минимум по α достигается на решении уравнения

$$\alpha_0 \gamma + \frac{1}{2} - \sqrt{\alpha_0 \gamma + \frac{1}{4}} = 1,$$

если последнее существует. Непосредственные вычисления дают $\alpha_0 = 2/\gamma$, причем $\sigma(R) \in [1, \lambda_{\max}(\alpha_0)]$. ■

5.1.3. Наилучшая оценка погрешности

Теперь можно применить результаты общей теории итерационных методов решения систем с симметризуемыми положительно определенными матрицами. В частности, явные двух и трехслойные схемы с оптимальными переменными параметрами (см. раздел 3.1)

гарантируют скорость сходимости в достаточно произвольной метрике $D_z = D_z^T > 0$ такой, что $D_z R = (D_z R)^T$:

$$\|z^k - z\|_{D_z} \leq \epsilon_k \|z^0 - z\|_{D_z},$$

где

$$\epsilon_k = \frac{2q_0^k}{1 + q_0^{2k}}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \left(\frac{2\Gamma}{\gamma} + \frac{1}{2} + \sqrt{\frac{2\Gamma}{\gamma} + \frac{1}{4}} \right)^{-1}.$$

5.2. ОПТИМИЗАЦИЯ МЕТОДА ДЛЯ СИСТЕМЫ С ПАРАМЕТРОМ

В разделе анализируется процедура симметризации и предобуславливания системы уравнений $L_\epsilon z = F$ с параметром $\epsilon \geq 0$.

5.2.1. Спектр оператора равносильной задачи

Рассмотрим зависимость собственных значений от параметра α оператора R_ϵ равносильной задачи

$$R_\epsilon z \equiv D^{-1} L_\epsilon D^{-1} L_\epsilon z = D^{-1} L_\epsilon D^{-1} F. \quad (5.3)$$

Теорема 5.2.1. Спектр $\sigma(R_\epsilon)$ оператора R в задаче (5.3) принадлежит множеству

$$\Lambda = \{1\} \cup \left\{ \alpha t + \frac{1}{2} + \frac{\epsilon^2 \alpha^2}{2} \pm \frac{1}{2} |\epsilon \alpha - 1| \sqrt{4\alpha t + (\epsilon \alpha + 1)^2} \right\},$$

где $t \in [\gamma, \Gamma]$.

Доказательство. Рассмотрим спектральную задачу $R_\epsilon z = \lambda z$:

$$R_\epsilon z \equiv \begin{pmatrix} I + \alpha A^{-1} B_0 & (1 - \alpha \epsilon) A^{-1} B \\ \alpha(1 - \alpha \epsilon) C^{-1} B^T & \alpha C^{-1} S_0 + \alpha^2 \epsilon^2 I \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} u \\ p \end{pmatrix}. \quad (5.4)$$

Для нахождения ее решения используем базис пространства Z , построенный в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (5.4) с $\lambda_k^{(1)} = 1$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\kappa^{-1} C^{-1} B^T g_j\}.$$

После применения к первому уравнению (5.4) оператора $C^{-1}B^T$ и замены $C^{-1}B^T g_j = p_j$ будем иметь

$$\begin{cases} (I + \alpha C^{-1}S_0)p_j - \kappa^{-1}(1 - \alpha\varepsilon)C^{-1}S_0p_j = \lambda p_j, \\ -\alpha(1 - \alpha\varepsilon)p_j + \kappa^{-1}(\alpha C^{-1}S_0 + \alpha^2\varepsilon^2 I)p_j = \lambda \kappa^{-1}p_j. \end{cases}$$

Перепишем эту систему в виде

$$\begin{cases} S_0p_j = \frac{\kappa(\lambda - 1)}{\kappa\alpha - (1 - \alpha\varepsilon)}Cp_j, \\ S_0p_j = \frac{\lambda + \kappa\alpha(1 - \alpha\varepsilon) - \alpha^2\varepsilon^2}{\alpha}Cp_j. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0p = tCp$ соответствует собственное значение $t_j, j = 1, \dots, N_p$ по теореме 1.3.1. Зафиксировав его, из полученной системы для λ и κ выведем соотношения

$$t_j = \frac{\kappa(\lambda - 1)}{\kappa\alpha - (1 - \alpha\varepsilon)} = \frac{\lambda + \kappa\alpha(1 - \alpha\varepsilon) - \alpha^2\varepsilon^2}{\alpha}.$$

Исключая из этих уравнений κ , приходим к выражению

$$\lambda_j^{(2,3)} = \frac{1}{2} + \alpha t_j + \frac{1}{2}\alpha^2\varepsilon^2 \pm \frac{1}{2}|\alpha\varepsilon - 1|\sqrt{(\alpha\varepsilon + 1)^2 + 4\alpha t_j}$$

и, соответственно,

$$\kappa_j^{(2,3)} = -\frac{1}{2\alpha} \left(1 + \alpha\varepsilon \pm \text{sign}(1 - \alpha\varepsilon)\sqrt{(\alpha\varepsilon + 1)^2 + 4\alpha t_j} \right) \neq 0$$

при $\alpha\varepsilon \neq 1$ и

$$\kappa_j^{(2,3)} = \pm \chi_2 / \chi_1 \sqrt{t_j} \neq 0, \quad \chi_1, \chi_2 > 0 \quad \text{при } \alpha\varepsilon = 1.$$

Последнее выражение для $\kappa_j^{(2,3)}$ нуждается в комментариях. Рассмотрим случай кратных собственных значений более подробно. Пусть значение параметра α таково, что выражение $\alpha\varepsilon - 1$ обращается в нуль. Тогда, используя

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^T g_j\},$$

после применения к первому уравнению (5.4) оператора $C^{-1}B^T$ и замены $C^{-1}B^T g_j = p_j$ получим

$$\begin{cases} (I + \alpha C^{-1}S_0)p_j = \lambda p_j, \\ \kappa^{-1}(\alpha C^{-1}S_0 + \alpha^2\varepsilon^2 I)p_j = \lambda \kappa^{-1}p_j, \end{cases}$$

откуда следует, что любой вектор вида

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^T g_j\}$$

является собственным; дополнительно имеем

$$\lambda_j^{(2)} = \lambda_j^{(3)} = \alpha t_j + 1.$$

Поэтому в данном случае, независимо от кратности собственного значения t_j , собственные векторы задачи (5.4) с одинаковыми g_j можно сделать ортогональными в следующей метрике пространства Z (см. теорему 1.3.2):

$$\begin{aligned}(z_1, z_2)_Z &= \chi_1(Au_1, u_2) + \chi_2(Cp_1, p_2), \\ z_i &= \{u_i, p_i\} \in Z, \quad \chi_i > 0, \quad i = 1, 2,\end{aligned}$$

положив $\kappa_j^{(2,3)} = \pm \chi_2 / \chi_1 \sqrt{t_j}$. Действительно,

$$\begin{aligned}(z_j^{(2)}, z_j^{(3)})_Z &= \chi_1(Ag_j, g_j) + \chi_2 \left(C \frac{C^{-1}B^T g_j}{\kappa_j^{(2)}}, \frac{C^{-1}B^T g_j}{\kappa_j^{(3)}} \right) = \\ &= \chi_1([A - t_j^{-1}BC^{-1}B^T]g_j, g_j) = 0.\end{aligned}$$

Из явного представления величин $\kappa_j^{(2)}$ и $\kappa_j^{(3)}$ следует, что для любого $\alpha > 0$ они будут различны. Это означает, что найденная система собственных векторов

$$\{z_i^{(1,2,3)}\}_{i=1}^{N_u+N_p}$$

задачи (5.4) удовлетворяет теореме 1.3.2, следовательно, все искомые собственные значения найдены.

Поэтому можно сделать вывод, что $\sigma(R_\epsilon)$ — спектр оператора R_ϵ в задаче (5.3) — может быть параметризован с помощью спектра оператора $C^{-1/2}S_0C^{-1/2}$. Если

$$\sigma(C^{-1/2}S_0C^{-1/2}) \subseteq [\gamma, \Gamma],$$

то $\sigma(R_\epsilon)$ принадлежит множеству Λ :

$$\Lambda = \{1\} \cup \left\{ \alpha t + \frac{1}{2} + \frac{\epsilon^2 \alpha^2}{2} \pm \frac{1}{2} |\epsilon \alpha - 1| \sqrt{4\alpha t + (\epsilon \alpha + 1)^2} \right\},$$

где $t \in [\gamma, \Gamma]$. ■

5.2.2. Минимизация числа обусловленности

Поскольку оператор R_ϵ является симметризуемым и положительно определенным по построению, рассмотрим вопрос о минимизации его спектрального числа обусловленности.

Теорема 5.2.2. При $\alpha_0 = 2/(\gamma + 2\varepsilon)$ спектральное число обусловленности оператора $R_\varepsilon = D^{-1}L_\varepsilon D^{-1}L_\varepsilon$ в задаче (5.3) принимает наименьшее значение. При этом спектр R_ε принадлежит отрезку:

$$\sigma(R_\varepsilon) \in \left[1, \alpha_0\Gamma + \frac{1}{2} + \varepsilon^2\alpha_0^2/2 + \frac{1}{2}(1 - \varepsilon\alpha_0)\sqrt{4\alpha_0\Gamma + (\varepsilon\alpha_0 + 1)^2} \right].$$

Доказательство. Обозначим границы спектра $\sup \sigma(R_\varepsilon)$ и $\inf \sigma(R_\varepsilon)$ через λ_{\max} и λ_{\min} соответственно. Кроме того, будем использовать обозначения $\lambda_{2,3}$ для точек спектра оператора R_ε , зависящих от параметра α :

$$\begin{aligned} \lambda_{2,3}(\alpha, \varepsilon, t) &= \frac{1}{2} + \alpha t + \frac{1}{2}\alpha^2\varepsilon^2 \pm \frac{1}{2}|\alpha\varepsilon - 1|\sqrt{\varphi_t}, \\ \varphi_t &= (\alpha\varepsilon + 1)^2 + 4\alpha t. \end{aligned}$$

Знак «+» относится к λ_2 .

В силу непрерывности функций $\lambda_{2,3}$ по параметру t и ограниченности $\sigma(R_\varepsilon)$ достаточно вместо $\sup(\inf)$ использовать процедуры $\max(\min)$.

Отметим справедливость неравенств

$$0 < \lambda_3(\alpha, \varepsilon, t) < \lambda_2(\alpha, \varepsilon, t). \quad (5.5)$$

Простой анализ показывает монотонность

$$0 < \frac{\partial \lambda_{2,3}}{\partial t} = \frac{\alpha}{\sqrt{\varphi_t}} (\sqrt{\varphi_t} \pm |\alpha\varepsilon - 1|). \quad (5.6)$$

Заметим, что справедливо представление

$$\lambda_2(\alpha, \varepsilon, 0) = \begin{cases} 1 & \text{при } 0 < \alpha < \frac{1}{\varepsilon}, \\ \alpha^2\varepsilon^2 \geq 1 & \text{при } \alpha \geq \frac{1}{\varepsilon}. \end{cases} \quad (5.7)$$

(Под выражениями $0 < \alpha < 1/\varepsilon$ и $\alpha > 1/\varepsilon$ при $\varepsilon = 0$ далее будем понимать $\alpha > 0$ и пустое множество соответственно, не делая каждый раз дополнительных оговорок.) После объединения соотношений (5.5)–(5.7), получаем

$$\begin{aligned} \text{cond}_2(R_\varepsilon) &= \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\max \left\{ 1, \max_t \lambda_{2,3}(\alpha, \varepsilon, t) \right\}}{\min \left\{ 1, \min_t \lambda_{2,3}(\alpha, \varepsilon, t) \right\}} = \\ &= \frac{\lambda_2(\alpha, \varepsilon, \Gamma)}{\min \{ 1, \lambda_3(\alpha, \varepsilon, \gamma) \}}. \end{aligned} \quad (5.8)$$

Имеют место следующие выражения:

$$0 < \frac{\partial \lambda_2(\alpha, \varepsilon, \Gamma)}{\partial \alpha} = \begin{cases} \frac{\varepsilon}{2\sqrt{\varphi\Gamma}}(\sqrt{\varphi\Gamma} + \alpha\varepsilon - 1)\left(\sqrt{\varphi\Gamma} + \alpha\varepsilon + 1 + \frac{2\Gamma}{\varepsilon}\right) \\ \quad \text{при } \alpha \geq \frac{1}{\varepsilon}, \\ -\frac{\varepsilon}{2\sqrt{\varphi\Gamma}}(\sqrt{\varphi\Gamma} - \alpha\varepsilon + 1)\left(\sqrt{\varphi\Gamma} - \alpha\varepsilon - 1 - \frac{2\Gamma}{\varepsilon}\right) \\ \quad \text{при } \alpha \leq \frac{1}{\varepsilon}. \end{cases} \quad (5.9)$$

Таким образом, $\lambda_2(\alpha, \varepsilon, \Gamma)$ монотонно возрастает в случае $\alpha > 0$.

Установим убывание функции $\lambda_2(\alpha, \varepsilon, \Gamma)/\lambda_3(\alpha, \varepsilon, \gamma)$ при $0 < \alpha < 1/\varepsilon$. Для этого достаточно показать, что функция

$$\begin{aligned} \psi(\alpha) &= \frac{\partial[\lambda_2(\alpha, \varepsilon, \Gamma)/\lambda_3(\alpha, \varepsilon, \gamma)]}{\partial \alpha} [\lambda_3(\alpha, \varepsilon, \gamma)]^2 = \\ &= \frac{\partial \lambda_2(\alpha, \varepsilon, \Gamma)}{\partial \alpha} \lambda_3(\alpha, \varepsilon, \gamma) - \lambda_2(\alpha, \varepsilon, \Gamma) \frac{\partial \lambda_3(\alpha, \varepsilon, \gamma)}{\partial \alpha} \end{aligned}$$

меньше нуля. Легко видеть, что

$$\frac{\partial \psi(\alpha)}{\partial \alpha} = \frac{\partial^2 \lambda_2(\alpha, \varepsilon, \Gamma)}{\partial \alpha^2} \lambda_3(\alpha, \varepsilon, \gamma) - \lambda_2(\alpha, \varepsilon, \Gamma) \frac{\partial^2 \lambda_3(\alpha, \varepsilon, \gamma)}{\partial \alpha^2} < 0$$

в силу неравенств (5.5) и того, что при $0 < \alpha < 1/\varepsilon$ справедливо

$$\begin{aligned} \frac{\partial^2 \lambda_2(\alpha, \varepsilon, \Gamma)}{\partial \alpha^2} &= \varepsilon^2 - \frac{\varepsilon^2(1 + \alpha\varepsilon) + 2\varepsilon\Gamma}{\sqrt{\varphi\Gamma}} - \frac{2\Gamma(1 - \alpha\varepsilon)(1 + \varepsilon\Gamma)}{\varphi\Gamma\sqrt{\varphi\Gamma}} < 0, \\ \frac{\partial^2 \lambda_3(\alpha, \varepsilon, \gamma)}{\partial \alpha^2} &= \varepsilon^2 + \frac{\varepsilon^2(1 + \alpha\varepsilon) + 2\varepsilon\gamma}{\sqrt{\varphi\gamma}} + \frac{2\gamma(1 - \alpha\varepsilon)(1 + \varepsilon\gamma)}{\varphi\gamma\sqrt{\varphi\gamma}} > 0. \end{aligned}$$

Отсюда сразу получаем, что $\psi(\alpha) < \psi(0) = 0$, а значит, функция $\lambda_2(\alpha, \varepsilon, \Gamma)/\lambda_3(\alpha, \varepsilon, \gamma)$ монотонно убывает по α на интервале $0 < \alpha < 1/\varepsilon$, т. е.

$$\frac{\partial[\lambda_2(\alpha, \varepsilon, \Gamma)/\lambda_3(\alpha, \varepsilon, \gamma)]}{\partial \alpha} < 0. \quad (5.10)$$

Из (5.8)–(5.10) следует, что минимум $\text{cond}(R_\varepsilon)$ достигается в точке пересечения функций $\lambda_2(\alpha, \varepsilon, \Gamma)/\lambda_3(\alpha, \varepsilon, \gamma)$ и $\lambda_2(\alpha, \varepsilon, \Gamma)$, т. е. в точке α_0 такой, что $\lambda_3(\alpha_0, \varepsilon, \gamma) = 1$, если последняя существует. Решая уравнение (решение единственно)

$$\frac{1}{2} + \alpha_0\gamma + \frac{1}{2}\alpha_0^2\varepsilon^2 - \frac{1}{2}|\alpha_0\varepsilon - 1|\sqrt{(\alpha_0\varepsilon + 1)^2 + 4\alpha_0\gamma} = 1,$$

получим

$$\alpha_0 = \frac{2}{\gamma + 2\varepsilon} < \frac{1}{\varepsilon}$$

и, соответственно,

$$\sigma(R_\varepsilon) \subseteq \left[1, \lambda_2\left(\frac{2}{\gamma + 2\varepsilon}, \varepsilon, \Gamma\right)\right].$$

Осталось лишь сделать замечание о том, что оставшийся интервал $\alpha \geq 1/\varepsilon$ не даст улучшения последней оценки, так как при $\alpha > \alpha_0$ имеем

$$\text{cond}_2(R_\varepsilon) = \lambda_2(\alpha, \varepsilon, \Gamma) > \lambda_2(\alpha_0, \varepsilon, \Gamma)$$

в силу (5.9). ■

5.2.3. Наилучшая оценка погрешности

Теперь из общей теории (см. раздел 3.1) в достаточно произвольной метрике

$$D_z = D_z^T > 0$$

такой, что

$$D_z R_\varepsilon = (D_z R_\varepsilon)^T,$$

следует сходимость метода сопряженных градиентов с оценкой

$$\|z^k - z\|_{D_z} \leq \epsilon_k \|z^0 - z\|_{D_z},$$

где

$$\begin{aligned} \epsilon_k &= \frac{2q_0^k}{1 + q_0^{2k}}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \\ \xi &= \left(\alpha_0 \Gamma + \frac{1}{2} + \varepsilon^2 \frac{\alpha_0^2}{2} + \frac{1}{2} (1 - \varepsilon \alpha_0) \sqrt{4\alpha_0 \Gamma + (\varepsilon \alpha_0 + 1)^2} \right)^{-1}, \\ \alpha_0 &= \frac{2}{\gamma + 2\varepsilon}. \end{aligned}$$

5.3. ОПТИМИЗАЦИЯ МЕТОДА ДЛЯ СЛУЧАЯ РАВНОСИЛЬНОЙ СИСТЕМЫ

В разделе анализируется процедура симметризации и предобуславливания задачи $\tilde{L}_0 z = \tilde{F}$ с параметром β , равносильной базовой системе:

$$\tilde{L}_0 z \equiv \begin{pmatrix} A - \beta B_0 & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f - \beta B C^{-1} B^T \varphi \\ \varphi \end{pmatrix} \equiv \tilde{F}. \quad (5.11)$$

5.3.1. Спектр оператора равносильной задачи

Рассмотрим зависимость собственных значений от параметра α оператора \tilde{R} задачи

$$\tilde{R}z \equiv D^{-1}\tilde{L}_0D^{-1}\tilde{L}_0z = D^{-1}\tilde{L}_0D^{-1}\tilde{F}. \quad (5.12)$$

Теорема 5.3.1. Спектр $\sigma(\tilde{R})$ оператора \tilde{R} в задаче (5.12) принадлежит множеству

$$\Lambda = \{1\} \cup \left\{ \alpha t + \frac{(\beta t - 1)^2}{2} \pm \frac{|\beta t - 1|}{2} \sqrt{4\alpha t + (\beta t - 1)^2} \right\},$$

где $t \in [\gamma, \Gamma]$.

Доказательство. Рассмотрим спектральную задачу

$$\tilde{R}z = \lambda z, \quad (5.13)$$

где

$$\tilde{R} = \begin{pmatrix} I + (\alpha - 2\beta)A^{-1}B_0 + \beta^2(A^{-1}B_0)^2 & A^{-1}B(I - \beta C^{-1}S_0) \\ \alpha C^{-1}B^T(I - \beta A^{-1}B_0) & \alpha C^{-1}S_0 \end{pmatrix}.$$

Для нахождения ее решения используем базис пространства Z , построенный в теореме 1.3.2. Вначале для векторов вида

$$z_k^{(1)} = \{h_k, 0\}, \quad k = 1, \dots, N_u - N_p,$$

непосредственной подстановкой убедимся, что каждый из них удовлетворяет соотношениям (5.13) с $\lambda_k^{(1)} = 1$. Оставшиеся собственные векторы будем искать в виде

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^Tg_j\}.$$

После применения к первому уравнению (5.13) матрицы $C^{-1}B^T$ и замены $C^{-1}B^Tg_j = p_j$ будем иметь

$$\begin{cases} p_j + (\alpha - 2\beta)C^{-1}S_0p_j + \beta^2(C^{-1}S_0)^2p_j - \\ \quad - \kappa^{-1}C^{-1}S_0(I - \beta C^{-1}S_0)p_j = \lambda p_j, \\ -\alpha(I - \beta C^{-1}S_0)p_j + \kappa^{-1}\alpha C^{-1}S_0p_j = \lambda \kappa^{-1}p_j. \end{cases}$$

Каждому собственному вектору p_j задачи $S_0p = tCp$ соответствует собственное значение t_j , $j = 1, \dots, N_p$ по теореме 1.3.1. Зафиксировав его, перепишем полученную систему в виде

$$\begin{cases} 1 + (\alpha - 2\beta)t_j + \beta^2t_j^2 - \kappa^{-1}t_j(1 - \beta t_j) = \lambda, \\ -\kappa^{-1}\alpha(1 - \beta t_j) + \alpha t_j = \lambda, \end{cases}$$

откуда для λ и κ имеем

$$\begin{cases} \beta(\kappa^{-1} + \beta)t_j^2 - (\kappa^{-1} + 2\beta - \alpha)t_j - (\lambda - 1) = 0, \\ \alpha(\kappa^{-1} + \beta)t_j - (\kappa^{-1}\lambda + \alpha) = 0. \end{cases}$$

Исключая из этих уравнений κ , получаем

$$\lambda_j^{(2,3)} = \alpha t_j + \frac{1}{2}(\beta t_j - 1)^2 \pm \frac{1}{2}|\beta t_j - 1|\sqrt{(\beta t_j - 1)^2 + 4\alpha t_j}$$

и, соответственно,

$$\kappa_j^{(2,3)} = \frac{1}{2\alpha}(\beta t_j - 1) \mp \frac{1}{2\alpha} \operatorname{sign}(\beta t_j - 1) \sqrt{(\beta t_j - 1)^2 + 4\alpha t_j} \neq 0$$

при $\beta t_j \neq 1$

и

$$\kappa_j^{(2,3)} = \pm \chi_2 / \chi_1 \sqrt{t_j} \neq 0, \quad \chi_1, \chi_2 > 0 \text{ при } \beta t_j = 1.$$

Проанализируем последнее выражение для $\kappa_j^{(2,3)}$. Рассмотрим для этого случай кратных собственных значений более подробно. Пусть значение параметра β таково, что выражение $(\beta t_j - 1)$ обращается в нуль при некотором t_j . Тогда, используя

$$z_j = \{g_j, -\kappa^{-1}C^{-1}B^T g_j\},$$

после применения к первому уравнению (5.13) оператора $C^{-1}B^T$ и замены $C^{-1}B^T g_j = p_j$ получим

$$\begin{cases} \alpha C^{-1}S_0 p_j = \lambda p_j, \\ \kappa^{-1}\alpha C^{-1}S_0 p_j = \lambda \kappa^{-1}p_j, \end{cases}$$

откуда следует, что любой вектор вида $z_j = \{g_j, -\kappa^{-1}C^{-1}B^T g_j\}$ является собственным; дополнительно имеем $\lambda_j^{(2)} = \lambda_j^{(3)} = \alpha t_j$. Поэтому в данном случае, независимо от кратности собственного значения t_j , собственные векторы задачи (5.13) с одинаковыми g_j можно сделать ортогональными в следующей метрике пространства Z (см. теорему 1.3.2):

$$\begin{aligned} (z_1, z_2)_Z &= \chi_1(Au_1, u_2) + \chi_2(Cp_1, p_2), \\ z_i &= \{u_i, p_i\} \in Z, \quad \chi_i > 0, \quad i = 1, 2, \end{aligned}$$

положив $\kappa_j^{(2,3)} = \pm \chi_2 / \chi_1 \sqrt{t_j}$. Действительно,

$$\begin{aligned} (z_j^{(2)}, z_j^{(3)})_Z &= \chi_1(Ag_j, g_j) + \chi_2 \left(C \frac{C^{-1}B^T g_j}{\kappa_j^{(2)}}, \frac{C^{-1}B^T g_j}{\kappa_j^{(3)}} \right) = \\ &= \chi_1([A - t_j^{-1}BC^{-1}B^T]g_j, g_j) = 0. \end{aligned}$$

Завершение доказательства, как в теореме 5.2.1. ■

5.3.2. Минимизация числа обусловленности

Поскольку оператор \tilde{R} является симметризуемым и положительно определенным по построению, рассмотрим вопрос о минимизации его спектрального числа обусловленности.

Теорема 5.3.2. *При*

$$\alpha_0 = \frac{2}{\gamma} - \frac{1}{\Gamma}, \quad \beta_0 = \frac{1}{\Gamma}$$

спектральное число обусловленности оператора $\tilde{R} = D^{-1}\tilde{L}_0D^{-1}\tilde{L}_0$ в задаче (5.12) принимает наименьшее значение. При этом спектр \tilde{R} принадлежит отрезку:

$$\sigma(\tilde{R}) \subseteq \left[1, \frac{2\Gamma}{\gamma} - 1\right].$$

Доказательство. Обозначим границы спектра $\sup \sigma(\tilde{R})$ и $\inf \sigma(\tilde{R})$ через λ_{\max} и λ_{\min} соответственно. Кроме того, будем использовать обозначения $\lambda_{2,3}$ для точек спектра оператора \tilde{R} , зависящих от параметров α и β :

$$\lambda_{2,3}(\alpha, \beta, t) = \alpha t + \frac{1}{2}(\beta t - 1)^2 \pm \frac{1}{2}|\beta t - 1|\sqrt{(\beta t - 1)^2 + 4\alpha t}$$

(напомним, что, как и ранее, знак «+» относится к λ_2).

В силу непрерывности функций $\lambda_{2,3}$ по параметру t и ограниченности $\sigma(\tilde{R})$ достаточно вместо $\sup(\inf)$ использовать процедуры $\max(\min)$.

Отметим справедливость неравенств

$$0 < \lambda_3(\alpha, \beta, t) < \lambda_2(\alpha, \beta, t) \quad \forall \beta, \forall \alpha > 0, \forall t \in [\gamma, \Gamma].$$

Принимая это во внимание, имеем

$$\text{cond}_2(\tilde{R}(\alpha, \beta)) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\max \left\{ 1, \max_t \lambda_2(\alpha, \beta, t) \right\}}{\min \left\{ 1, \min_t \lambda_3(\alpha, \beta, t) \right\}}.$$

Доказательство построим следующим образом: сначала вычислим $\text{cond}_2(\tilde{R}(\alpha_0, \beta_0))$, а затем покажем, что для всех остальных значений параметров α и β верна оценка

$$\text{cond}_2(\tilde{R}(\alpha, \beta)) > \text{cond}_2(\tilde{R}(\alpha_0, \beta_0)).$$

Итак, рассмотрим значение $\beta_0 = 1/\Gamma$. Имеют место соотношения

$$\begin{aligned}\frac{\partial[\lambda_2(\alpha, 1/\Gamma, t)]}{\partial t} &= -\frac{1}{2\sqrt{\varphi}} \left(\sqrt{\varphi} - \frac{t}{\Gamma} + 1 \right) \left[\frac{\sqrt{\varphi}}{\Gamma} - 2\alpha - \frac{1}{\Gamma} \left(\frac{t}{\Gamma} - 1 \right) \right], \\ \frac{\partial[\lambda_3(\alpha, 1/\Gamma, t)]}{\partial t} &= \frac{1}{2\sqrt{\varphi}} \left(\sqrt{\varphi} + \frac{t}{\Gamma} - 1 \right) \left[\frac{\sqrt{\varphi}}{\Gamma} + 2\alpha + \frac{1}{\Gamma} \left(\frac{t}{\Gamma} - 1 \right) \right], \\ \varphi &= \left(\frac{t}{\Gamma} - 1 \right)^2 + 4\alpha t,\end{aligned}$$

из которых следует, что $\partial[\lambda_{2,3}]/\partial t > 0$ при $\alpha_0 = 2/\gamma - 1/\Gamma > 1/\Gamma$, или

$$\begin{aligned}\max_t \lambda_2 \left(\frac{2}{\gamma} - \frac{1}{\Gamma}, \frac{1}{\Gamma}, t \right) &= \lambda_2 \left(\frac{2}{\gamma} - \frac{1}{\Gamma}, \frac{1}{\Gamma}, \Gamma \right) > 1, \\ \min_t \lambda_3 \left(\frac{2}{\gamma} - \frac{1}{\Gamma}, \frac{1}{\Gamma}, t \right) &= \lambda_3 \left(\frac{2}{\gamma} - \frac{1}{\Gamma}, \frac{1}{\Gamma}, \gamma \right) = 1,\end{aligned}$$

отсюда получаем

$$\text{cond}_2 \left(\tilde{R} \left(\frac{2}{\gamma} - \frac{1}{\Gamma}, \frac{1}{\Gamma} \right) \right) = \lambda_2 \left(\frac{2}{\gamma} - \frac{1}{\Gamma}, \frac{1}{\Gamma}, \Gamma \right) = 2\frac{\Gamma}{\gamma} - 1. \quad (5.14)$$

Теперь докажем, что для остальных значений α и β величина $\text{cond}_2(\tilde{R}(\alpha, \beta))$ строго больше, чем $2\Gamma/\gamma - 1$. Нам понадобятся следующие обозначения:

$$\psi_t = |1 - \beta t|, \quad \varphi_t = \psi_t^2 + 4\alpha t, \quad f(\alpha, \beta) = \frac{\lambda_2(\alpha, \beta, \Gamma)}{\min\{1, \lambda_3(\alpha, \beta, \gamma)\}}.$$

Имеет место соотношение

$$\text{cond}_2(\tilde{R}(\alpha, \beta)) \geq f(\alpha, \beta) = \max \left\{ \lambda_2(\alpha, \beta, \Gamma), \frac{\lambda_2(\alpha, \beta, \Gamma)}{\lambda_3(\alpha, \beta, \gamma)} \right\}.$$

Покажем, что функция $\lambda_2(\alpha, \beta, \Gamma)/\lambda_3(\alpha, \beta, \gamma)$ является убывающей по α , т. е. является отрицательным выражение

$$\frac{\partial}{\partial \alpha} \left[\frac{\lambda_2(\alpha, \beta, \Gamma)}{\lambda_3(\alpha, \beta, \gamma)} \right] = \frac{1}{2[\lambda_3(\alpha, \beta, \gamma)]^2 \sqrt{\varphi_\Gamma} \sqrt{\varphi_\gamma}} X,$$

где

$$\begin{aligned}X &= \psi_\gamma^2 \sqrt{\varphi_\Gamma} \sqrt{\varphi_\gamma} \Gamma + \psi_\Gamma \psi_\gamma^2 \sqrt{\varphi_\gamma} \Gamma + \gamma \psi_\Gamma^2 \psi_\gamma \sqrt{\varphi_\Gamma} + \\ &+ \gamma \psi_\Gamma^3 \psi_\gamma - \psi_\gamma^3 \sqrt{\varphi_\gamma} \Gamma - \psi_\gamma \psi_\Gamma^3 \Gamma - 2\alpha \gamma \Gamma \psi_\gamma \sqrt{\varphi_\Gamma} - \\ &- \gamma \psi_\Gamma^2 \sqrt{\varphi_\Gamma} \sqrt{\varphi_\gamma} - \gamma \psi_\Gamma^3 \sqrt{\varphi_\gamma} - 2\alpha \gamma \Gamma \psi_\Gamma \sqrt{\varphi_\Gamma}.\end{aligned}$$

Убедимся в том, что $X < 0$. Для этого перенесем отрицательные члены X вправо и возведем обе части полученного неравенства

в квадрат. Приводя подобные, имеем

$$\begin{aligned}
 & 4\alpha\gamma\Gamma\psi_\Gamma\psi_\gamma^4(\sqrt{\varphi_\Gamma} + \psi_\Gamma) < \\
 & < 4\alpha\gamma\Gamma\psi_\Gamma\psi_\gamma^3\sqrt{\varphi_\Gamma}(\sqrt{\varphi_\Gamma} + \psi_\Gamma) + 16\alpha^3\gamma^2\Gamma^2\psi_\gamma^2 + 4\alpha\gamma^3\Gamma^3\psi_\Gamma^6 + \\
 & \quad + \{\text{неотр. слагаемые}\}.
 \end{aligned}$$

Это неравенство справедливо при всех значениях α и β (так как $\varphi_\gamma > \psi_\gamma^2$) и из него следует, что сумма модулей отрицательных членов в X больше суммы положительных членов. Таким образом, функция

$$\lambda_2(\alpha, \beta, \Gamma) / \lambda_3(\alpha, \beta, \gamma)$$

убывает по α . Кроме того, справедливо

$$\frac{\partial[\lambda_2(\alpha, \beta, \Gamma)]}{\partial\alpha} = \Gamma \left(1 + \frac{|1 - \beta\Gamma|}{\sqrt{(1 - \beta\Gamma)^2 + 4\alpha\Gamma}} \right) > 0,$$

т. е. функция $\lambda_2(\alpha, \beta, \Gamma)$ возрастает по α , откуда сразу получаем, что минимум $f(\alpha, \beta)$ достигается при α_0 таком, что

$$\lambda_3(\alpha_0, \beta, \gamma) = 1$$

(такая точка единственна в силу монотонности λ_3 по α). Решая уравнение

$$\alpha_0\gamma + \frac{1}{2}(\beta\gamma - 1)^2 - \frac{1}{2}|\beta\gamma - 1|\sqrt{(\beta\gamma - 1)^2 + 4\alpha_0\gamma} = 1,$$

получим

$$\alpha_0 = \begin{cases} \frac{2}{\gamma} - \beta & \text{при } \beta \leq \frac{1}{\gamma}, \\ \beta & \text{при } \beta \geq \frac{1}{\gamma}. \end{cases}$$

При $\beta \leq 1/\gamma$ имеем

$$\begin{aligned}
 f(\alpha_0, \beta) &= \left(\frac{2}{\gamma} - \beta \right) \Gamma + \frac{1}{2}|1 - \beta\Gamma|^2 + \\
 &+ \frac{1}{2}|1 - \beta\Gamma|\sqrt{|1 - \beta\Gamma|^2 + 4\left(\frac{2}{\gamma} - \beta\Gamma\right)} \geq 2\frac{\Gamma}{\gamma} - 1,
 \end{aligned}$$

причем равенство имеет место тогда и только тогда, когда $\beta = 1/\Gamma$. Если же $\beta \geq 1/\gamma$, то

$$f(\alpha_0, \beta) = \beta^2\Gamma^2 \geq \frac{\Gamma^2}{\gamma^2} > 2\frac{\Gamma}{\gamma} - 1,$$

так как $\gamma < \Gamma$. Окончательно

$$\min_{\alpha > 0, \beta \neq 1/\Gamma} \text{cond}_2(\tilde{R}(\alpha, \beta)) \geq \min_{\alpha > 0, \beta \neq 1/\Gamma} f(\alpha, \beta) > 2\frac{\Gamma}{\gamma} - 1. \quad (5.15)$$

Объединяя (5.14) и (5.15), получаем утверждение теоремы. ■

5.3.3. Наилучшая оценка погрешности

Теперь из общей теории (см. раздел 3.1) в достаточно произвольной метрике $D_z = D_z^T > 0$ такой, что $D_z \tilde{R} = (D_z \tilde{R})^T$, следует сходимость метода сопряженных градиентов с оценкой

$$\|z^k - z\|_{D_z} \leq \epsilon_k \|z^0 - z\|_{D_z},$$

где

$$\epsilon_k = \frac{2q_0^k}{1 + q_0^{2k}}, \quad q_0 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \left(\frac{2\Gamma}{\gamma} - 1 \right)^{-1}.$$

5.4. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Рассматриваемые в главе методы, конечно, не относятся к классу релаксационных, но с последними их объединяет реализация, основанная на обращении только матриц A и C , и возможность получения неулучшаемых оценок погрешностей.

Для задач с седловыми операторами идея сочетания симметризации и предобусловливания предложена в работе [40].

Для дифференциальной задачи Стокса оптимизация такого подхода, основанная на использовании спектра Коссера [69, 140], проведена в [83, 84], обобщение этого результата на случай равносильной (регуляризированной) системы получено в [7].

Содержание настоящей главы — это изложение на алгебраическом языке [88] результатов [7, 83, 84] (см. также [8]).

Обратим внимание на серьезные ограничения в практическом применении рассмотренного подхода, связанные с большими вычислительными затратами на каждой итерации (двойными, по сравнению с релаксационными методами).

ЧАСТЬ II

ОБОБЩЕННЫЕ МЕТОДЫ

ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ

Глава содержит сведения о классах оптимизации, необходимую информацию из различных разделов анализа, а также доказательства ключевых результатов теории, которые лежат в основе исследования обобщенных методов для решения задач с седловыми операторами.

6.1. КЛАССЫ ОПТИМИЗАЦИИ

Разнообразие конструкций обобщенных методов, и, соответственно, возможных постановок оптимизационных задач необходимо приводит к понятию *класса оптимизации*.

Классом оптимизации (или постановкой оптимизации) K будем называть непустое подмножество множества всех четверок (A, B, Q, C) , таких что

$$A: U \rightarrow U, \quad B: P \rightarrow U, \quad Q: U \rightarrow U, \quad C: P \rightarrow P$$

являются линейными операторами над конечномерными эрмитовыми пространствами U и P , которые удовлетворяют следующим свойствам

$$A + A^* \geq 0, \quad Q = Q^* > 0, \quad C = C^* > 0, \quad \det \begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \neq 0. \quad (6.1)$$

Символ « $*$ » здесь и далее обозначает эрмитово сопряжение оператора. Условие невырожденности является необходимым и достаточным для существования и единственности решения системы линейных уравнений

$$\begin{cases} Au + Bp = f, \\ B^*u = \varphi, \end{cases} \quad (6.2)$$

при произвольных $f \in U$, $\varphi \in P$. Отметим, что определения пространств U и P неявно входят в определения операторов A , B , Q , C .

Пусть итерационный алгоритм представлен в форме метода простой итерации:

$$z^{k+1} = Tz^k + F, \quad z^0 \in Z \quad k = 0, 1, \dots,$$

где $Z = U \times P$, оператор $T: Z \rightarrow Z$ и вектор $F \in Z$ однозначно определяются зависимостью от (A, B, Q, C) и постоянных параметров $\bar{\alpha} = (\alpha_1, \dots, \alpha_n) \in \mathcal{D} \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$:

$$T \equiv T(\bar{\alpha}; A, B, Q, C), \quad F \equiv F(\bar{\alpha}; A, B, Q, C).$$

Если такая зависимость зафиксирована, то будем говорить, что алгоритм принадлежит классу \mathbb{K} тогда и только тогда, когда соответствующая четверка (A, B, Q, C) принадлежит \mathbb{K} .

Задачей асимптотической оптимизации алгоритма из класса оптимизации \mathbb{K} будем называть следующую задачу

$$q_{\mathbb{K}} = \inf_{\bar{\alpha} \in \mathcal{D}} \sup_{(A, B, Q, C) \in \mathbb{K}} \rho(T(\bar{\alpha}; A, B, Q, C)), \quad (6.3)$$

где $\rho(\cdot)$ — спектральный радиус линейного оператора. При этом $q_{\mathbb{K}}$ будем называть асимптотически оптимальным показателем сходимости, а параметры $\bar{\alpha} \in \mathcal{D}$, на которых достигается $q_{\mathbb{K}}$, если такие существуют, будем называть асимптотически оптимальными (или просто оптимальными) параметрами.

Наибольшую значимость при исследовании линейных задач с седловыми операторами имеют следующие классы оптимизации:

- Класс $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$ состоит из всех четверок (A, B, Q, C) , которые удовлетворяют неравенствам

$$A = A^* > 0, \quad \delta Q \leq A \leq \Delta Q, \quad \gamma C \leq B^* A^{-1} B \leq \Gamma C, \quad (6.4)$$

где $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$ — фиксированные числа.

- Класс $\mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$ состоит из всех четверок (A, B, Q, C) , которые удовлетворяют неравенствам

$$A = A^* > 0, \quad \delta Q \leq A \leq \Delta Q, \quad \gamma C \leq B^* Q^{-1} B \leq \Gamma C, \quad (6.5)$$

где $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$ — фиксированные числа.

- Класс $\mathbb{K}_3(\delta, \Delta, \gamma, \Gamma)$ состоит из всех четверок (A, B, Q, C) , которые удовлетворяют неравенствам

$$A = A^* \geq 0, \quad \delta Q P_B \leq A \leq \Delta Q, \quad \gamma C \leq B^* Q^{-1} B \leq \Gamma C, \quad (6.6)$$

где $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$ — фиксированные числа, P_B — Q -ортогональный ($Q P_B = P_B^* Q$) проектор на $\ker B^*$.

- Класс $\mathbb{K}_3(R, \gamma, \Gamma)$ состоит из всех четверок (A, B, Q, C) , которые удовлетворяют неравенствам

$$\begin{aligned} A &= Q + K, \quad K = -K^*, \\ (Ku, v) &\leq R \|u\|_Q \|v\|_Q \quad \forall u, v \in U, \\ \gamma C &\leq B^* Q^{-1} B \leq \Gamma C, \end{aligned} \quad (6.7)$$

где $0 \leq R$, $0 < \gamma \leq \Gamma$ — фиксированные числа.

Отметим, что при $\delta = \Delta = 1$ и $R = 0$ классы оптимизации \mathbb{K}_1 , \mathbb{K}_2 и \mathbb{K}_3 совпадают, таким образом, в этом случае можно ввести один класс

$$\mathbb{K}(\gamma, \Gamma) \equiv \mathbb{K}_1(1, 1, \gamma, \Gamma) (= \mathbb{K}_2(1, 1, \gamma, \Gamma), \mathbb{K}_3(0, \gamma, \Gamma)).$$

Исследованию алгоритмов из $\mathbb{K}(\gamma, \Gamma)$ посвящена вся первая часть этой книги. Случаи $0 < \delta < \Delta$ и $0 < R$ открывают новые возможности обобщения алгоритмов, представленных в первой части, и расширяют область их практического применения. Основанием служит важное предположение: *обращение оператора Q можно осуществлять значительно эффективнее, чем обращение оператора A* . При этом важно понять, почему теряет свою привлекательность самый эффективный из алгоритмов первой части — метод Узавы.

6.2. ЧТО ПРОИСХОДИТ С АЛГОРИТМОМ УЗАВЫ?

Что происходит с методом Узавы в случае, когда вспомогательная система уравнений с матрицей A решается неточно, т. е. когда норма невязки существенно отличается от машинной точности? Конечно, в данном случае речь не может идти о самой эффективной версии алгоритма, которая основана на методе сопряженных градиентов, так как матрица системы теряет необходимые для этого свойства — симметричность и положительную определенность. Поэтому попытаемся ответить на этот вопрос в более простой ситуации, когда метод Узавы представляет собой оптимальный одношаговый метод, причем оператор предобуславливания C выбран единичным. Это не сильно изменит смысл рассуждений, однако позволит существенно упростить выкладки. Итак, запишем простейшую версию точного метода Узавы для решения (6.2) в виде

$$\frac{p^{k+1} - p^k}{\tau} + B^* A^{-1} B p^k = B^* A^{-1} f - \varphi,$$

или в более удобной развернутой форме:

$$Au^{k+1} = f - Bp^k, \quad (6.8)$$

$$p^{k+1} = p^k + \tau (B^*u^{k+1} - \varphi). \quad (6.9)$$

Будем считать, что точные постоянные в матричном неравенстве

$$\gamma I \leq B^*A^{-1}B \leq \Gamma I$$

известны, параметр τ выбран оптимальным

$$\tau = \frac{2}{\gamma + \Gamma},$$

тогда из общей теории итерационных методов следует (см. раздел 3.1) оценка скорости сходимости алгоритма

$$\|p^k - p\| \leq q_1^k \|p^0 - p\|, \quad (6.10)$$

где $p^0 \in P$ — начальное приближение, $q_1 = (\Gamma - \gamma)/(\Gamma + \gamma)$, $\|\cdot\|$ — обыкновенная эрмитова норма.

Отметим полезное следствие этой оценки — убывание с такой же скоростью нормы невязки уравнения (6.9), т. е. справедливость неравенства

$$\|B^*v^{k+1}\| \leq q_1^k \|B^*v^1\|, \quad (6.11)$$

где $v^k = u^k - u$, $B^*v^k = B^*u^k - \varphi$.

Перейдем теперь к анализу неточного алгоритма Узавы. Если матрица A обращается неточно, то формула (6.8) изменится следующим образом

$$Au^{k+1} = f - Bp^k + \delta^k, \quad (6.12)$$

где δ^k — невязка уравнения на k -й итерации. Будем предполагать, что имеется возможность управлять ее величиной, т. е. решать уравнение до выполнения условия

$$\|\delta^k\| \leq \varepsilon \|B^*u^k - \varphi\|$$

с некоторым наперед выбранным параметром $\varepsilon > 0$. В этой ситуации оценим скорость убывания величины $\|B^*v^k\|$.

Приближения неточного метода Узавы удовлетворяют равенствам

$$Au^k = f - Bp^{k-1} + \delta^{k-1},$$

$$Au^{k+1} = f - Bp^k + \delta^k,$$

откуда следует

$$\begin{aligned} A(u^k - u^{k+1}) &= B(p^k - p^{k-1}) - \delta^k + \delta^{k-1} = \\ &= \tau B(B^*u^k - \varphi) - \delta^k + \delta^{k-1}. \end{aligned}$$

Перепишем последнее выражение в терминах ошибки v^k

$$v^{k+1} = v^k - \tau A^{-1} B B^* v^k - A^{-1} (\delta^k - \delta^{k-1})$$

и применим к его обеим частям оператор B^*

$$B^* v^{k+1} = (I - \tau B^* A^{-1} B) B^* v^k - B^* A^{-1} (\delta^k - \delta^{k-1}).$$

Из этого равенства следует оценка

$$\begin{aligned} \|B^* v^{k+1}\| &\leq q_1 \|B^* v^k\| + \|B^* A^{-1}\| (\|\delta^k\| + \|\delta^{k-1}\|) \leq \\ &\leq q_1 \|B^* v^k\| + \varepsilon \|B^* A^{-1}\| (\|B^* v^k\| + \|B^* v^{k-1}\|). \end{aligned} \quad (6.13)$$

Введем обозначение $\eta = \varepsilon \|B^* A^{-1}\|$ и решим разностное уравнение

$$\beta_{k+1} = q_1 \beta_k + \eta (\beta_k + \beta_{k-1})$$

в предположении, что β_1 и β_2 известны. В результате получим формулу

$$\beta_k = c_1 \mu_+^k + c_2 \mu_-^k, \quad \mu_{\pm} = \frac{q_1 + \eta}{2} \left(1 \pm \sqrt{1 + \frac{4\eta}{(q_1 + \eta)^2}} \right),$$

где постоянные c_1 и c_2 определяются по известным β_1 и β_2 . Из выражения для β_k следует неравенство

$$|\beta_k| \leq \max(|c_1|, |c_2|) (\mu_+^k + |\mu_-|^k) \leq 2 \max(|c_1|, |c_2|) \mu_+^k.$$

Применив к (6.13) полученную оценку, будем иметь

$$\|B^* v^k\| \leq c \hat{q}_1^k, \quad \hat{q}_1 = q_1 + \eta + \frac{q_1 + \eta}{2} \left(\sqrt{1 + \frac{4\eta}{(q_1 + \eta)^2}} - 1 \right),$$

причем постоянная c не зависит от номера итерации k . Это выражение является искомым. Оно означает, что скорость сходимости неточного метода Узава существенно зависит от точности обращения матрицы A . Кроме того, исследования, проведенные в работе [153], показали, что численные эксперименты адекватно описываются рассмотренной выше моделью, т. е. использованные теоретические оценки не являются слишком грубыми.

Суммируя все вышесказанное о неточном алгоритме Узава, приходим к выводу, что невозможность точного решения вспомогательной системы вида $Ax = b$ трансформирует очень эффективный алгоритм Узава — сопряженных градиентов в метод с плохими характеристиками сходимости. Другими словами, обобщать нужно не метод Узава, а метод MSOR и близкие к нему алгоритмы. Таким образом, становится актуальным анализ и оптимизация предобусловленных методов типа Эрроу—Гурвица, что приводит к новым постановкам задач и разработкам новых методов исследования.

6.3. НЕРЕГУЛЯРНЫЕ ЗАДАЧИ С СЕДЛОВОЙ ТОЧКОЙ

Другой важной причиной анализа обобщенных методов является необходимость численного решения нерегулярных задач с седловой точкой.

Определение 6.3.1. Задача с седловой точкой (6.2) называется нерегулярной, если выполнены условия

$$\ker(A + A^*) \neq \{0\}, \quad \ker(A + A^*) \cap \ker B^* = \{0\}, \quad \ker B = \{0\}.$$

Для полноты изложения приведем определение регулярной задачи с седловой точкой.

Определение 6.3.2. Задача с седловой точкой (6.2) называется регулярной, если выполнены условия

$$\ker(A + A^*) = \{0\}, \quad \ker B = \{0\}.$$

Обратим внимание, что определение нерегулярной задачи с седловой точкой не связано с понятием нерегулярности (вырожденности) системы уравнений, о чем говорит следующее утверждение.

Утверждение 6.3.1. Нерегулярная задача с седловой точкой (6.2) невырождена.

Доказательство. Рассмотрим вектор $\{u, p\} \in Z$ такой, что $Au + Bp = 0$, $B^*u = 0$, тогда $u \in \ker B^*$ и

$$0 = (Au + Bp, u) = (Au, u) + (Bp, u) = (Au, u) + (p, B^*u) = (Au, u).$$

Следовательно,

$$((A + A^*)u, u) = 0,$$

а так как $A + A^* \geq 0$, то

$$u \in \ker(A + A^*).$$

Таким образом,

$$u \in \ker(A + A^*) \cap \ker B^* = \{0\}.$$

Из последнего условия немедленно следует, что $u = 0$, $Bp = 0$, а так как $\ker B = \{0\}$, то и $p = 0$. Приходим к выводу, что седловой оператор имеет только тривиальное ядро, что и означает невырожденность системы (6.2). Утверждение доказано. ■

Для решения нерегулярных симметричных седловых задач метод Узавы, а также любой другой из методов, рассмотренных в первой части, в явном виде неприменимы, что следует из отсутствия

обратного к A оператора. Тем не менее, нерегулярные седловые задачи возникают в приложениях и построение методов их решения является актуальной задачей. Для того чтобы сохранить полученные результаты для нерегулярных седловых задач можно воспользоваться идеей регуляризации (см. [27]) и вместо (6.2) рассматривать равносильную задачу с седловой точкой

$$\begin{cases} A_\nu u + Bp = f + \nu BC^{-1}\varphi, \\ B^*u = \varphi, \end{cases} \quad (6.14)$$

где $A_\nu = A + \nu BC^{-1}B^*$, $C: P \rightarrow P$ — линейный оператор, $C = C^* > 0$, $\nu > 0$.

Утверждение 6.3.2. Пусть (6.2) — нерегулярная задача с седловой точкой и $\nu > 0$, тогда задача (6.14) является регулярной.

Доказательство. Если $(A_\nu + A_\nu^*)u = 0$, то

$$0 = (A_\nu u + A_\nu^* u, u) = ((A + A^*)u, u) + 2\nu(C^{-1}B^*u, B^*u).$$

Так как

$$((A + A^*)u, u) \geq 0, \quad (C^{-1}B^*u, B^*u) \geq 0,$$

отсюда сразу следует, что

$$(A + A^*)u = 0, \quad B^*u = 0$$

или, если задача (6.2) нерегулярна,

$$u \in \ker(A + A^*) \cap \ker B^* = \{0\}.$$

Таким образом,

$$\ker(A_\nu + A_\nu^*) = \{0\},$$

что в совокупности с равенством $\ker B = \{0\}$, имеющим место для (6.2), приводит к сформулированному результату. Утверждение доказано. ■

За счет несложного как с точки зрения теории, так и с точки зрения вычислений, преобразования (6.14), можно добиться регуляризации исходной задачи, что предоставляет формальную возможность использования для ее решения любых методов, пригодных для решения регулярных седловых задач. Однако на этом пути возникают новые препятствия. Рассмотрим, например, метод Узавы для задачи (6.14). Здесь на каждом шаге требуется вычисление оператора $B^*A_\nu^{-1}B$, т. е. эффективное обращение A_ν . В отличие от обращения оператора A в регулярных задачах, соответствующая процедура для A_ν , как правило, труднореализуема. Данную ситуацию усложняет необходимость выбора оптимального в некотором

смысле значения $\nu > 0$. Обобщенные методы, требующие на каждом шаге обращения спектрально-эквивалентного оператора $Q = Q^* > 0$ вместо A_ν , лишены перечисленных недостатков.

При использовании любого из обобщенных алгоритмов регуляризации вида (6.14) вводит в него дополнительный релаксационный параметр $\nu > 0$, который можно задействовать для оптимизации. При этом можно пойти в двух направлениях: рассматривать ν как независимый параметр алгоритма и оптимизировать по нему совместно с остальными параметрами, либо рассматривать ν как параметр задачи (6.14) и «оптимизировать» ее класс, т. е. выбирать ν таким образом, что полученная задача помещается в наилучший (с точки зрения оценок сходимости) из возможных классов оптимизации. Естественно, первый способ оптимизации с точки зрения результата более предпочтителен, однако он приводит к крайне сложным постановкам оптимизационных задач, поэтому в настоящей книге отдается предпочтение второму, более удобному для анализа, способу.

Остается понять — каким образом можно поставить задачу оптимизации для седловых задач в нерегулярном случае. Естественно, на этот вопрос не существует однозначного ответа. Мы предлагаем рассмотреть конструкцию класса оптимизации \mathbb{K}_{2s} для симметричных нерегулярных задач [26, 27]. Во-первых, мы пойдем по пути обобщения класса \mathbb{K}_2 на нерегулярные задачи, поэтому сохраним условия, на которые не влияет свойство нерегулярности:

$$A = A^* \geq 0, \quad A \leq \Delta Q, \quad \gamma C \leq B^* Q^{-1} B \leq \Delta C. \quad (6.15)$$

Во-вторых, расширим оставшееся условие $\delta Q \leq A$ таким образом, чтобы оно могло охватывать любые нерегулярные задачи. Из условия невырожденности следует, что

$$\ker Q^{-1/2} A Q^{-1/2} \cap \ker B^* Q^{-1/2} = \{0\}$$

и найдется $\delta > 0$ такое, что $Q^{-1/2} A Q^{-1/2} \geq \delta \tilde{P}_B$, где \tilde{P}_B — ортогональный проектор на $\ker B^* Q^{-1/2}$ или, что эквивалентно,

$$A \geq \delta Q P_B, \quad (6.16)$$

где P_B — Q -ортогональный проектор на $\ker B^*$, $P_B = Q^{-1/2} \tilde{P}_B Q^{1/2}$. Зафиксировав величины $0 < \delta < \Delta$, $0 < \gamma \leq \Gamma$ и определив класс как совокупность (A, B, Q, C) , удовлетворяющих условиям (6.15), (6.16), получаем конструкцию класса \mathbb{K}_{2s} . Из построения следует, что имеет место включение

$$\mathbb{K}_2(\delta, \Delta, \gamma, \Gamma) \subset \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma).$$

В рамках условий (6.15), (6.16) и при $\nu > 0$ для регуляризованной задачи (6.14) справедлива оценка

$$\begin{aligned} A_\nu &= A + \nu BC^{-1}B^* \geq \delta QP_B + \nu BC^{-1}B^* \geq \\ &\geq Q^{1/2}(\delta \tilde{P}_B + \nu Q^{-1/2}BC^{-1}B^*Q^{-1/2})Q^{1/2} \geq \\ &\geq Q^{1/2}(\min\{\delta, \nu\gamma\}I)Q^{1/2} = \min\{\delta, \nu\gamma\}Q, \end{aligned}$$

где последнее неравенство следует из того факта, что операторы

$$\tilde{P}_B \text{ и } Q^{-1/2}BC^{-1}B^*Q^{-1/2}$$

коммутируют. Добавляя естественную верхнюю оценку, приходим к неравенствам

$$0 < \min\{\delta, \nu\gamma\}Q \leq A_\nu \leq (\Delta + \nu\Gamma)Q$$

и импликации

$$\begin{aligned} (A, B, Q, C) \in \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma) &\Rightarrow \\ \Rightarrow (A_\nu, B, Q, C) \in \mathbb{K}_2(\min\{\delta, \nu\gamma\}, \Delta + \nu\Gamma, \gamma, \Gamma). \end{aligned} \quad (6.17)$$

Используя тот факт, что все оценки скорости сходимости, полученные в книге, монотонно зависят от величины $\text{cond}_2 Q^{-1}A$, получаем задачу «оптимизации» по параметру $\nu > 0$:

$$(\Delta + \nu\Gamma) / \min\{\delta, \nu\gamma\} \rightarrow \min_{\nu > 0},$$

решение которой единственно и достигается при $\nu_0 = \delta/\gamma$. В результате приходим к «оптимальной» импликации

$$\begin{aligned} (A, B, Q, C) \in \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma) &\Rightarrow \\ \Rightarrow (A_{\delta/\gamma}, B, Q, C) \in \mathbb{K}_2(\delta, \Delta + \delta\Gamma/\gamma, \gamma, \Gamma), \end{aligned}$$

при помощи которой легко распространять результаты, получаемые для класса \mathbb{K}_2 , на класс \mathbb{K}_{2s} .

6.4. ВСПОМОГАТЕЛЬНЫЕ УТВЕРЖДЕНИЯ

6.4.1. Сведения из анализа

Изучение обобщенных методов для решения задач с седловыми операторами требует привлечения значительно более обширного математического аппарата исследования, чем это требовалось для релаксационных методов. В этом разделе сосредоточены известные определения и утверждения из различных разделов анализа, которые являются основными для дальнейшего построения теории обобщенных методов решения задач с седловыми операторами.

Линейный анализ

Простые аналитические оценки спектра спектра линейного оператора в эрмитовом пространстве можно получить с помощью *лебегова множества* оператора.

Определение 6.4.1. Лебеговым множеством линейного оператора A в эрмитовом пространстве U называется множество

$$\Lambda(A) = \left\{ \frac{(Au, u)}{(u, u)} : u \in U \setminus \{0\} \right\} \subseteq \mathbb{C}.$$

Теорема 6.4.1 (Свойства лебегова множества). Пусть A — линейный оператор в конечномерном эрмитовом пространстве U , тогда

- 1) $\sigma(A) \subseteq \Lambda(A)$, где $\sigma(A)$ — спектр оператора A ;
- 2) $\Lambda(A)$ — ограниченное замкнутое выпуклое множество в \mathbb{C} ;
- 3) если A — нормальный, т. е.

$$AA^* = A^*A,$$

то

$$\Lambda(A) = \text{conv } \sigma(A),$$

где conv обозначает замкнутую выпуклую оболочку множества.

Доказательство указанных свойств $\Lambda(A)$ можно найти в [33, с. 107]. Нормальными операторами, в частности, являются сопряженные операторы ($A = A^*$), а также любые комплексные рациональные функции от них, множество полюсов которых не пересекается со спектром A .

Исследование задач, связанных с выпуклыми оболочками множеств, существенно упрощается при помощи теоремы Каратеодори [5, с. 211]:

Теорема 6.4.2 (Каратеодори). Пусть U — линейное пространство над \mathbb{R} размерности $n \in \mathbb{N}$ и $M \subseteq U$ — непустое подмножество. Тогда для любой точки u_0 из выпуклой оболочки M существуют $n + 1$ точка $u_1, \dots, u_{n+1} \in M$ и числа $\lambda_1, \dots, \lambda_{n+1} \in \mathbb{R}$, такие, что

$$\lambda_k \geq 0, \quad k = 1, \dots, n + 1,$$

$$\lambda_1 + \dots + \lambda_{n+1} = 1 \quad \text{и} \quad \lambda_1 u_1 + \dots + \lambda_{n+1} u_{n+1} = u_0.$$

Другими словами, теорема утверждает, что любая точка выпуклой оболочки всего множества M содержится в выпуклой оболочке не более чем $n + 1$ точки из M .

Комплексный анализ

Исследование спектров операторов часто сводится к анализу решений параметризованных алгебраических уравнений. Важную роль при этом играют оценки модулей решений этих уравнений. Наиболее простые и точные оценки получаются с использованием теоремы Шура-Кона [160, с. 491]:

Теорема 6.4.3 (Шур-Кон). Пусть

$$P(z) = \sum_{k=0}^n a_k z^k$$

-- многочлен степени $n \geq 2$ с коэффициентами из \mathbb{C} , $a_n \neq 0$. Все корни многочлена $P(z)$ лежат в $U = \{z \in \mathbb{C}: |z| < 1\}$ тогда и только тогда, когда $|a_n| > |a_0|$ и все корни взаимного многочлена

$$P^*(z) = \sum_{k=0}^{n-1} a_k^* z^k,$$

где

$$a_k^* = \overline{a_n} a_{k+1} - a_0 \overline{a_{n-k-1}}, \quad k = 0, \dots, n-1,$$

также лежат в U .

Теорема Шура-Кона носит конструктивный характер — для заданного многочлена степени $n \geq 2$ применение теоремы сводит задачу к анализу корней многочлена степени не более $n-1$, последовательное применение этой теоремы сводит задачу к многочлену первой степени. Для полноты изложения приведем также альтернативный способ записи взаимного многочлена:

$$P^*(z) = z^{-1}(\overline{a_n} P(z) - a_0 z^n \overline{P(1/\bar{z})}).$$

Минимаксный анализ

Общий подход к оптимизации алгоритмов для решения седловых задач состоит в сведении проблемы оптимизации к задаче

$$q = \min_{x \in X} \max_{y \in Y} f(x, y)$$

где f — вещественная функция, определенная и непрерывная на множестве $X \times Y$. Задачи такого вида называются *минимаксными*. Важную роль в исследовании таких задач играет следующая теорема:

Теорема 6.4.4 (Существование и непрерывность решения). Пусть $X \subset \mathbb{R}^n$, $n \in \mathbb{N}$, $Y \subset \mathbb{R}^m$, $m \in \mathbb{N}$ — непустые компактные

множества, $Z \subseteq \mathbb{R}^s$, $s \in \mathbb{N}$ — область, $g(x, z)$ — непрерывная функция в открытой окрестности $X \times Z$, $h(y, z)$ — непрерывная функции в открытой окрестности $Y \times Z$, причем множества

$$X_z = \{x \in X: g(x, z) \leq 0\}, \quad Y_z = \{y \in Y: h(y, z) \leq 0\}$$

являются непустыми для любого $z \in Z$. Тогда для любой непрерывной в открытой окрестности $X \times Y \times Z$ функции $f(x, y, z)$ существует решение задачи

$$\varphi(z) = \min_{x \in X_z} \max_{y \in Y_z} f(x, y, z),$$

причем $\varphi(z)$ — функция, непрерывная в Z .

Формулировка теоремы носит специфический характер и в этом виде не встречается в известной авторам литературе, поэтому приведем ее доказательство, опираясь на известные свойства непрерывных функций на компактных множествах [49, с. 124].

Доказательство. Для любых $x \in X$, $z \in Z$, в силу непрерывности f на компакте $\{x\} \times Y_z \times \{z\}$, существует $\max_{y \in Y_z} f(x, y, z)$, который обозначим $\psi(x, z)$. Зафиксируем некоторое компактное подмножество $V \subset Z$.

Для любого $\varepsilon > 0$, в силу равномерной непрерывности функций g , h , f на компактах $X \times V$, $Y \times V$ и $X \times Y \times V$ соответственно, существуют $\delta_1 \geq \delta_2 > 0$ такие, что

$$d_H(X_{z_1}, X_{z_2}) < \delta_1, \quad d_H(Y_{z_1}, Y_{z_2}) < \delta_1, \\ |f(x_1, y_1, z_1) - f(x_2, y_2, z_2)| < \varepsilon$$

для $x_{1,2} \in X$, $y_{1,2} \in Y$, $z_{1,2} \in V$ таких, что

$$\rho(x_1, x_2) < \delta_1, \quad \rho(y_1, y_2) < \delta_1, \quad \rho(z_1, z_2) < \delta_2.$$

Здесь $\rho(\cdot, \cdot)$ — стандартное евклидово расстояние между точками, $d_H(\cdot, \cdot)$ — хаусдорфово расстояние между множествами:

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \rho(a, b), \sup_{b \in B} \inf_{a \in A} \rho(a, b) \right\}.$$

Для любых $x_{1,2} \in X$, $z_{1,2} \in V$ таких, что

$$\rho(x_1, x_2) < \delta_1, \quad \rho(z_1, z_2) < \delta_2$$

имеет место неравенство

$$d_H(Y_{z_1}, Y_{z_2}) < \delta_1$$

и из ранее указанных оценок следует:

$$\begin{aligned} \max_{y \in Y_{z_2}} f(x_1, y, z_1) &< \max_{y \in Y_{z_2}} f(x_2, y, z_2) + \varepsilon, \\ \max_{y \in Y_{z_2}} f(x_2, y, z_2) &< \max_{y \in Y_{z_2}} f(x_1, y, z_1) + \varepsilon, \\ \max_{y \in Y_{z_1}} f(x_1, y, z_1) &< \max_{y \in Y_{z_2}} f(x_1, y, z_1) + \varepsilon, \\ \max_{y \in Y_{z_2}} f(x_1, y, z_1) &< \max_{y \in Y_{z_1}} f(x_1, y, z_1) + \varepsilon. \end{aligned}$$

Таким образом, справедливо

$$\begin{aligned} |\psi(x_1, z_1) - \psi(x_2, z_2)| &= \left| \max_{y \in Y_{z_1}} f(x_1, y, z_1) - \max_{y \in Y_{z_2}} f(x_2, y, z_2) \right| \leq \\ &\leq \left| \max_{y \in Y_{z_1}} f(x_1, y, z_1) - \max_{y \in Y_{z_2}} f(x_1, y, z_1) \right| + \\ &+ \left| \max_{y \in Y_{z_2}} f(x_1, y, z_1) - \max_{y \in Y_{z_2}} f(x_2, y, z_2) \right| < 2\varepsilon. \end{aligned}$$

Отсюда следует, что $\psi(x, z)$ непрерывна на компакте $X \times V$.

Для любого $z \in V$, в силу непрерывности $\psi(x, z)$ на компакте

$$X_z \times \{z\} \subseteq X \times V,$$

корректно определена функция $\varphi(z) = \min_{x \in X_z} \psi(x, z)$, для которой аналогичным образом получается оценка

$$\begin{aligned} |\varphi(z_1) - \varphi(z_2)| &= \left| \min_{x \in X_{z_1}} \psi(x, z_1) - \min_{x \in X_{z_2}} \psi(x, z_2) \right| \leq \\ &\leq \left| \min_{x \in X_{z_1}} \psi(x, z_1) - \min_{x \in X_{z_2}} \psi(x, z_1) \right| + \\ &+ \left| \min_{x \in X_{z_2}} \psi(x, z_1) - \min_{x \in X_{z_2}} \psi(x, z_2) \right| < 4\varepsilon. \end{aligned}$$

Итак, функция $\varphi(z)$ корректно определена и непрерывна на любом компактном подмножестве $V \subset Z$, а так как любая точка $z \in Z \subseteq \mathbb{R}^s$ содержится в некоторой компактной окрестности, то $\varphi(z)$ определена и непрерывна всюду в Z . Теорема доказана. ■

Важное замечание: при исследовании алгоритмов в этой части книги теорема применяется следующим образом:

$$\begin{aligned} y &= (s, t) \in \mathbb{R}^2, \quad z = (\delta, \Delta, \gamma, \Gamma) \in \mathbb{R}^4, \\ f(x, y, z) &\equiv f(s, t; x), \quad g(x, z) \equiv g(x), \\ h(y, z) &= \max\{\delta - s, s - \Delta, \gamma - t, t - \Gamma\}. \end{aligned}$$

В этом случае минимаксная задача может быть представлена в форме

$$\varphi(\delta, \Delta, \gamma, \Gamma) = \min_{x \in X} \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} f(s, t; x),$$

где x — вектор итерационных параметров (например, $x = (\alpha, \beta, \tau)$ для GMSOR). В дальнейшем упоминание о непрерывной зависимости значений задачи подразумевает именно такое применение теоремы.

При оптимизации некоторых алгоритмов для задач с седловыми операторами полезную роль играет следующее достаточное условие [37, с. 243]:

Лемма 6.4.1 (Достаточный признак оптимальности). Пусть вещественнозначные функции

$$f_1(x), \dots, f_k(x), \quad k > 1$$

определены в некоторой окрестности точки $x_0 \in \mathbb{R}^n$, $n < k$, дифференцируемы в этой точке и $f_1(x_0) = f_2(x_0) = \dots = f_k(x_0)$.

Если

$$0 \in] \operatorname{conv} \{v_1, \dots, v_k\} [,$$

где $v_i = \nabla f_i(x_0)$, conv — замкнутая выпуклая оболочка, $] \cdot [$ — множество внутренних точек, то x_0 является точкой строгого локального минимума функции $f(x) = \max \{f_1(x), \dots, f_k(x)\}$.

Нелинейный анализ

Для исследования непрерывных нелинейных задач с седловыми операторами успешно применяется техника линейаризации, базирующаяся на понятии обобщенной производной [139, с. 69]. Пусть Ω — область в \mathbb{R}^n , $n \in \mathbb{N}$, а $F : \Omega \rightarrow \mathbb{R}^n$ — непрерывное отображение. Обозначим через $\mathcal{D}(F)$ — множество точек в Ω , в которых F обладает производной по Фреше F' .

Определение 6.4.2. Обобщенной производной отображения F в точке $x \in [\mathcal{D}(F)]$ называется замкнутое выпуклое множество линейных операторов в \mathbb{R}^n :

$$\partial F(x) = \operatorname{conv} \left\{ A : A = \lim_{y \in \mathcal{D}(F) \rightarrow x} F'(y) \right\}.$$

Ключевую роль в дальнейших исследованиях играют условия существования обобщенной производной [139, с. 63] и обобщенная теорема о среднем [139, с. 72]:

Теорема 6.4.5 (Радемахер). Пусть F удовлетворяет условию Липшица, т. е. существует константа $L \geq 0$ такая, что для любых $x_0, x_1 \in \Omega$ в евклидовой норме выполнено неравенство

$$\|F(x_1) - F(x_0)\| \leq L \|x_1 - x_0\|$$

(условие Липшица), тогда

- 1) F обладает производной по Фреше почти всюду в Ω ;
- 2) F обладает обобщенной производной всюду в Ω ;
- 3) для любых $x \in \Omega$ и $A \in \partial F(x)$ имеет место неравенство

$$\|A\| = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} \leq L.$$

Теорема 6.4.6 (О среднем). Пусть F обладает обобщенной производной в каждой точке отрезка $[x_0, x_1] \subset \Omega$, тогда существует линейный оператор

$$A \in \text{conv}\{\partial F(x) : x \in [x_0, x_1]\}$$

такой, что $F(x_1) - F(x_0) = A(x_1 - x_0)$.

В совокупности теорема Радемахера и теорема о среднем позволяют осуществлять редукцию анализа сходимости алгоритмов для решения нелинейных задач к анализу алгоритмов для решения линейных задач.

6.4.2. Спектральные свойства одного пучка операторов

Рассмотрим операторный пучок χ :

$$\begin{aligned} \chi(\lambda) &= f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f &\in \mathbb{C}[\lambda, s], \quad g \in \mathbb{C}[\lambda, t], \quad h \in \mathbb{C}[\lambda], \end{aligned} \tag{6.18}$$

где L, G - линейные эрмитовы операторы в эрмитовом конечномерном пространстве U , $\lambda \in \mathbb{C}$, $f(\lambda, s)$, $g(\lambda, t)$, $h(\lambda)$ - многочлены с коэффициентами в \mathbb{C} . Число $\lambda \in \mathbb{C}$ называется *собственным значением* пучка операторов (6.18), если $\ker \chi(\lambda) \neq \{0\}$. Множество всех собственных значений пучка (6.18) называется *спектром* пучка операторов (6.18) и обозначается $\sigma(\chi)$.

Задача нахождения характеристик спектра пучка вида (6.18) часто возникает в исследовании алгоритмов для решения задач с деловыми операторами. Для последних исследование характеристик спектра проводится в следующих предположениях:

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad \delta_1, \delta_2 \in \mathbb{R}, \quad \delta_1 \leq \delta_2. \tag{6.19}$$

$$G = G^* \geq 0, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2], \quad \gamma_1, \gamma_2 \in \mathbb{R}, \quad 0 < \gamma_1 \leq \gamma_2. \tag{6.20}$$

Всюду далее будем считать, что $\chi(\lambda)$ — операторный пучок вида (6.18).

Теорема 6.4.7 (Нижняя оценка). Пусть $M \subset \mathbb{C}$ — конечное непустое множество, причем количество элементов в M не превосходит $\dim U$ и для любого $\lambda_0 \in M$ найдутся $s_0 \in [\delta_1, \delta_2]$, $t_0 \in [\gamma_1, \gamma_2]$ такие, что

$$f(\lambda_0, s_0) = 0 \quad \text{или} \quad f(\lambda_0, s_0)g(\lambda_0, t_0) + h(\lambda_0)t_0 = 0,$$

тогда существуют линейные операторы L и G в U , удовлетворяющие (6.19), (6.20) и такие, что $M \subseteq \sigma(\chi)$.

Доказательство. Пусть $\{e_i\}_{i=1}^{N_U}$ — произвольный ортонормированный базис в U , $N_U = \dim U$. Пусть

$$M = \{\lambda_i\}_{i=1}^{N_M},$$

где $N_M > 0$ — количество элементов в M , $N_M \leq N_U$. По условию существуют

$$s_i \in [\delta_1, \delta_2], \quad t_i \in \{0\} \cup [\gamma_1, \gamma_2]$$

такие, что

$$f(\lambda_i, s_i)g(\lambda_i, t_i) + h(\lambda_i)t_i = 0, \quad i = 1, \dots, N_M.$$

Зададим операторы L и G следующими соотношениями при $i = 1, \dots, N_U$:

$$Le_i = \begin{cases} s_i e_i, & \text{если } i < N_M, \\ s_{N_M} e_i, & \text{если } i \geq N_M, \end{cases} \quad Ge_i = \begin{cases} t_i e_i, & \text{если } i < N_M, \\ t_{N_M} e_i, & \text{если } i \geq N_M. \end{cases}$$

Несложно убедиться, что так определенные операторы L и G удовлетворяют (6.19), (6.20), причем при $i = 1, \dots, N_M$ выполнено

$$\chi(\lambda_i)e_i = (f(\lambda_i, s_i)g(\lambda_i, t_i) + h(\lambda_i)t_i)e_i = 0.$$

т. е. $\lambda_i \in \sigma(\chi)$. Теорема доказана. ■

Основное следствие этого результата состоит в том, что любая верхняя оценка $\sigma(\chi)$ на классе операторов (6.19), (6.20) содержит множество решений $\lambda(s, t)$, удовлетворяющих одному из уравнений

$$\begin{aligned} f(\lambda, s) &= 0, \\ f(\lambda, s)g(\lambda, t) + h(\lambda)t &= 0 \end{aligned} \tag{6.21}$$

при различных $s \in [\delta_1, \delta_2]$, $t \in [\gamma_1, \gamma_2]$. В дальнейшем будут указаны условия, при которых эта оценка является точной. О нетривиальности таких результатов свидетельствует пример (см. оценку для

метода GMSOR в доказательстве теоремы 7.2.1):

$$\begin{aligned} f(\lambda, s) &= (\lambda - 1)s + \tau, \quad g(\lambda, t) = \lambda - 1, \quad h(\lambda) = \lambda\tau\left(\beta + \frac{\tau}{\alpha}\right) - \tau\beta, \\ \delta_1 &= 1, \quad \delta_2 = 3, \quad \gamma_1 = \gamma_2 = 1, \quad \tau = \alpha = 1, \quad \beta = 0, \\ L &= \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

В этом случае все четыре собственные значения пучка χ могут быть найдены из уравнения

$$0 = \det \chi(\lambda) = 3\lambda(\lambda - 1)\left(\lambda^2 - \lambda + \frac{1}{3}\right), \quad (6.22)$$

в то же время алгебраические уравнения (6.21) соответствуют равенствам:

$$\begin{aligned} \lambda &= 1 - s^{-1}, \quad s \in [1/3, 1], \\ \lambda^2 - 2(1 - s)\lambda + (1 - s) &= 0, \quad s \in [1/3, 1]. \end{aligned} \quad (6.23)$$

Комплексно сопряженные корни (6.22) могут удовлетворять только второму уравнению (6.23) и только при условии совместности системы

$$\begin{cases} 2(1 - s) = 1, \\ (1 - s) = 1/3, \end{cases}$$

что невыполнимо при любых $s \in \mathbb{R}$. Противоречие приводит к тому факту, что, в общем случае, не все собственные значения χ описываются корнями (6.21).

Отметим, что при анализе свойств алгоритмов для решения седловых задач собственные значения пучка $\chi(\lambda)$, удовлетворяющие условию $g(\lambda, 0) = 0$, чаще всего исключаются из рассмотрения. В связи с этим все последующие утверждения построены таким образом, чтобы отбросить эти значения еще на этапе формулировки условий. Тем не менее, это не ограничивает общности оценок, которые, в случае необходимости, могут быть свободно расширены за счет добавления решений уравнения $g(\lambda, 0) = 0$.

Теорема 6.4.8 (Верхняя оценка). Пусть выполнены предположения (6.19), (6.20), $\lambda_0 \in \sigma(\chi)$ и $g(\lambda_0, 0) \neq 0$, тогда или найдется $s_0 \in [\delta_1, \delta_2]$ такое, что

$$f(\lambda_0, s_0) = 0,$$

или, если такого s_0 не существует, то

$$\text{conv}\{t^{-1}g(\lambda_0, t) \mid t \in [\gamma_1, \gamma_2]\} + \text{conv}\{f(\lambda_0, s)^{-1}h(\lambda_0) \mid s \in [\delta_1, \delta_2]\} \ni 0,$$

где conv обозначает замкнутую выпуклую оболочку множества в \mathbb{C} .

Доказательство. Пусть выполнены условия теоремы, тогда существует вектор $u \in U \setminus \{0\}$ такой, что $\chi(\lambda_0)u = 0$. Обозначим линейный оператор $f(\lambda_0, L)$ символом S . Если $f(\lambda_0, s_0) = 0$ для некоторого $s_0 \in [\delta_1, \delta_2]$, то теорема доказана.

Предположим, что $f(\lambda_0, s) \neq 0$ для любого $s \in [\delta_1, \delta_2]$, тогда

$$\sigma(S) = \{f(\lambda_0, s) \mid s \in \sigma(L)\} \subseteq \{f(\lambda_0, s) \mid s \in [\delta_1, \delta_2]\}, \quad (6.24)$$

и следовательно, $0 \notin \sigma(S)$, т. е. оператор S обратим. Из указанного предположения также следует, что $Gu \neq 0$. Действительно, если $Gu = 0$, то выполнено

$$0 = \chi(\lambda_0)u = f(\lambda_0, L)g(\lambda_0, G)u = Sg(\lambda_0, 0)u,$$

а так как по условию $g(\lambda_0, 0) \neq 0$, то $u \in \ker S$, что приводит к противоречию.

Представим вектор u в виде $u_0 + u_1$, где $u_0 \in H \equiv \ker G$, $u_1 \in K \equiv (\ker G)^\perp = \text{Im } G$, а спектральную задачу в виде

$$g(\lambda_0, G)u + S^{-1}h(\lambda_0)Gu = 0.$$

Умножим справа скалярно обе части уравнения на Gu , в затем разделим на $(Gu, Gu) \neq 0$, при этом воспользуемся равенствами

$$Gu = Gu_1, \quad (u, Gu) = (u_1, Gu_1),$$

тогда получим

$$\frac{(g(\lambda_0, G)u_1, Gu_1)}{(Gu_1, Gu_1)} + \frac{(S^{-1}h(\lambda_0)Gu_1, Gu_1)}{(Gu_1, Gu_1)} = 0. \quad (6.25)$$

Оператор $F = G|_K : K \rightarrow K$ самосопряжен в K и $\sigma(F) \subseteq [\gamma_1, \gamma_2]$, следовательно, оператор $F^{-1}g(\lambda_0, F)$ — нормальный и

$$\sigma(g(\lambda_0, F)) = g(\lambda_0, \sigma(F)),$$

таким образом, имеем

$$\begin{aligned} \frac{(g(\lambda_0, G)u_1, Gu_1)}{(Gu_1, Gu_1)} &= \frac{(F^{-1}g(\lambda_0, F)Fu_1, Fu_1)}{(Fu_1, Fu_1)} \subseteq \\ &\subseteq \text{conv} \{t^{-1}g(\lambda_0, t) \mid t \in [\gamma_1, \gamma_2]\} \end{aligned} \quad (6.26)$$

Так как оператор L самосопряженный, то оператор $S = f(\lambda_0, L)$ — нормальный, следовательно, оператор S^{-1} также нормальный, поэтому справедливо

$$\frac{(S^{-1}h(\lambda_0)Gu_1, Gu_1)}{(Gu_1, Gu_1)} \subseteq \text{conv} \{f(\lambda_0, s)^{-1}h(\lambda_0) \mid s \in [\delta_1, \delta_2]\}. \quad (6.27)$$

Из соотношений (6.25)–(6.27) немедленно следует утверждение теоремы. Теорема доказана. \blacksquare

Во многих задачах априорно известно, что спектр операторного пучка (6.18) является вещественным, тогда полезно следующее следствие теоремы 6.4.8.

Следствие 6.4.1. Пусть выполнены условия теоремы 6.4.8 и многочлены $f(\lambda, s)$ и $g(\lambda, t)$ имеют вещественные коэффициенты. Если $\lambda_0 \in \mathbb{R}$, тогда существуют $s_0 \in [\delta_1, \delta_2]$, $t_0 \in [\gamma_1, \gamma_2]$ такие, что

$$f(\lambda_0, s_0) = 0 \quad \text{или} \quad f(\lambda_0, s_0)g(\lambda_0, t_0) + h(\lambda_0)t_0 = 0.$$

Доказательство. Предположим, что для любого $s \in [\delta_1, \delta_2]$ выполнено неравенство $f(\lambda_0, s) \neq 0$, тогда по теореме 6.4.8 имеет место включение

$$\begin{aligned} & \text{conv} \{t^{-1}g(\lambda_0, t) \mid t \in [\gamma_1, \gamma_2]\} + \\ & + \text{conv} \{f(\lambda_0, s)^{-1}h(\lambda_0) \mid s \in [\delta_1, \delta_2]\} \ni 0. \end{aligned}$$

Здесь и далее выражение вида $f(\lambda_0, s)^{-1}$ имеет смысл обратной величины $-1/f(\lambda_0, s)$.

Из условия следует, что многочлен $P(s) \equiv f(\lambda_0, s)$ обладает вещественными коэффициентами, а значит образ отрезка $[\delta_1, \delta_2]$ под действием P является отрезком в $\mathbb{R} \setminus \{0\}$. Так как $h(\lambda_0)$ не зависит от s , то множество

$$\{f(\lambda_0, s)^{-1}h(\lambda_0) \mid s \in [\delta_1, \delta_2]\} = \{P(s)^{-1}h(\lambda_0) \mid s \in [\delta_1, \delta_2]\}$$

является отрезком в \mathbb{C} , а значит — выпуклым замкнутым множеством. Аналогичным образом доказывается, что $\{t^{-1}g(\lambda_0, t) \mid t \in [\gamma_1, \gamma_2]\}$ также является выпуклым замкнутым множеством.

Следовательно, имеет место включение:

$$\{t^{-1}g(\lambda_0, t) \mid t \in [\gamma_1, \gamma_2]\} + \{f(\lambda_0, s)^{-1}h(\lambda_0) \mid s \in [\delta_1, \delta_2]\} \ni 0,$$

поэтому найдутся $t_0 \in [\gamma_1, \gamma_2]$, $s_0 \in [\delta_1, \delta_2]$ такие, что справедливо равенство

$$t_0^{-1}g(\lambda_0, t_0) + f(\lambda_0, s_0)^{-1}h(\lambda_0) = 0,$$

которое эквивалентно следующему —

$$f(\lambda_0, s_0)g(\lambda_0, t_0) + h(\lambda_0)t_0 = 0.$$

Следствие доказано. ■

Отметим, что полученная в следствии оценка совпадает с нижней оценкой, полученной в теореме 6.4.7, следовательно, является точной.

Теоремы 6.4.7, 6.4.8 без дополнительных условий обобщаются на случай, когда h в определении $\chi(\lambda)$ зависит не только от λ , но и от s ,

т.е. $h(\lambda) \rightarrow h(\lambda, s) \in \mathbb{C}[\lambda, s]$; следствие 6.4.1 также обобщается на этот случай, если потребовать выполнения условия $h(\lambda, s) \in \mathbb{R}[\lambda, s]$.

Минимизация спектрального радиуса оператора перехода является основой асимптотической оптимизации стационарных алгоритмов. Величина спектрального радиуса оператора перехода алгоритмов такого типа непосредственно связана с величиной

$$\rho(\chi) = \sup_{\substack{\lambda \in \sigma(\chi) \\ g(\lambda, 0) \neq 0}} |\lambda|. \quad (6.28)$$

Оценки $\rho(\chi)$ можно вывести как следствие оценок спектра, полученных в теореме 6.4.8, однако целесообразно получение более компактных и удобных оценок при дополнительных ограничениях на вид операторного пучка (6.18).

Пусть $P \in \mathbb{C}[z_1, \dots, z_n]$, $n \in \mathbb{N}$, тогда обозначим символом $\deg_z P$ — степень многочлена $P|_{z_i = z_i^0, i \neq j} \in \mathbb{C}[z_j]$ при фиксированных $z_i = z_i^0, i \neq j$.

Теорема 6.4.9. Пусть выполнены предположения (6.19), (6.20). Также предположим, что $\deg_s f \leq 1$, $\deg_t g \leq 1$ и величина

$$\deg_\lambda (f(\lambda, s_1)f(\lambda, s_2)g(\lambda, t) + f(\lambda, s_3)h(\lambda)t) > 0$$

не зависит от $s_1, s_2, s_3 \in [\delta_1, \delta_2]$, $t \in [\gamma_1, \gamma_2]$, тогда

$$\rho(\chi) \leq \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\},$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим одному из уравнений

$$\begin{aligned} f(\lambda, s) &= 0, \\ f(\lambda, s)g(\lambda, t) + h(\lambda)t &= 0, \\ f(\lambda, \delta_1)f(\lambda, \delta_2)g(\lambda, t) + f(\lambda, s)h(\lambda)t &= 0. \end{aligned} \quad (6.29)$$

Доказательство. Для начала отметим, что из условия $\deg_t g \leq 1$ следует, что для любого $\lambda \in \mathbb{C}$ имеет место равенство

$$\text{conv} \{t^{-1}g(\lambda, t) \mid t \in [\gamma_1, \gamma_2]\} = \{t^{-1}g(\lambda, t) \mid t \in [\gamma_1, \gamma_2]\}.$$

Пусть $\lambda \in \sigma(\chi)$, $g(\lambda, 0) \neq 0$, тогда по теореме 6.4.8 с учетом сделанного замечания найдутся $s \in [\delta_1, \delta_2]$ и $t \in [\gamma_1, \gamma_2]$ такие, что

$$f(\lambda, s) = 0 \quad \text{или} \quad 0 \in t^{-1}g(\lambda, t) + N(\lambda),$$

где

$$N(\lambda) = \text{conv} \{f(\lambda, s)^{-1}h(\lambda) \mid s \in [\delta_1, \delta_2], f(\lambda, s) \neq 0\}.$$

Обозначим символом M множество значений $\lambda \in \mathbb{C}$, удовлетворяющих этой системе при каких либо $s \in [\delta_1, \delta_2]$ и $t \in [\gamma_1, \gamma_2]$. Отметим, что

$$\rho(\chi) \leq \sup_{\lambda \in M} |\lambda|.$$

Зафиксируем $\lambda \in \mathbb{C}$ и предположим, что $f(\lambda, s) \neq 0$ при любых $s \in [\delta_1, \delta_2]$, тогда функция $w(z) = f(\lambda, z)^{-1}h(\lambda)$ является дробно-линейной функцией, полюсы которой не лежат на отрезке $[\delta_1, \delta_2]$. Следовательно, образ отрезка $[\delta_1, \delta_2]$ под действием w является отрезком или дугой окружности в \mathbb{C} с концами в точках

$$w(\delta_1) = f(\lambda, \delta_1)^{-1}h(\lambda) \quad \text{и} \quad w(\delta_2) = f(\lambda, \delta_2)^{-1}h(\lambda).$$

Таким образом, любая точка $z \in N(\lambda)$ может быть представлена в виде

$$z = (\alpha/f(\lambda, s_1) + (1 - \alpha)/f(\lambda, s_2))h(\lambda) = \frac{f(\lambda, s_3)h(\lambda)}{f(\lambda, s_1)f(\lambda, s_2)},$$

где $s_1, s_2 \in [\delta_1, \delta_2]$, $s_1 \leq s_2$, $\alpha \in [0, 1]$, $s_3 = (1 - \alpha)s_1 + \alpha s_2$.

Из полученного представления следует, что $\lambda \in M$ тогда и только тогда, когда существуют

$$s_1, s_2 \in [\delta_1, \delta_2], \quad s_1 \leq s_2, \quad s_3 \in [s_1, s_2], \quad t \in [\gamma_1, \gamma_2]$$

такие, что выполнено

$$f(\lambda, s_1)f(\lambda, s_2)g(\lambda, t) + f(\lambda, s_3)h(\lambda)t = 0.$$

Из непрерывности левой части этого уравнения следует, что M — замкнутое множество. По условию теоремы величина

$$\deg_\lambda(f(\lambda, s_1)f(\lambda, s_2)g(\lambda, t) + f(\lambda, s_3)h(\lambda)t) > 0$$

не зависит от s_1, s_2, s_3, t , следовательно, в силу непрерывной зависимости корней многочлена от коэффициентов, множество M ограничено, а это означает справедливость равенства

$$\sup_M |\lambda| = \max_{\partial M} |\lambda|,$$

где ∂M — множество граничных точек M .

Предположим, что $\lambda \in \partial M$ такое, что $f(\lambda, s) \neq 0$ при любых $s \in [\delta_1, \delta_2]$. Покажем, что λ удовлетворяет включению

$$0 \in t^{-1}g(\lambda, t) + \partial N(\lambda)$$

при некотором $t \in [\gamma_1, \gamma_2]$. Предположим противное, тогда существуют $t \in [\gamma_1, \gamma_2]$ и $z \in N(\lambda)$ такие, что справедливо

$$0 = t^{-1}g(\lambda, t) + z.$$

В силу непрерывности функции $f(\mu, s)^{-1}h(\mu)$ в некоторой окрестности множества $\{\lambda\} \times [\delta_1, \delta_2]$, существуют $\varepsilon_0, \varepsilon_1 > 0$ такие, что $V_{\varepsilon_1}(z) \subset]N(\mu)[$ для любых $\mu \in V_{\varepsilon_0}(\lambda)$ (здесь и далее $V_a(b) = \{z \in \mathbb{C} \mid |z - b| < a\}$). Выберем $\varepsilon_2: 0 < \varepsilon_2 \leq \varepsilon_0$ таким образом, что неравенство

$$|t^{-1}g(\mu, t) - t^{-1}g(\lambda, t)| < \varepsilon_1$$

имеет место для любых $\mu \in V_{\varepsilon_2}(\lambda)$. Следовательно, для любых $\mu \in V_{\varepsilon_2}(\lambda)$ выполнено

$$-t^{-1}g(\mu, t) \in V_{\varepsilon_1}(-t^{-1}g(\lambda, t)) = V_{\varepsilon_1}(z) \subset]N(\mu)[,$$

т. е. $\mu \in M$, а значит $V_{\varepsilon_2}(\lambda) \subset M$, что противоречит предположению $\lambda \in \partial M$.

Таким образом, из условия $\lambda \in \partial M$ следует, что $f(\lambda, s) = 0$ для некоторого $s \in [\delta_1, \delta_2]$ или $0 \in t^{-1}g(\lambda, t) + \partial N(\lambda)$ для некоторого $t \in [\gamma_1, \gamma_2]$. Из представления

$$\begin{aligned} \partial N(\lambda) &= \{f(\lambda, s)^{-1}h(\lambda) \mid s \in [\delta_1, \delta_2]\} \cup \\ &\cup \{f(\lambda, s)f(\lambda, \delta_1)^{-1}f(\lambda, \delta_2)^{-1}h(\lambda) \mid s \in [\delta_1, \delta_2]\} \end{aligned}$$

и неравенства

$$\rho(\chi) \leq \sup_{\lambda \in M} |\lambda| = \max_{\lambda \in \partial M} |\lambda|$$

следует утверждение теоремы. Теорема доказана. ■

В некоторых практически важных случаях оценка, полученная в теореме 6.4.9, может быть уточнена.

Следствие 6.4.2. Пусть выполнены предположения (6.19), (6.20) и

- 1) $\deg_s f \leq 1, \deg_t g \leq 1$;
- 2) f, g, h — многочлены с вещественными коэффициентами;
- 3) $\deg_\lambda f = 1, \deg_\lambda g = 1, \deg_\lambda h \leq 1$ при $s \in [\delta_1, \delta_2], t \in [\gamma_1, \gamma_2]$;
- 4) $f(0, s)h(0)$ не зависит от s .

Тогда имеет место неравенство

$$\rho(\chi) \leq \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\},$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим одному из уравнений

$$\begin{aligned} f(\lambda, s) &= 0, \\ f(\lambda, s)g(\lambda, t) + h(\lambda)t &= 0. \end{aligned} \tag{6.30}$$

Доказательство. Заметим, что выполнены все условия теоремы 6.4.9, поэтому

$$\rho(\chi) \leq q_0 = \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\}, \tag{6.31}$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим одному из уравнений (6.29). Пусть

$$q_1 = \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\},$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим одному из уравнений (6.30). Легко заметить, что $0 \leq q_1 \leq q_0$. Докажем, что $q_0 = q_1$ в принятых предположениях. В случаях $q_0 = 0$ или $\delta_1 = \delta_2$ утверждение теоремы выполнено.

Пусть $q_0 > 0$ и $\delta_1 < \delta_2$. Зафиксируем $t_0 \in [\gamma_1, \gamma_2]$, $s_0 \in [\delta_1, \delta_2]$, при которых достигается максимум в (6.31), и запишем третье уравнение (6.29) в форме $R(q_0^{-1}\lambda, \alpha_0) = 0$, где

$$\alpha_0 = \frac{s_0 - \delta_1}{\delta_2 - \delta_1} \in [0, 1],$$

$R(\mu, \alpha) = f(q_0\mu, \delta_1)f(q_0\mu, \delta_2)g(q_0\mu, t_0) + f(q_0\mu, (1 - \alpha)\delta_1 + \alpha\delta_2)h(q_0\mu)t_0$. При любых $\alpha \in [0, 1]$ степень $\deg_\mu R$ постоянна и равна 3, а все корни уравнения $R(\mu, \alpha_0) = 0$ лежат в замкнутом единичном круге \overline{D} , где $D = \{\mu \in \mathbb{C} \mid |\mu| < 1\}$, причем существуют корни, лежащие на границе круга.

Допустим, что $q_1 < q_0$. При $\alpha \in \{0, 1\}$ уравнение $R(\mu, \alpha) = 0$ распадается на два уравнения

$$\begin{aligned} f(q_0\mu, (1 - \alpha)\delta_1 + \alpha\delta_2) &= 0, \\ f(q_0\mu, \alpha\delta_1 + (1 - \alpha)\delta_2)g(q_0\mu, t_0) + h(q_0\mu)t_0 &= 0, \end{aligned}$$

которые соответствуют уравнениям (6.30). Отсюда следует, что корни уравнения $R(\mu, \alpha) = 0$ при $\alpha \in \{0, 1\}$ не превосходят по модулю величины $q_1/q_0 < 1$, поэтому лежат в D .

Из условий 1), 2) следует, что $R(\mu, \alpha)$ является многочленом с вещественными коэффициентами, причем

$$R(\mu, \alpha) = (1 - \alpha)R(\mu, 0) + \alpha R(\mu, 1).$$

Если рассматривать $R(\mu, \alpha)$ как многочлен от μ с коэффициентами, зависящими от α , то из условия 4) и из того, что старший коэффициент R (при μ^3) совпадает со старшим коэффициентом многочлена $f(q_0\mu, \delta_1)f(q_0\mu, \delta_2)g(q_0\mu, t_0)$, следует, что старший коэффициент и свободный член R не зависят от α . Следовательно, имеет место следующее представление $R(\mu, \alpha)$:

$$R(\mu, \alpha) = a\mu^3 + b_\alpha\mu^2 + c_\alpha\mu + d.$$

где $a, d \in \mathbb{R}$, $a \neq 0$, $b_\alpha = (1 - \alpha)b_0 + \alpha b_1$, $c_\alpha = (1 - \alpha)c_0 + \alpha c_1$, $b_0, b_1, c_0, c_1 \in \mathbb{R}$.

Для оценки корней уравнения $R(\mu, \alpha) = 0$ при $\alpha \in [0, 1]$ воспользуемся теоремой Шура-Кона. Корни $R(\mu, 0)$ лежат в D , поэтому $|a| > |d|$ и взаимный многочлен имеет вид

$$R^*(\mu, \alpha) = (a^2 - d^2)\mu^2 + (ab_\alpha - dc_\alpha)\mu + (ac_\alpha - db_\alpha).$$

Так как корни многочленов $R^*(\mu, 0)$ и $R^*(\mu, 1)$ лежат в D , то неравенство

$$|ac_\alpha - db_\alpha| \leq \max\{|ac_0 - db_0|, |ac_1 - db_1|\} < |a^2 - d^2|$$

справедливо для любых $\alpha \in [0, 1]$ и, соответственно, можно записать

$$R^{**}(\mu, \alpha) = ((a^2 - d^2) - (ac_\alpha - db_\alpha))((a^2 - d^2) + (ac_\alpha - db_\alpha))\mu + (ab_\alpha - dc_\alpha).$$

Единственный корень многочлена $R^{**}(\mu, \alpha)$ представим формулой

$$\mu_\alpha = \frac{ab_\alpha - dc_\alpha}{(a^2 - d^2) + (ac_\alpha - db_\alpha)}$$

и является вещественной дробно-рациональной функцией аргумента α , знаменатель которой не обращается в 0 при $\alpha \in [0, 1]$. Таким образом, μ_α монотонно зависит от α на отрезке $[0, 1]$, причем корни многочленов $R^{**}(\mu, 0)$ и $R^{**}(\mu, 1)$ лежат в D , а значит справедлива оценка

$$|\mu_\alpha| \leq \max\{|\mu_0|, |\mu_1|\} < 1.$$

По теореме Шура-Кона все корни уравнения

$$R(\mu, \alpha_0) = 0$$

лежат в D , что противоречит наличию корня на границе D , таким образом, исходное допущение не верно и значит $q_0 = q_1$. Следствие доказано. ■

Отметим, что полученная в следствии оценка, в силу в теоремы 6.4.7, не может быть ослаблена, следовательно, является неулучшаемой.

6.4.3. Свойства некоторых классов функций

Исследование минимаксных задач, возникающих в процессе анализа сходимости алгоритмов для решения седловых задач, является сложной с технической точки зрения проблемой. Использование здесь общей теории минимаксных задач [37] затруднено наличием точек недифференцируемости целевых функций. Однако, большинство минимаксных постановок, рассматриваемых в этой части книги, обладает определенными свойствами, позволяющими применить

специальную технику исследования. Эти свойства характеризуются классами функций, которые вводятся далее.

Всюду в этом разделе будем считать, что I — непустой интервал в \mathbb{R} .

Рассмотрим следующие классы непрерывных вещественнозначных функций, определенных на I :

- класс $Y(I)$ состоит из функций f , непрерывных на I и обладающих тем свойством, что f монотонна на I , либо существует (зависящая от f) точка $x_f \in I$ такая, что f невозрастает на промежутке $(-\infty, x_f] \cap I$ и неубывает на промежутке $[x_f, -\infty) \cap I$;
- класс $YS(I)$ состоит из функций f , непрерывных на I и обладающих тем свойством, что f строго монотонна на I , либо существует (зависящая от f) точка $x_f \in I$ такая, что f строго убывает на промежутке $(-\infty, x_f] \cap I$ и строго возрастает на промежутке $[x_f, -\infty) \cap I$.

Отметим, что все функции из $Y(I)$, $YS(I)$ относятся к классу так называемых *унимодальных* функций [6, с. 29].

Введем операцию \uparrow над функциями

$$f_1: I \rightarrow \mathbb{R} \quad \text{и} \quad f_2: I \rightarrow \mathbb{R}$$

с помощью равенства

$$(f_1 \uparrow f_2)(x) = \max\{f_1(x), f_2(x)\}.$$

Теорема 6.4.10. *Справедливы следующие свойства классов $Y(I)$ и $YS(I)$:*

- 1) $YS(I) \subset Y(I)$.
- 2) Пусть $f \in Y(I)$ и $[a, b] \subset I$, $a \leq b$, тогда

$$\max_{x \in [a, b]} f(x) = \max\{f(a), f(b)\}.$$

- 3) Пусть $I_1, I_2 \subseteq \mathbb{R}$ — непустые интервалы, $f \in Y(I_1)$ и $\varphi: I_2 \rightarrow I_1$ — сюръективный гомеоморфизм, тогда

$$f \circ \varphi \in Y(I_2),$$

где функция $f \circ \varphi$ определена равенством

$$(f \circ \varphi)(x) = f(\varphi(x))$$

для всех $x \in I_2$; если кроме этого $f \in YS(I_1)$, то

$$f \circ \varphi \in YS(I_2).$$

- 4) Пусть $f \in \text{YS}(I)$, тогда f имеет в I не более одной точки локального минимума, если же такая точка существует, то является также точкой глобального минимума.
- 5) Пусть $f_1, f_2 \in \text{YS}(I)$, тогда $f_1 \uparrow f_2 \in \text{YS}(I)$.
- 6) Пусть $f_1, f_2 \in \text{Y}(I)$, тогда $f_1 \uparrow f_2 \in \text{Y}(I)$.

Доказательство. Пункт 1 следует непосредственно из определения классов.

2) Пусть $f \in \text{Y}(I)$ и $[a, b] \subset I$, $a \leq b$. Если существует точка $x \in [a, b]$ такая, что f невозрастает на $(-\infty, x] \cap I$ и неубывает на $[x, +\infty) \cap I$, то $f(a) \geq f(z)$ для любой $z \in [a, x]$ и $f(z) \leq f(b)$ для любой $z \in [x, b]$, следовательно, $f(z) \leq \max\{f(a), f(b)\}$ для $z \in [a, b]$. В противном случае f монотонна на отрезке $[a, b]$ и, следовательно, $f(z) \leq \max\{f(a), f(b)\}$ для $z \in [a, b]$.

3) Пусть $I_1, I_2 \subseteq \mathbb{R}$ — непустые интервалы, $f \in \text{YS}(I_1)$ и $\varphi: I_2 \rightarrow I_1$ — сюръективный гомеоморфизм, т. е. φ — непрерывное строго монотонное отображение интервала I_2 на интервал I_1 . Если существует точка $x_1 \in I_1$ такая, что f строго убывает на $(-\infty, x_1] \cap I_1$ и строго возрастает на $[x_1, +\infty) \cap I_1$, то определим

$$x_2 = \varphi^{-1}(x_1) \in I_2$$

(здесь φ^{-1} — обратное отображение). В этом случае $f \circ \varphi$ определена и непрерывна на I_2 , строго убывает на $(-\infty, x_2] \cap I_2$ и строго возрастает на $[x_2, +\infty) \cap I_2$, следовательно,

$$f \circ \varphi \in \text{YS}(I_2).$$

Если же f — строго монотонна на I_1 , то $f \circ \varphi$ также строго монотонна на I_2 , следовательно,

$$f \circ \varphi \in \text{YS}(I_2).$$

Для $f \in \text{Y}(I_1)$ доказательство аналогично.

4) Пусть $f \in \text{YS}(I)$, тогда либо f строго монотонна на I и, следовательно, не имеет точек локального минимума в I , либо существует точка $x \in I$ такая, что f строго убывает на $(-\infty, x] \cap I$ и строго возрастает на $[x, +\infty) \cap I$. В последнем случае точка x является точкой глобального минимума f , а на интервалах $(-\infty, x) \cap I$ и $(x, +\infty) \cap I$ строго монотонна, т. е. x — единственная точка локального минимума f на I .

5) Пусть $f_1, f_2 \in \text{YS}(I)$ и точки $x_1, x_2 \in \mathbb{R} \cup \{\pm\infty\}$ таковы, что f_i строго убывает на $(-\infty, x_i) \cap I$ и строго возрастает на $(x_i, +\infty) \cap I$ для $i = 1, 2$ (формально полагается $(-\infty, -\infty) = \emptyset$, $(+\infty, +\infty) = \emptyset$). Не ограничивая общности можно считать, что $x_1, x_2 \in I$ или совпадают с одной из концевых точек интервала I , а также, что $x_1 \leq x_2$.

Отметим, что в силу непрерывности f_1 и f_2 функция $f = f_1 \uparrow f_2$ также непрерывна на I .

Рассмотрим поведение функции f на интервале $(-\infty, x_1) \cap I$. Если $a, b \in (-\infty, x_1) \cap I$, $a < b$, тогда $f_1(a) > f_1(b)$, $f_2(a) > f_2(b)$, следовательно, справедливо

$$f(a) = \max\{f_1(a), f_2(a)\} > \max\{f_1(b), f_2(b)\} = f(b),$$

т. е. f строго убывает на указанном интервале. Аналогично доказывается, что f строго возрастает на интервале $(x_2, +\infty) \cap I$.

Рассмотрим теперь интервал $(x_1, x_2) \subseteq I$. Предположим существует точка $y \in (x_1, x_2)$ такая, что $f_1(y) = f_2(y)$. Так как f_1 строго возрастает и f_2 строго убывает на (x_1, x_2) , то $f(x) = f_2(x)$, если $x \in (x_1, y)$, и $f(x) = f_1(x)$, если $x \in (y, x_2)$. Если такой точки y не существует, то либо $x_1 = x_2$, в этом случае обозначим $y = x_1$, либо, в силу непрерывности f_1 и f_2 , функция f совпадает с f_1 на (x_1, x_2) , в этом случае обозначим $y = x_1$, или f совпадает с f_2 на (x_1, x_2) , и тогда обозначим $y = x_2$. Во всех случаях f строго убывает на $(-\infty, y) \cap (x_1, x_2)$ и строго возрастает на $(y, +\infty) \cap (x_1, x_2)$.

Таким образом, f строго убывает на $(-\infty, y) \cap I$ и строго возрастает на $(y, +\infty) \cap I$, следовательно, $f \in \text{YS}(I)$.

Пункт 6 доказывается аналогично пункту 5 с заменой строгих неравенств и монотонности на нестрогие. Теорема доказана. ■

Пусть M — некоторое множество вещественнозначных функций с общей областью определения, тогда через M^\uparrow будем обозначать множество, полученное из M замыканием относительно операции \uparrow , другими словами, совокупность всех функций вида

$$f_1 \uparrow \dots \uparrow f_k, \quad f_i \in M, \quad k \in \mathbb{N}.$$

Отметим, что из пунктов 5 и 6 теоремы следует, что

$$Y(I)^\uparrow = Y(I) \quad \text{и} \quad \text{YS}(I)^\uparrow = \text{YS}(I).$$

Для произвольных $a, b \in \mathbb{R}$, $|a| + |b| \neq 0$ определим интервалы $I_{a,b}$:

$$I_{a,b} = \begin{cases} \mathbb{R} & \text{при } a = 0. \\ (-b/a, +\infty) & \text{при } a \neq 0. \end{cases}$$

Введем следующее множество функций, определенных на интервале $I_{a,b}$:

$$\begin{aligned} F_{a,b} &= \left\{ h(x; a, b, c, d, e, f) \mid c, d, e, f \in \mathbb{R} \right\}, \\ h(x) &\equiv h(x; a, b, c, d, e, f) = \\ &= \max_{\pm} \left\{ \left| \frac{cx+d}{ax+b} \pm \sqrt{\left(\frac{cx+d}{ax+b} \right)^2 + \left(\frac{ex+f}{ax+b} \right)} \right| \right\}. \end{aligned}$$

Справедливы следующие свойства множества $F_{a,b}$:

Теорема 6.4.11. Пусть $a, b \in \mathbb{R}$, $|a| + |b| \neq 0$, тогда:

- 1) $F_{a,b} \subset Y(I_{a,b})$.
- 2) Пусть $f \in F_{a,b}^{\uparrow}$ и f обладает точкой строгого локального минимума в $I_{a,b}$, тогда $f \in YS(I_{a,b})$.

Прежде чем перейти к доказательству этой теоремы исследуем важный частный случай. Справедливо

Утверждение 6.4.1. Имеют место свойства:

- 1) $F_{0,1} \subset Y(\mathbb{R})$.
- 2) Пусть $f \in F_{0,1}^{\uparrow}$ и f обладает точкой строгого локального минимума в \mathbb{R} , тогда $f \in YS(\mathbb{R})$.

Доказательство. Пусть $h(x) \in F_{0,1}$, тогда существуют $a, b, c, d \in \mathbb{R}$ такие, что

$$h(x) = \max_{\pm} \left\{ \left| a + bx \pm \sqrt{b^2x^2 + 2cx + d} \right| \right\}.$$

Рассмотрим возможные значения a, b, c, d и одновременно вычислим множество M локальных минимумов функции h :

- 1) Пусть $c^2 - b^2d < 0$, тогда

$$h(x) = |a + bx| + \sqrt{b^2x^2 + 2cx + d}$$

для всех $x \in \mathbb{R}$. Производная

$$h'(x) = b \operatorname{sign}(a + bx) + \frac{b^2x + c}{\sqrt{b^2x^2 + 2cx + d}}$$

существует, отлична от нуля и непрерывна в точках $x \neq -a/b$. Несложно вычислить следующие пределы:

$$\lim_{x \rightarrow -\infty} h'(x) = -2|b|, \quad \lim_{x \rightarrow +\infty} h'(x) = 2|b|,$$

Значит $h'(x) < 0$ на $(-\infty, -a/b)$ и $h'(x) > 0$ на $(-a/b, +\infty)$, т. е. $h(x)$ строго убывает на $(-\infty, -a/b]$ и строго возрастает на $[-a/b, +\infty)$. а $M = \{-a/b\}$.

2) Пусть $c^2 - b^2d > 0$ и $b \neq 0$, тогда функция

$$h(x) = \begin{cases} |a + bx| + \sqrt{b^2x^2 + 2cx + d} & \text{при } b^2x^2 + 2cx + d \geq 0, \\ \sqrt{2(ab - c)x + (a^2 - d)} & \text{при } b^2x^2 + 2cx + d < 0, \end{cases}$$

обладает непрерывной производной

$$h'(x) = \begin{cases} b \operatorname{sign}(a + bx) + \frac{b^2x + c}{\sqrt{b^2x^2 + 2cx + d}} & \text{при } b^2x^2 + 2cx + d > 0, \\ \frac{ab - c}{\sqrt{2(ab - c)x + (a^2 - d)}} & \text{при } b^2x^2 + 2cx + d < 0, \end{cases}$$

в точках $x \in \mathbb{R}$ таких, что

$$b^2x^2 + 2cx + d \neq 0, \quad ax + b \neq 0.$$

Пусть x_1, x_2 — корни квадратного уравнения $b^2x^2 + 2cx + d = 0$, причем $x_1 < x_2$. Несложно вычислить следующие пределы:

$$\begin{aligned} \lim_{x \rightarrow -\infty} h'(x) &= -2|b|, & \lim_{x \rightarrow +\infty} h'(x) &= 2|b|, \\ \lim_{x \rightarrow x_1 - 0} h'(x) &= -\infty, & \lim_{x \rightarrow x_2 + 0} h'(x) &= +\infty. \end{aligned}$$

Следовательно, $h(x)$ строго убывает на $(-\infty, x_1]$ и строго возрастает на $[x_2, +\infty)$. На интервале $x \in (x_1, x_2)$, т. е. когда x удовлетворяет неравенству

$$b^2x^2 + 2cx + d < 0,$$

функция $h(x)$ дифференцируема, причем при $ab - c \neq 0$ выполнено неравенство $h'(x) \neq 0$, а при $ab - c = 0$ на всем интервале справедливо $h'(x) = 0$. Таким образом, $h(x)$ строго возрастает на (x_1, x_2) при $ab - c > 0$ и тогда $M = \{x_1\}$, $h(x)$ строго убывает на (x_1, x_2) при $ab - c < 0$ и тогда $M = \{x_2\}$, или же $h(x) = \operatorname{const}$ на (x_1, x_2) при $ab - c = 0$ и значит $M = [x_1, x_2]$.

3) Пусть $c^2 - b^2d = 0$ и $b \neq 0$, тогда справедливо представление

$$h(x) = |a + bx| + |c/b + bx|,$$

и несложно убедиться, что

$$M = [\min\{-a/b, -c/b^2\}, \max\{-a/b, -c/b^2\}].$$

4) Пусть $b = 0$, $c \neq 0$. В этом случае

$$h(x) = \begin{cases} |a| + \sqrt{2cx + d}, & \text{если } 2cx + d \geq 0, \\ \sqrt{a^2 - (2cx + d)}, & \text{если } 2cx + d < 0, \end{cases}$$

откуда несложно заключить, что $M = \{-d/(2c)\}$.

5) Если же $b = 0$ и $c = 0$, тогда $h \equiv \text{const}$ и $M = (-\infty, +\infty)$.

Отсюда следует, что либо $h(x) \equiv \text{const}$, либо $h(x)$ строго убывает на $(-\infty, x_1]$, постоянна на $M = [x_1, x_2]$ и строго возрастает на $[x_2, +\infty)$, где $x_1, x_2 \in \mathbb{R}$, $x_1 \leq x_2$, а также непрерывна на \mathbb{R} , следовательно, $h(x) \in Y(\mathbb{R})$. Пункт (1) доказан.

Докажем пункт 2. Пусть $g_1 : \mathbb{R} \rightarrow \mathbb{R}$ и $g_2 : \mathbb{R} \rightarrow \mathbb{R}$ такие непрерывные функции, множество локальных минимумов которых совпадает с множеством глобальных минимумов и образует отрезок либо все \mathbb{R} . Покажем, что и функция $g = g_1 \uparrow g_2$ обладает тем же свойством. Действительно, так как $g_1, g_2, g \in Y(\mathbb{R})$ и все точки глобального минимума g образуют отрезок, то в противном случае существовали бы точки локального минимума x_1, x_2 функции g такие, что $g(x_1) > g(x_2)$. Но тогда $g(x)$ есть константа в некоторой окрестности точки x_1 , так как она монотонна в ее окрестности. Отсюда следует, что либо $g_1(x)$, либо $g_2(x)$ тоже константа в некоторой окрестности x_1 , равная $g(x_1)$. Не ограничивая общности считаем, что это $g_1(x)$ и значит x_1 — точка ее локального минимума, а значит и глобального. Из неравенства

$$g(x_2) = \max\{g_1(x_2), g_2(x_2)\} \geq \max\{g_1(x_1), g_2(x_2)\} \geq g_1(x_1) = g(x_1)$$

получаем противоречие, т. е. наше предположение неверно.

Пусть функция $f \in F_{0,1}^\uparrow$ и f обладает точкой строгого локального минимума. По определению существует $k \in \mathbb{N}$ и $h_1, \dots, h_k \in F$ такие, что

$$f = h_1 \uparrow \dots \uparrow h_k,$$

причем множество локальных минимумов каждой h_i , $i = 1, \dots, k$ совпадает с множеством глобальных минимумов и образует отрезок. Следовательно, множество локальных минимумов f совпадает с множеством глобальных минимумов и образует отрезок либо все \mathbb{R} . Но f обладает точкой строгого локального минимума, следовательно, множество локальных минимумов f совпадает с этой точкой и $f \in YS(\mathbb{R})$. Утверждение доказано. ■

Перейдем к доказательству теоремы 6.4.11.

Доказательство. Из определения классов $F_{a,b}$ следует, что

$$I_{a,b} = I_{a/c, b/c}, \quad F_{a,b} = F_{a/c, b/c}$$

для любого $c \in \mathbb{R} \setminus \{0\}$. Таким образом, случай $a = 0, b \neq 0$ сводится к случаю $a = 0, b = 1$, который рассмотрен в утверждении 6.4.1.

Пусть $a \neq 0$, тогда замена переменных $y = ax + b$ позволяет свести рассмотрение класса $F_{a,b}$ к рассмотрению класса $F_{1,0}$, поэтому без ограничения общности считаем, что $a = 1, b = 0$.

1) Несложно заметить, что любая функция $f \in F_{1,0}$ представима в виде $h(x^{-1})$ при $x \in I_{1,0} = (0, +\infty)$, где

$$h(y) = \max_{\pm} \left\{ \left| c + dy \pm \sqrt{(c + dy)^2 + (e + fy)} \right| \right\} \in F_{0,1}.$$

Следовательно, $h \in Y(\mathbb{R})$, а значит $h \in Y(I_{1,0})$ и, в силу пункта 3 теоремы 6.4.10, справедливо $f = h \circ (x^{-1}) \in Y(I_{1,0})$.

2) Пусть $f = f_1 \uparrow \dots \uparrow f_k$, $f_i \in F_{1,0}$, $k \in \mathbb{N}$ и f обладает точкой строгого локального минимума в $I_{1,0}$. Представим f в виде

$$f(x) = h_1(x^{-1}) \uparrow \dots \uparrow h_k(x^{-1}),$$

где $h_i \in F_{0,1}$, и рассмотрим функцию

$$h(y) = h_1(y) \uparrow \dots \uparrow h_k(y) \in F_{0,1}^{\uparrow}.$$

Так как при $y > 0$ справедливо $h(y) = f(y^{-1})$, то h обладает точкой строгого локального минимума в $(0, +\infty)$. В силу утверждения 6.4.1, справедливо $h \in YS(\mathbb{R})$ и, следовательно, $h \in YS(I_{1,0})$. В силу пункта 3 теоремы 6.4.10, имеет место представление

$$f = h \circ (x^{-1}) \in YS(I_{1,0}).$$

Теорема доказана. ■

6.5. КЛЮЧЕВЫЕ ЭТАПЫ ПОСТРОЕНИЯ И АНАЛИЗА АЛГОРИТМОВ

Чтобы легче ориентироваться в многообразии конструкций обобщенных методов для решения задач (6.2) вида $Sz = F$, приведем общую схему их построения:

1) Выбирается «эффективный» предобусловливатель Q для матрицы A : в симметричном случае

$$Q = Q^*, \quad 0 < \delta Q \leq A \leq \Delta Q, \quad 0 < \delta \leq \Delta$$

или в несимметричном --

$$Q = (A + A^*)/2, \quad -RQ \leq (A - A^*)/2 \leq RQ, \quad 0 < R.$$

2) Выбирается «эффективный» предобусловливатель C для матриц $B^*A^{-1}B$ или $B^*Q^{-1}B$:

$$C = C^*, \quad 0 < \gamma C \leq B^*A^{-1}B \leq \Gamma C, \quad 0 < \gamma \leq \Gamma$$

или

$$C = C^*, \quad 0 < \gamma C \leq B^*Q^{-1}B \leq \Gamma C, \quad 0 < \gamma \leq \Gamma.$$

3) При помощи алгебраических операций над блоками A, A^*, Q, B, B^*, C строится предобусловливатель R для блочной матрицы S из (6.2):

$$R \equiv R(A, A^*, Q, B, B^*, C; \bar{\alpha}) = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}.$$

где R_{ij} — рациональные формы от A, A^*, Q, B, B^*, C и итерационных параметров $\bar{\alpha} \in \Omega \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$.

4) Решается система уравнений $R^{-1}Sz = R^{-1}F$, где F — правая часть (6.2):

- а) в общем случае, для решения $R^{-1}Sz = R^{-1}F$ используется один из традиционных итерационных методов (стационарные, с чебышевским ускорением, проекционные (метод сопряженных градиентов, метод градиентов, метод Ланцоша, обобщенный метод минимальных невязок (GMRES) и т. д.));
- б) если предобусловленный оператор $R^{-1}S$ имеет блочно треугольный вид, то система $R^{-1}Sz = R^{-1}F$ расщепляется, т. е. сводится к последовательному решению двух систем относительно одной неизвестной компоненты u или p , каждая из которых решается одним из традиционных итерационных методов.

После формулировки конкретного алгоритма его анализ проводится по специальной, удобной для восприятия схеме. Обозначим далее ее основные этапы:

1) Формулируется задача оптимизации, которая затем сводится к изучению характеристик спектра предобусловленного оператора $R^{-1}S$.

2) Исследование спектра $R^{-1}S$ сводится к анализу спектра линейного или квадратичного операторного пучка $\chi(\lambda)$:

- а) Если, например, оператор $R_{22} : P \rightarrow P$ обратим, то вывод пучка для предобусловленного оператора может быть осуществлен следующим образом: запишем задачу на собственные значения $R^{-1}Sz = \lambda z$ в развернутой форме

$$\begin{aligned} Au + Bp &= \lambda(R_{11}u + R_{12}p), \\ B^*u &= \lambda(R_{21}u + R_{22}p), \end{aligned}$$

тогда из второго уравнения имеем

$$p = R_{22}^{-1}(\lambda^{-1}B^* - R_{21})u.$$

что после подстановки в первое уравнение дает пучок:

$$\chi(\lambda)u \equiv \left[(R_{12}R_{22}^{-1}R_{21} - R_{11})\lambda^2 + \right. \\ \left. + (A - R_{12}R_{22}^{-1}B^* - BR_{22}^{-1}R_{21})\lambda + BR_{22}^{-1}B^* \right]u = 0.$$

- б) Приведем более специфический пример: большое число известных способов блочного преобусловливания симметричных седловых задач приводит к оператору

$$R = \begin{pmatrix} I & -\omega_1(D)BC^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \omega_2(D)Q & \omega_3(D)B \\ \alpha_1 B^* & \alpha_2 C \end{pmatrix},$$

где $D = AQ^{-1}$, $\alpha_{1,2} \in \mathbb{C}$, $\alpha_2 \neq 0$, $\omega_{1,2,3} \in \mathbb{C}[s]$, $\det R \neq 0$.

В этом случае задача на собственные значения может быть переписана в форме

$$\begin{pmatrix} A + \omega_1(D)BC^{-1}B^* & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} \omega_2(D)Q & \omega_3(D)B \\ \alpha_1 B^* & \alpha_2 C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix},$$

что, в свою очередь, приводит к пучку

$$\left[\alpha_2 \lambda (D - \lambda \omega_2(D))Q + ((1 - \alpha_1 \lambda)(I - \lambda \omega_3(D)) + \alpha_2 \lambda \omega_1(D))BC^{-1}B^* \right]u = 0.$$

Пусть $L = Q^{-1/2}AQ^{-1/2}$, $G = Q^{-1/2}BC^{-1}B^*Q^{-1/2}$, $v = Q^{1/2}u$, тогда имеем

$$\left[\alpha_2 \lambda (L - \lambda \omega_2(L)) + ((1 - \alpha_1 \lambda)(I - \lambda \omega_3(L)) + \alpha_2 \lambda \omega_1(L))G \right]v = 0,$$

или в удобной для анализа форме —

$$\left[f(\lambda, L)g(\lambda, t) + h(\lambda, L)G \right]v = 0,$$

где использованы обозначения:

$$f(\lambda, s) = \alpha_2(s - \lambda \omega_2(s)) \in \mathbb{C}[\lambda, s],$$

$$g(\lambda, t) = \lambda \in \mathbb{C}[\lambda, t],$$

$$h(\lambda, s) = (1 - \alpha_1 \lambda)(1 - \lambda \omega_3(s)) + \alpha_2 \lambda \omega_1(s) \in \mathbb{C}[\lambda, s].$$

Таким образом, спектральный анализ такого алгоритма можно свести к изучению спектра операторного пучка вида

$$\chi(\lambda) = f(\lambda, s)g(\lambda, t) + h(\lambda, s)t, \quad (6.32) \\ \deg_\lambda f \leq 1, \quad \deg_\lambda g \leq 1, \quad \deg_\lambda h \leq 2;$$

- в) Если для исследования алгоритма большее значение имеет не спектр оператора $R^{-1}S$, а например, спектр оператора

$$T = I - \tau R^{-1}S,$$

то задача на собственные значения $Tz = \lambda z$ может быть сведена к изучению пучка $\chi((1 - \lambda)/\tau)$, который также является квадратичным, а в случае предыдущего примера допускает представление (6.32);

3) Выводятся аналитические (в виде явных формул) оценки спектра $\chi(\lambda)$ в исследуемом классе задач \mathbb{K} :

$$\sigma(\chi; \bar{\alpha}) \subseteq \Lambda(\mathbb{K}; \bar{\alpha}) \subset \mathbb{C},$$

где $\Lambda(\mathbb{K}; \bar{\alpha})$ — множество, описываемое некоторыми аналитическими соотношениями. Здесь основополагающую роль играют результаты раздела 6.4.2:

- а) оценки вещественного спектра выводятся с применением следствия 6.4.1;
- б) оценки спектрального радиуса получаются с использованием теоремы 6.4.9 и следствия 6.4.2;
- 4) Оптимизируются оценки спектра относительно итерационных параметров. В большинстве рассматриваемых задач оптимизации процесс сводится к исследованию вариационной проблемы:

$$\min_{P \in \mathbb{R}[\lambda], \deg P \leq n, P(0)=1} \max_{\lambda \in \Lambda(\mathbb{K}; \bar{\alpha})} |P(\lambda)| \rightarrow \min, \quad \bar{\alpha} \in \Omega \quad (6.33)$$

решение которой можно точно найти в одном из следующих случаев:

- а) В случае *вещественного спектра*, в основном, используется одна из форм оценок

$$\begin{aligned} \Lambda(\mathbb{K}; \bar{\alpha}) &= [a, b] \text{ или } \Lambda(\mathbb{K}; \bar{\alpha}) = [-b, -a] \cup [a, b], \quad 0 < a \leq b, \\ a \equiv a(\bar{\alpha}) &= \min_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} |\lambda_{1,2,3}(t, s; \bar{\alpha})| > 0, \\ b \equiv b(\bar{\alpha}) &= \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} |\lambda_{1,2,3}(t, s; \bar{\alpha})|, \end{aligned}$$

где λ_i , $i = 1, 2, 3$ — корни уравнений

$$f(\lambda, s) = 0, \quad f(\lambda, s)g(\lambda, t) + h(\lambda, s)t = 0.$$

При этом (6.33) с помощью классической теории многочленов Чебышева [12, с. 69, с. 320] сводится к решению задачи

$$\bar{\alpha}_0 = \arg \min_{\bar{\alpha} \in \Omega} \frac{b(\bar{\alpha})}{a(\bar{\alpha})}.$$

Ее особенность заключается в простоте аналитического устройства функции $b(\bar{\alpha})/a(\bar{\alpha})$ и возможности применения стандартных методов минимизации;

- б) В случае оценок спектрального радиуса (здесь допустим комплексный спектр) имеем:

$$\Lambda(\mathbb{K}; \bar{\alpha}) = \{z \in \mathbb{C} \mid |1 - z| \leq \rho(\bar{\alpha})\},$$

$$\rho(\bar{\alpha}) = \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} |\lambda_{1,2,3}(t, s; \bar{\alpha})|,$$

где λ_i , $i = 1, 2, 3$ — корни уравнений

$$f(\lambda, s) = 0, \quad f(\lambda, s)g(\lambda, t) + h(\lambda, s)t = 0.$$

При этом (6.33) с помощью леммы Зарантонелло [183, с. 189] сводится к решению задачи

$$\bar{\alpha}_0 = \arg \min_{\bar{\alpha} \in \Omega} \rho(\bar{\alpha}). \quad (6.34)$$

Отметим, что в данном случае функция $\rho(\bar{\alpha})$ хотя и выражается явными аналитическими формулами, но, как показывает практика, не поддается известным методам исследования ввиду отсутствия всюду-дифференцируемости в области определения. Тем не менее, для решения (6.34) в этой книге успешно применяются результаты раздела 6.4.3;

5) Обосновываются, если требуется, точность и единственность оценок. Для исследования точности оценок, единственности оптимальных параметров, а также практической и теоретической проверки результатов, используется один из следующих наборов матриц:

- а) В классе $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$:

$$A = I, \quad B = \begin{pmatrix} \sqrt{t_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{t_{N_P}} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} s_1^{-1} & & 0 \\ & \ddots & \\ 0 & & s_{N_U}^{-1} \end{pmatrix},$$

$$C = I,$$

$$s_i \in [\delta, \Delta], \quad i = 1, \dots, N_U, \quad t_i \in [\gamma, \Gamma], \quad i = 1, \dots, N_P;$$

б) В классе $\mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$:

$$A = \begin{pmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_{N_U} \end{pmatrix}, \quad B = \begin{pmatrix} \sqrt{t_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{t_{N_P}} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix},$$

$$Q = I, \quad C = I,$$

$$s_i \in [\delta, \Delta], \quad i = 1, \dots, N_U, \quad t_i \in [\gamma, \Gamma], \quad i = 1, \dots, N_P;$$

в) В классе $\mathbb{K}_3(R, \gamma, \Gamma)$:

$$A = \begin{pmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_{N_U} \end{pmatrix}, \quad B = \begin{pmatrix} \sqrt{t_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{t_{N_P}} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix},$$

$$Q = I, \quad C = I, \quad s_i = 1 + r_i i,$$

$$r_i \in [-R, R], \quad i = 1, \dots, N_U, \quad t_i \in [\gamma, \Gamma], \quad i = 1, \dots, N_P;$$

Во всех этих случаях $\chi(\lambda)$ имеет диагональный вид и собственные значения λ могут быть найдены точно из уравнений:

$$f(\lambda_i, s_i) = 0, \quad i = 1, \dots, N_U - N_P, \quad \lambda_i \equiv \lambda_1(s_i),$$

$$f(\lambda_i, s_i)g(\lambda_i, t_i) + h(\lambda_i, s_i)t_i = 0, \quad i = 1, \dots, N_P, \quad \lambda_i \equiv \lambda_{2,3}(s_i, t_i).$$

6.6. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

С формальной точки зрения, обобщенные методы решения седловых задач отличаются от релаксационных методов, рассмотренных в первой части книги, не очень заметно: появляется только дополнительный предобусловливатель Q для оператора A , расположенного в левом верхнем углу матричной записи исходной задачи. Однако это нововведение носит качественный характер. С одной стороны, оно приводит к существенному расширению постановок задач оптимизации алгоритмов, связанному с различной формой задания априорной информации, а с другой, — к резкому усложнению процесса исследования каждой постановки, в силу увеличения количества независимых переменных в рассматриваемых минимаксных задачах.

Рассматриваемые классы оптимизации введены в [21, 22] и используются в целях классификации постановок задач.

Анализ простейшего варианта неточного алгоритма Узавы изложен в соответствии с работой [153].

Результаты раздела 6.4.2, посвященного спектральным свойствам пучка операторов специального вида, получены в [24] и имеют важное самостоятельное значение. Во-первых, они (теоремы 6.4.7, 6.4.8 и их следствия) позволяют компактно и единообразно получать оценки спектра для операторов перехода целого класса итерационных методов решения седловых задач. Для сравнения: теоремы 2.2.1, 2.3.1, 2.4.1 и 2.5.1 из первой части книги преследуют ту же цель для более простого случая модифицированных релаксационных алгоритмов, но это достигается более громоздким (хотя элементарным) способом. Во-вторых, теорема 6.4.9 и ее следствия (при дополнительных ограничениях на пучок операторов) позволяет избавиться от исследования корней многочленов третьей степени для получения оценок спектра в практически важных случаях. Ранее этот «неудобный» этап анализа просто отбрасывался за счет дополнительного предположения, что ядро оператора G инвариантно относительно оператора L (см., например, [34, 36, 94, 95, 97]).

Необходимость введения специальных классов функций $Y(I)$ и $YS(I)$ с последующим изучением их свойств последовала из решения задачи асимптотической оптимизации трехпараметрического метода 3MSOR [129]. Результаты раздела 6.4.3 получены в [21, 23] и являются необходимыми при решении рассматриваемых минимаксных задач в областях с двумя и более независимыми переменными.

БЛОЧНО ТРЕУГОЛЬНОЕ ПРЕДОБУСЛОВЛИВАНИЕ (GMSOR)

В главе изучается обобщенный модифицированный метод (GMSOR) для решения регулярных и нерегулярных симметричных, а также несимметричных седловых задач.

Выбор этого алгоритма в качестве стартового при исследовании обобщенных методов обусловлен сразу несколькими обстоятельствами. Во-первых, его релаксационные версии (методы MSOR и 3MSOR) также эффективны, как и метод Узава — сопряженных градиентов, поэтому выяснение количественных характеристик скорости сходимости обобщенного метода GMSOR во всевозможных постановках — интересная и актуальная задача. Во-вторых, он наиболее удобен для усвоения общей методологии исследования алгоритмов второй части книги, так как спектр его оператора перехода структурно разнообразен и имеет богатое наполнение, в том числе, комплекснозначное. И, наконец, неутрачиваемые результаты по асимптотической оптимизации метода GMSOR имеют настолько необычную форму, что могут доставить эстетическое удовольствие специалистам.

7.1. ФОРМУЛИРОВКА МЕТОДА И ЕГО СВОЙСТВА

Обобщенный модифицированный блочный метод SOR (метод GMSOR) для линейных задач с седловыми операторами (6.2) может быть записан в следующей форме:

$$\begin{cases} Q \frac{u^{k+1} - u^k}{\tau} + (A + \beta BC^{-1} B^*) u^k + B p^k = f + \beta BC^{-1} \varphi \\ -\alpha C \frac{p^{k+1} - p^k}{\tau} + B^* u^{k+1} = \varphi, \end{cases} \quad (7.1)$$

где $\alpha, \tau > 0$, $\beta \in \mathbb{R}$ — фиксированные итерационные параметры.

Запишем оператор перехода метода (7.1) в следующем виде:

$$T_{\text{GMSOR}} \equiv T(\alpha, \beta, \tau; A, B, Q, C) = I - \tau S, \quad (7.2)$$

где

$$S = \begin{pmatrix} Q & 0 \\ 0 & -\alpha C \end{pmatrix}^{-1} \begin{pmatrix} A + \beta BC^{-1}B^* & B \\ B^*(I - \tau Q^{-1}(A + \beta BC^{-1}B^*)) & -\tau B^*Q^{-1}B \end{pmatrix},$$

$$\alpha, \tau > 0, \quad \beta \in \mathbb{R}. \quad (7.3)$$

Важную роль в исследовании спектральных характеристик оператора T_{GMSOR} играет следующая

Лемма 7.1.1. Число $\mu \in \sigma(S)$ тогда и только тогда, когда $\mu \neq 0$ и существует вектор $u \in U \setminus \{0\}$ такой, что

$$\left[\mu^2 Q - \mu(A + (\beta + \tau/\alpha)BC^{-1}B^*) + \alpha^{-1}BC^{-1}B^* \right] u = 0. \quad (7.4)$$

Доказательство. Пусть $\mu \in \sigma(S)$, тогда существует вектор

$$z = \{u, p\} \in Z \setminus \{0\}$$

такой, что $Sz = \mu z$, или в развернутой форме

$$\begin{cases} (A + \beta BC^{-1}B^*)u + Bp = \mu Qu, \\ B^*(I - \tau Q^{-1}(A + \beta BC^{-1}B^*))u - \tau B^*Q^{-1}Bp = -\alpha \mu Cp. \end{cases} \quad (7.5)$$

Отметим, что $u \neq 0$, так как в противном случае из (7.5) будет следовать, что и $p = 0$. Кроме того, $\mu \neq 0$, в силу невырожденности исходной задачи с седловой точкой и, следовательно, оператора S . Применим к обеим частям второго уравнения (7.5) оператор BC^{-1} и подставим в полученное уравнение выражение Bp из первого уравнения (7.5), в результате получим

$$\left[\mu^2 Q - \mu(A + (\beta + \tau/\alpha)BC^{-1}B^*) + \alpha^{-1}BC^{-1}B^* \right] u = 0.$$

В обратную сторону: пусть $\mu \neq 0$ и $u \in U \setminus \{0\}$ удовлетворяет (7.4). Введем вектор $v = \mu Qu - (A + \beta BC^{-1}B^*)u$. Тогда из уравнения (7.4) следует

$$\mu v + \alpha^{-1}(1 - \tau\mu)BC^{-1}B^*u = 0.$$

Умножая скалярно обе части полученного равенства на произвольный вектор $w \in \ker B^*$, получаем $\mu(v, w) = 0$, а так как $\mu \neq 0$, то $v \in \text{Im } B$. Следовательно, существует $p \in P$ такой, что $v = Bp$. Несложно убедиться, что вектор $\{u, p\} \neq 0$ является собственным вектором S , соответствующим собственному значению μ . Лемма доказана. ■

Всюду далее в этой главе используется обозначение

$$\rho_K = \inf_{\substack{\alpha, \tau > 0, \\ \beta \in \mathbb{R}}} \sup_{(A, B, Q, C) \in K} \rho(T(\alpha, \beta, \tau; A, B, Q, C)),$$

где K — заданный класс оптимизации.

7.2. СИММЕТРИЧНАЯ РЕГУЛЯРНАЯ ЗАДАЧА: ОПТИМИЗАЦИЯ В КЛАССЕ \mathbb{K}_1

Для исследования оптимальных характеристик алгоритма сначала получим оценки спектрального радиуса оператора перехода. Справедлива

Теорема 7.2.1 (Оценка в классе \mathbb{K}_1). Пусть

$$0 < \delta \leq \Delta, \quad 0 < \gamma \leq \Gamma, \\ \mathbb{K}_1 = \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma), \quad \alpha, \tau > 0, \quad \beta \in \mathbb{R},$$

тогда имеют место неравенства

$$\inf_{\substack{\alpha, \tau > 0, \\ \beta \in \mathbb{R}}} \rho_1(\alpha, \beta, \tau) \leq q_{\mathbb{K}_1} \leq \rho_1(\alpha, 0, \tau),$$

где

$$\rho_1(\alpha, \beta, \tau) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{|1 - \tau \mu_1|, |1 - \tau \mu_2^{1,2}|\},$$

$\mu_1 = s$, $\mu_2^{1,2}$ — все корни квадратного уравнения

$$\mu^2 - \mu(s + (\beta + \tau \alpha^{-1})ts) + \alpha^{-1}ts = 0.$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$, положим

$$L = A^{-1/2}QA^{-1/2}, \quad G = A^{-1/2}BC^{-1}B^*A^{-1/2}, \\ \delta_1 = \Delta^{-1}, \quad \delta_2 = \delta^{-1}, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 7.1.1, число $\lambda \in \sigma(T)$ тогда и только тогда, когда $\lambda \neq 1$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G,$$

$$f(\lambda, s) = (1 - \lambda)s - \tau, \quad g(\lambda, t) = 1 - \lambda, \quad h(\lambda) = \lambda\tau(\beta + \tau\alpha^{-1}) - \tau\beta.$$

При $\beta = 0$ выполнены все условия следствия 6.4.2, следовательно, имеем

$$q_{\mathbb{K}_1} \leq \rho(T(\alpha, 0, \tau; A, B, Q, C)) \leq \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\},$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим уравнениям

$$(1 - \lambda)s - \tau = 0, \quad ((1 - \lambda)s - \tau)(1 - \lambda) + \lambda\tau^2\alpha^{-1}t = 0,$$

откуда после замены s на s^{-1} получаем верхнюю оценку.

С другой стороны, по теореме 6.4.7 имеет место неравенство

$$\max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)} \rho(T(\alpha, \beta, \tau; A, B, Q, C)),$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим уравнениям

$$(1 - \lambda)s - \tau = 0, \quad ((1 - \lambda)s - \tau)(1 - \lambda) + \tau(\lambda(\beta + \tau\alpha^{-1}) - \beta)t = 0,$$

откуда после замены s на s^{-1} получаем нижнюю оценку. Теорема доказана. ■

Определим функции

$$\begin{aligned} \lambda(\alpha, \beta, \tau; t, s) &= \max |1 - \tau\mu_2^{1,2}| = \\ &= \max_{\pm} \left| 1 - \tau \left(s + \left(\beta + \frac{\tau}{\alpha} \right) ts \right) \pm \sqrt{D(\alpha, \beta, \tau; t, s)} \right|, \\ D(\alpha, \beta, \tau; t, s) &= \tau^2 \left(s + \left(\beta + \frac{\tau}{\alpha} \right) ts \right)^2 - 4ts/\alpha, \end{aligned}$$

и рассмотрим вспомогательную задачу

$$q_1 = \min_{\substack{\alpha, \tau > 0, \\ \beta \in \mathbb{R}}} \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda(\alpha, \beta, \tau; t, s). \quad (7.6)$$

Ее решает

Теорема 7.2.2. Если $\xi = \gamma/\Gamma < 1$ и $\omega = \delta/\Delta < 1$, то задача (7.6) имеет единственное решение

$$\alpha_1 = \frac{\Gamma + \gamma}{2\delta} \frac{(1 - q_1^2)^2}{(2\omega - 1) + q_1^2} > 0, \quad \beta_1 = 0, \quad \tau_1 = \frac{1 - q_1^2}{\delta} > 0, \quad (7.7)$$

где q_1 является единственным на интервале $(0, 1)$ корнем уравнения

$$\frac{1 - \xi}{1 + \xi} q^3 - (1 + \omega) q^2 + (2\omega - 1) \frac{1 - \xi}{1 + \xi} q + (1 - \omega) = 0. \quad (7.8)$$

Доказательство. Обозначим левую часть (7.8) через $f(q)$, тогда выполнено:

$$\begin{aligned}\lim_{q \rightarrow -\infty} f(q) &= -\infty, \\ f(0) &> 0, \quad f\left(\sqrt{\frac{1-\omega}{1+\omega}}\right) > 0, \quad f(1) < 0, \\ \lim_{q \rightarrow +\infty} f(q) &= +\infty.\end{aligned}$$

Следовательно, кубическое уравнение $f(q) = 0$ имеет три различных вещественных корня, причем только один из них лежит на интервале $(0, 1)$. Таким образом, величина q_1 корректно определена и, более того, $q_1 > \sqrt{(1-\omega)/(1+\omega)}$, откуда следует, что $\alpha_1 > 0$, $\tau_1 > 0$.

Покажем, что

$$\lambda_1(t, s) \equiv \lambda(\alpha_1, \beta_1, \tau_1; t, s) = q_1$$

при любых $s = \delta, \Delta$, $t = \gamma, \Gamma$. Пусть

$$D_1(t, s) = D(\alpha_1, \beta_1, \tau_1; t, s),$$

тогда при $s = \delta$ имеем

$$\begin{aligned}\text{sign } D_1(t, \delta) &= \text{sign} \left(((1-x)(2\omega - 1 + q_1^2) + 1 - q_1^2)^2 - \right. \\ &\quad \left. - 4(1-x)(2\omega - 1 + q_1^2) \right) < 0,\end{aligned}\tag{7.9}$$

где

$$x = 1 - 2t(\gamma + \Gamma)^{-1} \in \left\{ -\frac{1-\xi}{1+\xi}, \quad \frac{1-\xi}{1+\xi} \right\}.$$

Это следует из неравенства для корней

$$x_{\pm} = 2 \frac{\omega - 1 \pm q_1}{2\omega - 1 + q_1^2}$$

квадратного многочлена относительно x , находящегося под знаком sign :

$$\begin{aligned}|x_{\pm}| &= \pm 2q_1(\omega - 1 \pm q_1) \left(q_1(2\omega - 1 + q_1^2) \frac{1-\xi}{1+\xi} \right)^{-1} \frac{1-\xi}{1+\xi} = \\ &= \pm \frac{2q_1(\omega - 1 \mp q_1)}{(1+\omega)q_1^2 - (1-\omega)} \frac{1-\xi}{1+\xi} > \frac{1-\xi}{1+\xi} \quad \text{при } q_1 \in \left(\sqrt{\frac{1-\omega}{1+\omega}}, 1 \right).\end{aligned}$$

Из неравенства (7.9) следует, что при $t = \gamma, \Gamma$ справедливо равенство

$$\lambda_1(t, \delta) = \sqrt{1 - \tau_1 \delta (1 + \beta_1 t)} = \sqrt{1 - \tau_1 \delta} = q_1.$$

При $s = \Delta$ имеет место неравенство

$$\operatorname{sign} D_1(t, \Delta) = \operatorname{sign} (((1-x)(2\omega-1+q_1^2)+1-q_1^2)^2 - 4\omega(1-x)(2\omega-1+q_1^2)) > 0, \quad (7.10)$$

которое следует из того, что корни

$$x_{\pm} = \pm 2 \frac{\sqrt{\omega(\omega-1+q_1^2)}}{2\omega-1+q_1^2}$$

многочлена относительно x , находящегося под знаком sign , либо чисто мнимые, либо при $q_1 \in [\sqrt{1-\omega}, 1)$ справедлива оценка

$$|x_{\pm}| = \frac{2q_1 \sqrt{\omega(\omega-1+q_1^2)}}{(1+\omega)q_1^2 - (1-\omega)} \frac{1-\xi}{1+\xi} < \frac{1-\xi}{1+\xi}.$$

Из неравенства (7.10) следует, что при $t = \gamma, \Gamma$ выполнено

$$\lambda_1(t, \Delta) - \left| 1 - \frac{\tau_1 \Delta}{2} \left(1 + \frac{\tau_1 t}{\alpha_1} \right) \right| = \tau_1 \sqrt{\frac{\Delta^2}{4} \left(1 + \frac{\tau_1 t}{\alpha_1} \right)^2 - \frac{\Delta t}{\alpha_1}}.$$

Возводя в квадрат обе части равенства и исключая τ_1 и α_1 , получим, что $\lambda_1(t, \Delta)$ является наибольшим из корней квадратного уравнения

$$\frac{1-\xi}{1+\xi} \lambda_1(t, \Delta) q_1^2 - (q_1^2 + \omega \lambda_1^2(t, \Delta)) + (2\omega-1) \frac{1-\xi}{1+\xi} \lambda_1(t, \Delta) + (1-\omega) = 0.$$

Введем обозначения $\lambda_{1,1}(t, \Delta)$ и $\lambda_{1,2}(t, \Delta)$ для корней этого уравнения. В силу (7.8), значение $\lambda_{1,1}(t, \Delta) = q_1$ удовлетворяет этому уравнению, поэтому, поделив уравнение на $\lambda_1(t, \Delta) - q_1$ с использованием (7.8), получаем второй корень

$$\lambda_{1,2}(t, \Delta) = \frac{q_1^2 + \omega - 1}{\omega q_1}.$$

Так как

$$\frac{q_1^2 + \omega - 1}{\omega q_1} < q_1$$

при любых значениях $q_1 \neq 0$, то отсюда следует, что

$$\lambda_1(\gamma, \Delta) = \lambda_1(\Gamma, \Delta) = q_1.$$

В задаче (7.6) проведем регулярную замену переменных

$$\alpha' = \tau^2/2\alpha, \quad \beta' = \tau(\beta + \tau/\alpha), \quad \tau' = \tau,$$

в результате получим функцию

$$\lambda'(\alpha', \beta', \tau'; t, s) = \max_{\pm} \left| 1 - \tau' s - \beta' t s \pm \sqrt{(\tau' s + \beta' t s)^2 - 2\alpha' t s} \right|,$$

при этом будем считать, что λ' определена при произвольных $\alpha', \beta', \tau', t, s \in \mathbb{R}$.

Если зафиксировать α', β', τ' и одну из переменных t, s , то $\lambda' \in F_{0,1} \subset Y(\mathbb{R})$ как функция оставшейся переменной. Отсюда (теорема 6.4.10, свойство 2) следует, что

$$\Lambda'(\alpha', \beta', \tau') \equiv \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \lambda'(\alpha', \beta', \tau'; t, s) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda'(\alpha', \beta', \tau'; t, s).$$

Докажем, что точка $(\alpha'_1, \beta'_1, \tau'_1)$ является точкой строгого локального минимума функции $\Lambda'(\alpha', \beta', \tau')$, для чего, в силу равенств

$$\begin{aligned} \lambda'(\alpha'_1, \beta'_1, \tau'_1; \gamma, \delta) &= \lambda'(\alpha'_1, \beta'_1, \tau'_1; \Gamma, \delta) = \\ &= \lambda'(\alpha'_1, \beta'_1, \tau'_1; \gamma, \Delta) = \lambda'(\alpha'_1, \beta'_1, \tau'_1; \Gamma, \Delta) = q_1 \end{aligned}$$

и леммы 6.4.1, достаточно доказать, что при $t \in \{\gamma, \Gamma\}$, $s \in \{\delta, \Delta\}$ существуют вектор-градиент

$$v(t, s) = \nabla' \lambda'|_{(\alpha'_1, \beta'_1, \tau'_1)}, \quad \text{где} \quad \nabla' = \left(\frac{\partial}{\partial \alpha'}, \frac{\partial}{\partial \beta'}, \frac{\partial}{\partial \tau'} \right),$$

и выполнено включение

$$0 \in \text{conv}\{v(\gamma, \delta), v(\Gamma, \delta), v(\gamma, \Delta), v(\Gamma, \Delta)\}. \quad (7.11)$$

Существование $v(t, s)$ при $t = \gamma, \Gamma$, $s = \delta, \Delta$ следует из аналитического вида λ'_1 и неравенств

$$D'_1(t, s) \neq 0, \quad 1 - \tau'_1 s - \beta'_1 t s \neq 0,$$

причем

$$\begin{aligned} v(\gamma, \delta) &= \frac{\delta \gamma}{q_1} \begin{pmatrix} 1 \\ -1 \\ -1/\gamma \end{pmatrix}, \quad v(\gamma, \Delta) = \frac{(1 - q_1) \Delta \gamma}{\sqrt{D'_1(\gamma, \Delta)}} \begin{pmatrix} -(1 - q_1)^{-1} \\ 1 \\ 1/\gamma \end{pmatrix}, \\ v(\Gamma, \delta) &= \frac{\delta \Gamma}{q_1} \begin{pmatrix} 1 \\ -1 \\ -1/\Gamma \end{pmatrix}, \quad v(\Gamma, \Delta) = \frac{(1 + q_1) \Delta \Gamma}{\sqrt{D'_1(\Gamma, \Delta)}} \begin{pmatrix} -(1 + q_1)^{-1} \\ 1 \\ 1/\Gamma \end{pmatrix}. \end{aligned}$$

Отметим, что все коэффициенты, вынесенные за скобки в полученных выражениях, положительны, а точки, заключенные в скобках, находятся в общем положении, так как $q_1 \in (0, 1)$ и $\gamma < \Gamma$. Поэтому условие (7.11) эквивалентно тому, что найдутся положительные константы C_1, C_2, C_3, C_4 такие, что справедливо

$$C_1 \begin{pmatrix} 1 \\ -1 \\ -1/\gamma \end{pmatrix} + C_2 \begin{pmatrix} 1 \\ -1 \\ -1/\Gamma \end{pmatrix} + C_3 \begin{pmatrix} -(1 - q_1)^{-1} \\ 1 \\ 1/\gamma \end{pmatrix} + C_4 \begin{pmatrix} -(1 + q_1)^{-1} \\ 1 \\ 1/\Gamma \end{pmatrix} = 0.$$

Несложно убедиться, что этому равенству удовлетворяют значения

$$C_1 = C_3 = 1, \quad C_2 = C_4 = \frac{1 + q_1}{1 - q_1} > 0.$$

Покажем теперь, что точка $(\alpha'_1, \beta'_1, \tau'_1)$ — единственная точка локального минимума функции $\Lambda'(\alpha', \beta', \tau')$ и, более того, является точкой ее глобального минимума. Рассмотрим произвольную точку $(\alpha'_0, \beta'_0, \tau'_0)$, отличную от $(\alpha'_1, \beta'_1, \tau'_1)$. Определим функцию

$$g(x; t, s) = \lambda'(\alpha'(x), \beta'(x), \tau'(x); t, s),$$

где

$$\begin{aligned} \alpha'(x) &= \alpha'_1(1 - x) + \alpha'_0x, \\ \beta'(x) &= \beta'_1(1 - x) + \beta'_0x, \\ \tau'(x) &= \tau'_1(1 - x) + \tau'_0x, \quad x \in \mathbb{R}. \end{aligned}$$

Если зафиксировать произвольные значения t, s , то $g \in F_{0,1}$ как функция от x ; а так как функция

$$h(x) \equiv \Lambda'(\alpha'(x), \beta'(x), \tau'(x)) = \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} g(x; t, s)$$

имеет строгий локальный минимум в точке $x = 0$, то $h \in \text{YS}(\mathbb{R})$ по теореме 6.4.11. Следовательно, справедливо неравенство

$$\Lambda'(\alpha'_1, \beta'_1, \tau'_1) = h(0) < h(1) = \Lambda'(\alpha'_0, \beta'_0, \tau'_0),$$

и точка $(\alpha'_1, \beta'_1, \tau'_1)$ — единственная точка локального минимума Λ' . Теорема доказана. ■

Полученные оценки немедленно приводят к основному результату.

Теорема 7.2.3 (Асимптотическая оптимизация в \mathbb{K}_1). Пусть

$$0 < \delta \leq \Delta, \quad 0 < \gamma < \Gamma,$$

тогда задача асимптотической оптимизации алгоритма GMSOR в классе $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$ имеет следующее решение:

$$q_{\mathbb{K}_1} = q_1, \quad \alpha_1 = \frac{\Gamma + \gamma}{2\delta} \frac{(1 - q_1^2)^2}{(2\omega - 1) + q_1^2} > 0, \quad \beta_1 = 0, \quad \tau_1 = \frac{1 - q_1^2}{\delta} > 0,$$

где q_1 является единственным на интервале $(0, 1)$ корнем уравнения

$$\frac{1 - \xi}{1 + \xi} q^3 - (1 + \omega) q^2 + (2\omega - 1) \frac{1 - \xi}{1 + \xi} q + (1 - \omega) = 0,$$

$$\omega = \delta/\Delta, \quad \xi = \gamma/\Gamma.$$

Доказательство. Пусть $\delta < \Delta$, тогда из неравенства

$$|1 - \tau_1 s| \leq \max \left\{ q_1^2, \left| 1 - \frac{1 - q_1^2}{\omega} \right| \right\} \leq q_1^2 < q_1 \quad (7.12)$$

и теоремы 7.2.2 следует, что

$$q_1 = \rho_1(\alpha_1, \beta_1, \tau_1) < \rho_1(\alpha, \beta, \tau)$$

для любых $\alpha, \tau > 0$, $\beta \in \mathbb{R}$, $(\alpha, \beta, \tau) \neq (\alpha_1, \beta_1, \tau_1)$. Таким образом, по теореме 7.2.1 имеем

$$\begin{aligned} q_1 = \rho_1(\alpha_1, \beta_1, \tau_1) &= \inf_{\substack{\alpha, \tau > 0, \\ \beta \in \mathbb{R}}} \rho_1(\alpha, \beta, \tau) \leq q_{\mathbb{K}_1} \leq \rho_1(\alpha_1, 0, \tau_1) = \\ &= \rho_1(\alpha_1, \beta_1, \tau_1) = q_1. \end{aligned}$$

Случай $\delta = \Delta$ следует из непрерывности минимаксной задачи. Теорема доказана. ■

Следствие 7.2.1. Имеют место асимптотические равенства

$$\begin{aligned} q_{\mathbb{K}_1} &= 1 - \omega + O(\omega^2) && \text{при } \omega \rightarrow 0, \xi = \text{const}, \\ q_{\mathbb{K}_1} &= 1 - \sqrt{\frac{4\omega}{2 - \omega}} \xi + O(\xi) && \text{при } \xi \rightarrow 0, \omega = \text{const}. \end{aligned}$$

Доказательство. При фиксированном $\omega \in \mathbb{C}$ функция $f(\xi) = q_{\mathbb{K}_1}(\omega, \xi)$ является алгебраической в $\mathbb{C} \setminus \{1\}$ и, значит, раскладывается в ряд Пуизо в окрестности точки $\xi = 0$ [101, с. 169]. Методом неопределенных коэффициентов несложно вычислить первые члены этого разложения. Аналогично рассматривается случай фиксированного ξ . Следствие доказано. ■

7.3. СИММЕТРИЧНАЯ РЕГУЛЯРНАЯ ЗАДАЧА: ОПТИМИЗАЦИЯ В КЛАССЕ \mathbb{K}_2

Для исследования задачи асимптотической оптимизации алгоритма GMSOR в классе \mathbb{K}_2 применим тот же подход, что был использован при изучении задачи в классе \mathbb{K}_1 .

Теорема 7.3.1 (Оценка в классе \mathbb{K}_2). Пусть $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$, $\mathbb{K}_2 = \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, $\alpha, \tau > 0$, $\beta \in \mathbb{R}$, тогда имеют место неравенства

$$\inf_{\substack{\alpha, \tau > 0, \\ \beta \in \mathbb{R}}} \rho_2(\alpha, \beta, \tau) \leq q_{\mathbb{K}_2} \leq \rho_2(\alpha, 0, \tau),$$

где

$$\rho_2(\alpha, \beta, \tau) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{|1 - \tau \mu_1| \cdot |1 - \tau \mu_2^{1,2}|\}.$$

$\mu_1 = s$, $\mu_2^{1,2}$ -- все корни квадратного уравнения

$$\mu^2 - \mu(s + (\beta + \tau\alpha^{-1})t) + \alpha^{-1}t = 0.$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, положим

$$L = Q^{-1/2}AQ^{-1/2}, \quad G = Q^{-1/2}BC^{-1}B^*Q^{-1/2},$$

$$\delta_1 = \delta, \quad \delta_2 = \Delta, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 7.1.1, число $\lambda \in \sigma(T)$ тогда и только тогда, когда $\lambda \neq 1$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G,$$

$$f(\lambda, s) = 1 - \lambda - \tau s,$$

$$g(\lambda, t) = 1 - \lambda,$$

$$h(\lambda) = \lambda\tau(\beta + \tau\alpha^{-1}) - \tau\beta.$$

При $\beta = 0$ выполнены все условия следствия 6.4.2, следовательно, имеем

$$q_{\mathbb{K}_2} \leq \rho(T(\alpha, 0, \tau; A, B, Q, C)) \leq \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\},$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим уравнениям

$$1 - \lambda - \tau s = 0, \quad (1 - \lambda - \tau s)(1 - \lambda) + \lambda\tau^2\alpha^{-1}t = 0,$$

откуда получаем верхнюю оценку.

С другой стороны, по теореме 6.4.7 имеет место неравенство

$$\max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)} \rho(T(\alpha, \beta, \tau; A, B, Q, C)),$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим уравнениям

$$1 - \lambda - \tau s = 0, \quad (1 - \lambda - \tau s)(1 - \lambda) + \tau(\lambda(\beta + \tau\alpha^{-1}) - \beta)t = 0,$$

откуда получаем нижнюю оценку. Теорема доказана. ■

Определим функции

$$\begin{aligned} \lambda(\alpha, \beta, \tau; t, s) &= \max |1 - \tau\mu_2^{1,2}| = \\ &= \max_{\pm} \left| 1 - \tau(s + (\beta + \tau/\alpha)t) \pm \sqrt{D(\alpha, \beta, \tau; t, s)} \right|, \end{aligned}$$

$$D(\alpha, \beta, \tau; t, s) = \tau^2((s + (\beta + \tau/\alpha)t)^2 - 4t/\alpha),$$

$$\Lambda(\alpha, \beta, \tau) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda(\alpha, \beta, \tau; t, s)$$

и рассмотрим вспомогательную задачу

$$q_2 = \min_{\alpha, \tau > 0, \beta \in \mathbb{R}} \Lambda(\alpha, \beta, \tau). \quad (7.13)$$

Ее решает

Теорема 7.3.2. Пусть $\xi = \gamma/\Gamma < 1$, $\omega = \delta/\Delta < 1$ и $\hat{q}_2 \in (0, 1)$ — корень уравнения

$$\frac{1-\xi}{1+\xi}q^3 - (1+\omega)q^2 + (2\omega-1)\frac{1-\xi}{1+\xi}q + (1-\omega) = 0. \quad (7.14)$$

Если выполнено условие

$$\omega > \frac{(1+\xi) - (1-\xi)\hat{q}_2}{2},$$

то задача (7.13) имеет единственное решение

$$\begin{aligned} q_2 = \hat{q}_2, \quad \alpha_2 &= \frac{\Gamma + \gamma}{2\delta\Delta} \frac{(1 - \hat{q}_2^2)^2}{(2\omega - 1) + \hat{q}_2^2} > 0, \\ \beta_2 = 0, \quad \tau_2 &= \frac{1 - \hat{q}_2^2}{\delta} > 0, \end{aligned} \quad (7.15)$$

и, кроме того, справедливо неравенство

$$q_2 \leq \Lambda(\alpha_2, \beta_2, \tau_2) = \hat{q}_2$$

для любых ω , ξ .

Доказательство. Обозначим левую часть (7.14) через $f(q)$, тогда выполнено:

$$\begin{aligned} \lim_{q \rightarrow -\infty} f(q) &= -\infty, \\ f(0) &> 0, \quad f\left(\sqrt{\frac{1-\omega}{1+\omega}}\right) > 0, \quad f(1) < 0, \\ \lim_{q \rightarrow +\infty} f(q) &= +\infty. \end{aligned}$$

Следовательно, кубическое уравнение $f(q) = 0$ имеет три различных вещественных корня, причем только один из них лежит на интервале $(0, 1)$. Таким образом, величина \hat{q}_2 корректно определена и, более того,

$$\hat{q}_2 > \sqrt{\frac{1-\omega}{1+\omega}},$$

откуда следует, что $\alpha_2 > 0$, $\tau_2 > 0$.

Покажем, что

при любых $s = \delta, \Delta$, $t = \gamma, \Gamma$. Обозначим через

$$D_2(t, s) = D(\alpha_2, \beta_2, \tau_2; t, s),$$

тогда при $s = \delta$ имеем

$$\begin{aligned} \text{sign } D_2(t, \delta) = \text{sign} \left(((1-x)(2\omega - 1 + \hat{q}_2^2) + \omega(1 - \hat{q}_2^2))^2 - \right. \\ \left. - 4\omega(1-x)(2\omega - 1 + \hat{q}_2^2) \right) < 0, \end{aligned} \quad (7.16)$$

где

$$x = 1 - 2t(\gamma + \Gamma)^{-1} \in \left\{ -\frac{1-\xi}{1+\xi}, \frac{1-\xi}{1+\xi} \right\}.$$

Это следует из неравенства для корней

$$x_{\pm} = \frac{(1-\omega)(\hat{q}_2^2 - 1) \pm 2\omega\hat{q}_2}{2\omega - 1 + \hat{q}_2^2}$$

квадратного многочлена относительно x , находящегося под знаком sign :

$$\begin{aligned} |x_{\pm}| &= \pm \hat{q}_2((1-\omega)(\hat{q}_2^2 - 1) \pm 2\omega\hat{q}_2) \left(\hat{q}_2(2\omega - 1 + \hat{q}_2^2) \frac{1-\xi}{1+\xi} \right)^{-1} \frac{1-\xi}{1+\xi} = \\ &= \pm \frac{\hat{q}_2((1-\omega)(\hat{q}_2^2 - 1) \pm 2\omega\hat{q}_2)}{(1+\omega)\hat{q}_2^2 - (1-\omega)} \frac{1-\xi}{1+\xi} > \frac{1-\xi}{1+\xi} \end{aligned}$$

при

$$\hat{q}_2 \in \left(\sqrt{\frac{1-\omega}{1+\omega}}, 1 \right).$$

Из неравенства (7.16) следует, что при $t = \gamma, \Gamma$ справедливо равенство

$$\lambda_2(t, \delta) = \sqrt{1 - \tau_2(\delta + \beta_2 t)} = \sqrt{1 - \tau_2 \delta} = \hat{q}_2.$$

При $s = \Delta$ имеет место неравенство

$$\begin{aligned} \text{sign } D_2(t, \Delta) = \text{sign} \left(((1-x)(2\omega - 1 + \hat{q}_2^2) + 1 - \hat{q}_2^2)^2 - \right. \\ \left. - 4\omega(1-x)(2\omega - 1 + \hat{q}_2^2) \right) > 0, \end{aligned} \quad (7.17)$$

которое следует из того, что корни

$$x_{\pm} = \pm 2 \frac{\sqrt{\omega(\omega - 1 + \hat{q}_2^2)}}{2\omega - 1 + \hat{q}_2^2}$$

многочлена относительно x , находящегося под знаком sign , либо чисто мнимые, либо $\hat{q}_2 \in [\sqrt{1-\omega}, 1)$ и справедлива оценка

$$|x_{\pm}| = \frac{2\hat{q}_2 \sqrt{\omega(\omega - 1 + \hat{q}_2^2)}}{(1+\omega)\hat{q}_2^2 - (1-\omega)} \frac{1-\xi}{1+\xi} < \frac{1-\xi}{1+\xi}.$$

Из неравенства (7.17) следует, что при $t = \gamma, \Gamma$ выполнено

$$\lambda_2(t, \Delta) - \left| 1 - \frac{\tau_2 \Delta}{2} \left(1 + \frac{\tau_2 t}{\alpha_2} \right) \right| = \tau_2 \sqrt{\frac{\Delta^2}{4} \left(1 + \frac{\tau_2 t}{\alpha_2} \right)^2 - \frac{\Delta t}{\alpha_2}}.$$

Возводя в квадрат обе части равенства и исключая τ_2 и α_2 , получим, что $\lambda_2(t, \Delta)$ является наибольшим из корней квадратного уравнения

$$\frac{1-\xi}{1+\xi} \lambda_2(t, \Delta) \hat{q}_2^2 - (\hat{q}_2^2 + \omega \lambda_2^2(t, \Delta)) + (2\omega - 1) \frac{1-\xi}{1+\xi} \lambda_2(t, \Delta) + (1-\omega) = 0.$$

Введем обозначения $\lambda_{2,1}(t, \Delta)$ и $\lambda_{2,2}(t, \Delta)$ для корней этого уравнения. В силу (7.14), значение $\lambda_{2,1}(t, \Delta) = \hat{q}_2$ удовлетворяет этому уравнению, поэтому, поделив уравнение на $\lambda_2(t, \Delta) - \hat{q}_2$ с использованием (7.14), получаем второй корень

$$\lambda_{2,2}(t, \Delta) = \frac{\hat{q}_2^2 + \omega - 1}{\omega \hat{q}_2}.$$

Так как

$$\frac{\hat{q}_2^2 + \omega - 1}{\omega \hat{q}_2} < \hat{q}_2$$

при любых значениях $\hat{q}_2 \neq 0$, то отсюда следует, что

$$\lambda_2(\gamma, \Delta) = \lambda_2(\Gamma, \Delta) = \hat{q}_2.$$

В задаче (7.6) проведем регулярную замену переменных

$$\alpha' = \tau^2/2\alpha, \quad \beta' = \tau(\beta + \tau/\alpha), \quad \tau' = \tau,$$

в результате получим функцию

$$\lambda'(\alpha', \beta', \tau'; t, s) = \max_{\pm} \left| 1 - \tau' s - \beta' t \pm \sqrt{(\tau' s + \beta' t)^2 - 2\alpha' t} \right|,$$

при этом будем считать, что λ' определена при произвольных $\alpha', \beta', \tau', t, s \in \mathbb{R}$.

Если зафиксировать α', β', τ' и одну из переменных t, s , то $\lambda' \in F_{0,1} \subset Y(\mathbb{R})$ как функция оставшейся переменной. Отсюда (теорема 6.4.10, свойство 2) следует, что

$$\Lambda'(\alpha', \beta', \tau') \equiv \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \lambda'(\alpha', \beta', \tau'; t, s) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda'(\alpha', \beta', \tau'; t, s).$$

Докажем, что при условии

$$\omega > \frac{(1+\xi) - (1-\xi)\hat{q}_2}{2}$$

точка $(\alpha'_2, \beta'_2, \tau'_2)$ является точкой строгого локального минимума функции $\Lambda'(\alpha', \beta', \tau')$, для чего, в силу равенств

$$\begin{aligned}\lambda'(\alpha'_2, \beta'_2, \tau'_2; \gamma, \delta) &= \lambda'(\alpha'_2, \beta'_2, \tau'_2; \Gamma, \delta) = \\ &= \lambda'(\alpha'_2, \beta'_2, \tau'_2; \gamma, \Delta) = \lambda'(\alpha'_2, \beta'_2, \tau'_2; \Gamma, \Delta) = \hat{q}_2\end{aligned}$$

и леммы 6.4.1, достаточно доказать, что при $t = \gamma, \Gamma$, $s = \delta, \Delta$ существуют вектор-градиент

$$v(t, s) = \nabla' \lambda' |_{(\alpha'_2, \beta'_2, \tau'_2)}, \quad \text{где} \quad \nabla' = \left(\frac{\partial}{\partial \alpha'}, \frac{\partial}{\partial \beta'}, \frac{\partial}{\partial \tau'} \right),$$

и выполнено включение

$$0 \in]\text{conv}\{v(\gamma, \delta), v(\Gamma, \delta), v(\gamma, \Delta), v(\Gamma, \Delta)\}]. \quad (7.18)$$

Существование $v(t, s)$ при $t = \gamma, \Gamma$, $s = \delta, \Delta$ следует из аналитического вида λ'_2 и неравенств

$$D'_2(t, s) \neq 0, \quad 1 - \tau'_2 s - \beta'_2 t \neq 0,$$

причем имеют место представления:

$$\begin{aligned}v(\gamma, \delta) &= \frac{\gamma}{\hat{q}_2} \begin{pmatrix} 1 \\ -1 \\ -\delta/\gamma \end{pmatrix}, \quad v(\gamma, \Delta) = \frac{(1 - \hat{q}_2)\gamma}{\sqrt{D'_2(\gamma, \Delta)}} \begin{pmatrix} -(1 - \hat{q}_2)^{-1} \\ 1 \\ \Delta/\gamma \end{pmatrix}, \\ v(\Gamma, \delta) &= \frac{\Gamma}{\hat{q}_2} \begin{pmatrix} 1 \\ -1 \\ -\delta/\Gamma \end{pmatrix}, \quad v(\Gamma, \Delta) = \frac{(1 + \hat{q}_2)\Gamma}{\sqrt{D'_2(\Gamma, \Delta)}} \begin{pmatrix} -(1 + \hat{q}_2)^{-1} \\ 1 \\ \Delta/\Gamma \end{pmatrix}.\end{aligned}$$

Отметим, что все коэффициенты, вынесенные за скобки в полученных выражениях, положительны, а точки, заключенные в скобках, находятся в общем положении, так как $\hat{q}_2 \in (0, 1)$ и $\gamma < \Gamma$. Поэтому условие (7.18) эквивалентно тому, что найдутся положительные константы C_1, C_2, C_3, C_4 такие, что справедливо

$$C_1 \begin{pmatrix} 1 \\ -1 \\ -\delta/\gamma \end{pmatrix} + C_2 \begin{pmatrix} 1 \\ -1 \\ -\delta/\Gamma \end{pmatrix} + C_3 \begin{pmatrix} -(1 - \hat{q}_2)^{-1} \\ 1 \\ \Delta/\gamma \end{pmatrix} + C_4 \begin{pmatrix} -(1 + \hat{q}_2)^{-1} \\ 1 \\ \Delta/\Gamma \end{pmatrix} = 0.$$

Такие значения существуют только при условии

$$\omega > ((1 + \xi) - (1 - \xi)\hat{q}_2)/2$$

и могут быть выбраны, например, следующим образом:

$$\begin{aligned}C_1 &= (1 + \xi) - (1 - \xi)\hat{q}_2 - 2\omega\xi, \\ C_2 &= 2\omega - (1 + \xi) + (1 - \xi)\hat{q}_2, \\ C_3 &= \omega(1 - \xi)(1 - \hat{q}_2), \\ C_4 &= \omega(1 - \xi)(1 + \hat{q}_2).\end{aligned}$$

Покажем теперь, что при условии

$$\omega > \frac{(1 + \xi) - (1 - \xi)\hat{q}_2}{2}$$

точка $(\alpha'_2, \beta'_2, \tau'_2)$ — единственная точка локального минимума функции $\Lambda'(\alpha', \beta', \tau')$ и, более того, является точкой ее глобального минимума. Рассмотрим произвольную точку $(\alpha'_0, \beta'_0, \tau'_0)$, отличную от $(\alpha'_2, \beta'_2, \tau'_2)$. Определим функцию

$$g(x; t, s) = \lambda'(\alpha'(x), \beta'(x), \tau'(x); t, s),$$

где

$$\alpha'(x) = \alpha'_1(1 - x) + \alpha'_0x,$$

$$\beta'(x) = \beta'_1(1 - x) + \beta'_0x,$$

$$\tau'(x) = \tau'_1(1 - x) + \tau'_0x, \quad x \in \mathbb{R}.$$

Если зафиксировать произвольные значения t, s , то $g \in F_{0,1}$ как функция от x ; а так как функция

$$h(x) \equiv \Lambda'(\alpha'(x), \beta'(x), \tau'(x)) = \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} g(x; t, s)$$

имеет строгий локальный минимум в точке $x = 0$, то $h \in \text{YS}(\mathbb{R})$ по теореме 6.4.11. Следовательно, имеет место неравенство

$$\Lambda'(\alpha'_2, \beta'_2, \tau'_2) = h(0) < h(1) = \Lambda'(\alpha'_0, \beta'_0, \tau'_0)$$

и точка $(\alpha'_2, \beta'_2, \tau'_2)$ — единственная точка локального минимума Λ' . Теорема доказана. ■

Докажем главный результат. Справедлива

Теорема 7.3.3 (Асимптотическая оптимизация в \mathbb{K}_2).

Пусть $\omega = \delta/\Delta \leq 1$, $\xi = \gamma/\Gamma < 1$, тогда при выборе параметров

$$\alpha_2 = \frac{\Gamma + \gamma}{2\delta\Delta} \frac{(1 - q_2^2)^2}{(2\omega - 1) + q_2^2} > 0, \quad \beta_2 = 0, \quad \tau_2 = \frac{1 - q_2^2}{\delta} > 0$$

справедливы соотношения

$$q_{\mathbb{K}_2} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)} \rho(T(\alpha_2, \beta_2, \tau_2; A, B, Q, C)) = q_2,$$

где q_2 является единственным на интервале $(0, 1)$ корнем уравнения

$$\frac{1 - \xi}{1 + \xi} q^3 - (1 + \omega) q^2 + (2\omega - 1) \frac{1 - \xi}{1 + \xi} q + (1 - \omega) = 0.$$

Если дополнительно выполнено неравенство

$$\omega \geq \frac{(1 + \xi) - (1 - \xi)q_2}{2},$$

то $q_{\mathbb{K}_2} = q_2$.

Доказательство. Пусть $\delta < \Delta$, тогда из неравенств

$$|1 - \tau_2 s| \leq \max \left\{ q_2^2, \left| 1 - \frac{1 - q_2^2}{\omega} \right| \right\} \leq q_2^2 < q_2$$

и теоремы 7.3.1 следует, что

$$q_{\mathbb{K}_2} \leq \rho_2(\alpha_2, 0, \tau_2) = \rho_2(\alpha_2, \beta_2, \tau_2) = q_2.$$

При условии $\omega \geq ((1 + \xi) - (1 - \xi)q_2)/2$ из непрерывности задачи и теоремы 7.3.2 следует

$$q_2 = \rho_2(\alpha_2, \beta_2, \tau_2) \leq \rho_2(\alpha, \beta, \tau)$$

для любых $\alpha, \tau > 0$, $\beta \in \mathbb{R}$, $(\alpha, \beta, \tau) \neq (\alpha_2, \beta_2, \tau_2)$. Таким образом, по теореме 7.3.1 имеем

$$\begin{aligned} q_2 = \rho_2(\alpha_2, \beta_2, \tau_2) &= \inf_{\substack{\alpha, \tau > 0, \\ \beta \in \mathbb{R}}} \rho_2(\alpha, \beta, \tau) \leq q_{\mathbb{K}_2} \leq \rho_2(\alpha_2, 0, \tau_2) = \\ &= \rho_2(\alpha_2, \beta_2, \tau_2) = q_2. \end{aligned}$$

Случай $\delta = \Delta$ следует из непрерывности минимаксной задачи. Теорема доказана. ■

Следствие 7.3.1. Имеет место асимптотика

$$q_{\mathbb{K}_2} = 1 - \sqrt{\frac{4\omega}{2 - \omega}} \xi + O(\xi) \quad \text{при } \xi \rightarrow 0, \omega = \text{const}.$$

Доказательство. Пусть $q_2 = q_2(\omega, \xi)$ — величина, определенная в теореме 7.3.3, тогда из соотношений

$$\lim_{\xi \rightarrow 0} ((1 + \xi) - (1 - \xi)q_2)/2 = (1 - \lim_{\xi \rightarrow 0} q_2)/2 = 0 < \omega$$

следует, что при фиксированном $\omega \in (0, 1]$ и достаточно малых значениях $\xi \in (0, 1]$ имеет место неравенство $\omega \geq ((1 + \xi) - (1 - \xi)q_2)/2$. Таким образом, при достаточно малых ξ по теореме 7.3.3 имеет место равенство $q_{\mathbb{K}_2} = q_2$.

При фиксированном $\omega \in \mathbb{C}$ функция $f(\xi) = q_2(\omega, \xi)$ является алгебраической в $\mathbb{C} \setminus \{1\}$ и, значит, раскладывается в ряд Пуизо в окрестности точки $\xi = 0$. Используя метод неопределенных коэффициентов, несложно получить первые члены разложения. ■

7.4. СИММЕТРИЧНАЯ НЕРЕГУЛЯРНАЯ ЗАДАЧА: ОЦЕНКА В КЛАССЕ \mathbb{K}_{2s}

Используя подход, предложенный в разделе 6.3, применим результаты, полученные для класса \mathbb{K}_2 , к случаю симметричных нерегулярных задач из класса \mathbb{K}_{2s} .

Имеет место

Теорема 7.4.1 (Оценка в классе \mathbb{K}_{2s}). Пусть $\omega = \delta/\Delta \leq 1$, $\xi = \gamma/\Gamma < 1$, тогда при выборе параметров

$$\alpha_{2s} = \frac{\Gamma + \gamma}{2\delta\Delta_s} \frac{(1 - q_{2s}^2)^2}{(2\omega_s - 1) + q_{2s}^2} > 0, \quad \beta_{2s} = \frac{\delta}{\gamma}, \quad \tau_{2s} = \frac{1 - q_{2s}^2}{\delta} > 0,$$

справедливы соотношения

$$q_{\mathbb{K}_{2s}} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma)} \rho(T(\alpha_{2s}, \beta_{2s}, \tau_{2s}; A, B, Q, C)) = q_{2s},$$

где

$$\omega_s = \frac{\delta}{\Delta_s}, \quad \Delta_s = \delta(\omega^{-1} + \xi^{-1}),$$

а q_{2s} является единственным на интервале $(0, 1)$ корнем уравнения

$$\frac{1 - \xi}{1 + \xi} q^3 - (1 + \omega_s) q^2 + (2\omega_s - 1) \frac{1 - \xi}{1 + \xi} q + (1 - \omega_s) = 0.$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma)$, тогда, в силу (6.17),

$$(A + \beta BC^{-1} B^*, B, Q, C) \in \mathbb{K}_2(\delta(\beta), \Delta(\beta), \gamma, \Gamma),$$

где

$$\beta > 0, \quad \delta(\beta) = \min\{\delta, \beta\gamma\}, \quad \Delta(\beta) = \Delta + \beta\Gamma.$$

По теореме 7.3.3 имеет место оценка

$$q_{\mathbb{K}_{2s}} \leq q_{\mathbb{K}_2}(\beta),$$

где $q_{\mathbb{K}_2}(\beta)$ — оптимальный показатель сходимости GMSOR в классе $\mathbb{K}_2(\delta(\beta), \Delta(\beta), \gamma, \Gamma)$.

Так как $q_{\mathbb{K}_2} = q_{\mathbb{K}_2}(\omega, \xi)$ монотонно зависит от $\omega \in (0, 1]$ при фиксированном $\xi \in (0, 1]$, то $q_{\mathbb{K}_2}(\beta)$ принимает наименьшее значение одновременно с величиной $\delta(\beta)/\Delta(\beta)$, т. е. при $\beta = \beta_{2s}$. Таким образом, справедливо неравенство $q_{\mathbb{K}_{2s}} \leq q_{2s}$ при $\alpha = \alpha_{2s}$, $\beta = \beta_{2s}$, $\tau = \tau_{2s}$. Теорема доказана. ■

Следствие 7.4.1. Имеют место асимптотические оценки

$$q_{\mathbb{K}_{2s}} \leq 1 - \omega + o(\omega) \quad \text{при} \quad \omega \rightarrow 0, \quad \xi = \text{const},$$

$$q_{\mathbb{K}_{2s}} \leq 1 - \sqrt{2}\xi + o(\xi) \quad \text{при} \quad \xi \rightarrow 0, \quad \omega = \text{const}.$$

Доказательство. При фиксированном $\omega \in \mathbb{C}$ функция $f(\xi) = q_{2s}(\omega, \xi)$ раскладывается в ряд Пуизо в окрестности точки $\xi = 0$. Методом неопределенных коэффициентов несложно вычислить первые члены этого разложения. Аналогично рассматривается случай фиксированного ξ . Следствие доказано. ■

7.5. НЕСИММЕТРИЧНАЯ РЕГУЛЯРНАЯ ЗАДАЧА: ОЦЕНКА В КЛАССЕ \mathbb{K}_3

Исследование для несимметричных задач осуществляется по той же схеме, что и для симметричных: сначала оценивается спектральный радиус, а затем проводится оптимизация полученной оценки. Существенное отличие состоит в большей сложности аналитического исследования оценок. В связи с этим анализ ограничивается изучением асимптотики оптимального показателя по параметру, характеризующему асимметрию задачи.

Теорема 7.5.1 (Оценка в классе \mathbb{K}_3). Пусть $0 < \gamma \leq \Gamma$, $0 \leq R$, $\mathbb{K}_3 = \mathbb{K}_3(R, \gamma, \Gamma)$, $\alpha, \tau > 0$, $\beta \in \mathbb{R}$, тогда имеют место неравенства

$$\inf_{\substack{\alpha, \tau > 0, \\ \beta \in \mathbb{R}}} \rho_3(\alpha, \beta, \tau) \leq q_{\mathbb{K}_3} \leq \hat{\rho}_3(\alpha, \beta, \tau),$$

где

$$\rho_3(\alpha, \beta, \tau) = \max_{\substack{s \in [1-iR, 1+iR] \\ t \in [\gamma, \Gamma]}} \{|1 - \tau\mu_1|, |1 - \tau\mu_2^{1,2}|\},$$

$$\hat{\rho}_3(\alpha, \beta, \tau) = \max_{\substack{s \in [1-iR, 1+iR] \\ t \in [\gamma, \Gamma]}} \{|1 - \tau\mu_1|, |1 - \tau\mu_2^{1,2}|, |1 - \tau\mu_3^{1,2,3}|\},$$

$\mu_1 = s$, $\mu_2^{1,2}$ — все корни квадратного уравнения

$$\mu^2 - \mu(s + t(\beta + \tau\alpha^{-1})) + \alpha^{-1}t = 0,$$

а $\mu_3^{1,2,3}$ — все корни кубического уравнения

$$\mu^3 - \mu^2(2 + (\beta + \tau\alpha^{-1})t) + \mu(1 + R^2 + \alpha^{-1}t + (\beta + \tau\alpha^{-1})ts) - \alpha^{-1}ts = 0.$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_3(R, \gamma, \Gamma)$, положим

$$L = -iQ^{-1/2}(A - A^*)Q^{-1/2}, \quad G = Q^{-1/2}BC^{-1}B^*Q^{-1/2}, \\ \delta_1 = -R, \quad \delta_2 = R, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 7.1.1, число $\lambda \in \sigma(T)$ тогда и только тогда, когда $\lambda \neq 1$ и $\lambda \in \sigma(\chi)$, где

Таким образом, выполнены все условия теоремы 6.4.9, следовательно, имеем

$$q_{\mathbb{K}_3} \leq \rho(T(\alpha, \beta, \tau; A, B, Q, C)) \leq \max_{\substack{s \in [1-iR, 1+iR] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\},$$

где $s = 1 + ir$ и максимум берется по всем $\lambda(s, t)$, удовлетворяющим уравнениям

$$\begin{aligned} 1 - \lambda - \tau s &= 0, \\ (1 - \lambda - \tau s)(1 - \lambda) + \lambda \tau^2 \alpha^{-1} t &= 0, \\ ((1 - \lambda - \tau)^2 + \tau^2 R^2)(1 - \lambda) + (1 - \lambda - \tau s)(\lambda \tau(\beta + \tau \alpha^{-1}) - \tau \beta) t &= 0, \end{aligned}$$

откуда получаем верхнюю оценку.

С другой стороны, по теореме 6.4.7 имеет место неравенство

$$\max_{\substack{s \in [1-iR, 1+iR] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_3(R, \gamma, \Gamma)} \rho(T(\alpha, \beta, \tau; A, B, Q, C)),$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим уравнениям

$$\begin{aligned} 1 - \lambda - \tau s &= 0, \\ (1 - \lambda - \tau s)(1 - \lambda) + \lambda \tau^2 \alpha^{-1} t &= 0, \end{aligned}$$

откуда получаем нижнюю оценку. Теорема доказана. \blacksquare

Получим асимптотически точные оценки оптимального показателя.

Теорема 7.5.2. Существует $R_0 = R_0(\gamma, \Gamma) \geq 0$ такое, что для любых $R \geq R_0$ имеют место неравенства

$$\frac{R}{\sqrt{1 + R^2}} \leq q_{\mathbb{K}_3} \leq \frac{2R^2 + 1}{2R^2 + 2}.$$

Доказательство. Положим

$$\alpha_0 = \frac{\gamma}{1 + R^2}, \quad \beta_0 = \frac{2}{\gamma}, \quad \tau_0 = \frac{1}{1 + R^2}.$$

Из теоремы 7.5.1 следует

$$q_{\mathbb{K}_3} \geq \min_{\tau > 0} \max_{s \in [1-iR, 1+iR]} \{ |1 - \tau s| \} = |1 - \tau_0(1 \pm iR)| = (1 + R^{-2})^{-1/2}.$$

Таким образом, получена нижняя оценка для $q_{\mathbb{K}_3}$.

По теореме 7.5.1 для доказательства верхней оценки достаточно показать, что существует R_0 такое, что для любых $R > R_0$, $s \in [1 - iR, 1 + iR]$, $t \in [\gamma, \Gamma]$ справедлива система неравенств

$$\begin{cases} |1 - \tau_0 \mu_1| \leq 1 - \tau_0/2 \\ |1 - \tau_0 \mu_2^{1,2}| \leq 1 - \tau_0/2 \\ |1 - \tau_0 \mu_3^{1,2,3}| \leq 1 - \tau_0/2 \end{cases} \quad (7.19)$$

Проанализируем неравенства последовательно.

Первое неравенство (7.19) имеет место для любых $R > R_1 = 0$, так как

$$|1 - \tau_0 \mu_1| = |1 - \tau_0 s| \leq |1 - \tau_0(1 \pm iR)| = (1 + R^{-2})^{-1/2} < 1 - \frac{\tau_0}{2}.$$

Величины $\mu_2^{1,2}$ из второго неравенства (7.19) удовлетворяют уравнению

$$\begin{aligned} \mu^2 - (1 - 3x + Ryi)\mu + (1 + R^2)x &= 0, \\ x = \frac{t}{\gamma} \in [1, \xi^{-1}], \quad y = \operatorname{Im} \frac{s}{R} &\in [-1, 1]. \end{aligned}$$

Для любых решений μ этого уравнения при фиксированных x, y и $R \rightarrow +\infty$ имеют место представления:

$$\begin{aligned} |1 - \tau_0 \mu| &= \sqrt{1 - \tau_0(2 \operatorname{Re} \mu - \tau_0(\operatorname{Im} \mu)^2) + O(\tau_0^2)}, \\ 2 \operatorname{Re} \mu - \tau_0(\operatorname{Im} \mu)^2 &= \left[(1 + 3x) \left(1 \pm \frac{y}{\sqrt{y^2 + 4x}} \right) - \right. \\ &\quad \left. - \frac{1}{4} \left(y \pm \sqrt{y^2 + 4x} \right)^2 \right] + o(1), \end{aligned}$$

причем выражение в квадратных скобках достигает минимального значения при $x = 1$, $y = \mp 1$. Таким образом, при достаточно больших R справедливы неравенства

$$2 \operatorname{Re} \mu - \tau_0(\operatorname{Im} \mu)^2 \geq \frac{5\sqrt{5} - 3}{2\sqrt{5}} + o(1) > 1,$$

откуда следует, что существует $R_2 > 0$ такое, что

$$|1 - \tau_0 \mu_{1,2}^2| \leq \sqrt{1 - \tau_0} < 1 - \frac{\tau_0}{2}$$

для любых $R > R_2$. Следовательно, при $R_0 \geq R_2$ имеет место второе неравенство (7.19).

Заметим, что величины

$$q_3^{1,2,3} = \frac{1 - \tau_0 \mu_3^{1,2,3}}{1 - \tau_0/2}$$

удовлетворяют уравнению

$$R(q) = aq^3 + bq^2 + c_y q + d_y = 0,$$

где

$$\begin{aligned} a &= \left(1 - \frac{\tau_0}{2}\right)^3, \quad b = -[3 - (2 + 3x)\tau_0] \left(1 - \frac{\tau_0}{2}\right)^2, \\ c_y &= \left[3 - (3 + 5x)\tau_0 + 3x\tau_0^2 \left(1 + \sqrt{\frac{1 - \tau_0}{\tau_0}} y i\right)\right] \left(1 - \frac{\tau_0}{2}\right), \\ d_y &= -\left(1 - (1 + 2x)\tau_0 + 2x\tau_0^2 \left(1 + \sqrt{\frac{1 - \tau_0}{\tau_0}} y i\right)\right). \end{aligned}$$

При $y = \pm 1$, $x \in [1, \xi^{-1}]$ и $R > R_2$ все корни этого уравнения лежат внутри единичного круга, так как в этом случае

$$\mu_3^1 = \mu_1, \quad \mu_3^2 = \bar{\mu}_2^1, \quad \mu_3^3 = \bar{\mu}_2^2.$$

Воспользуемся теоремой Шура-Кона для оценки корней при фиксированных $y \in (-1, 1)$, $x \geq 1$ и достаточно больших R . Так как a и b не зависят от y , а c_y и d_y зависят от y линейно, то получим

$$\begin{aligned} \max_{y \in [-1, 1]} |d_y| &= \max\{|d_{-1}|, |d_1|\} < |a|, \\ |c_y \bar{a} - d_y \bar{b}| &\leq \max_{y \in [-1, 1]} |c_y \bar{a} - d_y \bar{b}| = \\ &= \max_{y \in \{-1, 1\}} |c_y \bar{a} - d_y \bar{b}| < |a|^2 - |d_{\pm 1}|^2 \leq |a|^2 - |d_y|^2. \end{aligned}$$

Разложим функцию

$$f(\tau_0) = \frac{|(\bar{a}b - \bar{c}_y d_y)(|a|^2 - |d_y|^2) - (a\bar{b} - c_y \bar{d}_y)(c_y \bar{a} - d_y \bar{b})|}{(|a|^2 - |d_y|^2)^2 - |c\bar{a} - d\bar{b}|^2}$$

по формуле Тейлора в нуле

$$f(\tau_0) = 1 - F\tau_0 + o(\tau_0), \quad \tau_0 \rightarrow +0,$$

где

$$\begin{aligned} F &= \frac{(1 - y^2)(4 - 4x - 48x^2 - 16x^3 + 64x^4 + 108x^3 y^2 - 27x^2 y^2)}{2(2 - 4x - 16x^2 + 9xy^2)^2} \geq \\ &\geq \frac{y^2(1 - y^2)}{2(2 - y^2)^2} > 0. \end{aligned}$$

Таким образом, существует $R_{x,y}$ такое, что $f(\tau_0) = f((1 + R^2)^{-1}) < 1$ для любых $R > R_{x,y}$, а значит, все корни уравнения $R(q) = 0$ лежат внутри единичного круга.

В силу непрерывной зависимости корней уравнения $R(q) = 0$ от x, y и компактности множества $[1, \xi^{-1}] \times [-1, 1]$, можно выбрать $R_3 \geq R_2$ так, что все корни этого уравнения лежат внутри единичного круга при любых $(x, y) \in [1, \xi^{-1}] \times [-1, 1]$, $R > R_3$. При последнем условии справедлива оценка

$$|1 - \tau_0 \mu_3^{1,2,3}| = (1 - \tau_0/2)|q_3^{1,2,3}| < 1 - \tau_0/2.$$

Следовательно, при $R_0 \geq R_3$ имеет место третье неравенство (7.19).

Таким образом, система неравенств (7.19) справедлива при $R_0 = \max\{R_1, R_2, R_3\}$. Теорема доказана. ■

Следствие 7.5.1. Имеет место асимптотика

$$q_{K_3} = 1 - \frac{1}{2R^2} + O(R^{-4}) \text{ при } \xi = \text{const}, R \rightarrow +\infty.$$

Доказательство. Формула немедленно следует из теоремы 7.5.2 и представлений

$$\begin{aligned} \frac{R}{\sqrt{1+R^2}} &= 1 - \frac{1}{2R^2} + O(R^{-4}), \quad R \rightarrow +\infty, \\ \frac{2R^2+1}{2R^2+2} &= 1 - \frac{1}{2R^2} + O(R^{-4}), \quad R \rightarrow +\infty. \end{aligned}$$

Следствие доказано. ■

7.6. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Метод GMSOR, рассмотренный в этой главе, является наиболее общей формой классического метода Эрроу—Гурвица ($Q = I$, $C = I$, $\beta = 0$), который был предложен в [108].

Его изучение оказалось весьма сложной проблемой. В первую очередь, это было связано с основным направлением исследований. Дело в том, что по аналогии с симметричными положительно определенными задачами представлялось естественным сначала получить оценку погрешности в некоторой норме вида $\|z^k\| \leq q^k \|z^0\|$ с показателем, зависящим от итерационных параметров, например $q = q(\tau, \alpha)$, а затем провести оптимизацию показателя относительно параметров.

Видимо, впервые качественная оценка сходимости была получена в [180] для постановки задачи оптимизации в классе $K_2(\delta, \Delta, \gamma, \Gamma)$:

$$q_{K_2} \leq \sqrt{1 - \frac{1}{8}\omega^2\xi} < 1 - \frac{1}{16}\omega^2\xi, \quad \omega = \frac{\delta}{\Delta}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Существенно позднее в этом направлении в работе [124] было получено продвижение в решении задачи оптимизации в классе $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$:

$$q_{\mathbb{K}_1} \leq \frac{1}{2}((1 - \xi)\omega + \sqrt{(1 - \xi)^2\omega^2 + 4(1 - \omega)}) < 1 - \frac{1}{2}\omega\xi.$$

Далее эта оценка была уточнена в [202], но без качественного улучшения показателя.

Принципиально иной подход, основанный на оценке степеней оператора перехода, был предложен в [11], и получена следующая асимптотика оптимального показателя

$$q_{\mathbb{K}_1} \approx 1 - 2\omega\sqrt{\xi} \quad \text{при } \xi \rightarrow 0, \omega \rightarrow 1.$$

Наконец, в работе [135] (см. также [95]) было предложено решать задачу асимптотической оптимизации метода, т. е. минимизировать спектральный радиус, имея в виду, что для получения оценки погрешности всегда существует матричная норма, сколь угодно близкая к спектральному радиусу матрицы (см. [82], лемма 5.6.10). В указанных работах впервые было приведено кубическое уравнение для оптимального показателя сходимости $q_{\mathbb{K}_1}$ вида, представленного в формулировке теоремы 7.2.3 с асимптотикой

$$q_{\mathbb{K}_1} = 1 - \sqrt{\frac{4\omega}{2 - \omega}}\xi + O(\xi) \quad \text{при } \xi \rightarrow 0, \omega = \text{const}.$$

Однако оценки спектра оператора перехода там были получены в предположении, что ядро оператора $A^{-1/2}BC^{-1}B^*A^{-1/2}$ инвариантно относительно оператора $A^{-1/2}QA^{-1/2}$. Кроме этого, само решение задачи асимптотической оптимизации было получено более из интуитивных соображений, нежели строго обосновано.

Окончательные результаты по выводу и обоснованию асимптотически оптимальных итерационных параметров метода GMSOR в классах \mathbb{K}_1 и \mathbb{K}_2 были получены в [21, 128].

Класс оптимизации \mathbb{K}_{2s} был введен и проанализирован в [26] для задач с $\ker A \neq \{0\}$, и ослабление асимптотик показателей сходимости, видимо, является естественной компенсацией за расширение класса решаемых проблем и качественное ухудшение их спектральных свойств.

Оценка скорости сходимости метода GMSOR для несимметричных задач с седловыми операторами впервые получена в [125]:

$$q < 1 - \frac{\omega^3\xi}{24(1 + R)^2}.$$

Асимптотически точная оценка в классе \mathbb{K}_3 выведена в работе [22].

В работе [168] анализировалось блочно треугольное предобусловливание для системы L_ϵ с $\epsilon \geq 0$, основанное на методах GMRES [184] и BI-CGSTAB [192]. Полученные там результаты трудно сопоставить с точными оценками, так как вычислительные затраты на итерацию при таком подходе существенно выше, а оценки сходимости приводятся в других терминах.

БЛОЧНО ДИАГОНАЛЬНОЕ ПРЕДОБУСЛОВЛИВАНИЕ

Блочно диагональное предобусловливание седловых задач структурно распадается на два практически несвязанных направления исследований. Эту ситуацию удобно формализовать, представив предобусловливатель в виде

$$\begin{pmatrix} Q & 0 \\ 0 & \alpha C \end{pmatrix}, \quad \alpha \neq 0.$$

Если параметр α отрицательный, то метод простой итерации для предобусловленной системы является обобщением метода MJOR из первой части книги. В противном случае, предобусловленная система является симметризуемой, и для ее решения наиболее разумным является метод Ланцоша, использующий подпространства Крылова. Принципиальная разница в идейной конструкции алгоритмов приводит к качественному различию в постановках задач оптимизации: в первом случае минимизируется спектральный радиус оператора перехода, а во втором — число обусловленности.

Для удобства параметр α всюду считается положительным, а знак, разграничивающий указанные выше случаи, присутствует в явном виде.

8.1. ОБОБЩЕННЫЙ МЕТОД MJOR (GMJOR)

8.1.1. Формулировка метода

Обобщенный модифицированный блочный метод Якоби (метод GMJOR) для линейных задач с седловыми операторами (6.2) может быть записан в следующей форме:

$$\begin{cases} Q \frac{u^{k+1} - u^k}{\tau} + (A + \beta BC^{-1}B^*) u^k + Bp^k = f + \beta BC^{-1}\varphi, \\ -\alpha C \frac{p^{k+1} - p^k}{\tau} + B^* u^k = \varphi, \end{cases} \quad (8.1)$$

где $\alpha, \tau > 0$, $\beta \in \mathbb{R}$ — фиксированные итерационные параметры.

Запишем оператор перехода метода (8.1) в следующем виде:

$$T_{\text{GMJOR}} \equiv T(\alpha, \beta, \tau; A, B, Q, C) = I - \tau S, \quad (8.2)$$

где

$$S = \begin{pmatrix} Q & 0 \\ 0 & -\alpha C \end{pmatrix}^{-1} \begin{pmatrix} A + \beta BC^{-1} B^* & B \\ B^* & 0 \end{pmatrix}, \quad \alpha, \tau > 0, \beta \in \mathbb{R}. \quad (8.3)$$

8.1.2. Связь спектров операторов перехода GMJOR и GMSOR

Исследование спектральных характеристик оператора T_{GMJOR} может быть сведено к исследованию спектральных характеристик T_{GMSOR} при помощи следующей леммы:

Лемма 8.1.1. *Имеет место равенство*

$$\sigma(T_{\text{GMJOR}}(\alpha, \beta, \tau; A, B, Q, C)) = \sigma\left(T_{\text{GMSOR}}\left(\alpha, \beta - \frac{\tau}{\alpha}, \tau; A, B, Q, C\right)\right).$$

Доказательство. Пусть $\mu \in \sigma(S)$, тогда существует вектор $z = \{u, p\} \in Z \setminus \{0\}$ такой, что $Sz = \mu z$, или в развернутой форме

$$\begin{cases} (A + \beta BC^{-1} B^*)u + Bp = \mu Qu, \\ B^*u = -\alpha Cp. \end{cases} \quad (8.4)$$

Отметим, что $u \neq 0$, так как в противном случае из (8.4) будет следовать, что и $p = 0$. Кроме того, $\mu \neq 0$ в силу невырожденности исходной задачи с седловой точкой и, следовательно, оператора S . Применим к обеим частям второго уравнения (8.4) оператор BC^{-1} и подставим в полученное уравнение выражение Bp из первого уравнения (8.4), тогда придем к соотношению

$$[\mu^2 Q - \mu(A + \beta BC^{-1} B^*) + \alpha^{-1} BC^{-1} B^*]u = 0. \quad (8.5)$$

Рассмотрим произвольные $\mu \neq 0$ и $u \in U \setminus \{0\}$, удовлетворяющие (8.5). Введем вектор $v = \mu Qu - (A + \beta BC^{-1} B^*)u$. Тогда из уравнения (8.5) следует

$$\mu v + \alpha^{-1} BC^{-1} B^* u = 0.$$

Умножая скалярно обе части полученного равенства на произвольный вектор $w \in \ker B^*$, получаем $\mu(v, w) = 0$, а так как $\mu \neq 0$, то $v \in \text{Im } B$. Следовательно, существует $p \in P$ такой, что $v = Bp$. Несложно убедиться, что вектор $\{u, p\} \neq 0$ является собственным вектором S , соответствующим собственному значению μ .

Таким образом, $\mu \in \sigma(S)$ тогда и только тогда, когда $\mu \neq 0$ и найдется $u \in U \setminus \{0\}$ такой, что выполнено (8.5). Замена переменной β на $\beta + \tau/\alpha$ сводит задачу (8.5) к (7.4), откуда немедленно следует утверждение леммы. ■

8.1.3. Оптимизация метода в классах \mathbb{K}_1 , \mathbb{K}_2 , \mathbb{K}_3 , \mathbb{K}_{2s}

Лемма 8.1.1 позволяет естественным образом перенести все результаты, полученные для метода GMSOR, на случай метода GMJOR. Приведем их формулировки.

Теорема 8.1.1 (Асимптотическая оптимизация в \mathbb{K}_1). Пусть $\omega = \delta/\Delta \leq 1$, $\xi = \gamma/\Gamma < 1$, тогда задача асимптотической оптимизации алгоритма GMJOR в классе $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$ имеет следующее решение:

$$q_{\mathbb{K}_1} = q_1, \quad \alpha_1 = \frac{\Gamma + \gamma}{2\delta} \frac{(1 - q_1^2)^2}{(2\omega - 1) + q_1^2} > 0,$$

$$\beta_1 = \frac{\tau_1}{\alpha_1}, \quad \tau_1 = \frac{1 - q_1^2}{\delta} > 0,$$

где q_1 является единственным на интервале $(0, 1)$ корнем уравнения

$$\frac{1 - \xi}{1 + \xi} q^3 - (1 + \omega) q^2 + (2\omega - 1) \frac{1 - \xi}{1 + \xi} q + (1 - \omega) = 0.$$

Теорема 8.1.2 (Асимптотическая оптимизация в \mathbb{K}_2). Пусть $\omega = \delta/\Delta \leq 1$, $\xi = \gamma/\Gamma < 1$, тогда при выборе параметров

$$\alpha_2 = \frac{\Gamma + \gamma}{2\delta\Delta} \frac{(1 - q_2^2)^2}{(2\omega - 1) + q_2^2} > 0, \quad \beta_2 = \frac{\tau_2}{\alpha_2}, \quad \tau_2 = \frac{1 - q_2^2}{\delta} > 0$$

для оптимального показателя сходимости $q_{\mathbb{K}_2}$ метода GMJOR в классе $\mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$ имеет место оценка

$$q_{\mathbb{K}_2} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)} \rho(T(\alpha_2, \beta_2, \tau_2; A, B, Q, C)) = q_2,$$

где q_2 является единственным на интервале $(0, 1)$ корнем уравнения

$$\frac{1 - \xi}{1 + \xi} q^3 - (1 + \omega) q^2 + (2\omega - 1) \frac{1 - \xi}{1 + \xi} q + (1 - \omega) = 0.$$

Если дополнительно выполнено неравенство $\omega \geq ((1 + \xi) - (1 - \xi)q_2)/2$, то $q_{\mathbb{K}_2} = q_2$.

Теорема 8.1.3 (Оценка в классе \mathbb{K}_{2s}). Пусть $\omega = \delta/\Delta \leq 1$, $\xi = \gamma/\Gamma < 1$, тогда при выборе параметров

$$\alpha_{2s} = \frac{\Gamma + \gamma}{2\delta\Delta_s} \frac{(1 - q_{2s}^2)^2}{(2\omega_s - 1) + q_{2s}^2} > 0, \quad \beta_{2s} = \frac{\delta}{\gamma} + \frac{\tau_{2s}}{\alpha_{2s}}, \quad \tau_{2s} = \frac{1 - q_{2s}^2}{\delta} > 0$$

для оптимального показателя сходимости $q_{\mathbb{K}_{2s}}$ метода GMJOR в классе $\mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma)$ имеет место оценка

$$q_{\mathbb{K}_{2s}} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma)} \rho(T(\alpha_{2s}, \beta_{2s}, \tau_{2s}; A, B, Q, C)) = q_{2s},$$

где $\omega_s = \delta/\Delta_s$, $\Delta_s = \delta(\omega^{-1} + \xi^{-1})$, а q_{2s} является единственным на интервале $(0, 1)$ корнем уравнения

$$\frac{1-\xi}{1+\xi}q^3 - (1+\omega_s)q^2 + (2\omega_s-1)\frac{1-\xi}{1+\xi}q + (1-\omega_s) = 0.$$

Теорема 8.1.4 (Оценка в классе \mathbb{K}_3). Существует

$$R_0 = R_0(\gamma, \Gamma) \geq 0$$

такое, что для любых $R \geq R_0$ имеют место неравенства

$$\frac{R}{\sqrt{1+R^2}} \leq q_{\mathbb{K}_3} \leq \frac{2R^2+1}{2R^2+2},$$

где $q_{\mathbb{K}_3}$ — оптимальный показатель сходимости метода GMJOR в классе $\mathbb{K}_3(\gamma, \Gamma, R)$.

8.1.4. Случай $\beta = 0$

Случай $\beta = 0$ представляет самостоятельный интерес при изучении алгоритма GMSOR. Дело в том, что при указанном значении уменьшается вычислительная сложность метода на каждом шаге итерационного процесса, так как при этом можно избежать вычисления слагаемого $\beta BC^{-1}B^*u^k$ в первом уравнении (8.1).

Неизбежно возникает вопрос — насколько серьезно наличие ненулевого параметра β влияет на скорость сходимости алгоритма? Оказывается, если решается семейство задач в классах $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$ или $\mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$ при малых значениях $\xi = \gamma/\Gamma$, то условие $\beta = 0$ существенно (по порядку) снижает скорость сходимости алгоритма GMJOR. Продемонстрируем это на примере оценок в классе $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$.

Теорема 8.1.5 (Нижняя оценка в классе \mathbb{K}_1). Пусть $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$, $\mathbb{K}_1 = \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$, тогда для оптимального показателя скорости сходимости метода GMJOR в \mathbb{K}_1 при условии $\alpha, \tau > 0$, $\beta = 0$ имеет место неравенство

$$\inf_{\alpha, \tau > 0} \rho_1(\alpha, \tau) \leq q_{\mathbb{K}_1},$$

где

$$\rho_1(\alpha, \tau) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{|1 - \tau\mu_1|, |1 - \tau\mu_2^{1,2}|\},$$

$\mu_1 = s$, $\mu_2^{1,2}$ — все корни квадратного уравнения

$$\mu^2 - \mu s + \alpha^{-1}ts = 0.$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$, положим

$$L = A^{-1/2}QA^{-1/2}, \quad G = A^{-1/2}BC^{-1}B^*A^{-1/2}, \\ \delta_1 = \Delta^{-1}, \quad \delta_2 = \delta^{-1}, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу лемм 7.1.1 и 8.1.1, число $\lambda \in \sigma(T_{GMJOR}(\alpha, 0, \tau; A, B, Q, C))$ тогда и только тогда, когда $\lambda \neq 1$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G,$$

$$f(\lambda, s) = (1 - \lambda)s - \tau, \quad g(\lambda, t) = 1 - \lambda, \quad h(\lambda) = \frac{\tau^2}{\alpha}.$$

По теореме 6.4.7 имеет место неравенство

$$\max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|\lambda(s, t)|\} \leq \sup_{(A, B, Q, C) \in \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)} \rho(T_{GMJOR}(\alpha, 0, \tau; A, B, Q, C)),$$

где максимум берется по всем $\lambda(s, t)$, удовлетворяющим уравнениям

$$(1 - \lambda)s - \tau = 0, \quad ((1 - \lambda)s - \tau)(1 - \lambda) + \frac{\tau^2 t}{\alpha} = 0,$$

откуда после замены s на s^{-1} получаем нижнюю оценку. Теорема доказана. ■

Определим функции

$$\lambda(\alpha, \tau; t, s) \equiv \max_{\pm} |1 - \tau \mu_2^{1,2}| = \max_{\pm} \left| 1 - \frac{\tau s}{2} \pm \sqrt{D(\alpha, \tau; t, s)} \right|, \\ D(\alpha, \tau; t, s) = \tau^2 \left(\frac{s^2}{4} - \frac{ts}{\alpha} \right)$$

и рассмотрим вспомогательную задачу

$$q_1 = \min_{\alpha, \tau > 0} \rho_1(\alpha, \tau) = \min_{\alpha, \tau > 0} \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{|1 - \tau s|, \lambda(\alpha, \tau; t, s)\}. \quad (8.6)$$

Ее решает

Теорема 8.1.6. Если $\xi = \gamma/\Gamma < 1$ и $\omega = \delta/\Delta < 1$, то задача (8.6) имеет единственное решение

$$q_1 = \frac{2\omega - \omega\xi + \sqrt{(6\omega + 2\xi - \xi\omega)^2 - 16\omega(2\omega + \xi)}}{2(2\omega + \xi)}, \\ \alpha_1 = \frac{\Gamma\omega(1 + q_1)}{\Delta(q_1 - 1 + \omega)}, \quad \tau_1 = \frac{1 + q_1}{\Delta}.$$

Доказательство. Параметр $\tau_1 > 0$ является бoльшим корнем квадратного уравнения

$$f(\tau) \equiv (2\omega + \xi)(\Delta\tau)^2 - (6\omega + 2\xi - \xi\omega)(\Delta\tau) + 4\omega = 0,$$

причем

$$f\left(\frac{2-\xi}{\Delta}\right) = -\xi^2(2-\omega-\xi) < 0,$$

$$f\left(\frac{2-\omega}{\Delta}\right) = -2\omega^2(1-\omega) < 0,$$

$$f(2\Delta) = 2\omega\xi > 0,$$

откуда следует, что величина

$$\tau_1 \in \left(\frac{2-\omega}{\Delta}, 2\Delta\right), \quad q_1 = \Delta\tau_1 - 1 \in (1-\omega, 1) \cap (1-\xi, 1)$$

и $\alpha_1 > 0$. Кроме того, несложно убедиться, что имеют место соотношения:

$$q_1 = \tau_1\Delta - 1 = \lambda(\alpha_1, \tau_1; \gamma, \Delta) = \lambda(\alpha_1, \tau_1; \Gamma, \delta) > \lambda(\alpha_1, \tau_1; \gamma, \delta),$$

$$D(\alpha_1; \tau_1; \gamma, \Delta) > 0, \quad D(\alpha_1; \tau_1; \Gamma, \delta) < 0.$$

В задаче (8.6) проведем регулярную замену переменных

$$\alpha' = \tau^2/\alpha, \quad \tau' = \tau,$$

в результате получим функцию

$$\lambda'(\alpha', \tau'; t, s) \equiv \max_{\pm} \left| 1 - \tau's/2 \pm \sqrt{\tau'^2 s^2 / 4 - \alpha' t s} \right|,$$

при этом будем считать, что λ' определена при произвольных $\alpha', \tau', t, s \in \mathbb{R}$.

Если зафиксировать α', τ' и одну из переменных t, s , то $\lambda' \in F_{0,1} \subset Y(\mathbb{R})$ как функция оставшейся переменной. Отсюда (теорема 6.4.10, свойство 2) следует, что

$$\begin{aligned} \Lambda'(\alpha', \tau') &\equiv \\ &\equiv \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \{|1 - \tau's|, \lambda'(\alpha', \tau'; t, s)\} = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{|1 - \tau's|, \lambda'(\alpha', \tau'; t, s)\}. \end{aligned}$$

Докажем, что точка (α'_1, τ'_1) является точкой строгого локального минимума функции $\Lambda'(\alpha', \tau')$, для чего, в силу леммы 6.4.1, достаточно показать, что существуют производные

$$v_1 = \nabla'(\Delta\tau'_1 - 1), \quad v_2 = \nabla'\lambda'(\alpha_1, \tau_1; \gamma, \Delta), \quad v_3 = \nabla'\lambda'(\alpha_1, \tau_1; \Gamma, \delta),$$

$$\nabla' = \left(\frac{\partial}{\partial \alpha'}, \frac{\partial}{\partial \tau'} \right)$$

и выполнено включение $0 \in \operatorname{conv}\{v_1, v_2, v_3\}$. Существование v_1, v_2, v_3 следует из аналитического вида λ' и неравенств

$$D'(\alpha_1; \tau_1; \gamma, \Delta) > 0, \quad D'(\alpha_1; \tau_1; \Gamma, \delta) < 0, \quad 1 - \tau_1 \Delta / 2 \neq 0,$$

причем имеют место представления

$$v_1 = \Delta \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad v_2 = \frac{\Delta}{2\sqrt{D'(\alpha_1; \tau_1; \gamma, \Delta)}} \begin{pmatrix} -\gamma \\ 1 - q_1 \end{pmatrix}, \quad v_3 = \delta \begin{pmatrix} \Gamma \\ -1 \end{pmatrix}.$$

Отметим, что все коэффициенты, вынесенные за скобки в полученных выражениях, положительны, а точки, заключенные в скобках — обозначим их соответственно w_1, w_2, w_3 , — находятся в общем положении, поэтому условие $0 \in \operatorname{conv}\{v_1, v_2, v_3\}$ выполнено тогда и только тогда, когда существуют $C_1, C_2, C_3 > 0$ такие, что справедливо $C_1 w_1 + C_2 w_2 + C_3 w_3 = 0$. Несложно убедиться, что этому равенству удовлетворяют значения

$$C_1 = q_1 - 1 + \xi > 0, \quad C_2 = 1, \quad C_3 = \xi.$$

Покажем теперь, что точка (α'_1, τ'_1) — единственная точка локального минимума функции $\Lambda'(\alpha', \tau')$ и, более того, является точкой ее глобального минимума. Рассмотрим произвольную точку (α'_0, τ'_0) , отличную от (α'_1, τ'_1) . Определим функции

$$g_1(x) = |1 - \tau'(x)s|, \quad g_2(x; t, s) = \Lambda'(\alpha'(x), \tau'(x); t, s),$$

где

$$\alpha'(x) = \alpha'_1(1 - x) + \alpha'_0 x, \quad \tau'(x) = \tau'_1(1 - x) + \tau'_0 x, \quad x \in \mathbb{R}.$$

Если зафиксировать произвольные значения t, s , то $g_1, g_2 \in F_{0,1}$ как функции от x , а так как функция

$$h(x) \equiv \Lambda'(\alpha'(x), \tau'(x)) = \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \{g_1(x), g_2(x; t, s)\}$$

имеет строгий локальный минимум в точке $x = 0$, то $h \in \text{YS}(\mathbb{R})$ по теореме 6.4.11. Следовательно, справедливо неравенство

$$\Lambda'(\alpha'_1, \tau'_1) = h(0) < h(1) = \Lambda'(\alpha'_0, \tau'_0)$$

и точка (α'_1, τ'_1) — единственная точка локального минимума Λ' . Теорема доказана. ■

Следствие 8.1.1. Имеют место оценки

$$q_{K_1} \geq 1 - \xi, \quad q_{K_1} \geq 1 - \omega.$$

Доказательство. Оценки следуют из теорем 8.1.5, 8.1.6, неравенств $q_{k_1} \geq q_1$ и $q_1 \geq 1 - \xi$, $q_1 \geq 1 - \omega$ (последние два неравенства были получены при доказательстве теоремы 8.1.6). Следствие доказано. ■

Сравнивая полученную в следствии 8.1.1 оценку с оценкой теоремы 8.1.1, которая может быть представлена в виде

$$q_{k_1} = 1 - \sqrt{\frac{4\omega}{2-\omega}}\xi + O(\xi) \quad \text{при } \xi \rightarrow 0, \omega = \text{const},$$

приходим к выводу, что наличие ненулевого параметра β существенно ускоряет сходимость алгоритма GMJOR при $\xi \rightarrow 0$.

8.2. ОБОБЩЕННЫЙ МЕТОД ЛАНЦОША (GMLan)

8.2.1. Построение метода

Обобщенный модифицированный блочный метод Ланцоша (GMLan) для решения (6.2) строится как классический предобусловленный метод Ланцоша решения системы линейных уравнений следующего вида

$$\begin{aligned} Sz &\equiv \begin{pmatrix} Q & 0 \\ 0 & \alpha C \end{pmatrix}^{-1} \begin{pmatrix} A + \beta BC^{-1}B^* & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \\ &= \begin{pmatrix} Q^{-1}f + \beta Q^{-1}BC^{-1}\varphi \\ \alpha^{-1}C^{-1}\varphi \end{pmatrix} \equiv F, \end{aligned}$$

где $\alpha > 0$, $\beta \in \mathbb{R}$ — фиксированные параметры. Корректность применения метода Ланцоша для решения системы уравнений с матрицей S следует из симметричности S относительно скалярного произведения $(z, w)_M = (Mz, w)$, где

$$M = \begin{pmatrix} Q & 0 \\ 0 & \alpha C \end{pmatrix}.$$

Формальная схема вычислений по методу GMLan выглядит следующим образом:

$$\begin{aligned} r^k &= Sz^k - F, \quad z^k = z^{k-1} - \tau_k y^k, \quad k = 1, 2, \dots \\ y^1 &= r^0, \quad y^2 = (S - \nu_2 I)y^1, \quad y^k = (S - \nu_k I)y^{k-1} - \mu_k y^{k-2}, \quad k = 3, 4, \dots, \\ \tau_k &= \frac{(r^{k-1}, Sr^k)_M}{(Sy^k, Sy^k)_M}, \quad \nu_k = \frac{(S^2 y^{k-1}, Sy^{k-1})_M}{(Sy^{k-1}, Sy^{k-1})_M}, \quad \mu_k = \frac{(Sy^{k-1}, Sy^{k-1})_M}{(Sy^{k-2}, Sy^{k-2})_M}. \end{aligned}$$

Стандартная оценка показателя асимптотического скорости сходимости q алгоритма GMLan, следующая из оценки сходимости

классического алгоритма Ланцоша [30, с. 296] имеет вид

$$q \leq \sqrt{\frac{\text{cond}_2(S) - 1}{\text{cond}_2(S) + 1}}.$$

Наилучшая оценка такого рода достигается минимизацией величины (спектрального числа обусловленности) $\text{cond}_2(S)$ за счет выбора «оптимальных» значений $\alpha > 0$, $\beta \in \mathbb{R}$.

Лемма 8.2.2. Число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \neq 0$ и существует вектор $u \in U \setminus \{0\}$ такой, что

$$\left[\lambda^2 Q - \lambda(A + \beta BC^{-1} B^*) - \alpha^{-1} BC^{-1} B^* \right] u = 0. \quad (8.7)$$

Доказательство. Пусть $\lambda \in \sigma(S)$, тогда существует вектор $z = \{u, p\} \in Z \setminus \{0\}$ такой, что $Sz = \lambda z$, или в развернутой форме

$$\begin{cases} (A + \beta BC^{-1} B^*)u + Bp = \lambda Qu, \\ B^*u = \alpha \lambda Cp. \end{cases} \quad (8.8)$$

Отметим, что $u \neq 0$, так как в противном случае из (8.8) будет следовать, что и $p = 0$. Кроме того, $\lambda \neq 0$, в силу невырожденности исходной задачи с седловой точкой и, следовательно, оператора S . Выразим p из второго уравнения (8.8) и подставим полученное значение в первое уравнение, в результате получим

$$\left[\lambda^2 Q - \lambda(A + \beta BC^{-1} B^*) - \alpha^{-1} BC^{-1} B^* \right] u = 0.$$

В обратную сторону: пусть $\lambda \neq 0$ и $u \in U \setminus \{0\}$ удовлетворяет (8.7). Введем вектор

$$v = \lambda Qu - (A + \beta BC^{-1} B^*)u.$$

Тогда из уравнения (8.7) следует

$$\lambda v - \alpha^{-1} BC^{-1} B^* u = 0.$$

Умножая скалярно обе части полученного равенства на произвольный вектор

$$w \in \ker B^*,$$

получаем $\lambda(v, w) = 0$, а так как $\lambda \neq 0$, то

$$v \in \text{Im } B.$$

Следовательно, существует $p \in P$ такой, что $v = Bp$. Несложно убедиться, что вектор $\{u, p\} \neq 0$ является собственным вектором S , соответствующим собственному значению λ . Лемма доказана. ■

8.2.2. Оптимизация в классе \mathbb{K}_1

Сначала получим оценку распределения спектра оператора S .

Теорема 8.2.1 (Оценка в классе \mathbb{K}_1). Пусть $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$, $(A, B, Q, C) \in \mathbb{K}_1 = \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$, $\alpha > 0$, $\beta \in \mathbb{R}$, тогда

$$\sigma(S) \subseteq [\delta, \Delta] \cup \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \frac{s + \beta ts \pm \sqrt{(s + \beta ts)^2 + 4ts/\alpha}}{2} \right\}.$$

Доказательство. Пусть выполнены условия теоремы, положим

$$L = A^{-1/2}QA^{-1/2}, \quad G = A^{-1/2}BC^{-1}B^*A^{-1/2}, \\ \delta_1 = \Delta^{-1}, \quad \delta_2 = \delta^{-1}, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 8.2.2, число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \neq 0$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) = \lambda s - 1, \quad g(\lambda, t) = \lambda, \quad h(\lambda) = -(\beta\lambda + \alpha^{-1}).$$

Так как оператор S самосопряжен относительно скалярного произведения $(\cdot, \cdot)_M$, то $\sigma(S) \subset \mathbb{R}$ и выполнены все условия следствия 6.4.1, а значит

$$\sigma(S) \subseteq \bigcup_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \left\{ \lambda_1(s, t), \lambda_2^{1,2}(s, t) \right\},$$

где $\lambda_1(s, t)$ — корень уравнения

$$\lambda s - 1 = 0,$$

а $\lambda_2^{1,2}(s, t)$ — все корни уравнения

$$(\lambda s - 1)\lambda - (\beta\lambda + \alpha^{-1})t = 0,$$

откуда после замены $s \rightarrow s^{-1}$ немедленно следует утверждение теоремы. ■

Структура приведенной оценки позволяет несложным образом оценить спектральное число обусловленности исследуемого оператора.

Следствие 8.2.1. Пусть выполнены условия теоремы 8.2.1, тогда

$$\text{cond}_2(S) \leq M_1(\alpha, \beta)/m_1(\alpha, \beta),$$

где

$$M_1(\alpha, \beta) = \max_{t=\gamma, \Gamma} \left\{ 2\Delta, \Delta|1 + \beta t| + \sqrt{\Delta^2(1 + \beta t)^2 + 4t\frac{\Delta}{\alpha}} \right\},$$

$$m_1(\alpha, \beta) = \min_{t=\gamma, \Gamma} \left\{ 2\delta, -\delta|1 + \beta t| + \sqrt{\delta^2(1 + \beta t)^2 + 4t\frac{\delta}{\alpha}} \right\}.$$

Кроме того, при $\beta \geq 0$ справедливо:

$$M_1(\alpha, \beta) = \Delta(1 + \beta\Gamma) + \sqrt{\Delta^2(1 + \beta\Gamma)^2 + 4\Gamma\frac{\Delta}{\alpha}},$$

$$m_1(\alpha, \beta) = \min \left\{ 2\delta, -\delta(1 + \beta\gamma) + \sqrt{\delta^2(1 + \beta\gamma)^2 + 4\gamma\frac{\delta}{\alpha}} \right\}.$$

Доказательство. Из утверждения теоремы 8.2.1 и равенств

$$\max_{\pm} \left| (s + \beta ts) \pm \sqrt{(s + \beta ts)^2 + 4t\frac{s}{\alpha}} \right| = |s + \beta ts| + \sqrt{(s + \beta ts)^2 + 4t\frac{s}{\alpha}},$$

$$\min_{\pm} \left| (s + \beta ts) \pm \sqrt{(s + \beta ts)^2 + 4t\frac{s}{\alpha}} \right| = -|s + \beta ts| + \sqrt{(s + \beta ts)^2 + 4t\frac{s}{\alpha}},$$

справедливых при любых $t, s > 0$, $\alpha > 0$, $\beta \in \mathbb{R}$, следует оценка

$$\text{cond}_2(S) \leq \tilde{M}_1(\alpha, \beta) / \tilde{m}_1(\alpha, \beta),$$

где

$$\tilde{m}_1(\alpha, \beta) = \min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{ 2\delta, \lambda^-(t, s) \},$$

$$\tilde{M}_1(\alpha, \beta) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{ 2\Delta, \lambda^+(t, s) \},$$

$$\lambda^{\pm}(t, s) = \pm s|1 + \beta t| + \sqrt{s^2(1 + \beta t)^2 + 4t\frac{s}{\alpha}}.$$

Из неравенств

$$\text{sign} \frac{\partial \lambda^{\pm}(t, s)}{\partial s} = \text{sign} (2t/\alpha \pm |1 + \beta t| \lambda^{\pm}(t, s)) > 0$$

следует, что при фиксированном $t \in [\gamma, \Gamma]$ справедливы равенства

$$\min_{s \in [\delta, \Delta]} \lambda^-(t, s) = \lambda^-(t, \delta), \quad \max_{s \in [\delta, \Delta]} \lambda^+(t, s) = \lambda^+(t, \Delta),$$

а значит, корректны определения величин

$$\tilde{m}_1(\alpha, \beta) = \min_{t \in [\gamma, \Gamma]} \{ 2\delta, \lambda^-(t, \delta) \}, \quad \tilde{M}_1(\alpha, \beta) = \max_{t \in [\gamma, \Gamma]} \{ 2\Delta, \lambda^+(t, \Delta) \}.$$

При $1 + \beta t \neq 0$ из представления

$$\operatorname{sign} \frac{\partial \lambda^-(t, \delta)}{\partial t} = \operatorname{sign} (2/\alpha - \beta \operatorname{sign}(1 + \beta t) \lambda^-(t, \delta))$$

следует, что функция $\lambda^-(t, \delta)$ параметра $t \in (0, +\infty)$ монотонно возрастает при $\beta \geq 0$, $t \in (0, +\infty)$, монотонно возрастает при $\beta < 0$, $t \in (0, -1/\beta)$ и монотонна (неубывает или невозрастает) при $\beta < 0$, $t \in (-1/\beta, +\infty)$. Таким образом, при $\beta \geq -1/\Gamma$ имеем

$$\tilde{m}_1(\alpha, \beta) = \min_{t \in [\gamma, \Gamma]} \{2\delta, \lambda^-(t, \delta)\} = \min \{2\delta, \lambda^-(\gamma, \delta)\},$$

а при $\beta < -1/\Gamma$ справедливо

$$\tilde{m}_1(\alpha, \beta) = \min_{t \in [\gamma, \Gamma]} \{2\delta, \lambda^-(t, \delta)\} = \min \{2\delta, \lambda^-(\gamma, \delta), \lambda^-(\Gamma, \delta)\},$$

откуда следует, что $\tilde{m}_1(\alpha, \beta) = m_1(\alpha, \beta)$.

Так как $\lambda^+(t, \Delta) \in F_{0,1} \subset Y(\mathbb{R})$ как функция переменной t , то в силу пункта 2 теоремы 6.4.10 имеем

$$\max_{t \in [\gamma, \Gamma]} \{2\Delta, \lambda^+(t, \Delta)\} = \max_{t=\gamma, \Gamma} \{2\Delta, \lambda^+(t, \Delta)\} = M_1(\alpha, \beta),$$

а из неравенств

$$\operatorname{sign} \frac{\partial \lambda^+(t, \Delta)}{\partial t} > 0, \quad \lambda^+(t, \Delta) \geq 2\Delta(1 + \beta t) \geq 2\Delta,$$

при $\beta \geq 0$ следует представление для $M_1(\alpha, \beta)$. Следствие доказано. ■

Случай $\beta = 0$ представляет отдельный интерес, так как связан с уменьшением вычислительной сложности одного шага алгоритма (не требуется вычисление слагаемого $\beta BC^{-1}B^*$ у). Имеет место

Теорема 8.2.2 (Случай $\beta = 0$). Пусть выполнены условия теоремы 8.2.1 и $\beta = 0$, тогда

$$\min_{\alpha > 0} \frac{M_1(\alpha, 0)}{m_1(\alpha, 0)} = \frac{M_1(\alpha_1, 0)}{m_1(\alpha_1, 0)} = \frac{1}{2\omega} \left(1 + \sqrt{1 + 8\frac{\omega}{\xi}} \right),$$

где $\alpha_1 = \gamma/(2\delta)$, $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. В силу следствия 8.2.1, имеет место равенство

$$\frac{M_1(\alpha, 0)}{m_1(\alpha, 0)} = \omega^{-1} \frac{1 + \sqrt{1 + 4\Gamma/(\Delta\alpha)}}{\min \left\{ 2, -1 + \sqrt{1 + 4\gamma/(\delta\alpha)} \right\}}.$$

При $\alpha < \alpha_1$ выполнено

$$\frac{M_1(\alpha, 0)}{m_1(\alpha, 0)} = \omega^{-1} \frac{1 + \sqrt{1 + 4\Gamma/(\Delta\alpha)}}{2} > \frac{M_1(\alpha_1, 0)}{m_1(\alpha_1, 0)}.$$

При $\alpha > \alpha_1$ имеет место представление

$$\begin{aligned} \frac{M_1(\alpha, 0)}{m_1(\alpha, 0)} &= \omega^{-1} \frac{1 + \sqrt{1 + 4\Gamma/(\Delta\alpha)}}{-1 + \sqrt{1 + 4\gamma/(\delta\alpha)}} = \\ &= \frac{\Delta}{4\gamma} \left(\sqrt{\alpha} + \sqrt{\alpha + \frac{4\Gamma}{\Delta}} \right) \left(\sqrt{\alpha} + \sqrt{\alpha + \frac{4\gamma}{\delta}} \right), \end{aligned}$$

откуда следует, что $M_1(\alpha, 0)/m_1(\alpha, 0)$ — возрастающая функция при $\alpha > \alpha_1$ и справедлива оценка

$$M_1(\alpha, 0)/m_1(\alpha, 0) > M_1(\alpha_1, 0)/m_1(\alpha_1, 0).$$

Теорема доказана. ■

Перейдем к анализу общей ситуации.

Теорема 8.2.3 (Общий случай). Пусть выполнены условия теоремы 8.2.1, тогда

$$\min_{\alpha > 0, \beta \in \mathbb{R}} \frac{M_1(\alpha, \beta)}{m_1(\alpha, \beta)} = \frac{M_1(\alpha_1, \beta_1)}{m_1(\alpha_1, \beta_1)} = O_1,$$

где

$$\begin{aligned} \alpha_1 &= \gamma/(\delta(2 - \xi)), \quad \beta_1 = -1/\Gamma, \\ O_1 &= \begin{cases} 1/\omega, & \omega < \xi/(2 - \xi) \\ \sqrt{(2 - \xi)/(\omega\xi)}, & \omega \geq \xi/(2 - \xi) \end{cases}, \\ \omega &= \delta/\Delta, \quad \xi = \gamma/\Gamma. \end{aligned}$$

Доказательство. В силу следствия 8.2.1, имеет место представление

$$f(\alpha, \beta) \equiv \omega \frac{M_1(\alpha, \beta)}{m_1(\alpha, \beta)} = \frac{\max \{2, \lambda^+(\alpha, \beta; \gamma, \Delta), \lambda^+(\alpha, \beta; \Gamma, \Delta)\}}{\min \{2, \lambda^-(\alpha, \beta; \gamma, \delta), \lambda^-(\alpha, \beta; \Gamma, \delta)\}},$$

где

$$\lambda^\pm(\alpha, \beta; t, s) = \pm|1 + \beta t| + \sqrt{(1 + \beta t)^2 + 4t/(s\alpha)}.$$

Пусть $\alpha > 0$, $\beta \notin \{-1/\gamma, -1/\Gamma\}$, $s \in \{\delta, \Delta\}$, $t \in \{\gamma, \Gamma\}$, тогда

$$\text{sign} \frac{\partial \lambda^\pm(\alpha, \beta; t, s)}{\partial \beta} = \pm \text{sign}(1 + \beta t).$$

Отсюда следует, что при $\beta < -1/\gamma$ справедливо

$$f(\alpha, \beta) > f(\alpha, -1/\gamma),$$

а при $\beta > -1/\Gamma$ —

$$f(\alpha, \beta) > f(\alpha, -1/\Gamma).$$

Таким образом, оптимальное значение β расположено на отрезке $[-1/\gamma, -1/\Gamma]$.

Предположим, что при оптимальном выборе $\alpha > 0$, $\beta \in [-1/\gamma, -1/\Gamma]$ имеет место представление

$$f(\alpha, \beta) = \frac{\max \{2, \lambda^+(\alpha, \beta; t^+, \Delta)\}}{\min \{2, \lambda^-(\alpha, \beta; t^-, \delta)\}},$$

$$t^+ \rightarrow \min_{t=\gamma, \Gamma} \lambda^+(\alpha, \beta; t, \Delta), \quad t^- \rightarrow \max_{t=\gamma, \Gamma} \lambda^-(\alpha, \beta; t, \Delta),$$

причем

$$\lambda^+(\alpha, \beta; t^+, \Delta) \neq 2, \quad \lambda^-(\alpha, \beta; t^-, \delta) \neq 2.$$

Рассмотрим возможные ситуации.

- 1) $\lambda^+(\alpha, \beta; t^+, \Delta) > 2$, $\lambda^-(\alpha, \beta; t^-, \delta) < 2$. В этом случае имеем

$$\text{sign} \frac{\partial}{\partial \alpha} f(\alpha, \beta) = \text{sign} \frac{\partial}{\partial \alpha} \frac{\lambda^+(\alpha, \beta; t^+, \Delta)}{\lambda^-(\alpha, \beta; t^-, \delta)} < 0,$$

т. е. $\alpha > 0$ не является оптимальным.

- 2) $\lambda^+(\alpha, \beta; t^+, \Delta) < 2$, $\lambda^-(\alpha, \beta; t^-, \delta) < 2$. В этом случае —

$$\text{sign} \frac{\partial}{\partial \alpha} f(\alpha, \beta) = \text{sign} \frac{\partial}{\partial \alpha} \frac{2}{\lambda^-(\alpha, \beta; t^-, \delta)} > 0,$$

т. е. $\alpha > 0$ не является оптимальным.

- 3) $\lambda^+(\alpha, \beta; t^+, \Delta) > 2$, $\lambda^-(\alpha, \beta; t^-, \delta) > 2$. В этом случае —

$$\text{sign} \frac{\partial}{\partial \alpha} f(\alpha, \beta) = \text{sign} \frac{\partial}{\partial \alpha} \frac{\lambda^+(\alpha, \beta; t^+, \Delta)}{2} < 0,$$

т. е. $\alpha > 0$ не является оптимальным.

- 4) $\lambda^+(\alpha, \beta; t^+, \Delta) < 2$, $\lambda^-(\alpha, \beta; t^-, \delta) > 2$. Тогда $f(\alpha, \beta) = 2$ в некоторой окрестности оптимальных параметров и в силу, например, неограниченного возрастания $\lambda^+(\alpha, \beta; t^+, \Delta)$ при $\alpha \rightarrow +0$, можно выбрать другой набор оптимальных параметров, при котором $\lambda^+(\alpha, \beta; t^+, \Delta) = 2$ или $\lambda^-(\alpha, \beta; t^-, \delta) = 2$.

Из приведенных рассуждений следует, что оптимальные параметры достаточно искать среди параметров, удовлетворяющих хотя бы одному из условий:

$$\lambda^+(\alpha, \beta; t^+, \Delta) = 2, \quad \lambda^-(\alpha, \beta; t^-, \delta) = 2.$$

Разрешая каждое из них при $\beta \in [-1/\gamma, -1/\Gamma]$, $\alpha > 0$, приходим к тому, что оптимальные параметры связаны одним из следующих уравнений:

$$\alpha \Delta = \frac{\Gamma}{2 + \beta \Gamma}, \quad \alpha \delta = \frac{\gamma}{2 + \beta \gamma}.$$

Предположим теперь, что при оптимальном выборе параметров выполняются условия

$$\alpha = \alpha(\beta) = \frac{\Gamma}{\Delta(2 + \beta \Gamma)} > 0, \quad \beta \in \left(-\frac{1}{\gamma}, -\frac{1}{\Gamma}\right),$$

тогда имеют место соотношения:

$$\lambda^+(\gamma, \Delta; \alpha, \beta) < \lambda^+(\Gamma, \Delta; \alpha, \beta) = 2, \quad \lambda^-(\gamma, \delta; \alpha, \beta) < 2, \quad \lambda^-(\Gamma, \delta; \alpha, \beta) < 2,$$

$$f(\alpha, \beta) = \frac{2}{\min\{2, \lambda^-(\gamma, \delta; \alpha, \beta), \lambda^-(\Gamma, \delta; \alpha, \beta)\}},$$

$$\frac{d\lambda^-(\gamma, \delta; \alpha(\beta), \beta)}{d\beta} > 0, \quad \frac{d\lambda^-(\Gamma, \delta; \alpha(\beta), \beta)}{d\beta} > 0,$$

откуда следует, что значение β не является строгим локальным минимумом функции $f(\alpha(\beta), \beta)$, что либо противоречит оптимальности β , либо позволяет выбрать другое оптимальное значение β из некоторой окрестности. Аналогичные выводы имеют место и при $\alpha = \gamma/(\delta(2 + \beta\gamma)) > 0$, $\beta \in (-1/\gamma, -1/\Gamma)$.

Резюмируя полученные условия, приходим к тому, что поиск оптимальных значений можно свести к следующим четырем вариантам:

$$\beta \in \{-1/\gamma, -1/\Gamma\}, \quad \alpha \in \{\Gamma/(\Delta(2 + \beta\Gamma)), \gamma/(\delta(2 + \beta\gamma))\},$$

что немедленно приводит к искомым значениям $\beta = -1/\Gamma$, $\alpha = \gamma/(\delta(2 + \beta\gamma))$. Теорема доказана. ■

Отметим, что полученные в теореме значения оптимальных параметров в общем случае не являются единственными. Более подробный анализ доказательства позволяет сделать вывод о том, что единственность имеет место при выполнении неравенства

$$\omega \geq \frac{\xi}{2 - \xi}.$$

Следствие 8.2.2. Пусть выполнены условия теоремы 8.2.1, тогда для оптимального показателя скорости сходимости алгоритма GMLan справедливы асимптотические оценки

$$q_{K_1} \leq 1 - \sqrt{\frac{\omega}{2}} \sqrt{\xi} + O(\xi), \quad \omega = \text{const}, \quad \xi \rightarrow 0,$$

$$q_{K_1} \leq 1 - \omega + O(\omega^2), \quad \xi = \text{const}, \quad \omega \rightarrow 0,$$

где $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. Формулы немедленно следуют из теоремы 8.2.3, неравенства

$$q_{K_1} \leq \sqrt{\frac{O_1 - 1}{O_1 + 1}} < 1 - O_1^{-1} + O_1^{-2}/2, \quad O_1 \rightarrow +\infty$$

и разложения $O_1 = O_1(\omega, \xi)$ в ряды Тейлора по ξ и ω соответственно. Следствие доказано. ■

8.2.3. Оптимизация в классе \mathbb{K}_2

Действуя по аналогии с исследованием в классе \mathbb{K}_1 , получим сначала оценку распределения спектра предобусловленного оператора.

Теорема 8.2.4 (Оценка в классе \mathbb{K}_2). Пусть $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$, $(A, B, Q, C) \in \mathbb{K}_2 = \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, $\alpha > 0$, $\beta \in \mathbb{R}$, тогда

$$\sigma(S) \subseteq [\delta, \Delta] \cup \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \left(s + \beta t \pm \sqrt{(s + \beta t)^2 + 4t/\alpha} \right) / 2 \right\}.$$

Доказательство. Пусть выполнены условия теоремы, положим

$$L = Q^{-1/2} A Q^{-1/2}, \quad G = Q^{-1/2} B C^{-1} B^* Q^{-1/2}, \\ \delta_1 = \delta, \quad \delta_2 = \Delta, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 8.2.2, число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \neq 0$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) = \lambda - s, \quad g(\lambda, t) = \lambda, \quad h(\lambda) = -(\beta\lambda + \alpha^{-1}).$$

Так как оператор S самосопряжен относительно скалярного произведения $(\cdot, \cdot)_M$, то $\sigma(S) \subset \mathbb{R}$ и значит выполнены все условия следствия 6.4.1, а тогда

$$\sigma(S) \subseteq \bigcup_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \left\{ \lambda_1(s, t), \lambda_2^{1,2}(s, t) \right\},$$

где $\lambda_1(s, t)$ — корень уравнения

$$\lambda - s = 0,$$

а $\lambda_2^{1,2}(s, t)$ — все корни уравнения

$$(\lambda - s)\lambda - (\beta\lambda + \alpha^{-1})t = 0,$$

откуда немедленно следует утверждение теоремы. ■

Прежде, чем перейти к решению задачи оптимизации, оценим спектральное число обусловленности исследуемого оператора.

Следствие 8.2.3. Пусть выполнены условия теоремы 8.2.4. тогда

$$\text{cond}_2(S) \leq \frac{M_2(\alpha, \beta)}{m_2(\alpha, \beta)}.$$

где

$$M_2(\alpha, \beta) = \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \left\{ 2\Delta, |s + \beta t| + \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}} \right\},$$

$$m_2(\alpha, \beta) = \min_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \left\{ 2\delta, -|s + \beta t| + \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}} \right\}.$$

Кроме того, при $\beta \geq 0$ имеют место равенства

$$M_2(\alpha, \beta) = |\Delta + \beta\Gamma| + \sqrt{(\Delta + \beta\Gamma)^2 + \frac{4\Gamma}{\alpha}},$$

$$m_2(\alpha, \beta) = \min \left\{ 2\delta, -|\Delta + \beta\gamma| + \sqrt{(\Delta + \beta\gamma)^2 + \frac{4\gamma}{\alpha}} \right\}.$$

Доказательство. Из утверждения теоремы 8.2.4 и равенств

$$\max_{\pm} \left| (s + \beta t) \pm \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}} \right| = |s + \beta t| + \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}},$$

$$\min_{\pm} \left| (s + \beta t) \pm \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}} \right| = -|s + \beta t| + \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}},$$

справедливых при любых $t, s > 0, \alpha > 0, \beta \in \mathbb{R}$, следует оценка

$$\text{cond}_2(S) \leq \frac{\tilde{M}_1(\alpha, \beta)}{\tilde{m}_1(\alpha, \beta)},$$

где

$$\tilde{m}_2(\alpha, \beta) = \min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{2\delta, \lambda^-(t, s)\}, \quad \tilde{M}_2(\alpha, \beta) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{2\Delta, \lambda^+(t, s)\},$$

$$\lambda^{\pm}(t, s) = \pm |s + \beta t| + \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}}.$$

Пусть $s + \beta t \neq 0$, тогда

$$\text{sign} \frac{\partial \lambda^{\pm}(t, s)}{\partial s} = \pm \text{sign}(s + \beta t),$$

откуда следует, что при фиксированном $t \in [\gamma, \Gamma]$ справедливо

$$\min_{s \in [\delta, \Delta]} \lambda^-(t, s) = \min_{s=\delta, \Delta} \lambda^-(t, s), \quad \max_{s \in [\delta, \Delta]} \lambda^+(t, s) = \max_{s=\delta, \Delta} \lambda^+(t, s),$$

а значит, корректны определения

$$\tilde{m}_2(\alpha, \beta) = \min_{t \in [\gamma, \Gamma]} \{2\delta, \lambda^-(t, \delta), \lambda^-(t, \Delta)\},$$

$$\tilde{M}_2(\alpha, \beta) = \max_{t \in [\gamma, \Gamma]} \{2\Delta, \lambda^+(t, \delta), \lambda^+(t, \Delta)\}.$$

причем при $\beta \geq -\delta/\Gamma$ имеет место неравенство $s + \beta t \geq 0$, поэтому

$$\begin{aligned}\tilde{m}_2(\alpha, \beta) &= \min_{t \in [\gamma, \Gamma]} \{2\delta, \lambda^-(t, \Delta)\}, \\ \tilde{M}_2(\alpha, \beta) &= \max_{t \in [\gamma, \Gamma]} \{2\Delta, \lambda^+(t, \Delta)\}.\end{aligned}$$

При фиксированном $s \in \{\delta, \Delta\}$ и $s + \beta t \neq 0$ из представления

$$\text{sign} \frac{\partial \lambda^-(t, s)}{\partial t} = \text{sign} \left(\frac{2}{\alpha} - \beta \text{sign}(s + \beta t) \lambda^-(t, s) \right)$$

следует, что функция $\lambda^-(t, s)$ параметра $t \in (0, +\infty)$ монотонно возрастает при $\beta \geq 0$, $t \in (0, +\infty)$, монотонно возрастает при $\beta < 0$, $t \in (0, -s/\beta)$ и монотонна (неубывает или невозрастает) при $\beta < 0$, $t \in (-s/\beta, +\infty)$. Таким образом, при $\beta \geq -\delta/\Gamma$ имеем

$$\tilde{m}_2(\alpha, \beta) = \min_{t \in [\gamma, \Gamma]} \{2\delta, \lambda^-(t, \delta)\} = \min \{2\delta, \lambda^-(\gamma, \delta)\},$$

а при $\beta < -\delta/\Gamma$ справедливо

$$\tilde{m}_2(\alpha, \beta) = \min_{t \in [\gamma, \Gamma]} \{2\delta, \lambda^-(t, \delta), \lambda^-(t, \Delta)\} = \min_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \{2\delta, \lambda^-(t, s)\},$$

откуда следует, что $\tilde{m}_2(\alpha, \beta) = m_2(\alpha, \beta)$.

Так как $\lambda^+(t, s) \in F_{0,1} \subset Y(\mathbb{R})$ как функция переменной t при фиксированном s , то, в силу пункта 2 теоремы 6.4.10, получаем

$$\max_{t \in [\gamma, \Gamma]} \{2\Delta, \lambda^+(t, \delta), \lambda^+(t, \Delta)\} = \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \{2\Delta, \lambda^+(t, s)\} = M_2(\alpha, \beta),$$

а из неравенств

$$\text{sign} \frac{\partial \lambda^+(t, \Delta)}{\partial t} > 0, \quad \lambda^+(t, \Delta) \geq 2(\Delta + \beta t) \geq 2\Delta,$$

при $\beta \geq 0$ следует представление для $M_2(\alpha, \beta)$. Следствие доказано. \blacksquare

Случай $\beta = 0$, как упоминалось ранее, представляет отдельный интерес. Справедлива

Теорема 8.2.5 (Случай $\beta = 0$). Пусть выполнены условия теоремы 8.2.4 и $\beta = 0$, тогда

$$\min_{\alpha > 0} \frac{M_2(\alpha, 0)}{m_2(\alpha, 0)} = \frac{M_2(\alpha_2, 0)}{m_2(\alpha_2, 0)} = \frac{1}{2\omega} \left(1 + \sqrt{1 + 4\omega \frac{1 + \omega}{\xi}} \right),$$

где $\alpha_2 = \gamma/(\delta(\delta + \Delta))$, $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. В силу следствия 8.2.3, имеет место равенство

$$\frac{M_2(\alpha, 0)}{m_2(\alpha, 0)} = \frac{1 + \sqrt{1 + 4\Gamma/(\Delta^2\alpha)}}{\min \left\{ 2\omega, -1 + \sqrt{1 + 4\gamma/(\Delta^2\alpha)} \right\}}.$$

При $\alpha < \alpha_2$ справедливо

$$\frac{M_2(\alpha, 0)}{m_2(\alpha, 0)} = \frac{1 + \sqrt{1 + 4\Gamma/(\Delta^2\alpha)}}{2\omega} > \frac{M_2(\alpha_2, 0)}{m_2(\alpha_2, 0)}.$$

При $\alpha > \alpha_2$ имеет место представление

$$\begin{aligned} \frac{M_2(\alpha, 0)}{m_2(\alpha, 0)} &= \frac{1 + \sqrt{1 + 4\Gamma/(\Delta^2\alpha)}}{-1 + \sqrt{1 + 4\gamma/(\Delta^2\alpha)}} = \\ &= \frac{\Delta^2}{4\gamma} \left(\sqrt{\alpha} + \sqrt{\alpha + \frac{4\Gamma}{\Delta^2}} \right) \left(\sqrt{\alpha} + \sqrt{\alpha + \frac{4\gamma}{\Delta^2}} \right), \end{aligned}$$

откуда следует, что $M_2(\alpha, 0)/m_2(\alpha, 0)$ — возрастающая функция при $\alpha > \alpha_2$ и справедлива оценка

$$\frac{M_2(\alpha, 0)}{m_2(\alpha, 0)} > \frac{M_2(\alpha_2, 0)}{m_1(\alpha_2, 0)}.$$

Теорема доказана. ■

Перейдем к изучению общего случая.

Теорема 8.2.6 (Общий случай). Пусть выполнены условия теоремы 8.2.4, тогда

$$\min_{\substack{\alpha > 0, \\ \beta \in \mathbb{R}}} \frac{M_2(\alpha, \beta)}{m_2(\alpha, \beta)} = \frac{M_2(\alpha_2, \beta_2)}{m_2(\alpha_2, \beta_2)} = O_2,$$

где

$$\begin{aligned} \alpha_2 &= \min \left\{ \frac{2\gamma}{\delta(\Delta + \delta)(2 - \xi)}, \frac{2\Gamma}{\Delta(\Delta + \delta)} \right\}, \quad \beta_2 = -\frac{\delta + \Delta}{2\Gamma}, \\ O_2 &= \begin{cases} \frac{1}{\omega} & \text{при } \omega < \frac{\xi}{2 - \xi}, \\ \frac{1}{4\omega} \left(1 - \omega + \sqrt{(1 - \omega)^2 + 8\omega(1 + \omega) \left(\frac{2}{\xi} - 1 \right)} \right) & \text{при } \omega \geq \frac{\xi}{2 - \xi}, \end{cases} \\ \omega &= \frac{\delta}{\Delta}, \quad \xi = \frac{\gamma}{\Gamma}. \end{aligned}$$

Доказательство. В силу следствия 8.2.3, имеет место представление

$$\begin{aligned} f(\alpha, \beta) &\equiv \frac{M_2(\alpha, \beta)}{m_2(\alpha, \beta)} = \frac{\max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \{2\Delta, \lambda^+(\alpha, \beta; t, s)\}}{\min_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \{2\delta, \lambda^-(\alpha, \beta; t, s)\}} = \\ &= \frac{\max \{2\Delta, \lambda^+(\alpha, \beta; t^+, s^+)\}}{\min \{2\delta, \lambda^-(\alpha, \beta; t^-, s^-)\}}, \end{aligned}$$

где

$$\lambda^\pm(\alpha, \beta; t, s) = \pm|s + \beta t| + \sqrt{(s + \beta t)^2 + \frac{4t}{\alpha}},$$

а значения $t^\pm = t^\pm(\alpha, \beta) \in \{\gamma, \Gamma\}$, $s^\pm = s^\pm(\alpha, \beta) \in \{\delta, \Delta\}$ определяются из условий

$$\begin{aligned} \lambda^+(\alpha, \beta; t^+, s^+) &= \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \lambda^+(\alpha, \beta; t, s), \\ \lambda^-(\alpha, \beta; t^-, s^-) &= \min_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \lambda^-(\alpha, \beta; t, s). \end{aligned}$$

Пусть $\alpha > 0$, $\beta \notin \{-\delta/\gamma, -\delta/\Gamma, -\Delta/\gamma, -\Delta/\Gamma\}$, $s \in \{\delta, \Delta\}$, $t \in \{\gamma, \Gamma\}$, тогда

$$\text{sign} \frac{\partial \lambda^\pm(\alpha, \beta; t, s)}{\partial \beta} = \pm \text{sign}(s + \beta t).$$

Отсюда следует, что при $\beta < -\Delta/\gamma$ справедливо

$$f(\alpha, \beta) > f\left(\alpha, -\frac{\Delta}{\gamma}\right),$$

а при $\beta > -\delta/\Gamma$ —

$$f(\alpha, \beta) > f\left(\alpha, -\frac{\delta}{\Gamma}\right).$$

Таким образом, оптимальное значение β расположено на отрезке $[-\Delta/\gamma, -\delta/\Gamma]$.

Предположим, что

$$\lambda^+(\alpha, \beta; t^+, s^+) \neq 2\Delta, \quad \lambda^-(\alpha, \beta; t^-, s^-) \neq 2\delta$$

и рассмотрим возможные ситуации:

- 1) $\lambda^+(\alpha, \beta; t^+, s^+) > 2\Delta$, $\lambda^-(\alpha, \beta; t^-, s^-) < 2\delta$. В этом случае имеем

$$\begin{aligned} \text{sign} \frac{\partial}{\partial \alpha} f(\alpha, \beta) &= \text{sign} \frac{\partial}{\partial \alpha} \frac{\lambda^+(\alpha, \beta; t_1, s_1)}{\lambda^-(\alpha, \beta; t_2, s_2)} < 0, \\ t_{1,2} &= \gamma, \Gamma. \quad s_{1,2} = \delta, \Delta. \end{aligned}$$

т. е. $\alpha > 0$ не является оптимальным.

2) $\lambda^+(\alpha, \beta; t^+, s^+) < 2\Delta$, $\lambda^-(\alpha, \beta; t^-, s^-) < 2\delta$. В этом случае —

$$\text{sign } \frac{\partial}{\partial \alpha} f(\alpha, \beta) = \text{sign } \frac{\partial}{\partial \alpha} \frac{2\Delta}{\lambda^-(\alpha, \beta; t, s)} > 0,$$

$$t = \gamma, \Gamma, \quad s = \delta, \Delta,$$

т. е. $\alpha > 0$ не является оптимальным.

3) $\lambda^+(\alpha, \beta; t^+, s^+) > 2\Delta$, $\lambda^-(\alpha, \beta; t^-, s^+) > 2\delta$. В этом случае —

$$\text{sign } \frac{\partial}{\partial \alpha} f(\alpha, \beta) = \text{sign } \frac{\partial}{\partial \alpha} \frac{\lambda^+(\alpha, \beta; t^+, s^+)}{2\delta} < 0,$$

$$t = \gamma, \Gamma, \quad s = \delta, \Delta,$$

т. е. $\alpha > 0$ не является оптимальным.

4) $\lambda^+(\alpha, \beta; t^+, s^+) < 2\Delta$, $\lambda^-(\alpha, \beta; t^-, s^-) > 2\delta$. Тогда

$$f(\alpha, \beta) = \frac{\Delta}{\delta}$$

в некоторой окрестности оптимальных параметров и, в силу справедливости неравенств

$$\frac{\partial}{\partial \alpha} \lambda^\pm(\alpha, \beta; t, s) < 0 \text{ при } t = \gamma, \Gamma, \quad s = \delta, \Delta,$$

можно выбрать оптимальные параметры так, что будет выполнено

$$\lambda^+(\alpha, \beta; t^+, s^+) = 2\Delta \text{ или } \lambda^-(\alpha, \beta; t^-, s^-) = 2\delta.$$

Из приведенных рассуждений следует, что оптимальные параметры достаточно искать среди таких, которые удовлетворяют хотя бы одному из условий:

$$\lambda^+(\alpha, \beta; t^+, s^+) = 2\Delta, \quad \lambda^-(\alpha, \beta; t^-, s^-) = 2\delta.$$

В дальнейшем будем считать, что $\delta < \Delta$ и оптимальные параметры $\alpha_0 > 0$, $\beta_0 \in [-\Delta/\gamma, -\delta/\Gamma]$ выбраны таким образом, что для любых оптимальных $\alpha > 0$, $\beta \in \mathbb{R}$, $f(\alpha, \beta) = f(\alpha_0, \beta_0)$ имеет место неравенство

$$\frac{\lambda^+(\alpha_0, \beta_0; t_0^+, s_0^+)}{\lambda^-(\alpha_0, \beta_0; t_0^-, s_0^-)} \geq \frac{\lambda^+(\alpha, \beta; t^+, s^+)}{\lambda^-(\alpha, \beta; t^-, s^-)},$$

где $t_0^\pm = t^\pm(\alpha_0, \beta_0)$, $s_0^\pm = s^\pm(\alpha_0, \beta_0)$.

Предположим, что

$$|\delta + \beta_0 t_0^\pm| \neq |\Delta + \beta_0 t_0^\pm|$$

и найдутся $t_0 \in \{\gamma, \Gamma\}$, $s_0 \in \{\delta, \Delta\}$ такие, что $(t_0^+, s_0^+) \neq (t_0, s_0)$ и

$$\lambda^+(\alpha_0, \beta_0; t_0^+, s_0^+) = \lambda^+(\alpha_0, \beta_0; t_0, s_0) = 2\Delta,$$

тогда $t_0 \neq t_0^+$ и

$$\text{sign}(s_0^+ + \beta t_0^+) \neq \text{sign}(s_0 + \beta t_0),$$

откуда следует, что $s_0^+ \neq s_0$. Таким образом, не ограничивая общности, можно считать, что $s_0^+ = \delta$, $s_0 = \Delta$. При этом из

$$\lambda(\alpha_0, \beta_0; t_0, \Delta) = 2\Delta$$

следует условие

$$\alpha\beta\Delta = -1,$$

а из

$$\lambda^+(\alpha_0, \beta_0; t_0^+, \delta) = 2\Delta$$

в этом случае — равенство

$$|\Delta + \beta_0 t_0^+| = |\delta + \beta_0 t_0|,$$

что противоречит исходному предположению.

Предположим теперь, что

$$|\delta + \beta_0 t_0^\pm| \neq |\Delta + \beta_0 t_0^\pm|$$

и для любых $t = \gamma, \Gamma$, $s = \delta, \Delta$, $(t, s) \neq (t_0^+, s_0^+)$ имеют место соотношения

$$\lambda^+(\alpha_0, \beta_0; t_0^+, s_0^+) = 2\Delta > \lambda^+(\alpha_0, \beta_0; t, s),$$

тогда функции t^+ и s^+ постоянны в некоторой окрестности α_0 , β_0 и можно выбрать дифференцируемую в некоторой окрестности α_0 функцию g такую, что $\beta = g(\alpha)$, $\beta_0 = g(\alpha_0)$ так, чтобы добиться локального выполнения равенства

$$\lambda^+(\alpha, g(\alpha); t^+, s^+) = \lambda^+(\alpha, g(\alpha); t_0^+, s_0^+) = 2\Delta.$$

Отсюда и из условия выбора α_0 и β_0 следует, что

$$\left. \frac{d\lambda^-(\alpha, g(\alpha); t^-, s^-)}{d\alpha} \right|_{\alpha=\alpha_0} = 0, \quad \left. \frac{d\lambda^+(\alpha, g(\alpha); t^+, s^+)}{d\alpha} \right|_{\alpha=\alpha_0} = 0.$$

Раскрыв их, приходим к равенствам

$$\begin{aligned} \frac{\partial g}{\partial \alpha}(\alpha_0) &= \frac{-2}{\text{sign}(s_0^- + \beta_0 t_0^-) \alpha_0^2 \lambda^-(\alpha_0, \beta_0; t_0^-, s_0^-)} = \\ &= \frac{2}{\text{sign}(s_0^+ + \beta_0 t_0^+) \alpha_0^2 \lambda^+(\alpha_0, \beta_0; t_0^+, s_0^+)}, \end{aligned}$$

что противоречит условию

$$\lambda^-(\alpha_0, \beta_0; t_0^-, s_0^-) < \lambda^+(\alpha_0, \beta_0; t_0^+, s_0^+).$$

После аналогичного рассмотрения случая

$$\lambda^-(\alpha_0, \beta_0; t^-, s^-) = 2\delta, \quad |\delta + \beta_0 t^\pm| \neq |\Delta + \beta_0 t^\pm|.$$

получаем, что для некоторого $t \in \{\gamma, \Gamma\}$ имеет место равенство

$$|\delta + \beta_0 t| = |\Delta + \beta_0 t|,$$

откуда следует

$$\beta_0 \in \left\{ -\frac{\delta + \Delta}{2\gamma}, -\frac{\delta + \Delta}{2\Gamma} \right\}.$$

Резюмируя полученные условия, приходим к тому, что поиск оптимальных значений можно свести к следующим четырем вариантам: либо

$$\beta = -(\delta + \Delta)/(2\gamma), \quad \alpha \in \left\{ \frac{2\Gamma}{\Delta(\Delta + \delta)(2 - 1/\xi)}, \frac{2\gamma}{\delta(\Delta + \delta)} \right\}$$

либо

$$\beta = -(\delta + \Delta)/(2\Gamma), \quad \alpha \in \left\{ \frac{2\gamma}{\delta(\Delta + \delta)(2 - \xi)}, \frac{2\Gamma}{\Delta(\Delta + \delta)} \right\}.$$

Это, в свою очередь, приводит к искомым значениям α_2, β_2 . Теорема доказана. ■

Отметим, что полученные в теореме значения оптимальных параметров в общем случае *не являются единственными*. Более подробный анализ доказательства позволяет сделать вывод о том, что единственность имеет место при условии $\omega \geq \xi/(2 - \xi)$.

Следствие 8.2.4. Пусть выполнены условия теоремы 8.2.3, тогда для оптимального показателя скорости сходимости алгоритма GMLan справедливы асимптотические оценки

$$q_{K_2} \leq 1 - \sqrt{\frac{\omega}{1 + \omega}} \sqrt{\xi} + O(\xi), \quad \omega = \text{const}, \quad \xi \rightarrow 0,$$

$$q_{K_2} \leq 1 - \omega + O(\omega^2), \quad \xi = \text{const}, \quad \omega \rightarrow 0,$$

где $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. Формулы немедленно следуют из теоремы 8.2.6, неравенства

$$q_{K_2} \leq \sqrt{\frac{O_2 - 1}{O_2 + 1}} < 1 - O_2^{-1} + \frac{O_2^{-2}}{2},$$

где $O_2 \rightarrow +\infty$, и разложения $O_2 = O_2(\omega, \xi)$ в ряды Тейлора по ξ и ω соответственно. Следствие доказано. ■

8.2.4. Оценка в классе \mathbb{K}_{2s}

Используя подход, описанный в 6.3, применим ранее полученные результаты для изучения алгоритма в классе \mathbb{K}_{2s} .

Теорема 8.2.7 (Оценка в классе \mathbb{K}_{2s}). Пусть $\omega = \delta/\Delta \leq 1$, $\xi = \gamma/\Gamma \leq 1$, $(A, B, Q, C) \in \mathbb{K}_{2s} = \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma)$, тогда

$$\text{cond}_2 S(\alpha_{2s}, \beta_{2s}; A, B, Q, C) \leq O_{2s},$$

где

$$\alpha_{2s} = \min \left\{ \frac{2\gamma}{\delta(\Delta_s + \delta)(2 - \xi)}, \frac{2\Gamma}{\Delta_s(\Delta_s + \delta)} \right\}, \quad \beta_{2s} = \frac{\delta}{\gamma} - \frac{\delta + \Delta_s}{2\Gamma},$$

$$O_{2s} = \begin{cases} \frac{1}{\omega_s} & \text{при } \omega_s < \xi/(2 - \xi), \\ \frac{1}{4\omega_s} \left(1 - \omega_s + \sqrt{(1 - \omega_s)^2 + 8\omega_s(1 + \omega_s)\left(\frac{2}{\xi} - 1\right)} \right) & \\ \text{при } \omega_s \geq \frac{\xi}{2 - \xi}, \end{cases}$$

$$\Delta_s = \delta(\omega^{-1} + \xi^{-1}), \quad \omega_s = \frac{\delta}{\Delta_s}.$$

Доказательство. В силу (6.17), справедливо

$$(A + \nu BC^{-1}B^*, B, Q, C) \in \mathbb{K}_2(\delta(\nu), \Delta(\nu), \gamma, \Gamma),$$

где $\nu > 0$, $\delta(\nu) = \min\{\delta, \nu\gamma\}$, $\Delta(\nu) = \Delta + \nu\Gamma$. По теореме 8.2.6 имеет место оценка

$$O_{2s} \leq O_2(\nu),$$

где $O_2(\nu)$ — оптимальная оценка в классе $\mathbb{K}_2(\delta(\nu), \Delta(\nu), \gamma, \Gamma)$.

Так как $O_2 = O_2(\omega, \xi)$ монотонно зависит от $\omega \in (0, 1]$ при фиксированном $\xi \in (0, 1]$, то $O_2(\nu)$ принимает наименьшее значение одновременно с величиной $\delta(\nu)/\Delta(\nu)$, т. е. при $\nu = \delta/\gamma$. Таким образом, $O_{2s} \leq O_2$ при $\alpha = \alpha_{2s}$, $\beta = \beta_{2s}$. Теорема доказана. ■

Следствие 8.2.5. Пусть выполнены условия теоремы 8.2.7, тогда для оптимального показателя скорости сходимости алгоритма GMLan справедливы асимптотические оценки

$$q_{\mathbb{K}_{2s}} \leq 1 - \frac{4}{1 + \sqrt{17}}\xi + O(\xi^2) \quad \text{при } \xi \rightarrow 0, \omega = \text{const},$$

$$q_{\mathbb{K}_{2s}} \leq 1 - \omega + O(\omega^2) \quad \text{при } \omega \rightarrow 0, \xi = \text{const}.$$

Доказательство. Формулы немедленно следуют из теоремы 8.2.7, неравенства

$$q_{\kappa_{2s}} \leq \sqrt{\frac{O_{2s} - 1}{O_{2s} + 1}} < 1 - O_{2s}^{-1} + O_{2s}^{-2}/2,$$

где $O_{2s} \rightarrow +\infty$, и разложения $O_{2s} = O_{2s}(\omega, \xi)$ в ряды Тейлора по ξ и ω соответственно. Следствие доказано. ■

8.3. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Напомним происхождение асимптотики скорости сходимости метода Ланцоша (другое название — метод минимальных итераций). Если спектр предобусловленного оператора S принадлежит двум отрезкам

$$\sigma(S) \subseteq [c_1, c_2] \cup [c_3, c_4],$$

где

$$c_1 \leq c_2 < 0 < c_3 \leq c_4,$$

то для нормы невязки на k -й итерации справедлива оценка [30]

$$\|r^k\|_M \leq \left(T_{[\frac{k}{2}]} \left(\frac{M_1^2 + m_1^2}{M_1^2 - m_1^2} \right) \right)^{-1} \|r^0\|_M,$$

где

$$m_1 = \min \{-c_2, c_3\}, \quad M_1 = \max \{-c_1, c_4\}, \quad (8.9)$$

что эквивалентно асимптотическому убыванию невязки со скоростью геометрической прогрессии с показателем q :

$$q = \sqrt{\frac{M_1 - m_1}{M_1 + m_1}} = \sqrt{\frac{\text{cond}_2(S) - 1}{\text{cond}_2(S) + 1}}.$$

Хорошо известно, что такая же асимптотика скорости сходимости соответствует методу сопряженных градиентов для нормальных уравнений, полученных из предобусловленной системы $S^*Sz = S^*F$. В отечественной литературе для алгебраических систем с седловой точкой этот подход встречается под названием «методы, основанные на сочетании идей симметризации и предобусловливания» [40, 41]. В работе [169] отмечается, что на практике эффективность алгоритмов типа Ланцоша заметно выше, чем у метода сопряженных градиентов для нормальной системы, так как спектр матрицы S , как правило, расположен несимметрично относительно начала координат. Поэтому для симметричных задач рассматриваемого вида использование нормальных уравнений представляется малооправданным.

В любом случае повышение эффективности алгоритмов этого класса связано с минимизацией спектрального числа обусловленности оператора S .

Отметим, что принципиальное различие в подходах оптимизации (при различных знаках α) обусловлено появлением комплекснозначного спектра у предобусловленного оператора при отрицательных α . Комплекснозначный спектр значительно усложняет анализ. Например, в работе [11] было получено, что при $\beta = 0$, $\alpha = -1$ спектр предобусловленного оператора S принадлежит на комплексной плоскости пересечению трех прямоугольных областей, из которого удалено объединение еще двух. В результате полученное выражение для оптимального значения параметра τ имеет совершенно неконструктивный характер, а выразить величину спектрального радиуса в этом случае через ξ и ω чрезвычайно затруднительно.

Тем не менее в работе [94] при использовании предположения об инвариантности подпространств была решена в традиционной постановке (класс $\mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$) задача асимптотической оптимизации метода GMJOR, т. е. получено точное выражение для $q_{\mathbb{K}_1}$ как при $\beta = 0$, так и в общем случае. При $\beta = 0$ и $\Delta = \Gamma = 1$ (что не ограничивает общности результата) величина спектрального радиуса q равна $q_0 = \tau_0 - 1$, где τ_0 является взятым со знаком «+» корнем квадратного уравнения

$$(2\omega + \xi)\tau^2 - (6\omega + 2\xi - \omega\xi)\tau + 4\omega = 0.$$

Перейдем к случаю вариационных подходов, т. е. $\alpha > 0$. В работе [165] для $\beta = 0$, $\alpha = 1$ были получены величины M_1 и m_1 из (8.9), определяющие скорость сходимости метода Ланцоша:

$$M_1 = \frac{\Delta}{2} + \sqrt{\frac{\Delta^2}{4} + \Gamma\Delta};$$

$$m_1 = \begin{cases} \delta & \text{при } \gamma \geq 2\delta, \\ \sqrt{\frac{\delta^2}{4} + \gamma\delta} - \frac{\delta}{2} & \text{при } \gamma < 2\delta. \end{cases}$$

В доказательстве имелись неточности, которые были исправлены, в [18]. Там же эти результаты были обобщены на случай системы с матрицей L_ϵ (см. главу 4).

Примерно в то же самое время в работах [187, 198] были получены аналитические оценки применительно к предобусловленному методу сопряженных невязок PCR [109] (другое название — MINRES). асимптотика скорости сходимости которого совпадает с методом Ланцоша. Еще одной работой из этого цикла следует считать [169],

в которой изучалась регуляризация матрицы L_0 матрицей L_ϵ в рамках метода PCR.

Характерной особенностью указанных выше работ являлось отсутствие свободных параметров ($\beta = 0, \alpha = 1$), что не позволяло увеличить скорость сходимости методов. Они появились в работе [194], однако сложность используемой техники для оценок границ спектра не позволила получить минимум числа обусловленности.

Результаты, предшествующие приведенным в настоящей главе, имеются в [97], где проводится минимизация как относительно α , так и относительно β , в предположении об инвариантности подпространств (частный случай оптимизации в классе \mathbb{K}_1). Введем обозначение $O_1 = 1/\omega$ и $O_2 = 1/\xi$. Формулу для числа обусловленности можно записать в виде

$$\text{cond}_2(S) = \begin{cases} \sqrt{O_1(2K_2 - 1)} & \text{при } O_2 > (O_1 + 1)/2, \\ O_1 & \text{при } O_2 \leq (O_1 + 1)/2. \end{cases}$$

Отсюда видно, что величины O_1 и O_2 асимптотически влияют различным образом на эффективность алгоритмов. В первую очередь нужно стремиться к уменьшению O_1 . Далее, при фиксированном O_1 наилучшей является ситуация при $O_2 \leq (O_1 + 1)/2$, поскольку здесь и число обусловленности принимает наименьшее из всех возможных значение, и оптимальные параметры имеют простейший вид (например, $\alpha = 1$). В этой же работе рассмотрен частный случай оптимизации в классе \mathbb{K}_2 .

СИММЕТРИЗАЦИЯ СПЕЦИАЛЬНОГО ВИДА

В теории обобщенных методов решения блочных седловых задач важную роль играет *симметризация специального вида*, или, что то же самое, переформулировка исходной постановки к симметричной положительно определенной операторной форме относительно специально подобранного скалярного произведения.

Изложим базовую идею подхода. Пусть решается седловая задача (6.2) и в неравенстве $\delta Q \leq A \leq \Delta Q$ постоянная $\delta > 1$, тогда справедливо

$$0 < A - Q \leq \kappa A, \quad \kappa = \frac{\Delta - 1}{\Delta}.$$

Это дает возможность преобразовать исходную постановку к равносильному виду

$$Sz \equiv \begin{pmatrix} Q^{-1}A & Q^{-1}B \\ B^*Q^{-1}(A - Q) & B^*Q^{-1}B \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} Q^{-1}f \\ B^*Q^{-1}f - \varphi \end{pmatrix} \equiv \tilde{F} \quad (9.1)$$

и в пространстве $Z = U \times P$ ввести специальное скалярное произведение

$$\left[\begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ r \end{pmatrix} \right] = ((A - Q)u, v) + (p, r). \quad (9.2)$$

Непосредственные вычисления дают самосопряженность оператора S относительно скалярного произведения (9.2).

Затем определим предобуславливатель

$$R = \begin{pmatrix} I & 0 \\ 0 & B^*A^{-1}B \end{pmatrix},$$

также самосопряженный относительно скалярного произведения (9.2), и установим спектральную эквивалентность

$$\beta_1[Rz, z] \leq [Sz, z] \leq \beta_2[Rz, z]$$

с постоянными

$$\beta_1 = \left(1 + \frac{\kappa}{2} + \sqrt{\kappa + \frac{\kappa^2}{4}} \right)^{-1}, \quad \beta_2 = \frac{1 + \sqrt{\kappa}}{1 - \kappa}.$$

С точки зрения практических расчетов, это означает возможность использования предобусловленного метода сопряженных градиентов

для решения задачи (9.1), в котором вместо дополнения по Шуру $B^*A^{-1}B$ (в операторе R) применяется оператор C такой, что

$$\gamma C \leq B^*A^{-1}B \leq \Gamma C.$$

Теоретический аспект проблемы не менее важен: наличие свободных параметров в исходной задаче и в предобусловливателе позволяет формулировать и решать задачи асимптотической оптимизации в различных постановках с целью выяснения предельных возможностей подхода. Отметим также, что методы для рассматриваемой положительно определенной переформулировки исходной седловой задачи гарантируют стандартные оценки погрешности с точными показателями, определяемыми величинами β_1, β_2 .

Далее в главе анализируются различные обобщения симметризации специального вида.

9.1. ОБОБЩЕННЫЙ МЕТОД BRAMBLE–PASCIAK (GMBP)

9.1.1. Построение метода

Согласно классической теории итерационных методов, алгоритм Ричардсона для решения системы линейных уравнений $Sz = \tilde{F}$ с оператором перехода $T = I - \tau S$ может быть значительно ускорен чебышевским методом или методом сопряженных градиентов в случае, когда оператор S является самосопряженным и положительно определенным. Метод GMSOR (7.1) может быть представлен как метод Ричардсона для решения предобусловленной седловой задачи

$$Sz \equiv \begin{pmatrix} Q & 0 \\ \tau B^* & -\alpha C \end{pmatrix}^{-1} \begin{pmatrix} A + \beta BC^{-1}B^* & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \tilde{F},$$

$$\tilde{F} = \begin{pmatrix} Q^{-1}(f + \beta BC^{-1}\varphi) \\ \alpha^{-1}C^{-1}(B^*Q^{-1}(f + \beta BC^{-1}\varphi) - \varphi) \end{pmatrix}, \quad \alpha, \tau > 0, \quad \beta \in \mathbb{R}.$$

В общем случае оператор S не является ни самосопряженным, ни положительно определенным ни в каком скалярном произведении в Z , что следует из возможности наличия у S существенно комплексных собственных значений. Тем не менее, при определенных условиях, а именно

$$A = A^*, \quad \tau A + \tau \beta BC^{-1}B^* - Q > 0 \quad (9.3)$$

в пространстве Z можно определить скалярное произведение

$$[z_1, z_2] = ((\tau A + \tau \beta BC^{-1}B^* - Q)u_1, u_2) + \alpha(Cp_1, p_2), \quad (9.4)$$

$$z_1 = \{u_1, p_1\}, \quad z_2 = \{u_2, p_2\},$$

в котором оператор S является самосопряженным и положительно определенным (см. теорему 9.1.1). А это означает то, что для решения системы уравнений $Sz = \tilde{F}$ можно использовать метод сопряженных градиентов: пусть z^0 — начальное приближение к $z = \{u, p\}$, и $x^0 = y^0 = \tilde{F} - Sz^0$, тогда для $k = 0, 1, \dots$ таких, что $y^i \neq 0$ при $i = 0, \dots, k$

$$\begin{aligned} z^{k+1} &= z^k + \frac{[y^k, x^k]}{[Sx^k, x^k]} x^k, \\ y^{k+1} &= \tilde{F} - Sz^{k+1}, \\ x^{k+1} &= y^{k+1} - \frac{[Sy^{k+1}, x^k]}{[Sx^k, x^k]} x^k. \end{aligned}$$

При этом, если спектр оператора S принадлежит отрезку $[m, M]$, $0 < m \leq M < +\infty$, то для нормы ошибки на k -й итерации справедлива оценка

$$[z^k - z, z^k - z]^{1/2} \leq \left(T_k \left(\frac{M+m}{M-m} \right) \right)^{-1} [z^0 - z, z^0 - z]^{1/2}, \quad (9.5)$$

что эквивалентно асимптотическому убыванию ошибки со скоростью геометрической прогрессии с показателем q :

$$q = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}.$$

Теорема 9.1.1 (Обоснование сходимости метода). Пусть $\alpha, \tau > 0$, $\beta \in \mathbb{R}$ и выполнены условия (9.3), тогда билинейная форма (9.4) определяет скалярное произведение в Z , относительно которого оператор S является самосопряженным и положительно определенным.

Доказательство. При выполнении условий теоремы оператор

$$D = \begin{pmatrix} \tau A + \tau \beta BC^{-1} B^* - Q & 0 \\ 0 & \alpha C \end{pmatrix}$$

является самосопряженным и положительно определенным оператором в Z , откуда следует, что билинейная форма

$$[z_1, z_2] = (Dz_1, z_2), \quad z_1, z_2 \in Z$$

определяет скалярное произведение в Z .

По определению, оператор S является самосопряженным относительно скалярного произведения $[\cdot, \cdot]$, если для любых $z_1, z_2 \in Z$ выполнено равенство

$$[Sz_1, z_2] = [z_1, Sz_2],$$

т. е.

$$(DSz_1, z_2) = (Dz_1, Sz_2),$$

что, в свою очередь, эквивалентно условию

$$D^{1/2}SD^{-1/2} = D^{-1/2}S^*D^{1/2}.$$

Запишем оператор S в следующей форме:

$$S = \begin{pmatrix} Q^{-1}\tilde{A} & Q^{-1}B \\ \alpha^{-1}C^{-1}B^*Q^{-1}(\tau\tilde{A} - Q) & \tau\alpha^{-1}C^{-1}B^*Q^{-1}B \end{pmatrix},$$

где $\tilde{A} = A + \beta BC^{-1}B^*$, тогда

$$D^{1/2}SD^{-1/2} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix},$$

где

$$R_{11} = (\tau\tilde{A} - Q)^{-1/2}(\tau\tilde{A}Q^{-1}\tilde{A} - \tilde{A})(\tau\tilde{A} - Q)^{-1/2},$$

$$R_{12} = \alpha^{-1/2}(\tau\tilde{A} - Q)^{1/2}Q^{-1}BC^{-1/2},$$

$$R_{21} = \alpha^{-1/2}C^{-1/2}B^*Q^{-1}(\tau\tilde{A} - Q)^{1/2},$$

$$R_{22} = \tau\alpha^{-1}C^{-1/2}B^*Q^{-1}BC^{-1/2}.$$

Из того, что $R_{11} = R_{11}^*$, $R_{22} = R_{22}^*$ и $R_{12} = R_{21}^*$, следует равенство

$$D^{1/2}SD^{-1/2} = \left(D^{1/2}SD^{-1/2}\right)^*.$$

Положим

$$L = (\tau\tilde{A} - Q)^{1/2}, \quad G = \alpha^{-1/2}Q^{-1/2}BC^{-1/2},$$

тогда, используя представление для $D^{1/2}SD^{-1/2}$, получаем цепочку неравенств

$$\begin{aligned} \left(D^{1/2}SD^{-1/2}z, z\right) &= (L^{-1}(\tau\tilde{A}Q^{-1}\tilde{A} - \tilde{A})L^{-1}u, u) + \\ &+ 2(\tau^{1/2}Gp, \tau^{-1/2}Q^{-1/2}Lu) + \tau(Gp, Gp) \geq \\ &\geq (L^{-1}(\tau\tilde{A}Q^{-1}\tilde{A} - \tilde{A})L^{-1}u, u) - \\ &- \tau(Gp, Gp) - \tau^{-1}(Q^{-1/2}Lu, Q^{-1/2}Lu) + \tau(Gp, Gp) = \\ &= ((\tau\tilde{A}Q^{-1}\tilde{A} - \tilde{A})v, v) - \tau^{-1}(L^2Q^{-1}L^2v, v) = \tau^{-1}((\tau\tilde{A} - Q)v, v) \geq 0, \end{aligned}$$

где $z = \{u, p\} \in Z \setminus \{0\}$, $v = L^{-1}u$. Таким образом, справедлива оценка

$$[Sz, z] = (DSz, z) = \left(D^{1/2}SD^{-1/2} D^{1/2}z, D^{1/2}z\right) \geq 0$$

и, следовательно, оператор S неотрицательно определен. Оператор S невырожден и неотрицательно определен, а значит является положительно определенным. Теорема доказана. ■

9.1.2. Оптимизация метода в классе \mathbb{K}_1

Оценим границы спектра преобусловленного оператора S . Имеет место

Теорема 9.1.2. Пусть $(A, B, Q, C) \in \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$ и справедливы неравенства

$$\alpha > 0, \quad \tau > \delta^{-1}, \quad \beta > (\delta^{-1} - \tau) / \max\{\tau\Gamma, \alpha\},$$

тогда выполнены условия (9.3) и $\sigma(S) \subseteq [m_1, M_1] \subset \mathbb{R}$, где

$$m_1 = \frac{1}{2} \left(\Delta(1 + \nu\gamma) - \sqrt{(\Delta(1 + \nu\gamma))^2 - 4\alpha^{-1}\gamma\Delta} \right) > 0,$$

$$M_1 = \frac{1}{2} \left(\Delta(1 + \nu\Gamma) + \sqrt{(\Delta(1 + \nu\Gamma))^2 - 4\alpha^{-1}\Gamma\Delta} \right) \geq m_1,$$

$$\nu = \beta + \tau\alpha^{-1}.$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$, тогда $A = A^*$ и $\tau(A + \beta BC^{-1}B^*) - Q \geq (\tau + \min\{0, \beta\tau\Gamma\})A - Q > \delta^{-1}A - Q \geq 0$.

Таким образом, выполнены условия (9.3) и, в силу теоремы 9.1.1, справедливо $\sigma(S) \subset (0, +\infty)$. Положим

$$L = A^{-1/2}QA^{-1/2}, \quad G = A^{-1/2}BC^{-1}B^*A^{-1/2},$$

$$\delta_1 = \Delta^{-1}, \quad \delta_2 = \delta^{-1}, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 7.1.1, число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \neq 0$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G,$$

$$f(\lambda, s) = 1 - \lambda s, \quad g(\lambda, t) = \lambda, \quad h(\lambda) = \lambda(\beta + \tau\alpha^{-1}) - \alpha^{-1}.$$

Итак, выполнены все условия следствия 6.4.1 из теоремы 6.4.8 и, следовательно, для любого $\lambda \in \sigma(S)$ справедливо: либо $\lambda \in [\delta_1, \delta_2]$, либо существуют $s \in [\delta_1, \delta_2]$, $t \in [\gamma_1, \gamma_2]$ такие, что

$$\lambda^2 s - \lambda(1 + (\beta + \tau\alpha^{-1})t) + \alpha^{-1}t = 0.$$

После замены $s \rightarrow s^{-1}$ получаем оценку

$$\sigma(S) \subseteq \Lambda \equiv [\delta, \Delta] \cup \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \alpha^{-1}ts} \right\}, \quad \theta = s \frac{1 + \nu t}{2}.$$

Несложно убедиться, что

$$\lambda^2 - \lambda(1 + \nu ts) + \alpha^{-1}ts < 0$$

$$\text{при } s \in [\delta, \Delta], \quad t \in [\gamma, \Gamma], \quad \lambda = \tau^{-1},$$

откуда с учетом неравенства

$$\theta > \sqrt{\theta^2 - \alpha^{-1}ts} \geq 0$$

следует, что $\Lambda \subset (0, +\infty)$.

Определим функции

$$\lambda^\pm(t, s) = \theta \pm \sqrt{\theta^2 - \alpha^{-1}ts},$$

тогда, в силу условия $\nu > (\alpha\delta)^{-1}$, для любых $s \in [\delta, \Delta]$, $t \in [\gamma, \Gamma]$ существуют производные

$$\frac{\partial \lambda^-(t, s)}{\partial s} < 0, \quad \frac{\partial \lambda^+(t, s)}{\partial s} > 0, \quad \frac{\partial \lambda^-(t, s)}{\partial t} > 0, \quad \frac{\partial \lambda^+(t, s)}{\partial t} > 0.$$

Следовательно, корректны определения:

$$\min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \alpha^{-1}ts} \right\} = \lambda^-(\gamma, \Delta) = m_1,$$

$$\max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \alpha^{-1}ts} \right\} = \lambda^+(\Gamma, \Delta) = M_1$$

и $\sigma(S) \subseteq [\delta, \Delta] \cup [m_1, M_1]$.

Докажем неравенства

$$m_1 \leq \lambda^-(\gamma, \delta) \leq \delta, \quad \Delta \leq \lambda^+(\Gamma, \Delta) = M_1.$$

Неравенство $\lambda^-(\gamma, \delta) \leq \delta$ эквивалентно следующему:

$$\theta - \sqrt{\theta^2 - \alpha^{-1}\delta\gamma} \leq \delta,$$

т. е.

$$2\delta\theta - \delta^2 \geq \alpha^{-1}\delta\gamma,$$

что, в свою очередь, равносильно неравенству $\nu \geq (\alpha\delta)^{-1}$. Аналогично неравенство $\Delta \leq \lambda^+(\Gamma, \Delta)$ следует из выражения $\nu \geq (\alpha\Delta)^{-1}$, которое, в свою очередь, следует из оценки $\nu \geq \alpha^{-1}\delta^{-1}$. Таким образом, $[\delta, \Delta] \subseteq [m_1, M_1]$ и имеет место оценка $\sigma(S) \subseteq [m_1, M_1]$. Теорема доказана. ■

Оценка границ спектра совместно с оценкой скорости сходимости метода сопряженных градиентов позволяют доказать основной результат.

Теорема 9.1.3. Пусть выполнены условия теоремы 9.1.2, тогда имеет место оценка асимптотического показателя скорости сходимости алгоритма GMBP

$$q_1 \leq \frac{1 - \sqrt{\eta_1}}{1 + \sqrt{\eta_1}}, \quad \eta_1 = \frac{m_1}{M_1},$$

причем предельно наилучшее значение оценки достигается при

$$\alpha_1 = \frac{\sqrt{\Gamma\gamma}}{\delta}, \quad \beta_1 = 0, \quad \tau_1 = \delta^{-1} + \varepsilon, \quad \varepsilon \rightarrow 0.$$

Доказательство. Оценка для q_1 немедленно следует из неравенства (9.5) и оценок спектра, полученных в теореме 9.1.2.

При любых α, β, τ , удовлетворяющих условию теоремы 9.1.2, имеют место неравенства

$$\frac{\partial m_1}{\partial \nu} < 0, \quad \frac{\partial M_1}{\partial \nu} > 0.$$

Таким образом, при фиксированном $\alpha > 0$ функция η_1 является убывающей относительно аргумента $\nu = \beta + \tau/\alpha$, следовательно, предельно максимальное значение η_1 соответствует величине

$$\nu = \min_{\substack{\tau \geq \delta^{-1}, \\ \beta \geq (\delta^{-1} - \tau) / \max\{\tau\Gamma, \alpha\}}} \left(\beta + \frac{\tau}{\alpha} \right) = (\alpha\delta)^{-1},$$

где минимум достигается при $\tau = \delta^{-1}, \beta = 0$.

Пусть $\nu = (\alpha\delta)^{-1}$, тогда имеем

$$\eta_1 = \frac{1 + \xi x - \sqrt{(1 + \xi x)^2 - 4\omega\xi x}}{1 + x + \sqrt{(1 + x)^2 - 4\omega x}} = \frac{\xi^{1/2}}{4\omega} f(\xi x) f(x),$$

где

$$x = \frac{\Gamma}{\alpha\delta}, \quad f(x) = x^{1/2} + x^{-1/2} - \sqrt{(x^{1/2} + x^{-1/2})^2 - 4\omega},$$

$$\omega = \frac{\delta}{\Delta}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Из уравнения

$$0 = \frac{\partial \eta_1}{\partial x} = -\frac{\xi^{1/2} f(\xi x) f(x)}{8\omega x} \left(\frac{x-1}{\sqrt{(x+1)^2 - 4\omega x}} + \frac{\xi x - 1}{\sqrt{(\xi x + 1)^2 - 4\omega \xi x}} \right)$$

следует, что единственной точкой максимума функции η_1 при $x > 0$ является точка $x = \xi^{-1/2}$, которая соответствует значению $\alpha = \sqrt{\Gamma\gamma}/\delta$.

Окончательно приходим к выводу, что набор $\alpha = \sqrt{\Gamma\gamma}/\delta, \beta = 0, \tau = \delta^{-1}$ определяет предельный максимум η_1 . Теорема доказана. ■

Следствие 9.1.1. Пусть выполнены условия теоремы 9.1.2, тогда для алгоритма GMBP имеют место асимптотические оценки

$$\begin{aligned} q_1 &\leq 1 - 2\sqrt{\omega\xi} + O(\xi), \quad \omega = \text{const}, \quad \xi \rightarrow 0, \\ q_1 &\leq 1 - \frac{2\sqrt{\omega\xi}}{1 + \sqrt{\xi}} + O(\omega), \quad \xi = \text{const}, \quad \omega \rightarrow 0, \end{aligned}$$

где $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. Формулы немедленно следуют из теоремы 9.1.3 и выражения для предельно наилучшего значения

$$\eta_1 = \frac{(1 + \sqrt{\xi} - \sqrt{(1 + \sqrt{\xi})^2 - 4\omega\sqrt{\xi}})^2}{4\omega}.$$

Следствие доказано. ■

9.1.3. Оптимизация метода в классе \mathbb{K}_2

Следуя по аналогии с предыдущим разделом, оценим границы спектра предобусловленного оператора S . Имеет место

Теорема 9.1.4. Пусть $(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$ и справедливы неравенства

$$\alpha > 0, \quad \tau > \delta^{-1}, \quad \beta > (1 - \tau\delta)/\max\{\tau\Gamma, \alpha\delta\},$$

тогда выполнены условия (9.3) и $\sigma(S) \subseteq [m_2, M_2] \subset \mathbb{R}$, где

$$\begin{aligned} m_2 &= \frac{1}{2} \left(\Delta + \nu\gamma - \sqrt{(\Delta + \nu\gamma)^2 - 4\alpha^{-1}\gamma} \right) > 0, \\ M_2 &= \frac{1}{2} \left(\Delta + \nu\Gamma + \sqrt{(\Delta + \nu\Gamma)^2 - 4\alpha^{-1}\Gamma} \right) > 0, \\ \nu &= \beta + \tau\alpha^{-1}. \end{aligned}$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, тогда $A = A^*$ и имеют место неравенства

$$\tau(A + \beta BC^{-1}B^*) - Q \geq \tau A + (\min\{0, \beta\tau\Gamma\} - 1)Q > \tau A - \tau\delta Q \geq 0.$$

Таким образом, выполнены условия (9.3) и, в силу теоремы 9.1.1, справедливо $\sigma(S) \subset (0, +\infty)$. Положим

$$\begin{aligned} L &= A^{-1/2}QA^{-1/2}, \quad G = Q^{-1/2}BC^{-1}B^*Q^{-1/2}, \\ \delta_1 &= \delta, \quad \delta_2 = \Delta, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma, \end{aligned}$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 7.1.1, число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \neq 0$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) = s - \lambda, \quad g(\lambda, t) = \lambda, \quad h(\lambda) = \lambda(\beta + \tau\alpha^{-1}) - \alpha^{-1}.$$

Итак, выполнены все условия следствия 6.4.1 из теоремы 6.4.8 и, следовательно, для любого $\lambda \in \sigma(S)$ справедливо: либо $\lambda \in [\delta_1, \delta_2]$, либо существуют $s \in [\delta_1, \delta_2]$, $t \in [\gamma_1, \gamma_2]$ такие, что

$$\lambda^2 - \lambda(s + (\beta + \tau\alpha^{-1})t) + \alpha^{-1}t = 0.$$

В результате получаем оценку

$$\sigma(S) \subseteq \Lambda \equiv [\delta, \Delta] \cup \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \alpha^{-1}t} \right\}, \quad \theta = (s + \nu t)/2.$$

Несложно убедиться, что $\lambda^2 - \lambda(s + \nu t) + \alpha^{-1}t < 0$ при $s \in [\delta, \Delta]$, $t \in [\gamma, \Gamma]$, $\lambda = \tau^{-1}$, откуда с учетом неравенства $\theta > \sqrt{\theta^2 - \alpha^{-1}t} \geq 0$ следует, что $\Lambda \subset (0, +\infty)$.

Определим функции $\lambda^\pm(t, s) = \theta \pm \sqrt{\theta^2 - \alpha^{-1}t}$, тогда, в силу условия $\nu > (\alpha\delta)^{-1}$, для любых $s \in [\delta, \Delta]$, $t \in [\gamma, \Gamma]$ существуют производные

$$\frac{\partial \lambda^-(t, s)}{\partial s} < 0, \quad \frac{\partial \lambda^+(t, s)}{\partial s} > 0, \quad \frac{\partial \lambda^-(t, s)}{\partial t} > 0, \quad \frac{\partial \lambda^+(t, s)}{\partial t} > 0.$$

Следовательно, имеют место равенства

$$\min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \alpha^{-1}t} \right\} = \lambda^-(\gamma, \Delta) = m_2, \\ \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \alpha^{-1}t} \right\} = \lambda^+(\Gamma, \Delta) = M_2.$$

и $\sigma(S) \subseteq [\delta, \Delta] \cup [m_2, M_2]$.

Докажем неравенства

$$m_2 \leq \lambda^-(\gamma, \delta) \leq \delta \quad \text{и} \quad \Delta \leq \lambda^+(\Gamma, \Delta) = M_2.$$

Неравенство $\lambda^-(\gamma, \delta) \leq \delta$ эквивалентно следующему:

$$\theta - \sqrt{\theta^2 - \alpha^{-1}\delta} \leq \delta, \quad \text{т. е.} \quad 2\delta\theta - \delta^2 \geq \alpha^{-1}\gamma,$$

что, в свою очередь, равносильно неравенству $\nu\delta \geq \alpha^{-1}$. Аналогично неравенство $\Delta \leq \lambda^+(\Gamma, \Delta)$ следует из выражения $\nu\Delta \geq \alpha^{-1}$, которое, в свою очередь, следует из оценки $\nu\delta \geq \alpha^{-1}$. Таким образом,

$$[\delta, \Delta] \subseteq [m_2, M_2]$$

и имеет место оценка $\sigma(S) \subseteq [m_2, M_2]$. Теорема доказана. ■

Оценка границ спектра совместно с оценкой скорости сходимости метода сопряженных градиентов позволяют доказать основной результат.

Теорема 9.1.5. Пусть выполнены условия теоремы 9.1.4, тогда имеет место оценка асимптотического показателя скорости сходимости алгоритма GMBP

$$q_2 \leq \frac{1 - \sqrt{\eta_2}}{1 + \sqrt{\eta_2}}, \quad \eta_2 = m_2/M_2,$$

причем предельно наилучшее значение оценки достигается при

$$\alpha_2 = \frac{\sqrt{\Gamma\gamma}}{\Delta\delta}, \quad \beta_2 = 0, \quad \tau_2 = \delta^{-1} + \varepsilon, \quad \varepsilon \rightarrow 0.$$

Доказательство. Оценка для q_2 немедленно следует из неравенства (9.5) и оценок спектра, полученных в теореме 9.1.4.

При любых α , β , τ удовлетворяющих условию теоремы 9.1.4 имеют место неравенства

$$\frac{\partial m_2}{\partial \nu} < 0, \quad \frac{\partial M_2}{\partial \nu} > 0.$$

Таким образом, при фиксированном $\alpha > 0$ функция η_2 является убывающей относительно аргумента $\nu = \beta + \tau/\alpha$, следовательно, предельно максимальное значение η_2 соответствует величине

$$\nu = \min_{\substack{\tau \geq \delta^{-1}, \\ \beta \geq (1-\tau\delta)/\max\{\tau\Gamma, \alpha\delta\}}} \left(\beta + \frac{\tau}{\alpha} \right) = (\alpha\delta)^{-1},$$

где минимум достигается при $\tau = \delta^{-1}$, $\beta = 0$.

Пусть $\nu = (\alpha\delta)^{-1}$, тогда справедливо

$$\eta_2 = \frac{1 + \xi x - \sqrt{(1 + \xi x)^2 - 4\omega\xi x}}{1 + x + \sqrt{(1 + x)^2 - 4\omega x}} = \frac{\xi^{1/2}}{4\omega} f(\xi x) f(x),$$

где

$$x = \frac{\Gamma}{\alpha\delta\Delta}, \quad f(x) = x^{1/2} + x^{-1/2} - \sqrt{(x^{1/2} + x^{-1/2})^2 - 4\omega},$$

$$\omega = \frac{\delta}{\Delta}, \quad \xi = \frac{\gamma}{\Gamma}.$$

Из уравнения

следует, что единственной точкой максимума функции η_2 при $x > 0$ является точка $x = \xi^{-1/2}$, которая соответствует значению $\alpha = \sqrt{\Gamma\gamma}/(\Delta\delta)$.

Окончательно приходим к выводу, что набор

$$\alpha = \frac{\sqrt{\Gamma\gamma}}{\Delta\delta}, \quad \beta = 0, \quad \tau = \delta^{-1}$$

определяет предельный максимум η_2 . Теорема доказана. ■

Следствие 9.1.2. Пусть выполнены условия теоремы 9.1.4, тогда для алгоритма GMBP имеют место асимптотические оценки

$$\begin{aligned} q_2 &\leq 1 - 2\sqrt{\omega\xi} + O(\xi), \quad \omega = \text{const}, \quad \xi \rightarrow 0, \\ q_2 &\leq 1 - \frac{2\sqrt{\omega\xi}}{1 + \sqrt{\xi}} + O(\omega), \quad \xi = \text{const}, \quad \omega \rightarrow 0, \end{aligned}$$

где $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. Формулы немедленно следуют из теоремы 9.1.5 и выражения для предельно наилучшего значения

$$\eta_2 = \frac{(1 + \sqrt{\xi} - \sqrt{(1 + \sqrt{\xi})^2 - 4\omega\sqrt{\xi}})^2}{4\omega}.$$

Следствие доказано. ■

9.1.4. Оценка в классе \mathbb{K}_{2s}

Оценки в классе \mathbb{K}_{2s} получим согласно подходу, описанному в 6.3. Имеет место

Теорема 9.1.6. Пусть $(A, B, Q, C) \in \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma)$ и справедливы неравенства

$$\alpha > 0, \quad \tau > \delta^{-1}, \quad \beta > \frac{\delta}{\gamma} + \frac{1 - \tau\delta}{\max\{\tau\Gamma, \alpha\delta\}},$$

тогда выполнены условия (9.3) и $\sigma(S) \subseteq [m_{2s}, M_{2s}] \subset \mathbb{R}$, где

$$m_{2s} = \frac{1}{2}(\Delta_s + \nu\gamma - \sqrt{(\Delta_s + \nu\gamma)^2 - 4\alpha^{-1}\gamma}) > 0,$$

$$M_{2s} = \frac{1}{2}(\Delta_s + \nu\Gamma + \sqrt{(\Delta_s + \nu\Gamma)^2 - 4\alpha^{-1}\Gamma}) > 0,$$

$$\nu = \beta + \tau\alpha^{-1} - \frac{\delta}{\gamma}, \quad \Delta_s = \delta(\omega^{-1} + \xi^{-1}), \quad \xi = \frac{\delta}{\gamma}, \quad \omega = \frac{\delta}{\Delta}.$$

Доказательство. В силу (6.17), справедливо

$$(A + \delta/\gamma BC^{-1}B^*, B, Q, C) \in \mathbb{K}_2(\delta, \Delta_s, \gamma, \Gamma).$$

Тогда утверждение теоремы немедленно следует из теоремы 9.1.4. Теорема доказана. ■

Теорема 9.1.7. Пусть выполнены условия теоремы 9.1.6, тогда имеет место оценка асимптотического показателя скорости сходимости алгоритма GMBP

$$q_{2s} \leq \frac{1 - \sqrt{\eta_{2s}}}{1 + \sqrt{\eta_{2s}}}, \quad \eta_{2s} = m_{2s}/M_{2s},$$

причем предельно наилучшее значение оценки достигается при

$$\alpha_{2s} = \frac{\sqrt{\Gamma\gamma}}{\Delta_s\delta}, \quad \beta_{2s} = \frac{\delta}{\gamma}, \quad \tau_{2s} = \delta^{-1} + \varepsilon, \quad \varepsilon \rightarrow 0.$$

Доказательство. Дословно повторим доказательство теоремы 9.1.5 после формального переобозначения параметров $(\beta - \delta/\gamma) \rightarrow \beta$. Теорема доказана. ■

Следствие 9.1.3. Пусть выполнены условия теоремы 9.1.6, тогда для алгоритма GMBP имеют место асимптотические оценки

$$q_{2s} \leq 1 - 2\xi + O(\xi), \quad \omega = \text{const}, \quad \xi \rightarrow 0,$$

$$q_{2s} \leq 1 - \frac{2\sqrt{\omega\xi}}{1 + \sqrt{\xi}} + O(\omega), \quad \xi = \text{const}, \quad \omega \rightarrow 0,$$

где $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. Формулы немедленно следуют из теоремы 9.1.7 и выражения для предельно наилучшего значения

$$\eta_{2s} = \frac{\omega + \xi}{4\omega\xi} \left(1 + \sqrt{\xi} - \sqrt{(1 + \sqrt{\xi})^2 - 4\omega\xi \frac{\sqrt{\xi}}{\omega + \xi}} \right)^2.$$

Следствие доказано. ■

9.2. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Идея симметризации специального вида, ее обоснование, определение величин β_1, β_2 и приложение к смешанным (седловым) формулировкам для эллиптических задач были опубликованы в работе [123].

В работе [152] на примере дискретных уравнений Стокса описываемый подход сравнивался с предобусловленными алгоритмами Эрроу–Гурвица [180], сопряженных невязок (PCR-method) [198]

и многосеточным методом [199] с различными вариантами сглаживающих операторов. При анализе результатов численных экспериментов предпочтение отдавалось многосеточному методу. Среди алгоритмов одинакового типа упорядочивание по эффективности в явном виде не просматривалось. Однако анализируемый подход не был среди них худшим.

Такую же структуру и выводы имеет работа [18], посвященная решению жестких эллиптических задач с большими параметрами, но там метод сведения исходной задачи к системе с седловой точкой носил несколько другой характер.

В [130] впервые была сделана попытка оценить теоретические возможности рассматриваемой идеи. Для равносильной задачи

$$\begin{aligned}\tilde{S}z &\equiv \begin{pmatrix} Q^{-1}(A + \beta B_0) & Q^{-1}B \\ B^*Q^{-1}(\nu A - Q + \nu\beta B_0) & \nu B^*Q^{-1}B \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \\ &= \begin{pmatrix} Q^{-1}(f + \beta BC^{-1}\varphi) \\ \nu B^*Q^{-1}(f + \beta BC^{-1}\varphi) - \varphi \end{pmatrix},\end{aligned}$$

где $B_0 = BC^{-1}B^*$, была определена необходимая модификация скалярного произведения в пространстве $Z = U \times P$:

$$\left[\begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ r \end{pmatrix} \right] = ((\nu A - Q + \nu\beta B_0)u, v) + (p, r).$$

Отметим, что если введение параметра ν эквивалентно масштабированию постоянной δ в работе [123], то добавление в первое уравнение слагаемого вида $\beta B_0 u$ для такого подхода ранее не использовалось.

Для матрицы \tilde{S} был рассмотрен предобусловливатель с параметром $\alpha > 0$ вида

$$\tilde{R} = \begin{pmatrix} I & 0 \\ 0 & \alpha C \end{pmatrix}.$$

Введение предобусловливателя типа \tilde{R} в рассматриваемый алгоритм впервые предлагалось в [152], но без параметра α .

Анализ позволил уточнить границы спектра предобусловленного оператора:

$$M = \frac{1}{1 - \sqrt{\kappa}} \quad (= \beta_2), \quad m = \frac{1}{1 + \sqrt{\kappa}} \quad (> \beta_1).$$

Введем обозначение $\kappa = \beta + \nu/\alpha$. Оказывается, что число обусловленности в рассматриваемом алгоритме

$$\text{cond}_2(\tilde{R}^{-1}\tilde{S}) = \frac{M(\kappa)}{m(\kappa)}$$

принимает наименьшее значение при $\kappa = 1$. Любопытно, что оригинальная версия алгоритма [123] хорошо согласуется с этим результатом. Ей соответствует набор параметров

$$\alpha = 1, \quad \beta = 0, \quad \nu = 1,$$

и асимптотика оптимального числа обусловленности при больших значениях параметра Δ и $\delta = 1$:

$$\text{cond}_2 = \frac{(1 + \gamma)(1 + \Gamma)}{\gamma} \Delta + o(\Delta).$$

Следует также отметить, что идея симметризации специального вида седловых задач оказалась плодотворной для теоретического анализа алгоритмов и более сложной структуры [202]. Это позволило получить на основе единого подхода широкий круг оценок скорости сходимости различных методов. Однако результатов, связанных с оптимальным выбором параметров, там получить не удалось.

МОДЕЛЬНЫЕ СЕДЛОВЫЕ ОПЕРАТОРЫ

Использование симметричных седловых предобусловливателей, или *модельных седловых операторов*, при решении симметричных седловых задач преследует естественную цель построить итерационный метод с симметризуемым знакоопределенным оператором перехода. В этом случае эффективность численного алгоритма базируется на ускорении сходимости за счет применения чебышевских параметров или вариационных методов (типа сопряженных градиентов).

Однако на этом пути имеются, как минимум, две серьезные трудности. Первая — практическая; она связана с понятием *неконструктивного подхода*. Здесь имеется в виду необходимость обращения на каждом шаге седлового оператора, имеющего такую же структуру, как в исходной постановке. При этом предполагается, что обращение модельного оператора достигается все-таки проще (более эффективно). Этот трудноформализуемый момент имеет компенсацию в виде большей простоты технологии исследования и наглядности получаемых результатов. Не будет большим преувеличением сказать, что неконструктивный подход более важен с теоретической точки зрения, чем с практической. Другая трудность является органическим дополнением первой, так как связана с *конструктивным* построением легко обратимых модельных седловых операторов. При таком подходе, наоборот, резко усложняется теоретический анализ спектральных характеристик предобусловленного оператора, поэтому результаты их оптимизации носят особо ценный характер.

10.1. НЕКОНСТРУКТИВНЫЙ ПОДХОД

10.1.1. Построение методов

Решение задачи с седловым оператором можно свести к решению эквивалентной предобусловленной задачи

$$Sz \equiv \begin{pmatrix} \alpha Q & B \\ B^* & -\beta C \end{pmatrix}^{-1} \begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} \alpha Q & B \\ B^* & -\beta C \end{pmatrix}^{-1} F \equiv \tilde{F},$$

где $\alpha > 0$, $\beta \geq 0$. Это имеет смысл, так как (см. далее) при выполнении условий

$$A = A^*, \quad A > \alpha Q, \quad \beta > 0$$

оператор S является самосопряженным и положительно определенным относительно скалярного произведения

$$(z_1, z_2)_D = (Dz_1, z_2), \quad z_1, z_2 \in Z, \quad D = \begin{pmatrix} A - \alpha Q & 0 \\ 0 & \beta C \end{pmatrix}.$$

Это означает, что для решения системы уравнений $Sz = \tilde{F}$ можно использовать метод сопряженных градиентов: пусть z^0 — начальное приближение к $z = \{u, p\}$, и $x^0 = y^0 = \tilde{F} - Sz^0$, для $k = 0, 1, \dots$ таких, что $y^i \neq 0$ при $i = 0, \dots, k$

$$\begin{aligned} z^{k+1} &= z^k + \frac{(y^k, x^k)_D}{(Sx^k, x^k)_D} x^k, \\ y^{k+1} &= \tilde{F} - Sz^{k+1}, \\ x^{k+1} &= y^{k+1} - \frac{(Sy^{k+1}, x^k)_D}{(Sx^k, x^k)_D} x^k. \end{aligned}$$

При этом, если спектр оператора S принадлежит отрезку $[m, M]$, $0 < m \leq M < +\infty$, то для нормы ошибки на k -й итерации справедлива оценка

$$(z^k - z, z^k - z)_D^{1/2} \leq \left(T_k \left(\frac{M+m}{M-m} \right) \right)^{-1} (z^0 - z, z^0 - z)_D^{1/2}, \quad (10.1)$$

что эквивалентно асимптотическому убыванию ошибки со скоростью геометрической прогрессии с показателем q :

$$q = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}. \quad (10.2)$$

В случае $\beta = 0$ билинейная форма $(\cdot, \cdot)_D$ вырождена и формулы метода сопряженных градиентов применять нельзя. Однако имеет место представление

$$S = I + \begin{pmatrix} \alpha Q & B \\ B^* & 0 \end{pmatrix}^{-1} \begin{pmatrix} A - \alpha Q & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I + X & 0 \\ Y & I \end{pmatrix},$$

где X — самосопряженный оператор относительно скалярного произведения $((A - \alpha Q)u_1, u_2)$. Таким образом, процесс решения уравнения с оператором S можно разбить на две последовательные подзадачи:

$$(I + X)u = \tilde{f}, \quad p = \tilde{\varphi} - Yu.$$

Так как

$$\sigma(S) = \sigma(I + X) \cup \{1\} \subseteq [m, M] \subset (0, +\infty),$$

то для решения первой задачи можно применять метод сопряженных градиентов с показателем скорости сходимости q вида (10.2), вторая же задача решается явно.

Итак, при решении седловых задач с применением седловых предобусловливателей в качестве показателя скорости сходимости алгоритма в классе \mathbb{K} можно принять величину

$$q_{\mathbb{K}} = \sup_{(A, B, Q, C) \in \mathbb{K}} \frac{1 - \sqrt{\eta}}{1 + \sqrt{\eta}}, \quad (10.3)$$

$$\eta \equiv \eta(\alpha, \beta; A, B, Q, C) = \frac{\min_{\lambda \in \sigma(S)} \lambda}{\max_{\lambda \in \sigma(S)} \lambda}. \quad (10.4)$$

Следует учитывать, что задача оптимизации не может быть корректно сформулирована в общем виде из-за недостатка информации — это связано с тем, что обращение предобусловливателя, являющегося полноценным седловым оператором, в общем случае может являться достаточно трудоемкой задачей, т. е. «стоимость» одной итерации может быть достаточно высока. Поэтому при решении задачи оптимизации предобусловливателя необходимо учитывать компромисс между скоростью сходимости метода и эффективностью (трудоемкостью) одной итерации.

10.1.2. Оценка в классе \mathbb{K}_1

Обоснуем приведенные выше условия симметризуемости оператора S . Справедлива

Теорема 10.1.1. Пусть $A = A^* > \alpha Q$, $\beta > 0$, тогда билинейная форма $(z_1, z_2)_D$ определяет в пространстве Z скалярное произведение, относительно которого оператор S является самосопряженным.

Доказательство. Свойства скалярного произведения следуют из самосопряженности и положительной определенности оператора D . Оператор S может быть представлен в форме

$$S = I + P^{-1}D, \quad P = \begin{pmatrix} \alpha Q & B \\ B^* & -\beta C \end{pmatrix}, \quad P = P^*,$$

откуда следует цепочка равенств для любых $z_1, z_2 \in Z$:

$$\begin{aligned} (Sz_1, z_2)_D &= (DSz_1, z_2) = ((D + DP^{-1}D)z_1, z_2) = \\ &= (z_1, (D + DP^{-1}D)z_2) = (z_1, DSz_2) = (Dz_1, Sz_2) = \\ &= (z_1, Sz_2)_D. \end{aligned}$$

Таким образом, S самосопряжен относительно $(\cdot, \cdot)_D$. Теорема доказана. ■

Оценим распределение спектра изучаемого предобусловленного оператора. Имеет место

Теорема 10.1.2. Пусть $(A, B, Q, C) \in \mathbb{K}_1(\delta, \Delta, \gamma, \Gamma)$, $0 < \alpha \leq \delta$, $\beta \geq 0$, тогда

$$\sigma(S) \subseteq \Lambda \equiv \left[\frac{2\gamma + \beta - \sqrt{\beta^2 + 4\beta\gamma(1 - \alpha/\Delta)}}{2(\gamma + \alpha\beta/\Delta)}, \Delta/\alpha \right] \subset (0, +\infty).$$

Доказательство. В силу непрерывной зависимости спектра S от параметров $\alpha > 0$, $\beta \geq 0$, достаточно рассмотреть случай $0 < \alpha < \delta$, $\beta > 0$. При этом справедливо $A - \alpha Q > (\delta - \alpha)Q > 0$ и, следовательно, выполнены условия теоремы 10.1.1.

В силу теоремы 10.1.1 и невырожденности S , число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \in \mathbb{R} \setminus \{0\}$ и существует вектор $\{u, p\} \neq 0$ такой, что

$$\begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} \alpha Q & B \\ B^* & -\beta C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix}.$$

Выразив из второго уравнения компоненту $p = (\lambda\beta)^{-1}(\lambda - 1)C^{-1}B^*u$ и подставив полученное выражение в первое, получим

$$[\lambda^2(BC^{-1}B^* + \alpha\beta Q) - \lambda(2BC^{-1}B^* + \beta A) + BC^{-1}B^*]u = 0, \quad u \neq 0.$$

Таким образом, задача оценки спектра оператора S сводится к задаче нахождения спектра операторного пучка

$$\begin{aligned} \chi(\lambda) &= f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) &= \lambda\alpha s - 1, \quad g(\lambda, t) = \lambda\beta, \quad h(\lambda) = (\lambda - 1)^2, \\ L &= A^{-1/2}QA^{-1/2}, \quad G = A^{-1/2}BC^{-1}B^*A^{-1/2}. \end{aligned}$$

Пусть

$$\delta_1 = \Delta^{-1}, \quad \delta_2 = \delta^{-1}, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma, \quad (10.5)$$

тогда имеют место свойства

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

Итак, выполнены все условия следствия 6.4.1 из теоремы 6.4.8 и, следовательно, для любого $\lambda \in \sigma(S)$ справедливо: либо $\alpha\lambda s - 1 = 0$ для некоторого $s \in [\delta_1, \delta_2]$, либо существуют $s \in [\delta_1, \delta_2]$, $t \in [\gamma_1, \gamma_2]$ такие, что

$$\lambda^2(t + \alpha\beta s) - \lambda(2t + \beta) + t = 0.$$

После замены $s \rightarrow s^{-1}$ получаем оценку

$$\sigma(S) \subseteq \left[\frac{\delta}{\alpha}, \frac{\Delta}{\alpha} \right] \cup \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \frac{t}{t + \alpha\beta s^{-1}}} \right\},$$

$$\theta = \frac{t + \beta/2}{t + \alpha\beta s^{-1}}.$$

С учетом условия $\alpha < \delta$ имеют место неравенства

$$\lambda^-(t, s) \equiv \theta - \sqrt{\theta^2 - \frac{t}{t + \alpha\beta s^{-1}}} < 1 < \frac{s}{\alpha},$$

$$\lambda^+(t, s) \equiv \theta + \sqrt{\theta^2 - \frac{t}{t + \alpha\beta s^{-1}}} < \frac{s}{\alpha},$$

откуда следует, что

$$\sigma(S) \subseteq \Lambda \equiv \left[\min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda^-(t, s), \Delta/\alpha \right].$$

Несложно убедиться в справедливости неравенств

$$\frac{\partial \lambda^-(t, s)}{\partial s} \leq 0, \quad \frac{\partial \lambda^-(t, s)}{\partial t} \geq 0,$$

откуда следует

$$\min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda^-(t, s) = \lambda^-(\gamma, \Delta).$$

Таким образом, имеем $\Lambda = [\lambda^-(\gamma, \Delta), \Delta/\alpha]$. Теорема доказана. ■

Следствие 10.1.1. Пусть выполнены условия теоремы 10.1.2, тогда

$$q_{K_1} \leq \frac{1 - \sqrt{\eta_1}}{1 + \sqrt{\eta_1}}, \quad \eta_1 = \frac{2\gamma + \beta - \sqrt{\beta^2 + 4\beta\gamma(1 - \alpha/\Delta)}}{2(\gamma\Delta/\alpha + \beta)} \in (0, 1).$$

Доказательство. Формула немедленно следует из оценки спектра оператора S , полученной в теореме 10.1.2, и определения (10.3). Следствие доказано. ■

10.1.3. Оценка в классе K_2

Оценим распределение спектра предобусловленного оператора. Имеет место

Теорема 10.1.3. Пусть $(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, $0 < \alpha \leq \delta$, $\beta \geq 0$, тогда

$$\sigma(S) \subseteq \Lambda \equiv \left[\frac{2\gamma + \beta\Delta - \sqrt{\beta^2\Delta^2 + 4\beta\gamma(\Delta - \alpha)}}{2(\gamma + \alpha\beta)}, \frac{\Delta}{\alpha} \right] \subset (0, +\infty).$$

Доказательство. В силу непрерывной зависимости спектра S от параметров $\alpha > 0$, $\beta \geq 0$, достаточно рассмотреть случай $0 < \alpha < \delta$, $\beta > 0$. При этом справедливо

$$A - \alpha Q > (\delta - \alpha)Q > 0$$

и, следовательно, выполнены условия теоремы 10.1.1.

Повторяя рассуждения, проведенные в теореме 10.1.2, получаем, что задача оценки спектра оператора S сводится к задаче нахождения спектра операторного пучка

$$\begin{aligned} \chi(\lambda) &= f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) &= \lambda\alpha - s, \quad g(\lambda, t) = \lambda\beta, \quad h(\lambda) = (\lambda - 1)^2, \\ L &= Q^{-1/2}AQ^{-1/2}, \quad G = Q^{-1/2}BC^{-1}B^*Q^{-1/2}. \end{aligned}$$

Пусть

$$\delta_1 = \delta, \quad \delta_2 = \Delta, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда имеют место свойства

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

Итак, выполнены все условия следствия 6.4.1 из теоремы 6.4.8 и, следовательно, для любого $\lambda \in \sigma(S)$ справедливо: либо $\alpha\lambda - s = 0$ для некоторого $s \in [\delta_1, \delta_2]$, либо существуют $s \in [\delta_1, \delta_2]$, $t \in [\gamma_1, \gamma_2]$ такие, что

$$\lambda^2(t + \alpha\beta) - \lambda(2t + \beta s) + t = 0.$$

Таким образом, получаем оценку

$$\sigma(S) \subseteq \left[\frac{\delta}{\alpha}, \frac{\Delta}{\alpha} \right] \cup \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \theta \pm \sqrt{\theta^2 - \frac{t}{t + \alpha\beta}} \right\}, \quad \theta = \frac{t + \beta s/2}{t + \alpha\beta}.$$

С учетом условия $\alpha < \delta$ имеют место неравенства

$$\begin{aligned} \lambda^-(t, s) &\equiv \theta - \sqrt{\theta^2 - \frac{t}{t + \alpha\beta}} < 1 < \frac{s}{\alpha}, \\ \lambda^+(t, s) &\equiv \theta + \sqrt{\theta^2 - \frac{t}{t + \alpha\beta}} < \frac{s}{\alpha}, \end{aligned}$$

откуда следует, что

$$\sigma(S) \subseteq \Lambda \equiv \left[\min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda^-(t, s), \frac{\Delta}{\alpha} \right].$$

Несложно убедиться в справедливости неравенств

$$\frac{\partial \lambda^-(t, s)}{\partial s} \leq 0, \quad \frac{\partial \lambda^-(t, s)}{\partial t} \geq 0,$$

откуда следует

$$\min_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda^-(t, s) = \lambda^-(\gamma, \Delta).$$

Таким образом, имеем

$$\Lambda = \left[\lambda^-(\gamma, \Delta), \frac{\Delta}{\alpha} \right].$$

Теорема доказана. ■

Следствие 10.1.2. Пусть выполнены условия теоремы 10.1.3, тогда

$$q_{\mathbb{K}_2} \leq \frac{1 - \sqrt{\eta_2}}{1 + \sqrt{\eta_2}}, \quad \eta_2 = \frac{\alpha}{\Delta} \frac{2\gamma + \beta\Delta - \sqrt{\beta^2\Delta^2 + 4\beta\gamma(\Delta - \alpha)}}{2(\gamma + \alpha\beta)} \in (0, 1).$$

Доказательство. Формула немедленно следует из оценки спектра оператора S , полученной в теореме 10.1.3, и определения (10.3). Следствие доказано. ■

Рассмотрим вопрос выбора оптимальных параметров $\alpha > 0$ и $\beta \geq 0$ в классах \mathbb{K}_1 , \mathbb{K}_2 . Несложные вычисления показывают, что справедливы неравенства

$$\frac{\partial \eta_{1,2}}{\partial \beta} < 0, \quad \frac{\partial \eta_{1,2}}{\partial \alpha} > 0.$$

Другими словами, рост значения α при условии $\alpha < \delta$ и уменьшение значения β при $\beta \geq 0$ приводят к уменьшению $q_{1,2}$ и, потенциально, к увеличению скорости сходимости алгоритма. Не стоит забывать, однако, что уменьшение β приводит к ухудшению обусловленности седлового предобусловливателя, а значит, может увеличивать накладные расходы при вычислении на каждой итерации. Таким образом, при использовании модельных седловых предобусловливателей рекомендуется выбирать параметр $\alpha \approx \delta$, а параметр $\beta \geq 0$ — с учетом компромисса между скоростью сходимости основного алгоритма и эффективностью решения задачи с предобусловливателем на каждой итерации (например, скоростью сходимости возможных внутренних итераций).

10.1.4. Оценка в классе \mathbb{K}_{2s}

Имея оценки в классе \mathbb{K}_2 и следуя подходу, описанному в 6.3, получим результат для \mathbb{K}_{2s} .

Теорема 10.1.4. Пусть

$$A = A_0 + \nu BC^{-1}B^*, \quad (A_0, B, Q, C) \in \mathbb{K}_{2s}(\delta, \Delta, \gamma, \Gamma), \\ \nu > 0, \quad 0 < \alpha \leq \min\{\delta, \nu\gamma\}, \quad \beta \geq 0,$$

тогда

$$\sigma(S) \subseteq \Lambda \subset (0, +\infty),$$

где

$$\Lambda \equiv \left[\frac{2\gamma + \beta(\Delta + \nu\Gamma) - \sqrt{\beta^2(\Delta + \nu\Gamma)^2 + 4\beta\gamma(\Delta + \nu\Gamma - \alpha)}}{2(\gamma + \alpha\beta)}, \frac{\Delta + \nu\Gamma}{\alpha} \right], \\ q_{\mathbb{K}_{2s}} \leq \frac{1 - \sqrt{\eta_{2s}}}{1 + \sqrt{\eta_{2s}}},$$

$$\eta_{2s} = \frac{\alpha}{\Delta + \nu\Gamma} \frac{2\gamma + \beta(\Delta + \nu\Gamma) - \sqrt{\beta^2(\Delta + \nu\Gamma)^2 + 4\beta\gamma(\Delta + \nu\Gamma - \alpha)}}{2(\gamma + \alpha\beta)}.$$

Доказательство. Из (6.17) следует соотношение

$$(A, B, Q, C) \in \mathbb{K}_2(\min\{\delta, \nu\gamma\}, \Delta + \nu\Gamma, \gamma, \Gamma),$$

что в совокупности с теоремой 10.1.3 и следствием 10.1.2 приводит к требуемому результату. Теорема доказана. ■

При фиксированном $\nu > 0$ функция η_{2s} возрастает по параметру α при

$$0 < \alpha \leq \min\{\delta, \gamma\nu\},$$

таким образом, для получения оптимальной оценки в теореме 10.1.4 следует полагать $\alpha = \min\{\delta, \gamma\nu\}$. В то же время при фиксированном $\alpha \leq \delta$ имеет место неравенство

$$\frac{\partial \eta_{2s}}{\partial \nu} < 0,$$

откуда следует, что наилучшее значение ν находится в диапазоне $0 < \nu \leq \delta/\gamma$. Таким образом, при решении нерегулярных седловых задач с использованием модельных седловых операторов рекомендуется выбирать $\alpha = \gamma\nu$, а параметры $\nu \in (0, \delta/\gamma)$ и $\beta \geq 0$ — с учетом компромисса между скоростью сходимости основного алгоритма и эффективностью решения задачи с предобуславливателем на каждой итерации.

10.2. КОНСТРУКТИВНОЕ ПРЕДОБУСЛОВЛИВАНИЕ

10.2.1. Построение методов

Решение задачи с седловым оператором можно свести к решению равносильной предобусловленной задачи

$$\begin{aligned} Sz &\equiv \begin{pmatrix} \alpha Q & B \\ B^* & \alpha^{-1} B^* Q^{-1} B - \beta C \end{pmatrix}^{-1} \begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \\ &= \begin{pmatrix} \alpha Q & B \\ B^* & \alpha^{-1} B^* Q^{-1} B - \beta C \end{pmatrix}^{-1} F \equiv \tilde{F}, \end{aligned}$$

где $\alpha > 0$, $\beta \geq 0$. Из факторизованного представления

$$\begin{pmatrix} \alpha Q & B \\ B^* & \alpha^{-1} B^* Q^{-1} B - \beta C \end{pmatrix} = \begin{pmatrix} \alpha Q & 0 \\ B^* & I \end{pmatrix} \begin{pmatrix} \alpha Q & 0 \\ 0 & -\beta C \end{pmatrix}^{-1} \begin{pmatrix} \alpha Q & B \\ 0 & I \end{pmatrix}$$

следует, что оператор S допускает эффективную процедуру обращения.

Определим оператор

$$D = \begin{pmatrix} A - \alpha Q & 0 \\ 0 & \beta C - \alpha^{-1} B^* Q^{-1} B \end{pmatrix}.$$

В каждом из следующих случаев:

$$A = A^*, \quad A > \alpha Q, \quad \beta C > \alpha^{-1} B^* Q^{-1} B, \quad (10.6)$$

$$A = A^*, \quad A < \alpha Q, \quad \beta C < \alpha^{-1} B^* Q^{-1} B, \quad (10.7)$$

оператор D является самосопряженным и знакоопределенным, а значит, в пространстве Z можно ввести скалярное произведение по формуле

$$(z_1, z_2)_D = \pm (Dz_1, z_2), \quad z_1, z_2 \in Z, \quad (10.8)$$

где знак «+» относится к первому, а «-» — ко второму случаю. В дальнейшем будет показано, что в каждом из этих случаев оператор S является самосопряженным и положительно определенным в соответствующем скалярном произведении $(\cdot, \cdot)_D$, а это позволяет использовать для решения задачи с оператором S метод сопряженных градиентов.

10.2.2. Оценка в классе \mathbb{K}_2

Докажем приведенные ранее условия симметризуемости предобусловленного оператора S . Имеет место

Теорема 10.2.1. Пусть выполнено условие (10.6) или (10.7), тогда оператор S является самосопряженным относительно соответствующего скалярного произведения (10.8).

Доказательство. Свойства скалярного произведения следуют из самосопряженности и знакоопределенности оператора D . Оператор S может быть представлен в форме

$$S = I + P^{-1}D, \quad P = \begin{pmatrix} \alpha Q & B \\ B^* & \alpha^{-1}B^*Q^{-1}B - \beta C \end{pmatrix}, \quad P = P^*,$$

откуда следует цепочка равенств для любых $z_1, z_2 \in Z$: в случае (10.6) имеем

$$\begin{aligned} (Sz_1, z_2)_D &= (DSz_1, z_2) = ((D + DP^{-1}D)z_1, z_2) = \\ &= (z_1, (D + DP^{-1}D)z_2) = (z_1, DSz_2) = (Dz_1, Sz_2) = \\ &= (z_1, Sz_2)_D \end{aligned}$$

и, аналогично, в случае (10.7) —

$$\begin{aligned} (Sz_1, z_2)_D &= -(DSz_1, z_2) = -((D + DP^{-1}D)z_1, z_2) = \\ &= -(z_1, (D + DP^{-1}D)z_2) = -(z_1, DSz_2) = -(Dz_1, Sz_2) = \\ &= (z_1, Sz_2)_D. \end{aligned}$$

Теорема доказана. ■

Получим оценку распределения спектра исследуемого оператора. Справедлива

Теорема 10.2.2. Пусть $(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$. Тогда при α и β , удовлетворяющих условиям $0 < \alpha \leq \delta$, $\alpha\beta \geq \Gamma$, имеет место оценка

$$\begin{aligned} \sigma(S) &\subseteq \left[\Theta(\gamma, \Delta) - \sqrt{\Theta(\gamma, \Delta)^2 - \frac{\gamma}{\alpha\beta}}, \frac{\Delta}{\alpha} \right] \subseteq (0, +\infty), \\ \Theta(t, s) &= \frac{2t + \beta s - \alpha^{-1}ts}{2\alpha\beta}. \end{aligned}$$

Если α и β удовлетворяют условиям $\alpha \geq \Delta$, $\alpha\beta \leq \gamma$, то

$$\sigma(S) \subseteq \left[\frac{\delta}{\alpha}, \Theta(\Gamma, \delta) + \sqrt{\Theta(\Gamma, \delta)^2 - \frac{\Gamma}{\alpha\beta}} \right] \subseteq (0, +\infty).$$

Доказательство. В силу непрерывной зависимости спектра от параметров $\alpha, \beta > 0$, для получения оценок достаточно рассмотреть случаи $0 < \alpha < \delta$, $\alpha\beta > \Gamma$ и $\alpha > \Delta$, $\alpha\beta < \gamma$. В случае $0 < \alpha < \delta$, $\alpha\beta > \Gamma$ имеют место неравенства

$$A - \alpha Q \geq (\delta - \alpha)Q > 0, \quad \beta C - \alpha^{-1}B^*Q^{-1}B \geq (\beta - \alpha^{-1}\Gamma)C > 0,$$

а в случае $\alpha > \Delta$, $\alpha\beta < \gamma$ справедливо

$$A - \alpha Q \leq (\Delta - \alpha)Q < 0, \quad \beta C - \alpha^{-1}B^*Q^{-1}B \leq (\beta - \alpha^{-1}\gamma)C < 0.$$

Таким образом, из теоремы 10.2.1 и невырожденности S следует, что $\sigma(S) \subset \mathbb{R} \setminus \{0\}$, а значит число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \in \mathbb{R} \setminus \{0\}$ и существует вектор $\{u, p\} \neq 0$ такой, что

$$\begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} \alpha Q & B \\ B^* & \alpha^{-1} B^* Q^{-1} B - \beta C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix}.$$

Из первого уравнения следует, что

$$\lambda \neq 1 \quad \text{и} \quad Bp = (\lambda - 1)^{-1}(A - \lambda \alpha Q)u.$$

Подставив это соотношение во второе уравнение, умноженное слева на BC^{-1} , получим

$$[\lambda^2 \alpha \beta Q - \lambda(2BC^{-1}B^* + \beta A - \alpha^{-1}BC^{-1}B^* Q^{-1}A) + BC^{-1}B^*] u = 0, \\ u \neq 0.$$

Таким образом, задача оценки спектра оператора S сводится к задаче нахождения вещественного спектра операторного пучка, сопряженного к пучку

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) = \alpha\lambda - s, \quad g(\lambda, t) = \lambda(\beta - \alpha^{-1}t), \quad h(\lambda) = (\lambda - 1)^2, \\ L = Q^{-1/2}AQ^{-1/2}, \quad G = Q^{-1/2}BC^{-1}B^*Q^{-1/2}.$$

Пусть

$$\delta_1 = \delta, \quad \delta_2 = \Delta, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда имеют место свойства:

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

Итак, выполнены все условия следствия 6.4.1 из теоремы 6.4.8 и, следовательно, для любого $\lambda \in \sigma(S)$ справедливо: либо $\alpha\lambda - s = 0$ для некоторого $s \in [\delta_1, \delta_2]$, либо существуют $s \in [\delta_1, \delta_2]$, $t \in [\gamma_1, \gamma_2]$ такие, что

$$\alpha\beta\lambda^2 - \lambda(2t + \beta s - \alpha^{-1}ts) + t = 0.$$

Таким образом, получаем оценку

$$\sigma(S) \subseteq \Lambda \equiv \left[\frac{\delta}{\alpha}, \frac{\Delta}{\alpha} \right] \cup \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{\lambda^\pm(t, s)\},$$

$$\lambda^\pm(t, s) = \theta \pm \sqrt{\theta^2 - \frac{t}{\alpha\beta}}, \quad \theta = \frac{2t + \beta s - \alpha^{-1}ts}{2\alpha\beta}.$$

Из условий теоремы при $t \in [\gamma, \Gamma]$, $s \in [\delta, \Delta]$ следует неравенство

$$\theta - \sqrt{\frac{t}{\alpha\beta}} = (2\alpha^2\beta)^{-1}(\sqrt{\alpha\beta} - t)((\sqrt{\alpha\beta} + \sqrt{t})s - 2\alpha\sqrt{t}) \geq 0,$$

откуда справедливо включение $\Lambda \subseteq (0, +\infty)$. Также имеют место равенства

$$\begin{aligned}\operatorname{sign} \frac{\partial \lambda^{\pm}(t, s)}{\partial s} &= \pm \operatorname{sign}(\alpha \beta - t), \\ \operatorname{sign} \frac{\partial \lambda^{\pm}(t, s)}{\partial t} &= \pm \operatorname{sign} [(2 - \alpha^{-1}s)\lambda^{\pm}(t, s) - 1].\end{aligned}$$

В случае $0 < \alpha < \delta$, $\alpha\beta > \Gamma$ выполнены неравенства

$$\lambda^{-}(t, s) < \frac{s}{\alpha}, \quad \lambda^{+}(t, s) < \frac{s}{\alpha},$$

с учетом которых получаем

$$\frac{\partial \lambda^{-}(t, s)}{\partial s} < 0, \quad \frac{\partial \lambda^{-}(t, s)}{\partial t} > 0.$$

Следовательно, в этом случае справедлива оценка

$$\sigma(S) \subseteq \left[\lambda^{-}(\gamma, \Delta), \frac{\Delta}{\alpha} \right].$$

В случае $\alpha > \Delta$, $\alpha\beta < \gamma$ имеют место неравенства

$$\frac{s}{\alpha} < \lambda^{-}(t, s), \quad \frac{s}{\alpha} < \lambda^{+}(t, s),$$

с учетом которых получаем

$$\frac{\partial \lambda^{+}(t, s)}{\partial s} < 0, \quad \frac{\partial \lambda^{+}(t, s)}{\partial t} > 0.$$

Следовательно, в этом случае справедлива оценка

$$\sigma(S) \subseteq \left[\frac{\delta}{\alpha}, \lambda^{+}(\Gamma, \delta) \right].$$

Теорема доказана. ■

Оптимизация полученной оценки спектра приводит к основному результату. Имеет место

Теорема 10.2.3. Пусть выполнены условия теоремы 10.2.2. Тогда оптимальным выбором параметров являются значения

$$\alpha_0 = \Delta, \quad \beta_0 = \frac{\Gamma}{\Delta},$$

при этом

$$q_{\kappa_2} = \frac{1 - \sqrt{\eta_0}}{1 + \sqrt{\eta_0}}, \quad \eta_0 = 2\omega\xi[2 - \omega + \omega\xi + \sqrt{(2 - \omega + \omega\xi)^2 - 4\xi}]^{-1},$$

где $\omega = \delta/\Delta$, $\xi = \gamma/\Gamma$.

Доказательство. Рассмотрим случай $0 < \alpha \leq \delta$, $\alpha\beta \geq \Gamma$, тогда для величины η из (10.3) справедливо представление

$$\eta(\alpha, \beta) = \frac{1}{2} + \frac{\gamma}{\beta} \frac{2\alpha - \Delta}{2\alpha\Delta} - \sqrt{\left(\frac{1}{2} + \frac{\gamma}{\beta} \frac{2\alpha - \Delta}{2\alpha\Delta}\right)^2 - \frac{\alpha\gamma}{\beta\Delta^2}}.$$

и имеют место неравенства

$$\begin{aligned} \operatorname{sign} \frac{\partial \eta(\alpha, \beta)}{\partial \beta^{-1}} &= \operatorname{sign} \left(1 - \frac{\Delta}{\alpha} \left(2 - \frac{\Delta}{\alpha} \right) \eta(\alpha, \beta) \right) > 0, \\ \operatorname{sign} \frac{\partial \eta(\alpha, \Gamma\alpha^{-1})}{\partial \alpha} &= \operatorname{sign} \left(\frac{\alpha}{\Delta} - \eta(\alpha, \Gamma\alpha^{-1}) \right) > 0. \end{aligned}$$

Таким образом, выполнены равенства

$$\begin{aligned} \eta_1 &\equiv \min_{\substack{0 < \alpha \leq \delta, \\ \alpha\beta \geq \Gamma}} \eta(\alpha, \beta) = \min_{0 < \alpha \leq \delta} \eta(\alpha, \Gamma\alpha^{-1}) = \\ &= \eta(\delta, \Gamma\delta^{-1}) = \frac{1}{2} \left[1 - \xi + 2\omega\xi - \sqrt{(1 - \xi + 2\omega\xi)^2 - 4\omega^2\xi} \right]. \end{aligned}$$

В случае $\alpha \geq \Delta$, $0 < \alpha\beta \leq \gamma$ справедливо представление

$$\eta(\alpha, \beta) = \left[\frac{1}{2} + \frac{\Gamma}{\beta} \frac{2\alpha - \delta}{2\alpha\delta} + \sqrt{\left(\frac{1}{2} + \frac{\Gamma}{\beta} \frac{2\alpha - \delta}{2\alpha\delta}\right)^2 - \frac{\alpha\Gamma}{\beta\delta^2}} \right]^{-1}$$

и имеют место неравенства

$$\begin{aligned} \operatorname{sign} \frac{\partial \eta(\alpha, \beta)^{-1}}{\partial \beta^{-1}} &= \operatorname{sign} \left(\frac{\delta}{\alpha} \left(2 - \frac{\delta}{\alpha} \right) \eta(\alpha, \beta)^{-1} - 1 \right) > 0, \\ \operatorname{sign} \frac{\partial \eta(\alpha, \gamma\alpha^{-1})^{-1}}{\partial \alpha} &= \operatorname{sign} \left(\frac{\alpha}{\delta} - \eta(\alpha, \gamma\alpha^{-1})^{-1} \right) < 0. \end{aligned}$$

Таким образом, выполнены равенства

$$\begin{aligned} \eta_2 &\equiv \min_{\substack{\alpha \geq \Delta, \\ 0 < \alpha\beta \leq \gamma}} \eta(\alpha, \beta) = \min_{\alpha \geq \Delta} \eta(\alpha, \gamma\alpha^{-1}) = \\ &= \eta(\Delta, \Gamma\Delta^{-1}) = 2\omega\xi \left[2 - \omega + \omega\xi + \sqrt{(2 - \omega + \omega\xi)^2 - 4\xi} \right]^{-1}. \end{aligned}$$

Для завершения доказательства достаточно проверить неравенство $\eta_1 \leq \eta_2$: запишем его в эквивалентной форме

$$\begin{aligned} &\frac{2 - \omega + \omega\xi + \sqrt{(2 - \omega + \omega\xi)^2 - 4\xi}}{2\omega\xi} \leq \\ &\leq \frac{1 - \xi + 2\omega\xi + \sqrt{(1 - \xi + 2\omega\xi)^2 - 4\omega^2\xi}}{2\omega^2\xi}, \end{aligned}$$

что после элементарных преобразований дает

$$\begin{aligned} 2\omega - \omega^2 + \omega^2\xi + \sqrt{(2\omega - \omega^2 + \omega^2\xi)^2 - 4\omega^2\xi} &\leq \\ &\leq 1 - \xi + 2\omega\xi + \sqrt{(1 - \xi + 2\omega\xi)^2 - 4\omega^2\xi}. \end{aligned}$$

Последнее неравенство немедленно следует из соотношения

$$2\omega - \omega^2 + \omega^2\xi \leq 1 - \xi + 2\omega\xi,$$

которое имеет место при любых $\xi \in (0, 1]$. Теорема доказана. ■

Следствие 10.2.1. При выполнении условий теоремы 10.2.2 имеют место асимптотические равенства

$$\begin{aligned} q_{K_2} &= 1 - \sqrt{\frac{4\omega}{2-\omega}}\xi + O(\xi), & \omega &= \text{const}, & \xi &\rightarrow 0, \\ q_{K_2} &= 1 - \sqrt{\frac{4\xi}{1+\sqrt{1-\xi}}}\omega + O(\omega), & \xi &= \text{const}, & \omega &\rightarrow 0. \end{aligned}$$

Доказательство. Искомые формулы несложно получить разлагая представление для q_{K_2} из теоремы 10.2.3, в ряды Тейлора по ξ и ω соответственно. Следствие доказано. ■

10.2.3. Оценка в классе K_{2s}

Следуя подходу, приведенному в 6.3, получим оценку для класса K_{2s} . Имеет место

Теорема 10.2.4. Пусть

$$A = A_0 + \delta/\gamma BC^{-1}B^*, \quad (A_0, B, Q, C) \in K_{2s}(\delta, \Delta, \gamma, \Gamma),$$

тогда

$$q_{K_{2s}} \leq \frac{1 - \sqrt{\eta_0}}{1 + \sqrt{\eta_0}}, \quad \eta_0 = 2\omega\xi \left[2 - \omega + \omega\xi + \sqrt{(2 - \omega + \omega\xi)^2 - 4\xi} \right]^{-1},$$

где $\omega = (\Delta/\delta + \Gamma/\gamma)^{-1}$, $\xi = \gamma/\Gamma$.

Доказательство. Из (6.17) следует соотношение

$$(A, B, Q, C) \in K_2\left(\delta, \Delta + \frac{\delta}{\xi}, \gamma, \Gamma\right),$$

что в совокупности с выбором

$$\alpha_0 = \Delta + \delta\frac{\Gamma}{\gamma}, \quad \beta_0 = \Gamma\frac{\gamma}{\Delta\gamma + \delta\Gamma}$$

и теоремой 10.2.3 приводит к требуемому результату. Теорема доказана. ■

Следствие 10.2.2. При выполнении условий теоремы 10.2.4 имеет место асимптотическая оценка

$$q_{k_2} \leq 1 - \sqrt{2}\xi + O(\xi^2), \quad \omega = \text{const}, \quad \xi \rightarrow 0.$$

Доказательство. Искомую формулу несложно получить из разложения представления для q_{k_2} из теоремы 10.2.4 в ряд Тейлора. Следствие доказано. ■

10.3. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Видимо, впервые идея применения модельных седловых операторов была предложена в работе [40], где и приводились первые оценки скорости сходимости для неконструктивного подхода. Кроме того, там же предлагался конструктивный способ построения седловых предобусловливателей на основе блочно треугольной факторизации.

Большой интерес вызвали и в дальнейшем получили развитие результаты работы [166], где была проанализирована обобщенная спектральная задача, в которой предобусловливающий седловой оператор отличался от исходного только матрицей Q на месте матрицы A . Главный вывод работы — рекомендация использовать методы подпространств Крылова типа MINRES и GMRES [183, с. 134, с. 158], учитывающие специфику неконструктивного седлового предобусловливания.

Современное оформление идея конструктивного подхода получила в работе [113], где выведены оценки сходимости для алгоритмов с такими предобусловливателями. Впоследствии их удалось уточнить в работе [189]. Окончательные оценки при различном масштабировании постоянных в матричных неравенствах

$$\delta Q \leq A \leq \Delta Q, \quad \gamma C \leq B^* Q^{-1} B \leq \Gamma C$$

получены в [202].

Отличительной особенностью всех этих работ является отсутствие оптимизации предобусловливателя относительно каких-либо параметров. Впервые попытка проследить влияние свободных параметров на спектр предобусловленного оператора сделана в [194]. Один параметр вводился в предобусловливатель C , т. е. оператор C заменялся на αC , а вторым параметром был множитель β при слагаемом $BC^{-1}B^*$, которое добавлялось как в матрицу исходной задачи, так и в предобусловливатель. Однако оптимизация числа обусловленности также не производилась.

В предположении, что ядро оператора G инвариантно относительно оператора L , задачи минимизации числа обусловленности

были поставлены и решены при неконструктивном подходе в работах [34] и [35], а при конструктивном — в [36] и [98].

Результаты работы [36] представляют особый интерес. В частности, показано, что оптимальным значением параметра β является нуль и масштабирование матриц сильно влияет на спектральное число предобусловленного оператора. Если обозначить

$$O_1 = \text{cond}_2(Q^{-1}A), \quad O_2 = \text{cond}_2(C^{-1}B^*Q^{-1}B),$$

то в случае $1 \leq \delta < \Delta$, $\gamma < \Gamma \leq 1$ имеем

$$\text{cond}_2 S \approx O_1^2 O_2,$$

а при $\delta < \Delta \leq 1$, $1 \leq \gamma < \Gamma$ —

$$\text{cond}_2 S \approx O_1 O_2.$$

При отсутствии оптимизации ($\alpha = 1$, $\beta = 0$) полученные оценки совпадают с результатами [202].

МЕТОДЫ ПОПЕРЕМЕННЫХ СИММЕТРИЧНЫХ И КОСОСИММЕТРИЧНЫХ ИТЕРАЦИЙ

В общей теории итерационных методов решения линейных систем хорошо известна схема метода попеременных итераций, которая заключается в следующем. Пусть решается невырожденная система линейных уравнений

$$Sz = F,$$

причем матрица S может быть представлена как сумма двух (или более) матриц

$$S = S_1 + S_2,$$

обладающих определенными свойствами, а именно — возможностью эффективного обращения матриц вида $I + \alpha S_1$, $I + \alpha S_2$ при $\alpha \neq 0$. Тогда для численного решения исходной задачи можно использовать итерационный метод

$$\begin{cases} (I + \alpha S_1)z^{k+1/2} = (I - \alpha S_2)z^k + \alpha F, \\ (I + \alpha S_2)z^{k+1} = (I - \alpha S_1)z^{k+1/2} + \alpha F, \end{cases}$$

где z^k — вектор приближения на k -й итерации, z^0 — заданное начальное приближение.

Перечислим наиболее известные варианты используемых расщеплений: треугольное ($S_1 = L + D/2$, $S_2 = U + D/2$, где $S = L + D + U$ — представление матрицы в виде нижнетреугольной, верхнетреугольной и диагональной), симметрично-кососимметричное ($S_1 = (S + S^*)/2$, $S_2 = (S - S^*)/2$), метод переменных направлений (при условии $S_1 S_2 = S_2 S_1$).

При решении седловых задач можно использовать блочные варианты методов. Так блочный вариант попеременно-треугольного метода приводит к методу MSSOR (см. раздел 2.4). В настоящей главе рассматривается другой вариант блочного расщепления — соответствующий методу попеременных симметрично-кососимметричных итераций.

11.1. СТАЦИОНАРНЫЙ МЕТОД (GPHSSI)

11.1.1. Формулировка метода

Задача с седловой точкой (6.2) может быть представлена в эквивалентном виде

$$(H + J)z \equiv \begin{pmatrix} A & B \\ -B^* & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ -\varphi \end{pmatrix} \equiv F, \quad (11.1)$$

где $H = H^*$ — самосопряженный оператор, $J^* = -J$ — кососимметричный оператор.

Для решения задачи (11.1) рассмотрим итерационный процесс

$$\begin{cases} (M + H)z^{k+1/2} = (M - J)z^k + F, \\ (M + J)z^{k+1} = (M - H)z^{k+1/2} + F, \end{cases} \quad (11.2)$$

где

$$M = \begin{pmatrix} \alpha Q & 0 \\ 0 & \alpha^{-1}\beta C \end{pmatrix}, \quad \alpha > 0, \quad \beta > 0.$$

В англоязычной литературе этому методу присвоена аббревиатура GPHSSI — Generalized Preconditioned Hermitian-Skewhermitian Iterations.

11.1.2. Безусловная сходимость метода

Замечательная особенность метода GPHSSI заключается в том, что метод является безусловно сходящимся, т. е. сходится для любого начального приближения и независимо от выбора итерационных параметров. Справедлива

Теорема 11.1.1. Метод GPHSSI определен корректно и сходится для любого начального приближения.

Доказательство. Для того чтобы обосновать корректность метода, необходимо убедиться, что операторы $M + H$ и $M + J$ невырождены. Оператор

$$M + H = \begin{pmatrix} \alpha Q + (A + A^*)/2 & 0 \\ 0 & \alpha^{-1}\beta C \end{pmatrix}$$

является самосопряженным и положительно определенным при любых $\alpha > 0$ и $\beta > 0$, а следовательно, $M + H$ невырожден. Оператор

$$M + J = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \begin{pmatrix} \alpha Q + (A - A^*)/2 & B \\ B^* & -\alpha^{-1}\beta C \end{pmatrix}$$

невырожден, так как является произведением невырожденного оператора (слева) и регулярного седлового оператора (справа).

Так как $M = M^* > 0$, то оператор перехода метода можно представить в виде

$$T = (M + J)^{-1} M^{1/2} T_1 T_2 M^{-1/2} (M + J),$$

где

$$\begin{aligned} T_1 &= (I - M^{-1/2} H M^{-1/2})(I + M^{-1/2} H M^{-1/2})^{-1}, \\ T_2 &= (I - M^{-1/2} J M^{-1/2})(I + M^{-1/2} J M^{-1/2})^{-1}. \end{aligned}$$

Из этого представления следует, что

$$T_1 = T_1^*, \quad T_2 T_2^* = I \quad \text{и} \quad \sigma(T) = \overline{\sigma(T_2^{-1} T_1)},$$

причем

$$\begin{aligned} T_1 &= \begin{pmatrix} D & 0 \\ 0 & I \end{pmatrix}, \\ D &= \left(I - \frac{1}{2\alpha} Q^{-1/2} (A + A^*) Q^{-1/2} \right) \left(I + \frac{1}{2\alpha} Q^{-1/2} (A + A^*) Q^{-1/2} \right)^{-1} \end{aligned}$$

Предположим, что $z = \{0, p\}$ — собственный вектор оператора $T_2^{-1} T_1$, тогда

$$T_2^{-1} T_1 z = T_2^{-1} z,$$

т. е. z является собственным вектором оператора T_2 . Из равенства

$$M^{-1/2} J M^{-1/2} (I + T_2) = (I - T_2)$$

имеем, что z — собственный вектор оператора $M^{-1/2} J M^{-1/2}$ и $M^{-1/2} z$ — собственный вектор J , однако из равенства

$$\begin{aligned} J M^{-1/2} z &= \begin{pmatrix} (A - A^*)/2 & B \\ -B^* & \alpha^{-1} \beta C \end{pmatrix} \begin{pmatrix} 0 \\ \alpha^{1/2} \beta^{-1/2} C^{-1/2} p \end{pmatrix} = \\ &= \begin{pmatrix} \alpha^{1/2} \beta^{-1/2} B C^{-1/2} p \\ \alpha^{-1/2} \beta^{1/2} C^{1/2} p \end{pmatrix} \end{aligned}$$

следует, что первая компонента вектора $J M^{-1/2} z$ не равна нулю. Таким образом, для любого собственного вектора $\{u, p\}$ оператора $T_2^{-1} T_1$ справедливо неравенство $u \neq 0$.

Пусть $z = \{u, p\} \neq 0$ — произвольный собственный вектор $T_2^{-1} T_1$, соответствующий собственному значению $\lambda \in \mathbb{C}$, тогда

$$\begin{aligned} |\lambda|^2 \|z\|^2 &= (T_2^{-1} T_1 z, T_2^{-1} T_1 z) = (T_1 z, T_1 z) = \\ &= (Du, u) + (p, p) < (u, u) + (p, p) = \|z\|^2, \end{aligned}$$

где последнее неравенство следует из условий $u \neq 0$ и

$$\sigma(D) = \left\{ \frac{1 - s/(2\alpha)}{1 + s/(2\alpha)} \mid s \in \sigma(Q^{-1/2}AQ^{-1/2}) \subset (0, +\infty) \right\} \subset (-1, 1).$$

Следовательно, $|\lambda| < 1$ и $\rho(T) < 1$. Теорема доказана. ■

11.1.3. Оптимизация в классе \mathbb{K}_2

Пусть $(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, тогда $A = A^* > 0$ и оператор перехода в методе GPHSSI примет вид

$$T = T(\alpha, \beta; A, B, Q, C) = I - 2S,$$

где

$$S = \begin{pmatrix} \alpha Q & B \\ -B^* & \alpha^{-1}\beta C \end{pmatrix}^{-1} \begin{pmatrix} I + \alpha^{-1}AQ^{-1} & 0 \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} A & B \\ -B^* & 0 \end{pmatrix}.$$

Лемма 11.1.1. Число $\mu \in \sigma(S)$ тогда и только тогда, когда $\mu \neq 0$ и существует вектор $u \in U \setminus \{0\}$ такой, что

$$\begin{aligned} & [\mu^2(\alpha Q + A)Q^{-1}(\beta Q + BC^{-1}B^*) - \\ & - \mu(\beta A + 2\alpha BC^{-1}B^* + AQ^{-1}BC^{-1}B^*) + \alpha BC^{-1}B^*]u = 0. \end{aligned} \quad (11.3)$$

Доказательство. Пусть $\mu \in \sigma(S)$, тогда существует вектор

$$z = \{u, p\} \in Z \setminus \{0\}$$

такой, что $Sz = \mu z$, или в развернутой форме

$$\begin{cases} Au + Bp = \mu(I + \alpha^{-1}AQ^{-1})(\alpha Qu + Bp), \\ -B^*u = \mu(-B^*u + \alpha^{-1}\beta Cp). \end{cases} \quad (11.4)$$

Отметим, что $u \neq 0$, так как в противном случае из (11.4) будет следовать, что и $p = 0$. Кроме того, $\mu \neq 0$ в силу невырожденности S . Выразив p из второго уравнения (11.4) и подставив его в первое уравнение (11.4), получим (11.3).

В обратную сторону, пусть $\mu \neq 0$ и $u \in U \setminus \{0\}$ удовлетворяет (11.3). Из (11.3) следует, что $\mu \neq 1$, так как при $\mu = 1$ уравнение (11.3) сводится к $\alpha\beta Qu = 0$, или $u = 0$, что противоречит определению u .

Определим компоненту

$$p = \frac{\mu - 1}{\mu\alpha^{-1}\beta} C^{-1}B^*u.$$

Тогда вектор $\{u, p\}$ удовлетворяет второму уравнению (11.4) и

$$BC^{-1}B^*u = \frac{\mu\alpha^{-1}\beta}{\mu-1}Bp.$$

Подставив это выражение в (11.3), получим первое уравнение (11.4). Лемма доказана. ■

Теорема 11.1.2. Пусть

$$0 < \delta \leq \Delta, \quad 0 < \gamma \leq \Gamma, \\ (A, B, Q, C) \in \mathbb{K}_2 = \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma), \quad \alpha, \beta > 0,$$

тогда

$$\rho(T(\alpha, \beta; A, B, Q, C)) \leq \rho(\alpha, \beta), \quad q_{\mathbb{K}_2} = \inf_{\alpha, \beta > 0} \rho(\alpha, \beta),$$

где

$$\rho(\alpha, \beta) = \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ |1 - 2\mu_1|, |1 - 2\mu_2^{1,2}| \right\},$$

$\mu_1 = s/(s + \alpha)$, $\mu_2^{1,2}$ — корни квадратного уравнения

$$\mu^2(\alpha + s)(\beta + t) - \mu(\beta s + 2\alpha t + st) + \alpha t = 0.$$

Доказательство. Пусть $(A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, положим

$$L = Q^{-1/2}AQ^{-1/2}, \quad G = Q^{-1/2}BC^{-1}B^*Q^{-1/2}, \\ \delta_1 = \delta, \quad \delta_2 = \Delta, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma,$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 11.1.1, число $\lambda \in \sigma(T)$ тогда и только тогда, когда $\lambda \neq 1$ и $\lambda \in \sigma(\chi)$, где

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) = (1 - \lambda)(\alpha + s) - 2s, \quad g(\lambda, t) = (1 - \lambda)(\beta + t), \quad h(\lambda) = 4\alpha\lambda.$$

Таким образом, выполнены все условия следствия 6.4.2, следовательно,

$$\rho(T) \leq \max_{\substack{s \in [\delta_1, \delta_2] \\ t \in [\gamma_1, \gamma_2]}} \{|1 - 2\mu(s, t)|\} = \rho(\alpha, \beta),$$

где максимум берется по всем $\mu(s, t) = (1 - \lambda(s, t))/2$, удовлетворяющим уравнениям

$$\mu(\alpha + s) - s = 0, \quad \mu^2(\alpha + s)(\beta + t) - \mu(\beta s + 2\alpha t + st) + \alpha t = 0.$$

В силу теоремы 6.4.7, полученная оценка неулучшаема в классе $\mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$, поэтому

$$q_{\mathbb{K}_2} = \inf_{\alpha, \beta > 0} \rho(\alpha, \beta).$$

Теорема доказана. ■

Используя обозначения теоремы 11.1.2, определим следующие функции:

$$\begin{aligned} \lambda(t, s) &\equiv \lambda(t, s; \alpha, \beta) = \max_{1,2} \left\{ \left| 1 - 2\mu_2^{1,2} \right| \right\}, \\ D(t, s) &\equiv D(t, s; \alpha, \beta) = (\beta s + 2\alpha t + st)^2 - 4\alpha t(\alpha + s)(\beta + t), \\ \Lambda(\alpha, \beta) &= \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \lambda(t, s), \end{aligned}$$

где $D(t, s)$ — дискриминант соответствующего квадратного уравнения, и докажем вспомогательные утверждения.

Лемма 11.1.2. *Имеет место равенство*

$$\max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda(t, s) = \Lambda(\alpha, \beta).$$

Доказательство. Простая проверка показывает, что $\lambda \in F_{1, \alpha\beta}^\uparrow$ как функция t при фиксированных $\alpha, \beta, s > 0$ или же $\lambda \in F_{1, \alpha}^\uparrow$ как функция s при фиксированных $\alpha, \beta, t > 0$. Используя свойство 2 теоремы 6.4.10, получаем

$$\begin{aligned} \max_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda(t, s) &= \max_{s \in [\delta, \Delta]} \max_{t \in [\gamma, \Gamma]} \lambda(t, s) = \max_{s \in [\delta, \Delta]} \max_{t=\gamma, \Gamma} \lambda(t, s) = \\ &= \max_{t=\gamma, \Gamma} \max_{s \in [\delta, \Delta]} \lambda(t, s) = \max_{t=\gamma, \Gamma} \max_{s=\delta, \Delta} \lambda(t, s) = \Lambda(\alpha, \beta). \end{aligned}$$

Лемма доказана. ■

Лемма 11.1.3. *Для любого $\alpha, \beta > 0$ справедливо неравенство*

$$\Lambda(\alpha, \sqrt{\gamma\Gamma}) \leq \Lambda(\alpha, \beta).$$

Доказательство. Отметим некоторые свойства величины

$$\lambda(t, s; \alpha, \beta) = \begin{cases} \left| \frac{\beta - t}{\beta + t} \right| \frac{\alpha}{\alpha + s} + \sqrt{\left(\frac{\beta - t}{\beta + t} \frac{\alpha}{\alpha + s} \right)^2 - \frac{\alpha - s}{\alpha + s}} & \text{при } D(t, s; \alpha, \beta) > 0, \\ \sqrt{\frac{\alpha - s}{\alpha + s}} & \text{при } D(t, s; \alpha, \beta) \leq 0 \end{cases}$$

при фиксированных $\alpha > 0$, t и s :

- 1) $\lambda(t, s; \alpha, \beta)$ не зависит от β при $D(t, s; \alpha, \beta) \leq 0$;
- 2) $\lambda(t, s; \alpha, \beta)$ строго возрастает в области $D(t, s; \alpha, \beta) > 0$ при строгом возрастании величины $\left| \frac{\beta - t}{\beta + t} \right|$.

Отсюда следует, что минимум $\Lambda(\alpha, \beta)$ достигается в точке, которая является решением следующей задачи:

$$\beta_0 = \arg \min_{t=\gamma, \Gamma} \left| \frac{\beta - t}{\beta + t} \right|.$$

Несложно убедиться, что β_0 удовлетворяет уравнению

$$\frac{\beta_0 - \gamma}{\beta_0 + \gamma} = - \frac{\beta_0 - \Gamma}{\beta_0 + \Gamma},$$

откуда следует, что $\beta_0 = \sqrt{\gamma\Gamma}$. Лемма доказана. ■

Получим основной результат. Справедлива

Теорема 11.1.3 (Асимптотическая оптимизация в K_2).
Задача асимптотической оптимизации метода GPHSSI в классе $K_2(\delta, \Delta, \gamma, \Gamma)$ имеет решение

$$\alpha_2 = \frac{\sqrt{\delta\Delta}}{2} \left(\sqrt[4]{\xi} + \frac{1}{\sqrt[4]{\xi}} \right) \left(1 - \frac{1}{2}\omega\theta^2 + \theta\sqrt{1 - \omega + \frac{1}{4}\omega^2\theta^2} \right)^{1/2},$$

$$\beta_2 = \sqrt{\gamma\Gamma}, \quad q_{K_2} = \sqrt{\frac{\alpha_2 - \delta}{\alpha_2 + \delta}},$$

где

$$\omega = \frac{\delta}{\Delta} \leq 1, \quad \xi = \frac{\gamma}{\Gamma} \leq 1, \quad \theta = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Если одно из неравенств $\omega \leq 1$ и $\xi \leq 1$ выполняется строго, то решение единственно.

Доказательство. Для доказательства теоремы достаточно рассмотреть случай $\omega < 1$ или $\xi < 1$. Из теоремы 11.1.2 и лемм 11.1.2, 11.1.3 следует, что

$$q_{K_2} = \min_{\alpha > 0} \max \left\{ \left| \frac{\alpha - \delta}{\alpha + \delta} \right|, \left| \frac{\alpha - \Delta}{\alpha + \Delta} \right|, \Lambda(\alpha, \sqrt{\gamma\Gamma}) \right\},$$

причем при любых $t = \gamma, \Gamma$, $s = \delta, \Delta$, $\alpha > 0$ справедливо

$$\lambda(t, s; \alpha, \sqrt{\gamma\Gamma}) = \left| \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}} \frac{\alpha}{\alpha + s} + \sqrt{\left(\frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}} \frac{\alpha}{\alpha + s} \right)^2 - \frac{\alpha - s}{\alpha + s}} \right|.$$

При фиксированных t, s функция $f(\alpha) \equiv \lambda(t, s; \alpha, \sqrt{\gamma\Gamma})$ принадлежит $F_{1,s}$, причем $f(\alpha)$ строго убывает в интервале $(0, \alpha_0)$ и строго возрастает в $(\alpha_0, +\infty)$, где $\alpha_0 = s(\xi^{1/4} + \xi^{-1/4})/2$. Следовательно, $\Lambda(\alpha, \sqrt{\gamma\Gamma}) \in \text{YS}(0, +\infty)$ как функция α .

Несложно убедиться, что α_2 является единственным решением следующей задачи:

$$\begin{cases} \lambda(\gamma, \delta; \alpha, \sqrt{\gamma\Gamma}) = \lambda(\gamma, \Delta; \alpha, \sqrt{\gamma\Gamma}), \\ D(\gamma, \delta; \alpha, \sqrt{\gamma\Gamma}) \leq 0, \\ D(\gamma, \Delta; \alpha, \sqrt{\gamma\Gamma}) \geq 0. \end{cases}$$

Отсюда имеем, что величина $\lambda(\gamma, \delta; \alpha, \sqrt{\gamma\Gamma})$ строго убывает при $\alpha \in (0, \alpha_2)$, а величина $\lambda(\gamma, \Delta; \alpha, \sqrt{\gamma\Gamma})$ строго возрастает при $\alpha \in (\alpha_2, +\infty)$. Таким образом, α_2 является строгим локальным минимумом функции $\Lambda(\alpha, \sqrt{\gamma\Gamma})$, а значит и единственным ее глобальным минимумом.

Для любого $s \in [\delta, \Delta]$ выполнены неравенства

$$\frac{\alpha_2 - \Delta}{\alpha_2 + \Delta} \leq 1 - 2 \frac{s}{\alpha_2 + s} \leq \frac{\alpha_2 - \delta}{\alpha_2 + \delta} = q_{\mathbb{K}_2}^2.$$

Если $\alpha_2 \geq \Delta$, то справедливы соотношения

$$-q_{\mathbb{K}_2} < 0 \leq \frac{\alpha_2 - \Delta}{\alpha_2 + \Delta} \leq \frac{\alpha_2 - \delta}{\alpha_2 + \delta} = q_{\mathbb{K}_2}^2 < q_{\mathbb{K}_2} < 1.$$

В случае $\alpha_2 < \Delta$ нижняя оценка следует из неравенства $\alpha_2 \geq \sqrt{\delta\Delta}$:

$$-1 < -q_{\mathbb{K}_2} < -q_{\mathbb{K}_2}^2 \leq -\frac{\sqrt{\Delta} - \sqrt{\delta}}{\sqrt{\Delta} + \sqrt{\delta}} \leq \frac{\alpha_2 - \Delta}{\alpha_2 + \Delta} \leq \frac{\alpha_2 - \delta}{\alpha_2 + \delta} = q_{\mathbb{K}_2}^2 < q_{\mathbb{K}_2} < 1.$$

Отсюда немедленно получаем, что

$$q_{\mathbb{K}_2} = \Lambda(\alpha_2, \beta_2).$$

Докажем единственность оптимальных параметров. Предположим, что существует другой набор оптимальных параметров — (α_1, β_1) . В силу леммы 11.1.3, параметры $(\alpha_1, \sqrt{\gamma\Gamma})$ тоже являются оптимальными и, следовательно, $\alpha_1 = \alpha_2$. Если $\beta_1 > \beta_2$, то справедливо

$$\left| \frac{\beta_1 - \gamma}{\beta_1 + \gamma} \right| > \left| \frac{\beta_2 - \gamma}{\beta_2 + \gamma} \right|, \quad D(\gamma, \Delta; \alpha_2, \beta_1) > D(\gamma, \Delta; \alpha_2, \beta_2) \geq 0,$$

и, в силу свойств $\lambda(t, s; \alpha, \beta)$ (см. доказательство леммы 11.1.3), имеет место неравенство

$$\lambda(\gamma, \Delta; \alpha_2, \beta_1) > \lambda(\gamma, \Delta; \alpha_2, \beta_2).$$

Аналогично, если $\beta_1 < \beta_2$, то выполнено

$$\lambda(\Gamma, \Delta; \alpha_2, \beta_1) > \lambda(\Gamma, \Delta; \alpha_2, \beta_2).$$

Таким образом, приходим к оценке

$$\begin{aligned} \Lambda(\alpha_1, \beta_1) = \Lambda(\alpha_2, \beta_1) &\geq \max\{\lambda(\gamma, \Delta; \alpha_2, \beta_1), \lambda(\Gamma, \Delta; \alpha_2, \beta_1)\} > \\ &> \Lambda(\alpha_2, \beta_2), \end{aligned}$$

что противоречит определению α_1, β_1 . Теорема доказана. ■

Следствие 11.1.1. Имеют место асимптотики

$$\begin{aligned} q_{K_2} &= 1 - \sqrt{\frac{2\sqrt{\xi}}{1 + \sqrt{\xi}}} \sqrt{\omega} + O(\omega) \quad \text{при } \omega \rightarrow 0, \\ q_{K_2} &= 1 - \sqrt{\frac{4\omega}{2 - \omega}} \sqrt[4]{\xi} + O(\sqrt{\xi}) \quad \text{при } \xi \rightarrow 0, \omega = \text{const}. \end{aligned}$$

Доказательство. Формулы немедленно следуют из разложения в ряд Пуансо представления для q_{K_2} , полученного в теореме 11.1.3. Следствие доказано. ■

11.2. HSS-ПРЕДОБУСЛОВЛИВАНИЕ

11.2.1. Построение методов

Метод GPHSSI, рассмотренный выше, можно рассматривать как метод Ричардсона ($T = I - 2S$) для решения предобусловленной седловой задачи $Sz = F$. При определенных условиях, а именно:

$$A = A^* > \alpha Q > 0$$

оператор S является симметризуемым и обладает вещественным положительным спектром. Это позволяет эффективно ускорить метод GPHSSI при помощи методов чебышевского типа или проекционных методов (GMRES, BICGSTAB и т. п., см. [183]).

Важную роль в анализе и оптимизации итерационных алгоритмов указанного типа играет задача нахождения или оценки многочлена с единичным младшим коэффициентом, минимально уклоняющегося от нуля на спектре оператора S , или, формально,

$$\theta_n(T) = \min_{\substack{P_n(x) \in \mathbb{R}[x], \\ \deg P_n \leq n, P_n(0)=1}} \max_{x \in \sigma(T)} |P_n(x)|, \quad n \in \mathbb{N}.$$

Например, в случае, когда спектр оператора лежит на отрезке $[a, b]$, $a > 0$ хорошо известна (в некотором смысле неулучшаемая) оценка

(см. [61, с. 84])

$$\theta_n(T) \leq \min_{\substack{P_n(x) \in \mathbb{R}[x], \\ \deg P_n \leq n, P_n(0)=1}} \max_{x \in [a,b]} |P_n(x)| = \frac{2q^n}{1+q^{2n}} \leq 2q^n, \quad (11.5)$$

где

$$q \equiv q\left(\frac{a}{b}\right) = \frac{1 - \sqrt{a/b}}{1 + \sqrt{a/b}} < 1,$$

причем решение указанной минимаксной задачи единственно и дается нормированным многочленом Чебышева I-рода на отрезке $[a, b]$:

$$C_k^{a,b}(x) = \prod_{j=1}^n \left(1 - \frac{2x}{b+a+(b-a)\cos(\pi(j-1/2)/k)} \right).$$

Несложно заметить, что чем большее значение принимает величина a/b , тем лучше оценка (11.5). В этом случае задача оптимизации оценки спектра на параметризованном семействе операторов сводится к нахождению оператора (т. е. набора итерационных параметров), у которого величина отношения a/b принимает наибольшее значение.

Имеет место

Теорема 11.2.1 (Обоснование методов). Пусть $\alpha > 0$, $\beta > 0$, $A = A^* > \alpha Q$, тогда существует оператор W такой, что $W^{-1}SW$ — самосопряженный оператор.

Доказательство. Определим операторы

$$\begin{aligned} R_1 &= M^{-1/2} \begin{pmatrix} \alpha Q & B \\ B^* & -\alpha^{-1}\beta C \end{pmatrix} M^{-1/2} = \begin{pmatrix} I & G \\ G^* & -I \end{pmatrix}, \\ R_2 &= M^{-1/2} \begin{pmatrix} I + \alpha^{-1}AQ^{-1} & 0 \\ 0 & I \end{pmatrix} M^{1/2} = \begin{pmatrix} I + L & 0 \\ 0 & I \end{pmatrix}, \\ D &= M^{-1/2} \begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} M^{-1/2} = \begin{pmatrix} L & G \\ G^* & 0 \end{pmatrix}, \end{aligned}$$

где

$$\begin{aligned} L &= \alpha^{-1}Q^{-1/2}AQ^{-1/2}, \\ G &= \beta^{-1/2}Q^{-1/2}BC^{-1/2}, \end{aligned}$$

тогда справедливо представление

$$S = M^{-1/2}R_1^{-1}R_2^{-1}DM^{1/2}.$$

Так как $L = L^* > I$, то выполнены соотношения

$$R_2 > 0, \quad D - R_1 > 0, \quad R_2(D - R_1) = (D - R_1)R_2.$$

Положим

$$W = M^{-1/2} R_1^{-1} R_2^{-1/2} (D - R_1)^{1/2},$$

тогда из представления

$$\begin{aligned} W^{-1} S W &= (D - R_1)^{-1/2} R_2^{-1/2} D R_1^{-1} R_2^{-1/2} (D - R_1)^{1/2} = \\ &= R_2^{-1/2} (D - R_1)^{-1/2} D R_1^{-1} (D - R_1)^{1/2} R_2^{-1/2} = \\ &= R_2^{-1/2} (D - R_1)^{1/2} R_1^{-1} (D - R_1)^{1/2} R_2^{-1/2} + R_2^{-1} \end{aligned}$$

следует самосопряженность $W^{-1} S W$. Теорема доказана. \blacksquare

11.2.2. Оценка спектра в классе \mathbb{K}_2

Спектр преобусловленного оператора имеет интересную особенность — он сосредоточен в трех отрезках I_1, I_2, I_3 . Оценим их границы. Сначала потребуется

Лемма 11.2.4. Для любых $\alpha \neq 0, \beta \neq 0, (A, B, Q, C) \in \mathbb{K}_2(\delta, \Delta, \gamma, \Gamma)$ имеет место оценка

$$\sigma(S) \cap \mathbb{R} \subseteq \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \left\{ \lambda_1(s, t), \lambda_2^{1,2}(s, t) \right\},$$

где $\lambda_1 = s/(s + \alpha)$, $\lambda_2^{1,2}$ — корни квадратного уравнения

$$\lambda^2(\alpha + s)(\beta + t) - \lambda(\beta s + 2\alpha t + st) + \alpha t = 0.$$

Доказательство. Положим

$$\begin{aligned} L &= Q^{-1/2} A Q^{-1/2}, \quad G = Q^{-1/2} B C^{-1} B^* Q^{-1/2}, \\ \delta_1 &= \delta, \quad \delta_2 = \Delta, \quad \gamma_1 = \gamma, \quad \gamma_2 = \Gamma, \end{aligned}$$

тогда

$$L = L^*, \quad \sigma(L) \subseteq [\delta_1, \delta_2], \quad G = G^*, \quad \sigma(G) \subseteq \{0\} \cup [\gamma_1, \gamma_2].$$

В силу леммы 11.1.1, число $\lambda \in \sigma(S)$ тогда и только тогда, когда $\lambda \neq 0$ и $\lambda \in \sigma(\chi)$, где

$$\begin{aligned} \chi(\lambda) &= f(\lambda, L)g(\lambda, G) + h(\lambda)G, \\ f(\lambda, s) &= \lambda(\alpha + s) - s, \quad g(\lambda, t) = \lambda(\beta + t), \quad h(\lambda) = \alpha(1 - \lambda). \end{aligned}$$

Таким образом, выполнены все условия следствия 6.4.1, откуда и следует утверждение леммы. Лемма доказана. \blacksquare

Границы отрезков, которым принадлежит спектр S , определяет

Теорема 11.2.2. Пусть $0 < \alpha \leq \delta$, $\beta > 0$, тогда

$$\sigma(S) \subseteq I_1 \cup I_2 \cup I_3 \subset (0, +\infty), \quad (11.6)$$

где

$$\begin{aligned} I_1 &= \left[\frac{\Delta}{2(\Delta + \alpha)} \left(1 + 2\frac{\alpha}{\Delta} \frac{\gamma}{\gamma + \beta} - \sqrt{1 - \frac{\alpha^2}{\Delta^2} \frac{4\beta\gamma}{(\gamma + \beta)^2}} \right), \right. \\ &\quad \left. \frac{\delta}{2(\delta + \alpha)} \left(1 + 2\frac{\alpha}{\delta} \frac{\Gamma}{\Gamma + \beta} - \sqrt{1 - \frac{\alpha^2}{\delta^2} \frac{4\beta\Gamma}{(\Gamma + \beta)^2}} \right) \right], \\ I_2 &= \left[\frac{\delta}{\delta + \alpha}, \frac{\Delta}{\Delta + \alpha} \right], \\ I_3 &= \left[\frac{\delta}{2(\delta + \alpha)} \left(1 + 2\frac{\alpha}{\delta} \frac{\gamma}{\gamma + \beta} + \sqrt{1 - \frac{\alpha^2}{\delta^2} \frac{4\beta\gamma}{(\gamma + \beta)^2}} \right), \right. \\ &\quad \left. \frac{\Delta}{2(\Delta + \alpha)} \left(1 + 2\frac{\alpha}{\Delta} \frac{\Gamma}{\Gamma + \beta} + \sqrt{1 - \frac{\alpha^2}{\Delta^2} \frac{4\beta\Gamma}{(\Gamma + \beta)^2}} \right) \right]. \end{aligned}$$

Доказательство. В силу теоремы 11.2.1, при $\alpha < \delta$ существует оператор W такой, что $W^{-1}SW$ — самосопряжен. Следовательно, выполнено равенство

$$\sigma(S) = \sigma(W^{-1}SW) \subset \mathbb{R}. \quad (11.7)$$

Таким образом, в силу непрерывной зависимости спектра от $\alpha > 0$, оператор S обладает вещественным спектром при $\alpha \leq \delta$.

В этом случае из леммы 11.2.4 следует оценка

$$\sigma(S) = \sigma(S) \cap \mathbb{R} \subseteq \bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \{ \lambda_1(x(s)), \lambda_2^\pm(x(s), y(t)) \},$$

где

$$\begin{aligned} \lambda_1(x) &= \frac{1}{1+x}, \quad \lambda_2^\pm(x, y) = \frac{1}{2} \frac{1}{1+x} \left[1 + \frac{2x}{1+y} \pm \sqrt{1 - \frac{4x^2y}{(1+y)^2}} \right], \\ x(s) &= \frac{\alpha}{s} \in \left[\frac{\alpha}{\Delta}, \frac{\alpha}{\delta} \right] \subset (0, 1], \quad y(t) = \frac{\beta}{t} \in \left[\frac{\beta}{\Gamma}, \frac{\beta}{\gamma} \right] \subset (0, +\infty). \end{aligned}$$

При $x \in (0, 1]$ и $y \in (0, +\infty)$ справедливо $\lambda_1(x) \in (0, +\infty)$, а из неравенств

$$0 < \frac{4x^2y}{(1+y)^2} \leq \frac{4y}{(1+y)^2} = 1 - \left(\frac{1-y}{1+y} \right)^2 \leq 1$$

следует, что и $\lambda_2^\pm(x, y) \in (0, +\infty)$.

Несложно убедиться, что в области $(x, y) \in (0, 1) \times (0, +\infty)$ имеют место неравенства

$$\frac{\partial \lambda_1}{\partial x} \leq 0, \quad \frac{\partial \lambda_2^-}{\partial x} \geq 0, \quad \frac{\partial \lambda_2^+}{\partial x} \leq 0, \quad \frac{\partial \lambda_2^\pm}{\partial y} \leq 0,$$

$$\lambda_2^-(x, y) \leq \lambda_1(x) \leq \lambda_2^+(x, y).$$

откуда следует, что

$$\bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda_2^-(x(s), y(t)) = [\lambda_2^-(\alpha/\Delta, \beta/\gamma), \lambda_2^-(\alpha/\delta, \beta/\Gamma)] = I_1,$$

$$\bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda_2^+(x(s), y(t)) = [\lambda_2^+(\alpha/\delta, \beta/\gamma), \lambda_2^+(\alpha/\Delta, \beta/\Gamma)] = I_3,$$

$$\bigcup_{\substack{s \in [\delta, \Delta] \\ t \in [\gamma, \Gamma]}} \lambda_1(x(s)) = [\lambda_1(\alpha/\delta), \lambda_1(\alpha/\Delta)] = I_2.$$

Теорема доказана. ■

11.2.3. Анализ сходимости методов чебышевского типа и GMRES

Оценим норму многочлена, наименее уклоняющегося от нуля на спектре оператора S . Сначала потребуется

Лемма 11.2.5. Пусть $0 < \alpha < \delta$, тогда

$$\sigma(I - 2S) \subseteq [-b, -a] \cup [a, b],$$

где

$$a = \frac{\delta - \alpha}{\delta + \alpha} > 0,$$

$$b = \frac{\alpha}{\Delta + \alpha} \max \left\{ -\frac{\gamma - \beta}{\gamma + \beta} + \sqrt{\frac{\Delta^2}{\alpha^2} - \frac{4\beta\gamma}{(\beta + \gamma)^2}}, \right.$$

$$\left. \frac{\Gamma - \beta}{\Gamma + \beta} + \sqrt{\frac{\Delta^2}{\alpha^2} - \frac{4\beta\Gamma}{(\beta + \Gamma)^2}} \right\} < 1.$$

Кроме того, для любого $\beta > 0$ имеет место неравенство

$$b \geq \frac{\alpha}{\Delta + \alpha} \left(\frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}} + \sqrt{\frac{\Delta^2}{\alpha^2} - \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}} \right),$$

где $\xi = \gamma/\Gamma$, причем равенство достигается только при $\beta = \sqrt{\gamma\Gamma}$.

Доказательство. Воспользуемся результатом теоремы 11.2.2: из (11.6) следует, что $\sigma(I - 2S) \subseteq [-b, -a] \cup [a_0, b]$, где a , b — величины, определенные в формулировке леммы, и

$$a_0 = 1 - \frac{\delta}{(\delta + \alpha)} \left(1 + 2 \frac{\alpha}{\delta} \frac{\Gamma}{\beta + \Gamma} - \sqrt{1 - \frac{\alpha^2}{\delta^2} \frac{4\beta\Gamma}{(\beta + \Gamma)^2}} \right).$$

Для получения требуемой оценки достаточно доказать неравенство $a \leq a_0$, которое эквивалентно следующему

$$1 - \frac{\alpha}{\delta} \leq \frac{\alpha}{\delta} - 2 \frac{\alpha}{\delta} \frac{\Gamma}{\beta + \Gamma} + \sqrt{1 - \frac{\alpha^2}{\delta^2} \frac{4\beta\Gamma}{(\beta + \Gamma)^2}},$$

т. е.

$$1 - x + y \leq \sqrt{1 - x^2 + y^2}, \quad x = \frac{\alpha}{\delta}, \quad y = \frac{\alpha}{\delta} \frac{\Gamma - \beta}{\Gamma + \beta}.$$

Последнее неравенство имеет место при любых $x \in (0, 1]$, $y \in [-x, x]$.

Докажем теперь неравенство для величины b , для чего представим b в виде

$$b = \max \left\{ f \left(\frac{\Gamma - \beta}{\Gamma + \beta} \right), f \left(\frac{\beta - \gamma}{\beta + \gamma} \right) \right\},$$

где

$$f(z) = \frac{\alpha}{\Delta + \alpha} \left(z + \sqrt{\frac{\Delta^2}{\alpha^2} - 1 + z^2} \right).$$

В области $\Delta^2/\alpha^2 - 1 + z^2 > 0$ производная $\partial f/\partial z > 0$ и, таким образом, $f(z)$ монотонно возрастает. Отсюда немедленно следует, что b как функция β обладает строгим глобальным минимумом в точке $\beta > 0$, удовлетворяющей уравнению

$$\frac{\Gamma - \beta}{\Gamma + \beta} = \frac{\beta - \gamma}{\beta + \gamma},$$

которое имеет единственное решение $\beta = \sqrt{\gamma\Gamma}$. Лемма доказана. ■

Имеет место

Теорема 11.2.3. Пусть $0 < \alpha < \delta$, $\beta > 0$, a и b — величины, определенные в лемме 11.2.5, тогда справедлива оценка спектра

$$\sigma(S(I - S)) \subseteq [\sigma_0, \sigma_1] \equiv \left[\frac{1 - b^2}{4}, \frac{1 - a^2}{4} \right]$$

и для любого $n \in \mathbb{N}$ выполнено неравенство

$$\theta_{2n}(S) \leq \max_{x \in [\sigma_0, \sigma_1]} |C_n^{\sigma_0, \sigma_1}(x(1 - x))| = \frac{2q^{2n}}{1 + q^{4n}},$$

где

$$q = \sqrt{\frac{1 - \sqrt{\varphi}}{1 + \sqrt{\varphi}}}, \quad \varphi = \frac{\sigma_0}{\sigma_1},$$

причем

$$\varphi \leq \frac{\omega}{2} \left(\frac{1 + \alpha/\delta}{1 + \alpha/\Delta} \right)^2 \left(1 - \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}} \sqrt{1 - \frac{\alpha^2}{\Delta^2} \frac{4\sqrt{\xi}}{(1 + \sqrt{\xi})^2}} \right), \quad (11.8)$$

$$\xi = \frac{\gamma}{\Gamma}, \quad \omega = \frac{\delta}{\Delta}$$

и равенство в последней оценке достигается только при $\beta = \sqrt{\gamma\Gamma}$.

Доказательство. Оценка спектра немедленно следует из леммы 11.2.5 с учетом того факта, что

$$4S(I - S) = I - (I - 2S)^2.$$

Используя (11.5), получаем неравенства

$$\begin{aligned} \theta_{2n}(S) &\leq \min_{\substack{P_n(x) \in \mathbb{R}[x], \\ \deg P_n \leq n, P_n(0)=1}} \max_{x \in \sigma(S)} |P_n(x - x^2)| \leq \\ &\leq \min_{\substack{P_n(x) \in \mathbb{R}[x], \\ \deg P_n \leq n, P_n(0)=1}} \max_{y = \frac{1 - (1 - 2x)^2}{4} \in [\sigma_0, \sigma_1]} |P_n(y)| \leq \frac{2q_0^n}{1 + q_0^{2n}}, \end{aligned}$$

где $q_0 = (1 - \sqrt{\varphi})/(1 + \sqrt{\varphi}) = q^2$. Неравенство (11.8) следует из леммы 11.2.5 и равенства

$$\max_{\beta > 0} \varphi = \max_{\beta > 0} \frac{1 - b^2}{1 - a^2} = \frac{1 - (\min_{\beta > 0} b)^2}{1 - a^2}.$$

Теорема доказана. ■

Лемма 11.2.6. Пусть $y^k = z^k - z \in Z$ — вектор ошибки итерационного метода на k -й итерации, $r^k = Sy^k \in Z$ — соответствующий вектор невязки и имеет место представление

$$y^k = P_k(S)y^0, \quad P_k \in \mathbb{R}[x], \quad k \in \mathbb{N},$$

тогда выполнены неравенства

$$\begin{aligned} \|y^k\|_M &\leq \sqrt{\left(1 + \frac{\Gamma}{\beta}\right) \frac{\delta + \alpha}{\delta - \alpha}} \max_{x \in \sigma(S)} |P_k(x)| \|y^0\|_M, \\ \|r^k\| &\leq \sqrt{\frac{\max\{\alpha^2 \lambda_{\max}(Q), \beta \lambda_{\max}(C)\}}{\min\{\alpha^2 \lambda_{\min}(Q), \beta \lambda_{\min}(C)\}}} \left(1 + \frac{\Gamma}{\beta}\right) \frac{\delta + \alpha}{\delta - \alpha} \times \\ &\times \max_{x \in \sigma(S)} |P_k(x)| \|r^0\|. \end{aligned}$$

Доказательство. Воспользуемся представлением

$$\Lambda = W^{-1}SW, \quad \Lambda = \Lambda^* \quad \sigma(\Lambda) = \sigma(S),$$

полученным в теореме 11.2.1, где

$$W = M^{-1/2}W_1 = M^{-1/2} \begin{pmatrix} I & G \\ G^* & -I \end{pmatrix}^{-1} \begin{pmatrix} (L-I)(L+I)^{-1} & 0 \\ 0 & I \end{pmatrix}^{1/2},$$

$$L = \alpha^{-1}Q^{-1/2}AQ^{-1/2}, \quad L = L^*, \quad \sigma(L) \subseteq \left[\frac{\delta}{\alpha}, \frac{\Delta}{\alpha} \right],$$

$$G = \beta^{-1/2}Q^{-1/2}BC^{-1/2}, \quad \sigma(GG^*) \subseteq \{0\} \cup \left[\frac{\gamma}{\beta}, \frac{\Gamma}{\beta} \right].$$

Имеет место оценка спектра

$$\sigma \begin{pmatrix} (L-I)(L+I)^{-1} & 0 \\ 0 & I \end{pmatrix} \subseteq [(\delta - \alpha)/(\delta + \alpha), 1],$$

кроме того, используя лемму 11.2.4, несложно получить:

$$\sigma \begin{pmatrix} I & G \\ G^* & -I \end{pmatrix} \subseteq \left[-\sqrt{1 + \frac{\Gamma}{\beta}}, -1 \right] \cup \left[1, \sqrt{1 + \frac{\Gamma}{\beta}} \right].$$

Следовательно, справедливы неравенства для норм

$$\|W_1\| \leq 1, \quad \|W_1^{-1}\| \leq \sqrt{\left(1 + \frac{\Gamma}{\beta}\right) \frac{\delta + \alpha}{\delta - \alpha}}.$$

Из условия теоремы имеем представление

$$M^{1/2}y^k = W_1 P_k(\Lambda) W_1^{-1} M^{1/2}y^0,$$

откуда получаем первую оценку для вектора ошибки

$$\begin{aligned} \|y^k\|_M &= \|M^{1/2}y^k\| \leq \|W_1 P_k(\Lambda) W_1^{-1}\| \|M^{1/2}y^0\| \leq \\ &\leq \|W_1\| \|W_1^{-1}\| \|P_k(\Lambda)\| \|y^0\|_M \leq \\ &\leq \sqrt{\left(1 + \frac{\Gamma}{\beta}\right) \frac{\delta + \alpha}{\delta - \alpha}} \max_{x \in \sigma(S)} |P_k(x)| \|y^0\|_M. \end{aligned}$$

Представление для вектора невязки имеет вид

$$r^k = M^{-1/2}W_1 P_k(\Lambda) W_1^{-1} M^{1/2}r^0,$$

откуда получаем вторую оценку —

$$\begin{aligned} \|r^k\| &\leq \|M\|^{1/2} \|M^{-1}\|^{1/2} \|W_1\| \|W_1^{-1}\| \|P_k(\Lambda)\| \|r^0\| \leq \\ &\leq \sqrt{\frac{\max\{\alpha^2 \lambda_{\max}(Q), \beta \lambda_{\max}(C)\}}{\min\{\alpha^2 \lambda_{\min}(Q), \beta \lambda_{\min}(C)\}}} \left(1 + \frac{\Gamma}{\beta}\right) \frac{\delta + \alpha}{\delta - \alpha} \times \\ &\times \max_{x \in \sigma(S)} |P_k(x)| \|r^0\|. \end{aligned}$$

Лемма доказана. ■

Классический чебышевский итерационный метод решения системы линейных уравнений $Sz = F$, $\sigma(S) \subseteq [a, b]$, $a > 0$ может быть представлен в форме [76, с. 269]

$$\begin{aligned} z^k &= P_k(S)z^0 + S^{-1}(I - P_k(S))F, \quad k \in \mathbb{N}, \\ P_k(x) &= \prod_{j=1}^k (1 - \tau_j x), \end{aligned} \tag{11.9}$$

где z^0 — произвольное начальное приближение, а итерационные параметры имеют вид

$$\tau_j = \frac{2}{b + a + (b - a) \cos(\pi(j - 1/2)/k)}.$$

Конкретная схема реализации метода может быть различной (например, с перемешиванием параметров, предложенная В. И. Лебедевым и С. А. Финогоновым [62]).

Алгоритм, принимающий в расчет кластеризацию спектра S , строится применением метода Чебышева для решения равносильной задачи $S(I - S)z = F - SF$. Формально его можно представить в виде (11.9), используя $\tau_j \in \mathbb{C}$ из представления

$$\begin{aligned} (1 - \tau_{2k}x)(1 - \tau_{2k+1}x) &= 1 - \rho_k x(1 - x), \\ \rho_k &= \frac{2}{\sigma_1 + \sigma_0 + (\sigma_1 - \sigma_0) \cos(\pi(j - 1/2)/k)}, \end{aligned} \tag{11.10}$$

где $0 < \sigma_0 \leq \sigma_1 < +\infty$ — величины, определенные в теореме 11.2.3. Независимо от способа реализации метода имеет место тождество

$$P_{2k}(x) = \prod_{j=1}^{2k} (1 - \tau_j x) \equiv C_k^{\sigma_0, \sigma_1}(x(1 - x)). \tag{11.11}$$

Справедлива

Теорема 11.2.4. Пусть $0 < \alpha < \delta$, $\beta > 0$, σ_0, σ_1 — величины, определенные в теореме 11.2.3, и для решения предобусловленной

системы уравнений $Sz = F$ используется модифицированный алгоритм Чебышева (11.9) с параметрами (11.10), тогда имеет место оценка нормы погрешности

$$\|y^{2k}\|_M \leq C_{MCh} \frac{2q^{2k}}{1+q^{4k}} \|y^0\|_M,$$

где

$$C_{MCh} = \sqrt{\left(1 + \frac{\Gamma}{\beta}\right) \frac{\delta + \alpha}{\delta - \alpha}}, \quad q = \sqrt{\frac{\sqrt{\sigma_1} - \sqrt{\sigma_0}}{\sqrt{\sigma_1} + \sqrt{\sigma_0}}}.$$

Доказательство. Оценка немедленно следует из теоремы 11.2.3, леммы 11.2.6 и тождества (11.11). Теорема доказана. ■

Метод GMRES [183, с. 158] является одним из наиболее универсальных итерационных методов решения несимметричных задач. Ключевой характеристикой метода является свойство минимальности невязки

$$\|r^k\| = \min_{P_k \in \mathbb{R}, \deg P_k \leq k, P_k(0)=1} \|P_k(S)r^0\|. \quad (11.12)$$

Справедлива

Теорема 11.2.5. Пусть $0 < \alpha < \delta$, $\beta > 0$, σ_0, σ_1 — величины, определенные в теореме 11.2.3, и для решения предобусловленной системы уравнений $Sz = F$ используется алгоритм GMRES, тогда имеет место оценка нормы невязки

$$\|r^{2k+1}\| \leq \|r^{2k}\| \leq C_{GMRES} \frac{2q^{2k}}{1+q^{4k}} \|r^0\|,$$

где

$$C_{GMRES} = \sqrt{\frac{\max\{\alpha^2 \lambda_{\max}(Q), \beta \lambda_{\max}(C)\}}{\min\{\alpha^2 \lambda_{\min}(Q), \beta \lambda_{\min}(C)\}}} \left(1 + \frac{\Gamma}{\beta}\right) \frac{\delta + \alpha}{\delta - \alpha},$$

$$q = \sqrt{\frac{\sqrt{\sigma_1} - \sqrt{\sigma_0}}{\sqrt{\sigma_1} + \sqrt{\sigma_0}}}.$$

Доказательство. Оценки немедленно следуют из теоремы 11.2.3, леммы 11.2.6 и свойства (11.12). Теорема доказана. ■

Практическое применение HSS-предобусловливателя требует на каждом шаге алгоритма решения вспомогательных систем уравнений с матрицами

$$\begin{pmatrix} \alpha Q & B \\ B^* & -\alpha^{-1} \beta C \end{pmatrix}, \quad I + \alpha^{-1} A Q^{-1}.$$

При отсутствии эффективных прямых методов уравнение с первой матрицей можно решать приближенно (внутренними итерациями), например, методом Узавы — сопряженных градиентов, который сводится к решению уравнения с матрицей $C^{-1}B^*Q^{-1}B + \beta I$. Для нахождения приближенного решения уравнения со второй матрицей может быть использован метод сопряженных градиентов в классической форме. Оценки показателей асимптотической скорости сходимости указанных методов имеют соответственно вид [76, с. 270]:

$$q_1 = \frac{1 - \sqrt{\varphi_1}}{1 + \sqrt{\varphi_1}}, \quad \varphi_1 = \frac{\beta + \gamma}{\beta + \Gamma},$$

$$q_2 = \frac{1 - \sqrt{\varphi_2}}{1 + \sqrt{\varphi_2}}, \quad \varphi_2 = \frac{\alpha + \delta}{\alpha + \Delta}.$$

Таким образом, с точки зрения обусловленности решаемых уравнений и скорости сходимости внутренних итераций предпочтительно выбирать по возможности наибольшие значения $\alpha \in (0, \delta)$, $\beta > 0$.

Выбор $\alpha \approx \delta$ хорошо согласуется с асимптотиками оценок теорем 11.2.4 и 11.2.5, так как асимптотический показатель q в этих оценках монотонно убывает по α . С другой стороны, большие значения

$$\beta \geq \sqrt{\gamma\Gamma}$$

приводят к снижению скорости сходимости внешних итераций, что требует принятия компромиссного решения при выборе β .

В случае, когда оператор S может быть эффективно вычислен, «идеальным» выбором параметров является $\beta = \sqrt{\gamma\Gamma}$ и большое значение $\alpha \approx \delta$.

В случае рассматриваемых алгоритмов при

$$\beta = \sqrt{\gamma\Gamma} \quad \text{и} \quad \alpha \in (0, \delta)$$

из (11.8) следует, что асимптотический показатель скорости сходимости имеет представление

$$q = 1 - c\sqrt{\omega}^4\sqrt{\xi} + O(\sqrt{\xi}),$$

$$\omega = \text{const}, \quad \xi \rightarrow 0, \quad c \geq 1.$$

Порядок этой оценки совпадает с оптимальным порядком асимптотического показателя сходимости, полученного в предыдущем разделе для стационарного метода GPHSSI (отметим, что для GPHSSI оптимальное значение β то же самое — $\sqrt{\gamma\Gamma}$). И хотя метод GPHSSI проще в реализации и менее требователен к ресурсам на каждой итерации, чем методы рассмотренные в этом разделе, использование метода HSS в качестве предобусловливателя надежнее с точки зрения наличия оценок сходимости по норме.

11.3. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Общая теория методов переменных итераций, свойств многочленов Чебышева, итерационного метода Чебышева и вариантов его реализации изложена в [61, с. 247], [76, с. 269–283]. Теорию метода GMRES и, в том числе, свойство (11.12) можно найти в [80, с. 218], [183, с. 193].

При реализации метода переменных симметричных и кососимметричных итераций требуется решение вспомогательных систем с матрицами $M + H$, $M + J$, $M = M^* > 0$, $H = H^* > 0$, $J = -J^*$. В случае, когда $M = I$ (или задача может быть сведена к эквивалентной задаче такого вида) данные системы могут быть эффективно (приближенно) решены классическим методом сопряженных градиентов и методом сопряженных градиентов, обобщенным для несимметричного случая [183, с. 186].

Идея метода попеременных симметричных и кососимметричных итераций (HSSI) для седловых задач, по-видимому, впервые предложена в [115] (позднее те же самые результаты опубликованы в [116]), где приведена схема построения метода для частного случая $Q = I$, $\beta = \alpha^2$ и обоснована его безусловная сходимость.

В [114] была проведена оптимизация скорости сходимости метода для решения дискретного аналога уравнения Пуассона. Несмотря на то, что полученные в работе результаты носят крайне ограниченный характер, здесь впервые было отмечено свойство кластеризации спектра предобусловленного оператора при малых значениях $\alpha > 0$.

В работе [201] была рассмотрена модификация метода HSSI, соответствующая частному случаю $Q = A$, и для нее получены оптимальные значения параметров. Этот случай соответствует релаксационным методам, рассматриваемым в первой части книги, и данный результат может быть выведен предложенными там способами.

Окончательное обобщение, оценка скорости сходимости и оптимизация GPHSSI для симметричных седловых задач были проведены в работе [23].

В работе [118] впервые предпринята попытка анализа итераций метода как предобусловливателя в методе GMRES, получено (неточное) условие вещественности спектра предобусловленного оператора для симметричных задач в виде неравенства

$$\alpha < \frac{\delta}{2}$$

и отмечено, что практическая сходимость метода GMRES для решения дискретной задачи Стокса наиболее эффективна при значениях

$\alpha \approx \delta$. Последнее подтверждается полученными в этом разделе теоремами.

Ключевые результаты в исследовании HSS-предобусловливателя для решения симметричных седловых задач, в том числе — вещественность спектра и диагонализуемость предобусловленной задачи при $\alpha < \delta$, а также оценки спектра и скорости сходимости алгоритмов чебышевского типа и GMRES, получены в работе [25].

НЕЛИНЕЙНЫЕ ЗАДАЧИ И БЛОЧНО ТРЕУГОЛЬНОЕ ПРЕДОБУСЛОВЛИВАНИЕ

Принципиально важные этапы исследования алгоритмов для решения линейных седловых задач могут быть обобщены и адаптированы для анализа нелинейных проблем. Причем это касается не только вопросов сходимости итерационных методов, но и существования и единственности решений рассматриваемых задач.

В главе для иллюстрации возможностей предлагаемого подхода рассмотрены блочно треугольные методы (типа GMSOR) решения седловых задач с нелинейными операторами двух принципиально различных типов. В первом случае нелинейный оператор является кососимметричным возмущением (как линейным, так и нелинейным) линейного самосопряженного оператора, а во втором — обладает свойством сильной монотонности. Постановки задач такого рода часто возникают при дискретизации уравнений в частных производных в процессе математического моделирования сложных физических процессов.

12.1. УРАВНЕНИЯ С КОСОСИММЕТРИЧНЫМ ВОЗМУЩЕНИЕМ

12.1.1. Постановка задачи

Широкое применение на практике имеет класс седловых задач специального вида

$$\begin{cases} Au + Ku + N(u)u + Bp = f, \\ B^*u = \varphi, \end{cases} \quad (12.1)$$

где $A: U \rightarrow U$ — линейный самосопряженный положительно определенный оператор, $K: U \rightarrow U$ — линейный кососимметричный оператор, а $N: U \times U \rightarrow U$ — билинейный оператор обладающий свойством кососимметрии:

$$(N(u)v, v) = 0 \quad \forall u, v \in U. \quad (12.2)$$

Метод GMSOR может быть несложным образом обобщен на этот класс задач:

$$\begin{cases} Q \frac{u^{k+1} - u^k}{\tau} + (A + K + \beta BC^{-1} B^*) u^k + \\ + N(u^k) u^k + B p^k = f + \beta BC^{-1} \varphi, \\ - \alpha \tau C \frac{p^{k+1} - p^k}{\tau} + B^* u^{k+1} = \varphi, \end{cases} \quad (12.3)$$

где $\tau > 0$, $\beta \geq 0$, $\alpha > 0$ — фиксированные параметры, $u^0 \in U$, $p^0 \in P$ — заданные начальные приближения, $k = 0, 1, \dots$

Всюду далее будем предполагать известной априорную информацию об операторах из (12.3):

$$0 < \delta Q \leq A \leq \Delta Q, \quad (12.4)$$

$$0 < \gamma C \leq B^* Q^{-1} B \leq \Gamma C, \quad (12.5)$$

$$|(Ku, v)| \leq R_K \|u\|_Q \|v\|_Q \quad \forall u, v \in U, \quad (12.6)$$

$$|(N(u)v, w)| \leq R_N \|u\|_Q \|v\|_Q \|w\|_Q \quad \forall u, v, w \in U, \quad (12.7)$$

где $0 < \delta \leq \Delta$, $0 < \gamma \leq \Gamma$, $0 \leq R_K$, $0 \leq R_N$ — константы.

Покажем, что задача (12.1) разрешима. Имеет место

Теорема 12.1.1 (Существование решения). Пусть

$$R_N \|\varphi\|_{C^{-1}} < \delta \gamma^{1/2},$$

тогда для любого $f \in U$ существует решение задачи (12.1).

Доказательство. Случай $R_N = 0$ соответствует линейной невырожденной задаче с седловым оператором, поэтому решение существует и единственно при любых $f \in U$, $\varphi \in P$. Рассмотрим случай $R_N > 0$.

Обозначим через P_H ортогональный проектор пространства U на подпространство $H = \ker B^*$, а через u_0 — единственное решение задачи

$$B^* u = \varphi, \quad P_H u = 0.$$

Так как

$$\sigma(Q^{-1/2} B C^{-1} B^* Q^{-1/2}) \subseteq \{0\} \cup [\gamma, \Gamma]$$

и $P_H u_0 = 0$, то имеет место неравенство

$$\begin{aligned} \|u_0\|_Q &= (Qu_0, u_0)^{1/2} \leq \gamma^{-1/2} (BC^{-1} B^* u_0, u_0)^{1/2} = \\ &= \gamma^{-1/2} \|\varphi\|_{C^{-1}} = \frac{\delta - \varepsilon}{R_N}, \end{aligned}$$

где

$$\varepsilon = \delta - \gamma^{-1/2} R_N \|\varphi\|_{C^{-1}},$$

причем из условия теоремы следует, что $\varepsilon > 0$.

Положим $u = u_0 + v$, $v \in H$, тогда задача (12.1) равносильна задаче

$$Av + Kv + N(u_0)v + N(v)u_0 + N(v)v + Bp - \tilde{f} = 0, \quad v \in H, \quad p \in P, \quad (12.8)$$

где

$$\tilde{f} = f - Au_0 - Ku_0 - N(u_0, u_0).$$

Решение (12.8) сводится к последовательному нахождению нуля v_0 непрерывного векторного поля $F: H \rightarrow H$

$$F(v) = P_H(Av + Kv + N(u_0)v + N(v)u_0 + N(v)v - \tilde{f})$$

и p_0 , удовлетворяющего уравнению

$$Bp = -(I - P_H)(Av_0 + Kv_0 + N(u_0)v_0 + N(v_0)u_0 + N(v_0)v_0 - \tilde{f}).$$

Последнее уравнение совместно, так как $\text{Im}(I - P_H) = \text{Im } B$.

В силу свойств операторов K и $N(\cdot)$, для любого

$$v \in H, \quad \|v\|_Q > \varepsilon^{-1} \|\tilde{f}\|_{Q^{-1}}$$

имеет место неравенство

$$\begin{aligned} (F(v), v) &= (Av, v) + (N(v)u_0, v) - (\tilde{f}, v) \geq \\ &\geq \left((\delta - R_N \|u_0\|_Q) \|v\|_Q - \|\tilde{f}\|_{Q^{-1}} \right) \|v\|_Q \geq \\ &\geq \left(\varepsilon \|v\|_Q - \|\tilde{f}\|_{Q^{-1}} \right) \|v\|_Q > 0, \end{aligned}$$

откуда следует, что F гомотопно тождественному векторному полю I ($Iv = v \quad \forall v \in H$) на «сфере»

$$\left\{ v \in H \mid \|v\|_Q = \varepsilon^{-1} \|\tilde{f}\|_{Q^{-1}} + \rho \right\},$$

для любого $\rho > 0$, поэтому [54, с. 20] принимает нулевое значение в некоторой точке множества

$$\|v\|_Q \leq \varepsilon^{-1} \|\tilde{f}\|_{Q^{-1}}.$$

Теорема доказана. ■

Отметим, что в общем случае решение (12.1) может быть неединственным.

12.1.2. Оценка скорости сходимости

Предположим, что существует решение $\{u, p\}$ задачи (12.1), удовлетворяющее неравенству

$$R_N \|u\|_Q < \delta.$$

Обозначим через

$$y^k \equiv \{v^k, r^k\} = \{u^k - u, p^k - p\}$$

ошибку приближения в алгоритме (12.3), который, с учетом этого, можно записать в следующем виде:

$$\begin{cases} Q \frac{v^{k+1} - v^k}{\tau} + (A_1(u) + \beta BC^{-1} B^*) v^k + \\ \quad + K_1(u, v^k) v^k + B r^k = 0, \\ -\alpha C(r^{k+1} - r^k) + B^* v^{k+1} = 0, \end{cases} \quad (12.9)$$

где

$$A_1(u)v = Av + (N(v)u + N^*(v)u)/2,$$

$$K_1(u, v^k)v = Kv + N(u)v + N(v^k)v + (N(v)u - N^*(v)u)/2,$$

а оператор $N^* : U \times U \rightarrow U$ однозначно определяется равенством

$$(N(v)u, w) = (v, N^*(w)u) \quad \forall u, v, w \in U.$$

Кроме того, с учетом (12.4) имеем неравенства

$$0 < \delta_1 Q = (\delta - R_N \|u\|_Q) Q \leq A_1(u) \leq (\Delta + R_N \|u\|_Q) Q = \Delta_1 Q.$$

Зафиксируем $\varkappa \in (0, 1)$ и введем оператор

$$D \equiv D(u, \alpha, \tau, \beta) = \begin{pmatrix} Q - \tau(\varkappa A_1(u) + \beta BC^{-1} B^*) & 0 \\ 0 & \alpha \tau C \end{pmatrix}.$$

Справедлива

Лемма 12.1.1. Пусть

$$0 < \tau < (\varkappa \Delta_1 + \beta \Gamma)^{-1},$$

тогда $D = D^* > 0$ и функционал

$$\|z\|_D = \sqrt{(Dz, z)}$$

определяет норму в пространстве $Z = U \times P$.

Доказательство. Самосопряженность D следует из его явного вида. При указанных ограничениях на τ справедливы неравенства

$$D \geq (1 - \tau(\kappa\Delta_1 + \beta\Gamma))Q > 0 > 0,$$

откуда немедленно следуют свойства нормы для функционала $\|z\|_D$. Лемма доказана. ■

Пусть выполнены условия леммы 12.1.1, тогда для любых $v, w \in U$ справедлива оценка

$$\begin{aligned} |(K_1(u, v^k)v, w)| &\leq \left(R_K + R_N \left(2\|u\|_Q + \|v^k\|_Q \right) \right) \|v\|_Q \|w\|_Q \leq \\ &\leq \left(R_K + R_N \left(2\|u\|_Q + (1 - \tau(\kappa\Delta_1 + \beta\Gamma))^{-1/2} \|y^k\|_D \right) \right) \times \\ &\times \|v\|_Q \|w\|_Q = R \|v\|_Q \|w\|_Q, \end{aligned}$$

с постоянной R , где $R = R(u, y^k, \tau, \beta)$.

Оператор перехода в (12.9) на k -й итерации имеет вид

$$T(u, v^k) = \begin{pmatrix} I - \tau Q^{-1} A_2(u, v^k) & -\tau Q^{-1} B \\ \alpha^{-1} C^{-1} B^* (I - \tau Q^{-1} A_2(u, v^k)) & I - \frac{\tau}{\alpha} C^{-1} B^* Q^{-1} B \end{pmatrix},$$

где

$$A_2(u, v^k) = A_1(u) + \beta B C^{-1} B^* + K_1(u, v^k).$$

Определим при фиксированных u и v^k линейные операторы (их зависимость от u и v^k в обозначениях будем опускать в целях компактности), затем

$$\begin{aligned} L &= Q^{-1/2} A_1(u) Q^{-1/2}, \quad G = Q^{-1/2} B C^{-1/2}, \\ J &= Q^{-1/2} K_1(u, v^k) Q^{-1/2}, \end{aligned}$$

тогда справедлива

Лемма 12.1.2. Оператор L — самосопряженный, J — кососимметричный и имеют место следующие неравенства:

$$\delta_1 I \leq L \leq \Delta_1 I, \quad \gamma I \leq G^* G \leq \Gamma I, \quad \|J\| \leq R.$$

Доказательство. Симметричность L и кососимметричность J следуют из явного вида операторов. Из равенств

$$\begin{aligned} \frac{(Lw, w)}{(w, w)} &= \frac{(A_1(u)(Q^{-1/2}w), Q^{-1/2}w)}{(Q(Q^{-1/2}w), Q^{-1/2}w)} \quad \forall w \in U \setminus \{0\}, \\ \frac{(G^* G p, p)}{(p, p)} &= \frac{(B^* Q^{-1} B (C^{-1/2}p), C^{-1/2}p)}{(C(C^{-1/2}p), C^{-1/2}p)} \quad \forall p \in P \setminus \{0\}, \\ \frac{(Jw_1, w_2)}{\|w_1\| \|w_2\|} &= \frac{(K_1(u, v^k)(Q^{-1/2}w_1), Q^{-1/2}w_2)}{\|Q^{-1/2}w_1\|_Q \|Q^{-1/2}w_2\|_Q} \quad \forall w_1, w_2 \in U \setminus \{0\} \end{aligned}$$

следуют двусторонние оценки во второй части утверждения леммы. Лемма доказана. ■

Введем оператор

$$M = \begin{pmatrix} I - \tau(\kappa L + \beta GG^*) & 0 \\ 0 & I \end{pmatrix}$$

и рассмотрим представление оператора перехода

$$T(u, v^k) = D^{-1/2} T_1 T_2 D^{1/2},$$

где

$$T_1 = M^{1/2} \begin{pmatrix} I & -\sqrt{\frac{\tau}{\alpha}} G \\ \sqrt{\frac{\tau}{\alpha}} G^* I & -\frac{\tau}{\alpha} G^* G \end{pmatrix} M^{1/2},$$

$$T_2 = M^{-1/2} \begin{pmatrix} I - \tau(L + \beta GG^* + J) & 0 \\ 0 & I \end{pmatrix} M^{-1/2}.$$

В следующих леммах (леммы 12.1.3–12.1.6) будет получена оценка нормы оператора перехода $T(u, v^k)$.

Лемма 12.1.3. При выполнении условий леммы 12.1.1 операторы $M^{-1/2}$ и $D^{-1/2}$ корректно определены и имеет место оценка

$$\|T(u, v^k)\|_D \leq \|T_1\| \|T_2\|.$$

Доказательство. Действительно, в силу леммы 12.1.1, $D = D^* > 0$, $M = M^* > 0$ и справедливо

$$\begin{aligned} \|T(u, v^k)\|_D^2 &= \sup_{z \in Z \setminus \{0\}} \frac{(DT(u, v^k)z, T(u, v^k)z)}{(Dz, z)} = \\ &= \sup_{z \in Z \setminus \{0\}} \frac{(T_1 T_2 D^{1/2} z, T_1 T_2 D^{1/2} z)}{(D^{1/2} z, D^{1/2} z)} = \\ &= \sup_{z \in Z \setminus \{0\}} \frac{(T_1 T_2 z, T_1 T_2 z)}{(z, z)} = \|T_1 T_2\|^2 \leq \|T_1\|^2 \|T_2\|^2. \end{aligned}$$

Лемма доказана. ■

Оценим норму оператора T_1 . Имеет место

Лемма 12.1.4. Пусть выполнены условия леммы 12.1.1, тогда $\|T_1\| = \rho(T_3)$, где

$$T_3 = \begin{pmatrix} I & \sqrt{\frac{\tau}{\alpha}} G \\ \sqrt{\frac{\tau}{\alpha}} G^* & \frac{\tau}{\alpha} G^* G - I \end{pmatrix} \begin{pmatrix} I - \tau(\kappa L + \beta GG^*) & 0 \\ 0 & I \end{pmatrix}.$$

Доказательство. Представим оператор T_1 в виде

$$T_1 = M^{1/2}T_3M^{-1/2}E,$$

где

$$E = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}.$$

Отсюда следует, что справедлива цепочка равенств

$$\begin{aligned} \|T_1\|^2 &= \sup_{z \in Z \setminus \{0\}} \frac{(T_1z, T_1z)}{(z, z)} = \\ &= \sup_{z \in Z \setminus \{0\}} \frac{(M^{1/2}T_3M^{-1/2}Ez, M^{1/2}T_3M^{-1/2}Ez)}{(z, z)} = \\ &= \sup_{z \in Z \setminus \{0\}} \frac{(M^{-1/2}T_3^*MT_3M^{-1/2}z, z)}{(E^{-1}z, E^{-1}z)} = \\ &= \rho \left(M^{-1/2}T_3^*M^{1/2}M^{1/2}T_3M^{-1/2} \right) = \\ &= \rho \left(M^{1/2}T_3M^{-1/2} \right)^2 = \rho(T_3)^2, \end{aligned}$$

так как $E^{-1}(E^{-1})^* = I$, а оператор $M^{1/2}T_3M^{-1/2}$ является самосопряженным. Лемма доказана. ■

Лемма 12.1.5. Пусть выполнены условия леммы 12.1.1, тогда

$$\rho(T_3) \leq \max_{\substack{s \in [\delta_1, \Delta_1] \\ t \in [\gamma, \Gamma]}} \left\{ |1 - \tau \kappa s|, \tau \left| \theta - \frac{t}{2\alpha} \right| + \sqrt{1 - 2\tau\theta + \tau^2 \left(\theta - \frac{t}{2\alpha} \right)^2} \right\},$$

где $\theta = (\kappa s + \beta t)/2$.

Доказательство. Так как $M^{1/2}T_3M^{-1/2}$ — самосопряженный оператор, то $\sigma(T_3) \subseteq \mathbb{R}$.

Задача на собственные значения $T_3z = \lambda z$ имеет вид

$$\begin{pmatrix} I - \tau(\kappa L + \beta GG^*) & \sqrt{\tau/\alpha}G \\ \sqrt{\tau/\alpha}G^*(I - \tau(\kappa L + \beta GG^*)) & \tau/\alpha G^*G - I \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} u \\ p \end{pmatrix},$$

где $\lambda \in \mathbb{R}$, $\{u, p\} \neq 0$. Умножая первое уравнение этой системы на $\sqrt{\tau/\alpha}G^*$ слева и вычитая результат из второго, получим уравнение

$$\lambda \sqrt{\frac{\tau}{\alpha}} G^* u - (\lambda + 1)p = 0.$$

Если $\lambda = -1$, то $p = 0$, и из первого уравнения следует

$$((I - \tau \kappa L)u, u) + (u, u) = 0,$$

откуда имеем $u = 0$. Таким образом, $\lambda \neq -1$ и исключая из первого уравнения вектор

$$p = \frac{\lambda}{\lambda + 1} \sqrt{\frac{\tau}{\alpha}} G^* u,$$

приходим к равенству

$$[\lambda^2 I - \tau \lambda ((\alpha^{-1} - \beta) G G^* - \kappa L) - I + \tau (\kappa L + \beta G G^*)] u = 0.$$

Итак, задача оценки спектра оператора T_3 сводится к задаче нахождения вещественного спектра λ , отличного от -1 , операторного пучка

$$\chi(\lambda) = f(\lambda, L)g(\lambda, G G^*) + h(\lambda)G G^*,$$

$$f(\lambda, s) = \lambda - 1 + \tau \kappa s, \quad g(\lambda, t) = \lambda + 1, \quad h(\lambda) = \tau(\beta - (\alpha^{-1} - \beta)\lambda).$$

В силу следствия 6.4.1 из теоремы 6.4.8 имеем: либо $\lambda - 1 + \tau \kappa s = 0$ для некоторого $s \in [\delta_1, \Delta_1]$, либо существуют $s \in [\delta_1, \Delta_1]$, $t \in [\gamma, \gamma]$ такие, что

$$\lambda^2 - \tau \lambda ((\alpha^{-1} - \beta)t - \kappa s) - 1 + \tau(\kappa s + \beta t) = 0,$$

откуда и следует утверждение леммы. Лемма доказана. ■

Следствие 12.1.1. Пусть выполнены условия леммы 12.1.5 и справедливо неравенство

$$\kappa \delta_1 + (\beta - \alpha^{-1})\Gamma \geq 0,$$

тогда имеет место оценка

$$\|T_1\| \leq \max \left\{ 1 - \frac{\tau}{2\alpha} \gamma, 1 - \tau \kappa \delta_1 \right\} < 1.$$

Доказательство. Из лемм 12.1.4 и 12.1.5 следует неравенство

$$\begin{aligned} \|T_1\| &= \rho(T_3) \leq \\ &\leq \max_{\substack{s \in [\delta_1, \Delta_1] \\ t \in [\gamma, \Gamma]}} \left\{ |1 - \tau \kappa s|, \tau \left| \theta - \frac{t}{2\alpha} \right| + \sqrt{1 - 2\tau \theta + \tau^2 \left(\theta - \frac{t}{2\alpha} \right)^2} \right\}, \end{aligned}$$

где $\theta = (\kappa s + \beta t)/2$. Оценим выражения под знаком \max . Из неравенства в условии имеем, что

$$\theta - \frac{t}{2\alpha} \geq 0 \quad \text{и} \quad \tau \theta < \frac{1}{2}$$

при любых $t \in [\gamma, \Gamma]$, $s \in [\delta_1, \Delta_1]$ и, следовательно, справедливо

$$\begin{aligned} & \tau \left| \theta - \frac{t}{2\alpha} \right| + \sqrt{1 - 2\tau\theta + \tau^2 \left(\theta - \frac{t}{2\alpha} \right)^2} \leq \\ & \leq \tau \left(\theta - \frac{t}{2\alpha} \right) + \sqrt{1 - 2\tau\theta + \tau^2 \theta^2} = 1 - \frac{\tau}{2\alpha} t \leq 1 - \frac{\tau}{2\alpha} \gamma. \end{aligned}$$

Кроме того, выполнено $1 - \tau\kappa s > 0$ и, значит, верна оценка

$$|1 - \tau\kappa s| = 1 - \tau\kappa s \leq 1 - \tau\kappa\delta_1.$$

Лемма доказана. ■

Укажем условия, при которых норма оператора T_2 не превосходит 1. Справедлива

Лемма 12.1.6. Пусть выполнены условия леммы 12.1.1 и, дополнительно, выполнено неравенство

$$\tau \leq \frac{2(1 - \kappa)\delta_1(1 - \tau(\kappa\Delta_1 + \beta\Gamma))^3}{2((1 - \kappa)(1 - \tau(\kappa\Delta_1 + \beta\Gamma)))^2\delta_1\Delta_1 + R^2},$$

тогда $\|T_2\| = 1$.

Доказательство. Обозначим $F = I - \tau(\kappa L + \beta GG^*)$. В силу леммы 12.1.1, оператор F является самосопряженным и положительно определенным. Несложно заметить, что

$$\|T_2\| = \max \left\{ 1, \left\| I - \tau F^{-1/2}((1 - \kappa)L + J)F^{-1/2} \right\| \right\}.$$

Таким образом, для справедливости $\|T_2\| = 1$ достаточно показать, что

$$\left\| I - \tau F^{-1/2}((1 - \kappa)L + J)F^{-1/2} \right\| \leq 1.$$

Пусть $\mu \in (0, 1)$, тогда, если

$$0 < \tau \leq \frac{1 - \mu}{(1 - \kappa)\Delta_1 \|F\|^{-1}},$$

то справедливо неравенство

$$\left\| (1 - \mu)I - \tau(1 - \kappa)F^{-1/2}LF^{-1/2} \right\| \leq 1 - \mu - \tau(1 - \kappa) \|F\|^{-1} \delta_1$$

и, кроме того, имеет место оценка

$$\begin{aligned} \left\| \mu I - \tau F^{-1/2}JF^{-1/2} \right\|^2 &= \left\| \mu^2 I - \tau^2 F^{-1/2}JF^{-1}JF^{-1/2} \right\| \leq \\ &\leq \mu^2 + \tau^2 R^2 \|F^{-1}\|^2, \end{aligned}$$

следовательно, получаем

$$\begin{aligned} & \left\| I - \tau F^{-1/2} ((1 - \kappa)L + J) F^{-1/2} \right\| \leq \\ & \leq 1 - \mu - \tau(1 - \kappa) \|F\|^{-1} \delta_1 + \sqrt{\mu^2 + \tau^2 R^2 \|F^{-1}\|^2} \leq \\ & \leq 1 - \tau(1 - \kappa) \|F\|^{-1} \delta_1 + \frac{1}{2} \mu^{-1} \tau^2 R^2 \|F^{-1}\|^2. \end{aligned}$$

Отсюда следует, что при выполнении неравенства

$$\tau \leq \mu \frac{2(1 - \kappa) \|F\|^{-1} \delta_1}{R^2 \|F^{-1}\|^2}$$

справедлива оценка $\|T_1\| \leq 1$.

Выберем «наилучшее» $\mu \in (0, 1)$ из условия

$$\mu \frac{2(1 - \kappa) \|F\|^{-1} \delta_1}{R^2 \|F^{-1}\|^2} = \frac{1 - \mu}{(1 - \kappa) \Delta_1 \|F\|^{-1}}$$

и подставим в последнее неравенство, тогда получим следующее ограничение на τ :

$$\tau \leq \frac{2(1 - \kappa) \delta_1 \|F\|}{2(1 - \kappa)^2 \delta_1 \Delta_1 + R^2 \|F\|^2 \|F^{-1}\|^2}.$$

Учитывая, что $\|F\| \geq 1 - \tau(\kappa \Delta_1 + \beta \Gamma) > 0$ и

$$\|F\| \|F^{-1}\| \leq \frac{1 - \tau \kappa \delta_1}{1 - \tau(\kappa \Delta_1 + \beta \Gamma)} < (1 - \tau(\kappa \Delta_1 + \beta \Gamma))^{-1},$$

приходим к утверждению леммы. Лемма доказана. ■

Справедлива основная

Теорема 12.1.2. Пусть выполнены следующие условия:

$$\begin{aligned} & R_N \|u\|_Q < \delta, \quad 0 < \kappa < 1, \\ & \tau < \min \left\{ \frac{1}{\kappa \Delta_1 + \beta \Gamma}, \frac{2(1 - \kappa) \delta_1 (1 - \tau(\kappa \Delta_1 + \beta \Gamma))^3}{2((1 - \kappa)(1 - \tau(\kappa \Delta_1 + \beta \Gamma)))^2 \delta_1 \Delta_1 + R_0^2} \right\}, \\ & \beta \geq 0, \quad \kappa \delta_1 + (\beta - \alpha^{-1}) \Gamma \geq 0, \end{aligned}$$

где u — первая компонента решения задачи (12.1),

$$R_0 = \left(R_K + R_N \left(2\|u\|_Q + \frac{\|y^0\|_D}{\sqrt{1 - \tau(\kappa \Delta_1 + \beta \Gamma)}} \right) \right),$$

а y^0 — вектор начальной ошибки. Тогда алгоритм (12.3) сходится в норме $\|\cdot\|_D$ со скоростью геометрической прогрессии с показателем q таким, что

$$q \leq \max \left\{ |1 - \tau \kappa \delta_1|, \tau \theta + \sqrt{1 - 2\tau \left(\theta + \frac{\gamma}{2\alpha} \right) + \tau^2 \theta^2} \right\} < 1,$$

где $\theta = (\kappa \delta_1 + (\beta - \alpha^{-1})\gamma)/2$.

Доказательство. В силу леммы 12.1.5, имеет место оценка

$$\rho_3 = \max_{\substack{s \in [\delta_1, \Delta_1] \\ t \in [\gamma, \Gamma]}} \left\{ |1 - \tau \kappa s|, \tau \left| \tilde{\theta} - \frac{t}{2\alpha} \right| + \sqrt{1 - 2\tau \tilde{\theta} + \tau^2 \left(\tilde{\theta} - \frac{t}{2\alpha} \right)^2} \right\},$$

где $\tilde{\theta} = (\kappa s + \beta t)/2$. Используя свойства классов $F_{0,1}$ и неравенства

$$\tau < (\kappa \Delta_1 + \beta \Gamma)^{-1}, \quad \beta \geq 0, \quad \kappa \delta_1 + (\beta - \alpha^{-1})\Gamma \geq 0,$$

несложно убедиться, что

$$\rho_3 = \max \left\{ |1 - \tau \kappa \delta_1|, \tau \theta + \sqrt{1 - 2\tau \left(\theta + \frac{\gamma}{2\alpha} \right) + \tau^2 \theta^2} \right\}.$$

Из лемм 12.1.3, 12.1.4 следует оценка

$$\|T(u, v^k)\|_D \leq \|T_1\| \|T_2\| \leq \rho(T_3) \|T_2\| \leq \rho_3 \|T_2\|.$$

При $k = 0$, в силу равенства $R(u, y^0, \tau, \beta) = R_0$ и леммы 12.1.6, имеет место неравенство

$$\|T(u, v^0)\|_D \leq \rho_3,$$

откуда имеем

$$\|y^1\|_D = \|T(u, v^0)y^0\|_D \leq \|T(u, v^0)\|_D \|y^0\|_D \leq \rho_3 \|y^0\|_D.$$

Так как $\rho_3 < 1$, то $\|y^1\|_D \leq \|y^0\|_D$, $R(u, y^1, \tau, \beta) \leq R_0$, поэтому, в силу леммы 12.1.6, получаем

$$\|y^2\|_D = \|T(u, v^1)y^1\|_D \leq \|T(u, v^1)\|_D \|y^1\|_D \leq \rho_3 \|y^1\|_D \leq \rho_3^2 \|y^0\|_D.$$

Продолжая цепочку рассуждений, для любого $k \in \mathbb{N}$, выводим неравенство

$$\|y^k\|_D \leq \rho_3^k \|y^0\|_D.$$

Таким образом, имеет место сходимость со скоростью геометрической прогрессии с показателем $q \leq \rho_3$. Теорема доказана. ■

Следствие 12.1.2 (Асимптотика оценки). Пусть выполнены условия теоремы 12.1.2, тогда существует постоянная

$$c_1 = c_1(\delta, \Delta, R_K, R_N, u^0, p^0, f, \varphi) > 0$$

такая, что

$$q \leq 1 - c_1 \xi, \quad \xi = \gamma/\Gamma.$$

Если, кроме того, величины $R_K, R_N = O(\delta)$ при $\delta \rightarrow 0$, то существует постоянная

$$c_2 = c_2(R_K, R_N, u^0, p^0, f, \varphi) > 0$$

такая, что

$$q \leq 1 - c_2 \omega_1 \xi, \quad \omega_1 = (\delta - R_N \|u\|_Q) / (\Delta - R_N \|u\|_Q).$$

Доказательство. Положим

$$\kappa = \frac{1}{2}, \quad \alpha_0 = \frac{2\Gamma}{\delta_1}, \quad \beta_0 = 0, \quad \tau_0 = \frac{\delta_1}{\delta_1 \Delta_1 + 8R_0^2},$$

где

$$R_0 = \left(R_K + 2R_N \left(\|u\|_Q + \|y^0\|_D \right) \right),$$

тогда справедливо $1 - \tau_0(\kappa\Delta_1 + \beta\Gamma) \geq 1/2$, откуда, используя следствие 12.1.1 и теорему 12.1.2, получаем оценку для q :

$$q < 1 - \frac{1}{1 + 8\omega_1 (R_0 \delta_1^{-1})^2} \omega_1 \xi \equiv 1 - c_0 \xi,$$

что немедленно приводит к требуемым неравенствам. Следствие доказано. ■

Отметим важное следствие теоремы о сходимости. Справедливо

Следствие 12.1.3 (Единственность решения). Пусть $f \in U$, $\varphi \in P$ удовлетворяют неравенству

$$R_N(2\delta + \Delta + R_K)\gamma^{-1/2} \|\varphi\|_{C^{-1}} + R_N \|f\|_{Q^{-1}} < \delta^2,$$

тогда решение задачи (12.1) существует и единственно.

Доказательство. Случай $R_N = 0$ соответствует линейной невырожденной задаче с седловым оператором, поэтому решение существует и единственно при любых $f \in U$, $\varphi \in P$. Рассмотрим случай $R_N > 0$.

Из условия следует, что выполнено неравенство

$$\gamma^{-1/2} \|\varphi\|_{C^{-1}} < \frac{\delta}{R_N}.$$

Следовательно (см. доказательство теоремы 12.1.1) существует решение $z = \{u, p\}$ задачи (12.1) и вектор $u_0 \in U$ такие, что

$$\|u - u_0\|_Q \leq \varepsilon^{-1} \|f - Au_0 - Ku_0 - N(u_0, u_0)\|_{Q^{-1}}, \quad \|u_0\|_Q \leq \frac{\delta - \varepsilon}{R_N},$$

где $\varepsilon = \delta - \gamma^{-1/2} R_N \|\varphi\|_{C^{-1}} > 0$. Тогда справедливы неравенства

$$\begin{aligned} \|u\|_Q &\leq \|u\|_Q + \varepsilon^{-1} \times \\ &\quad \times (\|f\|_{Q^{-1}} + \|Au_0\|_{Q^{-1}} + \|Ku_0\|_{Q^{-1}} + \|N(u_0, u_0)\|_{Q^{-1}}) \leq \\ &\leq \varepsilon^{-1} ((\varepsilon + \Delta + R_K + R_N \|u_0\|_Q) \|u_0\|_Q + \|f\|_{Q^{-1}}) \leq \\ &\leq \varepsilon^{-1} ((\delta + \Delta + R_K) \|u_0\|_Q + \|f\|_{Q^{-1}}) \leq \\ &\leq \varepsilon^{-1} \left(\frac{(\delta + \Delta + R_K)(\delta - \varepsilon)}{R_N} + \|f\|_{Q^{-1}} \right) = \\ &= \varepsilon^{-1} ((\delta + \Delta + R_K) \gamma^{-1/2} \|\varphi\|_{C^{-1}} + \|f\|_{Q^{-1}}) < \\ &< \varepsilon^{-1} \left(\frac{\delta^2}{R_N} - \delta \gamma^{-1/2} \|\varphi\|_{C^{-1}} \right) = \frac{\delta}{R_N}. \end{aligned}$$

Таким образом, выполнено неравенство $R_N \|u\|_Q < \delta$, что в силу теоремы 12.1.2, является достаточным для сходимости алгоритма (12.3) к решению z независимо от начального приближения при соответствующем выборе итерационных параметров. А так как одна итерация (12.3) переводит произвольное решение в себя, то отсюда следует, что z — единственное решение задачи (12.1). Следствие доказано. ■

12.2. УРАВНЕНИЯ С СИЛЬНО МОНОТОННЫМ ОПЕРАТОРОМ

12.2.1. Постановка задачи

Пусть Ω — некоторая область в U и в области $\Omega \times P$ решается задача

$$\begin{cases} A(u) + Bp = f, \\ B^*u = \varphi. \end{cases} \quad (12.10)$$

Будем предполагать, что $A(u) : \Omega \rightarrow U$ определен всюду в Ω и удовлетворяет условию сильной монотонности:

$$\delta \|u_1 - u_2\|_Q^2 \leq (A(u_1) - A(u_2), u_1 - u_2) \quad \forall u_1, u_2 \in \Omega, \quad (12.11)$$

а также условию Липшица:

$$\|A(u_1) - A(u_2)\|_{Q^{-1}} \leq \Delta \|u_1 - u_2\|_Q \quad \forall u_1, u_2 \in \Omega, \quad (12.12)$$

где $0 < \delta \leq \Delta$. Кроме того, будем предполагать, что известны величины $0 < \gamma \leq \Gamma$ из неравенств

$$0 < \gamma C \leq B^* Q^{-1} B \leq \Gamma C. \quad (12.13)$$

Потребуем также, чтобы задача (12.10) имела хотя бы одно решение в области $\Omega \times P$, что, вообще говоря, не следует из указанных условий.

Обобщим алгоритм GMSOR для нахождения решения задач такого класса:

$$\begin{cases} Q \frac{u^{k+1} - u^k}{\tau} + A(u^k) + \beta B C^{-1} B^* u^k + B p^k = \tilde{f}, \\ -\alpha \tau C \frac{p^{k+1} - p^k}{\tau} + B^* u^{k+1} = \varphi, \end{cases} \quad (12.14)$$

где $\tilde{f} = f + \beta B C^{-1} \varphi$. Отметим, что для того, чтобы указанный алгоритм был определен корректно необходимо также указать множество допустимых начальных приближений $z^0 \in Z$.

Теорема 12.2.1. Если задача (12.10) при условиях (12.11)–(12.13) имеет решение в $\Omega \times P$, то оно единственно в $\Omega \times P$.

Доказательство. Пусть $z_1 = \{u_1, p_1\} \in \Omega \times P$, $z_2 = \{u_2, p_2\} \in \Omega \times P$ — решения задачи (12.10), тогда

$$\begin{cases} A(u_1) - A(u_2) + B(p_1 - p_2) = 0, \\ B^*(u_1 - u_2) = 0. \end{cases} \quad (12.15)$$

Умножим скалярно первое уравнение (12.15) на $u_1 - u_2$, второе — на $p_1 - p_2$ и вычтем одно из другого. В результате получим

$$(A(u_1) - A(u_2), u_1 - u_2) = 0.$$

Отсюда и из условия (12.11) следует неравенство

$$0 \geq \delta \|u_1 - u_2\|_Q^2$$

при $\delta > 0$, поэтому $u_1 = u_2$ и $B(p_1 - p_2) = 0$. Так как $\ker B = \{0\}$, то и $p_1 = p_2$. Теорема доказана. ■

Теорема 12.2.2. Пусть $\Omega = U$ и выполнены условия (12.11)–(12.13), тогда существует решение задачи (12.10).

Доказательство. Обозначим через P_H ортогональный проектор пространства U на подпространство $H \equiv \ker B^*$, а через u_0 единственное решение задачи

$$B^* u = \varphi, \quad P_H u = 0.$$

Так как

$$\sigma(Q^{-1/2}BC^{-1}B^*Q^{-1/2}) \subseteq \{0\} \cup [\gamma, \Gamma]$$

и $P_H u_0 = 0$, то имеет место неравенство

$$\|u_0\|_Q = (Qu_0, u_0)^{1/2} \leq \gamma^{-1/2} (BC^{-1}B^*u_0, u_0)^{1/2} = \gamma^{-1/2} \|\varphi\|_{C^{-1}}.$$

Положим $u = u_0 + v$, тогда задача (12.10) равносильна следующей —

$$A(u_0 + v) + Bp - f = 0, \quad v \in H = \ker B^*, \quad p \in P. \quad (12.16)$$

Решение (12.16) сводится к последовательному нахождению нуля v_0 непрерывного векторного поля $F: H \rightarrow H$

$$F(v) = P_H(A(u_0 + v) - f)$$

и p_0 , удовлетворяющего уравнению

$$Bp = -(I - P_H)(A(u_0 + v_0) - f).$$

Последнее уравнение совместно, так как $\text{Im}(I - P_H) = \text{Im } B$.

Для любых $v \in H$ таких, что

$$\|v\|_Q > \delta^{-1} \|f - A(u_0)\|_{Q^{-1}},$$

имеет место неравенство

$$\begin{aligned} (F(v), v) &= (A(u_0 + v) - A(u_0), v) - (f - A(u_0), v) \geq \\ &\geq \delta \|v\|_Q^2 - \|f - A(u_0)\|_{Q^{-1}} \|v\|_Q > 0, \end{aligned}$$

откуда следует, что поле F гомотопно тождественному векторному полю I ($Iv = v \forall v \in H$) на «сфере»

$$\{v \in H \mid \|v\|_Q = \delta^{-1} \|f - A(u_0)\|_{Q^{-1}} + \rho\}$$

для любого $\rho > 0$, поэтому [54, с. 20] принимает нулевое значение в некоторой точке множества

$$\|v\|_Q \leq \delta^{-1} \|f - A(u_0)\|_{Q^{-1}}.$$

Теорема доказана. ■

12.2.2. Вспомогательные факты и утверждения

Для анализа алгоритма воспользуемся техникой обобщенных производных (см. определение 6.4.2).

Лемма 12.2.7. Оператор $A(u)$ обладает обобщенной производной всюду в Ω .

Доказательство. В силу теоремы Радемахера (теорема 6.4.5), множество $\mathcal{D}(A)$ всюду плотно в Ω и, следовательно, $\partial A(u)$ определена во всех $u \in [\mathcal{D}(A)] = \Omega$. Лемма доказана. ■

Лемма 12.2.8. Пусть $A_0 \in \text{conv} \{\partial A(u) : u \in \Omega\}$, тогда имеют место следующие неравенства:

$$\delta Q \leq A_0, \quad \left\| Q^{-1/2} A_0 Q^{-1/2} \right\| \leq \Delta.$$

Доказательство. Пусть $u_0 \in \mathcal{D}(A)$, тогда $A_0 = A'(u_0)$ и следовательно,

$$A_0 u = \lim_{\lambda \rightarrow +0} \lambda^{-1} (A(u_0 + \lambda u) - A(u_0)) \quad \forall u \in U,$$

откуда, в силу (12.11), справедливо

$$(A_0 u, u) = \lim_{\lambda \rightarrow +0} \lambda^{-1} (A(u_0 + \lambda u) - A(u_0), u) \geq \delta \|u\|_Q^2 = \delta (Qu, u) \quad \forall u \in U.$$

С другой стороны, имеется оценка

$$\begin{aligned} \left\| Q^{-1/2} A_0 Q^{-1/2} u \right\| &= \left\| A_0 Q^{-1/2} u \right\|_{Q^{-1}} = \\ &= \lim_{\lambda \rightarrow +0} \lambda^{-1} \left\| A(u_0 + \lambda Q^{-1/2} u) - A(u_0) \right\|_{Q^{-1}} \leq \\ &\leq \Delta \left\| Q^{-1/2} u \right\|_Q = \Delta \|u\| \quad \forall u \in U. \end{aligned}$$

Если $A_0 \in \partial A(u_0)$ и существует последовательность $u_k \in \mathcal{D}(A)$ такая, что

$$\lim_{k \rightarrow +\infty} u_k = u_0, \quad A_0 = \lim_{k \rightarrow +\infty} A'(u_k),$$

то указанные выше неравенства сохраняются по непрерывности.

Пусть $A_0 \in \text{conv} \{\partial A(u) : u \in \Omega\}$, тогда, в силу теоремы Каратеодоре (теорема 6.4.2), существует число $k \in \mathbb{N}$, $k \leq (\dim U)^2 + 1$, точки $u_1, \dots, u_k \in \Omega$ и операторы $A_1 \in \partial A(u_1), \dots, A_k \in \partial A(u_k)$ такие, что

$$A_0 = \sum_{i=1}^k \lambda_i A_i$$

для некоторых $\lambda_i > 0$, удовлетворяющих условию $\lambda_1 + \dots + \lambda_k = 1$. Отсюда следуют неравенства

$$A_0 \geq \sum_{i=1}^k (\lambda_i \delta Q) = \delta Q,$$

$$\left\| Q^{-1/2} A_0 Q^{-1/2} \right\| \leq \sum_{i=1}^k \lambda_i \left\| Q^{-1/2} A_i Q^{-1/2} \right\| \leq \sum_{i=1}^k \lambda_i \Delta = \Delta.$$

Лемма доказана. ■

12.2.3. Оценка скорости сходимости

Представим алгоритм (12.14) в виде $z^{k+1} = T(z^k)$, где $z = \{u, p\}$ и

$$T(z) = \begin{pmatrix} Q & 0 \\ 0 & \alpha\tau C \end{pmatrix}^{-1} \begin{pmatrix} Q & -\tau B \\ \tau B^* & \alpha\tau C - \tau^2 B^* Q^{-1} B \end{pmatrix} \times \\ \times \begin{pmatrix} u - \tau Q^{-1}(A(u) + \beta BC^{-1} B^* u) \\ p \end{pmatrix}.$$

Пусть $T_0 \in \text{conv} \{\partial T(z): z \in \Omega \times P\}$, тогда справедливо представление

$$T_0 = \begin{pmatrix} Q & 0 \\ 0 & \alpha\tau C \end{pmatrix}^{-1} \begin{pmatrix} Q & -\tau B \\ \tau B^* & \alpha\tau C - \tau^2 B^* Q^{-1} B \end{pmatrix} \times \\ \times \begin{pmatrix} I - \tau Q^{-1}(A_0 + \beta BC^{-1} B^*) & 0 \\ 0 & I \end{pmatrix},$$

где $A_0 \in \text{conv} \{\partial A(u): u \in \Omega\}$.

Определим линейные операторы

$$L = \frac{1}{2} Q^{-1/2} (A_0 + A_0^*) Q^{-1/2}, \\ G = Q^{-1/2} B C^{-1/2}, \\ J = \frac{1}{2} Q^{-1/2} (A_0 - A_0^*) Q^{-1/2},$$

тогда T_0 можно переписать в форме $T_0 = M^{-1/2} T_1 M^{1/2}$, где

$$M = \begin{pmatrix} Q & 0 \\ 0 & \alpha\tau C \end{pmatrix}, \\ T_1 = \begin{pmatrix} D & -\sqrt{\tau/\alpha} G \\ \sqrt{\tau/\alpha} G^* D & I - \tau/\alpha G^* G \end{pmatrix}, \\ D = I - \tau(L + \beta G G^* + J).$$

Кроме того, из условия (12.13) и леммы 12.2.8 следуют неравенства:

$$0 < \delta I \leq L \leq \Delta I, \quad 0 < \gamma I \leq G^* G \leq \Gamma I, \quad \|J\| = R \leq \Delta.$$

Оценим величину $\|T_0\|_M$. Справедлива

Лемма 12.2.9. Пусть $\mu \in (0, 1)$ и выполнены неравенства

$$\beta \geq 0, \quad \alpha > \frac{\Gamma}{\delta}, \quad 0 < \tau < \min \left\{ \frac{\alpha}{2\Gamma}, \quad 2\mu \left(\delta - \frac{\Gamma}{\alpha} \right) R^{-2}, \quad \frac{2(1-\mu)}{\delta + \Delta + \beta\Gamma} \right\},$$

тогда имеет место оценка

$$\|T_0\|_M \leq \max \left\{ \left(1 + \frac{\tau\Gamma}{\alpha}\right) \left(1 - \tau\delta + \frac{1}{2}\mu^{-1}\tau^2 R^2\right), \right. \\ \left. \sqrt{1 - \frac{\tau\gamma}{\alpha} + 2\left(\frac{\tau\gamma}{\alpha}\right)^2}, \sqrt{1 - \frac{\tau\Gamma}{\alpha} + 2\left(\frac{\tau\Gamma}{\alpha}\right)^2} \right\} = q < 1.$$

Доказательство. В силу представления $T_0 = M^{-1/2}T_1M^{1/2}$, имеет место равенство

$$\|T_0\|_M^2 = \|T_1\|^2 = \sup_{z \in Z \setminus \{0\}} \frac{(T_1 z, T_1 z)}{(z, z)}.$$

Для любого $z \in Z$ справедливы неравенства

$$(T_1 z, T_1 z) = \left(\left(D^* D + \frac{\tau}{\alpha} D^* G G^* D \right) u, u \right) - \\ - 2 \left(\sqrt{\frac{\tau}{\alpha}} G^* D u, \frac{\tau}{\alpha} G^* G p \right) + \left(\left(\left(I - \frac{\tau}{\alpha} G^* G \right)^2 + \frac{\tau}{\alpha} G^* G \right) p, p \right) \leq \\ \leq \left(\left(D^* D + \frac{2\tau}{\alpha} D^* G G^* D \right) u, u \right) + \\ + \left(\left(\left(I - \frac{\tau}{\alpha} G^* G \right)^2 + \frac{\tau}{\alpha} G^* G + \left(\frac{\tau}{\alpha} G^* G \right)^2 \right) p, p \right),$$

откуда следует оценка

$$\|T_1\| \leq \max \left\{ \left(1 + \frac{\tau\Gamma}{\alpha}\right) \|D\|, \max_{t \in [\gamma, \Gamma]} \sqrt{1 - \frac{\tau t}{\alpha} + 2\left(\frac{\tau t}{\alpha}\right)^2} \right\}.$$

Пусть $\mu \in (0, 1)$ и справедливо неравенство

$$\tau \leq 2 \frac{1 - \mu}{\delta + \Delta + \beta\Gamma},$$

тогда можно оценить $\|D\|$:

$$\|D\| \leq \|(1 - \mu) - \tau(L + \beta G G^*)\| + \|\mu - \tau J\| \leq \\ \leq \max \left\{ |(1 - \mu) - \tau\delta|, |(1 - \mu) - \tau(\Delta + \beta\Gamma)| \right\} + \sqrt{\mu^2 + \tau^2 R^2} \leq \\ \leq (1 - \mu) - \tau\delta + \sqrt{\mu^2 + \tau^2 R^2} \leq 1 - \tau\delta + \frac{1}{2}\mu^{-1}\tau^2 R^2.$$

Таким образом, если выполнены условия леммы, то имеют место оценки:

$$\left(1 + \frac{\tau\Gamma}{\alpha}\right) \|D\| \leq \left(1 + \frac{\tau\Gamma}{\alpha}\right) \left(1 - \tau\delta + \frac{1}{2}\mu^{-1}\tau^2 R^2\right) < 1$$

и

$$1 - \frac{\tau\Gamma}{\alpha} + 2 \left(\frac{\tau\Gamma}{\alpha} \right)^2 < 1,$$

которые требовалось доказать. Лемма доказана. ■

Справедлива основная

Теорема 12.2.3. Пусть $\mu \in (0, 1)$ и выполнены неравенства

$$\beta \geq 0, \alpha > \frac{\Gamma}{\delta}, 0 < \tau < \min \left\{ \frac{\alpha}{2\Gamma}, 2\mu \left(\delta - \frac{\Gamma}{\alpha} \right) R^{-2}, \frac{2(1-\mu)}{\delta + \Delta + \beta\Gamma} \right\},$$

тогда для любого начального приближения $z^0 = \{u^0, p^0\} \in \Omega \times P$ такого, что

$$\{z \in Z : \|z - \tilde{z}\|_M \leq \|z^0 - \tilde{z}\|_M\} \subseteq \Omega \times P,$$

где $\tilde{z} \in \Omega$ — решение задачи (12.10), алгоритм (12.14) корректно определен и сходится к \tilde{z} со скоростью геометрической прогрессии в норме $\|\cdot\|_M$ с показателем

$$q = \max \left\{ \left(1 + \frac{\tau\Gamma}{\alpha} \right) \left(1 - \tau\delta + \frac{1}{2}\mu^{-1}\tau^2 R^2 \right), \sqrt{1 - \frac{\tau\gamma}{\alpha} + 2 \left(\frac{\tau\gamma}{\alpha} \right)^2}, \sqrt{1 - \frac{\tau\Gamma}{\alpha} + 2 \left(\frac{\tau\Gamma}{\alpha} \right)^2} \right\}.$$

Доказательство. Множество

$$S \equiv \{z \in Z : \|z - \tilde{z}\|_M \leq \|z^0 - \tilde{z}\|_M\} \subseteq \Omega \times P$$

является выпуклым, так как является шаром в норме $\|\cdot\|_M$. Поэтому для любого $z \in S$ имеет место включение $[\tilde{z}, z] \subset M_0$ и, по теореме о среднем (теорема 6.4.6), справедливо

$$T(z) - \tilde{z} = T(z) - T(\tilde{z}) = T_0(z - \tilde{z})$$

для некоторого $T_0 \in \text{conv} \{\partial T(z_0) : z_0 \in [\tilde{z}, z]\}$.

В силу леммы 12.2.9, имеет место оценка

$$\|T_0\|_M \leq q < 1,$$

и следовательно, выполнено

$$\|T(z) - \tilde{z}\|_M \leq \|T_0\|_M \|z - \tilde{z}\|_M \leq q \|z - \tilde{z}\|_M.$$

Отсюда следует, что $T(z) \in S$, значит, алгоритм корректно определен для любых начальных приближений из S и сходится со скоростью геометрической прогрессии с показателем, не превышающим q . Теорема доказана. ■

Следствие 12.2.1 (Асимптотика оценки). Пусть выполнены условия теоремы 12.2.3, тогда существует набор итерационных параметров, при котором справедливо

$$q \leq 1 - \frac{1}{16} \omega^2 \xi \quad \text{при } \xi \rightarrow 0, \omega = \text{const},$$

если же $R = 0$ (т. е. все производные оператора $A(u)$ симметричны), то существует набор итерационных параметров, при котором —

$$q \leq 1 - \frac{1}{16} \omega \xi \quad \text{при } \xi \rightarrow 0, \omega = \text{const}.$$

Доказательство. Положим $\alpha_0 = 2\Gamma/\delta$, $\beta_0 = 0$, $\mu = 1/2$ и выберем τ_0 из условия

$$\tau_0 = \min \left\{ \frac{1}{\delta}, \frac{1}{2} \delta R^{-2}, \frac{1}{\delta + \Delta} \right\}.$$

В общем случае $R = \Delta$ и, значит, имеем $\tau_0 = \omega/(2\Delta)$. Подставляя эти значения в оценку теоремы 12.2.3, получаем, что при фиксированном ω и достаточно малых $\xi > 0$ справедливо

$$q < 1 - \frac{\tau_0 \gamma}{2\alpha_0} \left(1 - \frac{2\tau_0 \gamma}{\alpha_0} \right) \leq 1 - \frac{\tau_0 \gamma}{4\alpha_0} = 1 - \frac{1}{16} \omega^2 \xi.$$

Если $R = 0$ (т. е. все производные оператора $A(u)$ симметричны), то получаем $\tau_0 = 1/(\delta + \Delta)$, значит при фиксированном ω и достаточно малых $\xi > 0$ справедливы оценки

$$q < 1 - \frac{\tau_0 \gamma}{4\alpha_0} = 1 - \frac{1}{8(1 + \omega)} \omega \xi \leq 1 - \frac{1}{16} \omega \xi.$$

Следствие доказано. ■

12.3. БИБЛИОГРАФИЯ И КОММЕНТАРИИ

Неиссякаемым источником постановок седловых задач с нелинейным кососимметрическим возмущением служит численное решение как стационарных, так и нестационарных уравнений Навье—Стокса, описывающих течения вязкой несжимаемой жидкости. Различные способы их дискретизации по пространству и по времени [13, 31, 32, 158], а также широкий спектр предобуславливающих операторов, учитывающих ту или иную специфику постановки, обозначают не поддающиеся оценкам масштабы приложения развиваемой теории итерационных методов.

В свою очередь, седловые задачи с сильно монотонным оператором — порождение интенсивно развивающейся области математического моделирования связанного, в первую очередь, с постановками задач в форме вариационных неравенств.

Отметим недавние результаты [66, 67] по применению и обоснованию трехпараметрического алгоритма типа GMSOR для моделирования течений бингамовской жидкости в регуляризованной постановке; в седловых операторах таких задач одновременно присутствуют слагаемые как кососимметричного, так и сильно монотонного типа.

Результаты главы получены в [20, 21].

Актуальность разработки такого рода алгоритмов обоснована простотой реализации и более широкой областью сходимости по сравнению с методами ньютоновского типа.

ЧАСТЬ III

ПРИЛОЖЕНИЕ К ГИДРОДИНАМИКЕ

Итерационные методы решения седловых задач, которым была посвящена основная часть книги, содержат в своей основе два предобусловливающих оператора и несколько скалярных параметров, выбор которых существенно влияет на вычислительную эффективность. Как уже говорилось в предисловии, внимание к исследованию алгоритмов такого рода было инициировано интересом к решению уравнений типа Навье—Стокса. Поэтому основная идея приложения — сделать теоретические результаты исследований более доступными для практики. Следует заметить, что в дискретном аналоге гидродинамических уравнений Стокса матрица A является векторным сеточным оператором Лапласа, которому сопутствуют заданные граничные условия (часто, типа Дирихле и однородные). Теория решения таких матричных задач хорошо разработана (см., например, [41, 64, 76, 159, 183, 200], хорошее изложение современных подходов имеется в [29], полезно также знакомство с [44]), поэтому построение предобусловливателей для них, как правило, не вызывает затруднений.

Совершенно по-другому обстоит дело с дополнением по Шуру

$$S_0 = B^T A^{-1} B$$

для оператора исходной седловой задачи, порожденной дискретизацией уравнений Стокса.

Сначала необходимо принять решение, требуется ли для S_0 предобусловливатель вообще или можно ограничиться тождественным оператором? Ответ на поставленный вопрос следует из оценки константы Ладыженской в inf-sup-неравенстве. С этой целью в первой части приложения собрана информация о точных формулах и асимптотических свойствах этой важной постоянной в зависимости от геометрических характеристик области. Особое внимание фиксируется на двух факторах: анизотропии области и наличию угловых точек на границе, так как именно они существенно влияют на значение константы Ладыженской. Отметим, что этот раздел связан исключительно с дифференциальными свойствами задачи.

Далее требуется построить дискретный аналог задачи Стокса, удовлетворяющий условию Ладыженской—Бабушки—Брецци, что необходимо для корректности получающейся алгебраической седловой задачи. Таким построениям также посвящена обширная литература

(см., например, [127, 158] и цитированную там литературу), что свидетельствует о наличии удовлетворительной теории в этой области. После того, как зафиксирована какая-либо аппроксимация возникает трудный вопрос: насколько ее LBB-константа близка к константе Ладыженской? Источник этой трудности — некомпактность дифференциального прообраза оператора S_0 , а проявляется это свойство в наличии дополнительных точек сгущения собственных значений и в негладкости собственных функций. Чтобы получить представление о влиянии анизотропии области и угловых точек на сходимость по параметру дискретизации решения спектральной задачи для оператора S_0 , во второй части приложения приведены численные эксперименты для областей следующего вида: концентрическое кольцо (анизотропия искусственно убрана, угловых точек нет), квадрат (анизотропии нет, имеются особенности у собственных функций в угловых точках), вытянутый прямоугольник (анизотропия резко выделена, особенности у собственных функций в угловых точках ослаблены). Отметим, что для дискретизаций компактных операторов, порожденных седловыми задачами, необходимая теория сходимости имеется в [119] (см. также [120]).

Наконец, в третьей части приложения приводятся численные эксперименты по использованию трехпараметрического метода GMSOR для решения модельных задач для уравнений Навье—Стокса. Здесь во главу угла поставлен актуальный вопрос о влиянии члена $\beta BC^{-1}B^T u^k$ на сходимость метода. Необходимость в подобном анализе возникла, с одной стороны, из-за неулучшаемых результатов, полученных для линейных симметричных задач — $\beta = 0$ (т. е. слагаемое не увеличивает скорость сходимости метода и, казалось бы, введение его в алгоритм не дает преимуществ), а с другой стороны, из-за отсутствия аналогичных результатов об оптимальных характеристиках метода при решении нелинейных седловых задач.

Конечно, важной для приложений и теоретически интересной является тематика построения предобуславливающих операторов для дискретизаций уравнений Навье—Стокса на основе многосеточных методов. Она находится несколько в стороне от научных интересов авторов, поэтому рекомендуется знакомство с основными идеями работ [121, 193], а также см. [71] и цитированную там литературу.

INF-SUP НЕРАВЕНСТВО И СМЕЖНЫЕ ВОПРОСЫ

Введем обозначения функциональных пространств

$$P = \left\{ p \mid p \in L_2(\Omega), \int_{\Omega} p d\Omega = 0 \right\}, \quad U = (W_2^1(\Omega))^s,$$

где $s = 2, 3$ — размерность задачи, $\Omega \subset \mathbb{R}^s$. Их определения и свойства можно найти, например, в [4, 61]. Будем предполагать, что область Ω является односвязной ограниченной с липшицевой границей $\partial\Omega$ [4, с. 12].

Рассмотрим вопрос о наилучшем значении постоянной \mathcal{L} в inf — sup неравенстве, т. е. величине $\mathcal{L} = \mathcal{L}(\Omega)$, зависящей только от области Ω и определяемой выражением

$$\inf_{p \in P} \sup_{u \in U} \frac{|(p, \operatorname{div} u)|}{\|u\|_U \|p\|_P} \geq \mathcal{L}, \quad (13.1)$$

где

$$\|u\|_U = (\operatorname{grad} u, \operatorname{grad} u)^{1/2}, \quad \|p\|_P = (p, p)^{1/2}, \quad (\varphi, \psi) = \int_{\Omega} \varphi \psi d\Omega.$$

Это неравенство играет принципиальную роль при исследовании существования решений линейных задач гидродинамики и теории упругости. Осознание его важности для теории уравнений в частных производных традиционно связывают с именами О. А. Ладыженской и И. Нечаса и относят к 60-м годам XX века. Одно из последних (по времени) доказательств его справедливости, причем изложенное в современной терминологии, приведено в [122].

Обобщению его на более широкий класс областей, функциональных пространств и эквивалентности формулировок, включая использование комплекснозначных функций, посвящена работа [53]. В частности, там имеется пример области, для которой inf-sup-неравенство не выполнено: если $\Omega \subset \mathbb{C}$, то в полярных координатах (ρ, φ) границу $\partial\Omega$ можно задать как совокупность трех кривых $-\rho = -1/\ln(\varphi)$, $\rho = -1/\ln(-\varphi)$, $\rho = 1$ (область с внешним «клювом»).

Основной интерес к неравенству (13.1) возник примерно в середине 70-х годов XX века, когда для обоснования сходимости методов дискретизации для уравнений типа Навье—Стокса потребовался его

конечномерный аналог: пусть определены конечномерные пространства $U_h \subset U$ и $P_h \subset P$, тогда существует такая постоянная $\mathcal{L}_0 > 0$, не зависящая от параметра дискретизации области h , что справедливо неравенство

$$\inf_{p_h \in P_h} \sup_{u_h \in U_h} \frac{|(p_h, \operatorname{div} u_h)|}{\|u_h\|_U \|p_h\|_P} = \mathcal{L}_h \geq \mathcal{L}_0 \quad \forall h > 0. \quad (13.2)$$

Начало этого этапа связывают с работой [126]. В настоящее время за неравенством (13.2) прочно установилось название *условие Ладыженской–Бабушки–Брецици*. Конструированию пространств P_h и U_h , для которых справедливо LBB-условие посвящено большое количество работ (см., например, [127, 158] и цитированную там литературу). Отметим, что во второй части приложения будут затронуты некоторые важные аспекты, связанные с близостью величин \mathcal{L} и \mathcal{L}_0 . Учитывая исключительный вклад О. А. Ладыженской в развитие этой области математики, будем называть постоянные \mathcal{L} и \mathcal{L}_0 *константой Ладыженской* и *LBB-константой*, соответственно.

13.1. О ЗАДАЧЕ СТОКСА И СПЕКТРЕ КОССЕРА

Рассмотрим оператор давления S_0 , ассоциированный с первой краевой задачей Стокса

$$\begin{aligned} -\Delta u + \operatorname{grad} p &= f && \text{в } \Omega, \\ \operatorname{div} u &= 0 && \text{в } \Omega, \\ u &= 0 && \text{на } \partial\Omega. \end{aligned} \quad (13.3)$$

Перепишем (13.3) в следующем равносильном виде:

$$\begin{aligned} u &= (\Delta)_0^{-1}(\operatorname{grad} p - f), \\ S_0 p &\equiv \operatorname{div} (\Delta)_0^{-1} \operatorname{grad} p = \operatorname{div} (\Delta)_0^{-1} f, \end{aligned} \quad (13.4)$$

где выражение $(\Delta)_0^{-1}g$ для $g \in U^{-1}$ обозначает вектор-функцию $v \in U$, такую что $\Delta v = g$. Оператор $S_0 : P \rightarrow P$ часто называют дополнением по Шуру для оператора задачи (13.3), или оператором давления, ассоциированным с первой краевой задачей Стокса. Видимо, одно из первых его исследований было проведено в работе [144]. Там были установлены замечательные свойства оператора S_0 : самосопряженность, положительная определенность, дискретность и ограниченность спектра, наличие полной ортонормированной в P счетной системы собственных функций. При этом предполагалась достаточная гладкость границы области Ω .

В развернутом виде спектральную задачу $S_0 p = \lambda p$ можно записать как

$$\begin{aligned} -\Delta \mathbf{u} + \operatorname{grad} p &= 0 & \text{в } \Omega, \\ \operatorname{div} \mathbf{u} &= \lambda p & \text{в } \Omega, \\ \mathbf{u} &= 0 & \text{на } \partial\Omega, \end{aligned} \quad (13.5)$$

что устанавливает близость со спектром пучка операторов теории упругости, или спектром Коссера [68, 69]:

$$\begin{aligned} \Delta \mathbf{u} + \omega \operatorname{grad} \operatorname{div} \mathbf{u} &= 0 & \text{в } \Omega, \\ \mathbf{u} &= 0 & \text{на } \partial\Omega \end{aligned} \quad (13.6)$$

при $\omega = -\lambda^{-1}$. Суммируя результаты отмеченных выше исследований, нужно подчеркнуть, что в случае гладкой границы, единственным «неприятным» свойством оператора S_0 является некомпактность, которое связано с наличием особых точек спектра: $\lambda = 1$ — изолированное значение бесконечной кратности и $\lambda = 1/2$ — точка сгущения конечнократных собственных значений (за исключением области круговой формы, где оно также бесконечнократно).

Покажем, что это свойство не влияет на определение величины \mathcal{L} . Для этого потребуется связь минимального собственного значения оператора S_0 с постоянной \mathcal{L} из (13.1). Рассмотрим следующую цепочку равенств, справедливых для произвольной функции $q \in W_2^1(\Omega) \cap P$:

$$\begin{aligned} (S_0 q, q) &= (\operatorname{div} (\Delta)_0^{-1} \operatorname{grad} q, q) = -((\Delta)_0^{-1} \operatorname{grad} q, \operatorname{grad} q) = \\ &= -((\Delta)_0^{-1} \operatorname{grad} q, \Delta (\Delta)_0^{-1} \operatorname{grad} q) = (-\Delta v, v)|_v = (\Delta)_0^{-1} \operatorname{grad} q = \\ &= \|v\|_U^2 = \sup_{\mathbf{u} \in U} \frac{(-\Delta v, \mathbf{u})^2}{\|\mathbf{u}\|_U^2} = \sup_{\mathbf{u} \in U} \frac{(\operatorname{grad} q, \mathbf{u})^2}{\|\mathbf{u}\|_U^2} = \sup_{\mathbf{u} \in U} \frac{(q, \operatorname{div} \mathbf{u})^2}{\|\mathbf{u}\|_U^2}. \end{aligned}$$

Множество функций из $W_2^1(\Omega)$ всюду плотно в P (см. [45]), откуда следует корректность замыкания в равенствах:

$$(S_0 q, q) = \|v\|_U^2 = \sup_{\mathbf{u} \in U} \frac{(q, \operatorname{div} \mathbf{u})^2}{\|\mathbf{u}\|_U^2}. \quad (13.7)$$

Таким образом, равенства (13.7) справедливы для произвольной $q \in P$ и $v = \Delta_0^{-1} \operatorname{grad} q \in U$. В частности, из (13.1) и (13.7) следует, что

$$\lambda_{\min}(S_0) = \inf_{p \in P} \frac{(S_0 p, p)}{\|p\|_P^2} = \inf_{p \in P} \sup_{\mathbf{u} \in U} \frac{(p, \operatorname{div} \mathbf{u})^2}{\|\mathbf{u}\|_U^2 \|p\|_P^2} = \mathcal{L}^2,$$

где $\lambda_{\min}(S_0)$ — минимальное собственное значение оператора S_0 . Легко заметить, что $\lambda_{\max}(S_0) = 1$, поэтому интересует оценка только для нижней границы спектра. Этот интерес обусловлен асимптотической

связью $\lambda_{\min}(S_0) \approx \gamma/\Gamma$, существенно влияющей на эффективность всех рассмотренных методов из основной части книги. Например, метод Узаы – сопряженных градиентов, примененный непосредственно к задаче (13.4), имеет асимптотическую скорость сходимости с показателем $q_0 = (1 - \mathcal{L})/(1 + \mathcal{L})$.

13.2. НЕРАВЕНСТВА ФРИДРИХСА И КОРНА В ДВУХМЕРНОМ СЛУЧАЕ

Пусть функция $h(z) = f(x_1, x_2) + ig(x_1, x_2)$ (здесь $z = x_1 + ix_2$) аналитична в ограниченной области $\Omega \subset \mathbb{R}^2$, тогда неравенство

$$\int_{\Omega} f^2 d\Omega \leq \mathcal{F} \int_{\Omega} g^2 d\Omega, \quad (13.8)$$

при условии $\int_{\Omega} f d\Omega = 0$, называют неравенством Фридрихса [155]. Здесь постоянная \mathcal{F} зависит только от формы области. Как и выше, будем иметь в виду ее наилучшее (наименьшее) значение.

Рассмотрим вектор-функцию $\mathbf{u} = (u_1, u_2)^T$, $u_i \in W_2^1(\Omega)$, $i = 1, 2$ и определим квадратичные функционалы

$$D(\mathbf{u}) = \int_{\Omega} \sum_{i,j=1}^2 \left| \frac{\partial u_i}{\partial x_j} \right|^2 d\Omega,$$

$$E(\mathbf{u}) = \frac{1}{4} \int_{\Omega} \sum_{i,j=1}^2 |e_{ij}(\mathbf{u})|^2 d\Omega, \quad e_{ij}(\mathbf{u}) = \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}.$$

Если \mathbf{u} удовлетворяет ограничениям

$$\int_{\Omega} \left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i} \right) d\Omega = 0, \quad 1 \leq i, j \leq 2,$$

то неравенство

$$D(\mathbf{u}) \leq \mathcal{K} E(\mathbf{u}) \quad (13.9)$$

называют неравенством Корна второго рода [163, 175]. В данном случае также речь идет о наилучшем значении \mathcal{K} , зависящем только от формы области $\Omega \subset \mathbb{R}^2$. Отметим, что неравенства (13.8) и (13.9) известны в различных эквивалентных формулировках, которые можно найти, например, в [164].

Введем обозначение $\mathcal{B} = -\omega_{\min}$ для постоянной, связанной с нижней границей (так как $\omega < 0$) конечнократного спектра Коссера (13.6), которую иногда называют постоянной из неравенства Бабушки–Азиза [110], и приведем соотношения, связывающие

величины $\mathcal{F}, \mathcal{K}, \mathcal{B}$ [164] между собой и с константой Ладыженской \mathcal{L} в случае, когда область Ω является двумерной и односвязной:

$$\mathcal{K} = 2(1 + \mathcal{F}), \quad \mathcal{B} = \mathcal{F} + 1, \quad \mathcal{L}^2 = \mathcal{B}^{-1}. \quad (13.10)$$

Обратим внимание, что первые три постоянные связаны между собой линейно, а величина \mathcal{L}^2 имеет обратную с ними зависимость. Поэтому оценки сверху для $\mathcal{F}, \mathcal{K}, \mathcal{B}$ порождают оценки снизу для \mathcal{L} и наоборот.

В двумерном случае равенства (13.10) играют очень важную роль; для их аналогов в случае трех пространственных переменных (за исключением неравенства Фридрихса) представляет интерес работа [195].

13.3. ТОЧНЫЕ ЗНАЧЕНИЯ И ОЦЕНКИ СНИЗУ КОНСТАНТЫ ЛАДЫЖЕНСКОЙ

Несмотря на указанную выше взаимосвязь постоянных в функциональных неравенствах (13.1), (13.8), (13.9) и, соответственно, наличие своего специфического инструментария для их исследований, точные значения констант являются большой редкостью. Приведем известные примеры:

- круг произвольного радиуса [155]

$$\mathcal{L} = \frac{1}{\sqrt{2}};$$

- эллипс с полуосями a и b ($a \leq b$) [155]

$$\mathcal{L} = \sqrt{\frac{a^2}{a^2 + b^2}};$$

- область, ограниченная «улиткой Паскаля»,

$$\Omega = \{(r, \varphi) \mid 0 < r < a(1 + \varepsilon \cos \varphi), \quad 0 \leq \varphi \leq 2\pi\},$$

где a, ε — постоянные, $0 \leq \varepsilon \leq 1$ [162], $\mathcal{L} = \sqrt{2 - \varepsilon^2}/2$.

Эти формулы были получены на основе неравенств Фридрихса и Корна.

Приведем формулы для решения спектральной задачи (13.5) для области концентрической кольцевой формы [134, 138]

$$\Omega = \{(r, \varphi) \mid R_1 < r < R_2, \quad 0 \leq \varphi \leq 2\pi\}$$

в зависимости от величины отношения радиусов $x = R_2/R_1 > 1$.

Все собственные значения λ задачи (13.5) принадлежат множеству

$$\Lambda = \{1\} \cup \left\{ \frac{1}{2} \left(1 \pm \sqrt{\frac{x^2 - 1}{x^2 + 1} \frac{1}{\ln x}} \right) \right\} \cup \\ \cup \left\{ \frac{1}{2} \left(1 \pm \frac{(x^{m+1} - x^{m-1})\sqrt{m^2 - 1}}{\sqrt{(x^{2(m+1)} - 1)(x^{2(m-1)} - 1)}} \right), m = \pm 2, \pm 3, \dots \right\}$$

и все собственные функции $p_m(r, \varphi)$, соответствующие $\lambda_m \neq 1$, имеют вид

$$Cr \left(1 \mp \left[\frac{R_1}{r} \right]^2 \sqrt{\frac{x^4 - 1}{4 \ln x}} \right) e^{im\varphi}, \quad |m| = 1, \\ Cr^k \left(1 \mp \left[\frac{R_1}{r} \right]^{2k} x^{k-1} \sqrt{\frac{k-1}{k+1} \frac{x^{2(k+1)} - 1}{x^{2(k-1)} - 1}} \right) e^{im\varphi}, \quad k = |m| > 1,$$

с некоторой произвольной постоянной $C \neq 0$.

Отсюда следует, что для области, имеющей форму концентрического кольца, значение константы Ладыженской равно

$$\mathcal{L} = \sqrt{\frac{1}{2} \left(1 - \sqrt{\frac{x^2 - 1}{x^2 + 1} \frac{1}{\ln x}} \right)}, \quad x = \frac{R_2}{R_1} > 1. \quad (13.11)$$

Отметим, что в работе [147] на основе неравенства Корна второго рода была сделана попытка решить спектральную задачу, эквивалентную $S_0 p = \lambda p$, и определить соответствующее значение \mathcal{K} . Однако, результат получился ошибочным. Поэтому, в качестве научного курьеза, любопытно ознакомиться с комментариями в работе [167] по поводу расхождения примерно в два раза численных (верных!) и аналитических (неверных!) значений константы Ладыженской при различных значениях R_2/R_1 . При этом, конечно, результаты расчетов [167] являются хорошей иллюстрацией к формуле (13.11).

В случае трех пространственных переменных также известны точные значения константы Ладыженской для некоторых областей:

- шар произвольного радиуса [140]

$$\mathcal{L} = \frac{1}{\sqrt{3}};$$

- эллипсоид с полуосями a, b, c ($a \leq b \leq c$) [141]

$$\mathcal{L} = \sqrt{\frac{a^2 b^2}{a^2 b^2 + b^2 c^2 + c^2 a^2}};$$

- для конечнократных значений спектра Коссера в шаровом слое [69, 142] известно квадратное уравнение

$$\left(\omega_n + \frac{2n+1}{n}\right) \left(\omega_n + \frac{2n+1}{n+1}\right) = \\ = \frac{(2n-1)(2n+3)}{4} \frac{(R_2^2 - R_1^2)\omega_n^2}{(R_2^{2n+3} - R_1^{2n+3})(R_2^{-2n+1} - R_1^{-2n+1})}.$$

Здесь R_2, R_1 ($R_2 > R_1$) радиусы сфер, ограничивающие слой, $n = 1, 2, \dots$ — номер собственного значения. В данном случае, если ω_1 — минимальное (отрицательное) решение этого уравнения при $n = 1$, то $\mathcal{L} = \sqrt{-\omega_1^{-1}}$.

Эти значения были получены на основе исследований спектра пучка операторов теории упругости.

Двухмерные периодические каналы

Приведем значения константы Ладыженской для двухмерных каналов прямоугольной формы в плоском и цилиндрическом случаях, когда по одному из направлений имеется периодическая зависимость функций u и p . Следует отметить, что для течений вязкой несжимаемой жидкости в каналах граничные условия в общем случае носят более сложный характер по сравнению с однородными условиями первого рода (13.3). Поэтому представляет интерес комбинированный вариант граничных условий: периодические по одному из направлений и первого рода — по другому. Такой выбор продиктован следующими обстоятельствами: во-первых, практической направленностью; во-вторых, традиционностью формы возмущающих течений; в-третьих, возможностью теоретического анализа, и, наконец, качественной близостью получаемого решения с решениями, соответствующими широкому набору других вариантов граничных условий.

В случае декартовых координат, т. е. области прямоугольной формы

$$\Omega = \{(x, y) | 0 < x < l_1, 0 < y < l_2\},$$

и периодичности по x имеем

$$\mathcal{L} = \sqrt{\frac{1}{2} \left(1 - \frac{z}{\sinh z}\right)}, \quad z = \frac{\pi l_2}{l_1}. \quad (13.12)$$

Эта формула хорошо известна и может быть получена различными способами [107, 157, 167].

Для двухмерного аксиально-симметричного канала цилиндрической формы (т. е. для круглой трубы) с периодическими условиями по z

$$\Omega = \{(z, r) | 0 < z < L, 0 < r < R\}$$

значение константы Ладыженской определяется формулой [86, 96]

$$\mathcal{L} = \sqrt{\frac{1}{y I_0(y) I_1(y)} \int_0^y \xi I_1^2(\xi) d\xi}, \quad y = \frac{\pi R}{L}, \quad (13.13)$$

где $I_0(y), I_1(y)$ — модифицированные функции Бесселя [104].

Оценки снизу для константы Ладыженской

Как правило, в априорных и апостериорных оценках решений (например, для уравнений Стокса см. [74, 182] и цитированные там работы) константа Ладыженской фигурирует в знаменателе. Поэтому первостепенное значение имеют ее оценки снизу для различных областей.

Имеется достаточно общая оценка. Если область $\Omega \subset \mathbb{R}^s$ имеет диаметр R и является звездной относительно шара Q радиуса R_Q , то справедливо неравенство [19]

$$\mathcal{L}(\Omega) \geq C_s \left(\frac{R_Q}{R} \right)^{s+1}$$

с постоянной C_s , не зависящей от Ω .

Для небольших значений размерности s этот результат может быть качественно улучшен. Наиболее развитый аппарат для этого построен при исследовании неравенств Корна и Фридрихса. Приведем здесь некоторые полезные оценки сверху для постоянных из этих неравенств. Напомним, что из (13.10) следует равенство $\mathcal{L} = \sqrt{2/\mathcal{K}}$.

Пусть ограниченная область $\Omega \subset \mathbb{R}^s$ с границей $\partial\Omega$ имеет диаметр R , звездна относительно шара Q радиуса R_Q и γ — наименьшее расстояние между Q и $\partial\Omega$. Тогда постоянная Корна \mathcal{K} в неравенстве (13.9) в зависимости от размерности s имеет следующий асимптотический вид:

$$\begin{aligned} D(\mathbf{u}) &\leq C_2 \left(\frac{R}{R_Q} \right)^2 \left(\ln \frac{3R}{R_Q} + \frac{R_Q^2}{\gamma^2} \right) E(\mathbf{u}), \quad s = 2, \\ D(\mathbf{u}) &\leq C_3 \left(\frac{R}{R_Q} \right)^3 \left(1 + \frac{R_Q^2}{\gamma^2} \right) E(\mathbf{u}), \quad s = 3, \end{aligned}$$

где постоянные C_s , $s = 2, 3$ не зависят от области Ω . Зависимости такого рода для постоянной в неравенстве Корна изучались в работах [52, 176].

Пусть Ω_1, Ω_2 — ограниченные области в \mathbb{R}^s , имеющие непустое пересечение, и для каждой из них справедливо неравенство Корна второго рода с постоянными \mathcal{K}_1 и \mathcal{K}_2 соответственно. Тогда неравенство Корна справедливо для области $\Omega = \Omega_1 \cup \Omega_2$ и имеет место оценка сверху

$$\mathcal{K} \leq \min\{K_1 + |\Omega_1|\mathcal{M}, K_2 + |\Omega_2|\mathcal{M}\},$$

где

$$\mathcal{M} = \frac{(\sqrt{\mathcal{K}_1} + \sqrt{\mathcal{K}_2})^2}{|\Omega_1 \cap \Omega_2|},$$

а обозначения $|\Omega_1|, |\Omega_2|, |\Omega_1 \cap \Omega_2|$ введены для Лебеговых мер соответствующих областей. Приведенное утверждение есть усиление результата работы [147], полученное в [167].

В случае правильного n -угольника ($s = 2$) на основе неравенства Фридрикса в [164] была получена следующая оценка

$$\mathcal{F} \leq \frac{1 + \sin(\pi/n)}{1 - \sin(\pi/n)},$$

откуда в силу (13.10), имеем неравенство

$$\mathcal{L}^2 \geq \frac{1 - \sin(\pi/n)}{2}.$$

13.4. АНИЗОТРОПИЯ ОБЛАСТИ

Под анизотропией области будем понимать существенную разницу в линейных масштабах по различным направлениям. Самым простым примером анизотропной (вытянутой) области является прямоугольник

$$\Omega = \{(y_1, y_2) | 0 < y_i < l_i, \quad i = 1, 2\},$$

в котором параметр растяжения обозначим через

$$\ell = \max\left\{\frac{l_1}{l_2}, \frac{l_2}{l_1}\right\}.$$

В этом случае имеет место асимптотическое поведение константы Ладыженской

$$\mathcal{L} = O(\ell^{-1}) \quad \text{при} \quad \ell \rightarrow \infty.$$

Этот факт сначала был установлен экспериментально [9, 107], затем обоснован теоретически [72, 138] с последовательным улучшением постоянной в оценке снизу. В настоящий момент времени наиболее

точная оценка снизу получена в [188] и двусторонние ограничения для константы Ладыженской имеют вид

$$\frac{\sin(\pi/8)}{\ell} \leq \mathcal{L} \leq \frac{\pi}{2\sqrt{3}\ell}. \quad (13.14)$$

Оценка сверху, видимо, является предельно точной, чему имеется экспериментальное подтверждение [167].

Вытянутые области часто встречаются в приложениях, поэтому установление асимптотики в простейшем случае повлекли за собой попытки обобщить этот результат в пространствах большей размерности. Приведем некоторые результаты из работ [149, 150].

Пусть Ω — некоторая фиксированная область из \mathbb{R}^s ($s \geq 2$) и $\mathcal{L}(\Omega)$ — соответствующее ей значение константы Ладыженской. С помощью вектора $l \in \mathbb{R}^s$ с компонентами

$$1 = l_1 \leq l_i \leq l_s, \quad 1 \leq i \leq s$$

определим вытянутую (stretched) область

$$\Omega_l = \left\{ y \in \mathbb{R}^s \mid \left(\frac{y_1}{l_1}, \dots, \frac{y_s}{l_s} \right)^T \in \Omega \right\},$$

характеризуемую параметром $\ell = l_s$. Определим также диаметр Ω в направлении e_i как

$$d_i = \sup\{h \mid x, x + he_i \in \Omega\}, \quad i = 1, \dots, s.$$

Тогда для константы Ладыженской области Ω_l справедливы оценки

$$\frac{\mathcal{L}(\Omega)}{\ell} \leq \mathcal{L}(\Omega_l) \leq \frac{C}{\ell},$$

где

$$C^2 = \frac{d_1 \dots d_s d_1^2}{d^{s+2}},$$

и Q_d — наибольший s -мерный куб со стороной d , содержащийся в Ω . Отметим, что границы в приведенных оценках не зависят от l_2, \dots, l_{s-1} .

Другая оценка связана с трехмерными каналами. Пусть

$$\Omega_l = \omega \times (0, \ell), \quad \omega \subset \mathbb{R}^2.$$

Обозначим через C_ω наилучшее значение постоянной в двухмерном неравенстве

$$\int_\omega |f(y)| dy \leq C_\omega \left(\int_\omega |Df(y)|^2 dy \right)^{1/2}, \quad f|_{\partial\omega} = 0.$$

Здесь $D = (D_1, D_2)^T$, $D_i = \partial/\partial y_i$, $i = 1, 2$. Тогда для константы Ладыженской области Ω_l справедлива оценка сверху

$$\mathcal{L}(\Omega_l) \leq \frac{C_\omega}{\ell} \sqrt{\frac{12}{|\omega|}},$$

где $|\omega|$ — мера Лебега области ω .

В частности, если ω — круг единичного радиуса, то $C_\omega = \sqrt{\pi/8}$, и оценка приобретает вид

$$\mathcal{L}(\Omega_l) \leq \frac{1}{\ell} \sqrt{\frac{3}{2}}.$$

13.5. УГЛОВЫЕ ТОЧКИ НА ГРАНИЦЕ

К ухудшению константы Ладыженской (с вычислительной точки зрения) может приводить не только анизотропия области, но и наличие угловых точек на границе области. Приведем оценку [87, 96] для плоских областей типа криволинейных многоугольников, возможно, содержащих разрезы, но не имеющих самопересекающихся (т. е. с точками возврата) границ.

Определим, что область Ω принадлежит классу $\Theta_2(\mathbb{R}^2)$ [148] тогда и только тогда, когда выполнены следующие условия:

- 1) Ω — ограничена и односвязна;
- 2) граница области $\partial\Omega$ состоит из конечного числа гладких жордановых дуг $\Gamma_1, \dots, \Gamma_N$, концы которых A_j ($A_{N+1} = A_1$) называются вершинами Ω ;
- 3) в окрестности A_j область Ω локально диффеоморфна некоторой окрестности нуля плоского сектора Γ_{A_j} с внутренним углом $0 < \varphi_j \leq 2\pi$.

В данном случае принципиально важным является сингулярное поведение собственных функций оператора S_0 в окрестности углов A_j .

Рассмотрим задачу (13.5) на собственные значения: $S_0 p = \lambda p$ и зафиксируем ее некоторое решение: пусть $0 < \lambda < 1/2$ — собственное значение, а p — отвечающая ему собственная функция. Изучению подлежит показатель особенности μ у собственной функции вида

$$p = r^{\mu-1} q(\varphi)$$

в угловой точке A_j в зависимости от собственного значения λ и величины угла φ_j . Согласно работам [50, 51] для этого требуется рассмотреть задачу (13.5) как уравнение $(S_0 - \lambda E)p = 0$ в угловой

точке с фиксированным λ . Выведем уравнение

$$(1 - 2\lambda)^2 \sin^2 \mu \varphi_j - \mu^2 \sin^2 \varphi_j = 0, \quad (13.15)$$

связывающее интересующие параметры.

Запишем (13.5) в полярной системе координат [63]:

$$\begin{cases} -\nabla^2 v_r + \frac{v_r}{r^2} + \frac{2}{r^2} \frac{\partial v_\varphi}{\partial \varphi} + \frac{\partial p}{\partial r} = 0, \\ -\nabla^2 v_\varphi + \frac{v_\varphi}{r^2} - \frac{2}{r^2} \frac{\partial v_r}{\partial \varphi} + \frac{1}{r} \frac{\partial p}{\partial \varphi} = 0, \\ \frac{1}{r} \left(\frac{\partial}{\partial r} (r v_r) + \frac{\partial v_\varphi}{\partial \varphi} \right) = \lambda p. \end{cases} \quad (13.16)$$

Здесь

$$\nabla^2 = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2}{\partial \varphi^2} = \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \varphi^2},$$

$$\mathbf{u} = (v_r, v_\varphi)^T.$$

Будем искать главную часть локального решения (v_r, v_φ, p) в виде

$$v_r = r^\mu u_r(\varphi), \quad v_\varphi = r^\mu u_\varphi(\varphi), \quad p = r^{\mu-1} q(\varphi). \quad (13.17)$$

Отметим, что необходимым и достаточным условием принадлежности вектора \mathbf{u} к пространству U и соответственно собственной функции p к P является выполнение неравенства

$$\operatorname{Re}[2(\mu - 1) + 1] > -1, \quad \text{т. е.} \quad \operatorname{Re} \mu > 0.$$

После подстановки в (13.16) вида (13.17) получим

$$\begin{cases} u_r'' + (\mu^2 - 1)u_r - 2u_\varphi' - (\mu^2 - 1)q = 0, \\ u_\varphi'' + (\mu^2 - 1)u_\varphi + 2u_r' - q' = 0, \\ (\mu + 1)u_r + u_\varphi' = \lambda q. \end{cases} \quad (13.18)$$

Здесь штрихом обозначено дифференцирование по φ .

Напомним необходимые для (13.18) граничные условия:

$$u_r(0) = u_r(\varphi_j) = 0, \quad u_\varphi(0) = u_\varphi(\varphi_j) = 0. \quad (13.19)$$

Умножим первое уравнение (13.18) на $(\mu + 1)$, продифференцируем второе и результаты сложим. Использование третьего уравнения (13.18) для исключения в полученном выражении u_r и u_φ даст

$$(\lambda - 1) [q'' + (\mu - 1)^2 q] = 0.$$

Поскольку нас интересуют только собственные функции, отвечающие $\lambda < 1/2$, откуда получаем общее решение $q(\varphi)$, зависящее от постоянных C_1 и C_2 как от параметров:

$$q(\varphi) = C_1 \cos(\mu - 1)\varphi + C_2 \sin(\mu - 1)\varphi. \quad (13.20)$$

Это дает возможность найти явную зависимость от C_1 и C_2 функции u_r . Подставим в первое уравнение (13.18) формулу (13.20) и учтем граничные условия для u_r из (13.19). Будем иметь

$$\begin{aligned} u_r(\varphi) = & \left(- \left[C_2 \frac{\sin(\mu - 1)\varphi_j}{\sin(\mu + 1)\varphi_j} - 2C_1 \frac{\sin \mu \varphi_j \sin \varphi_j}{\sin(\mu + 1)\varphi_j} \right] \times \right. \\ & \times \sin(\mu + 1)\varphi - C_1 \cos(\mu + 1)\varphi + C_1 \cos(\mu - 1)\varphi + \\ & \left. + C_2 \sin(\mu - 1)\varphi \right) \frac{\mu - 1 + 2\lambda}{4\mu}. \end{aligned} \quad (13.21)$$

Аналогичная процедура для $u_\varphi(\varphi)$ со вторым уравнением (13.18) дает

$$\begin{aligned} u_\varphi(\varphi) = & \left(\left[C_1 \frac{\sin(\mu - 1)\varphi_j}{\sin(\mu + 1)\varphi_j} + 2C_2 \frac{\sin \mu \varphi_j \sin \varphi_j}{\sin(\mu + 1)\varphi_j} \right] \times \right. \\ & \times \sin(\mu + 1)\varphi - \\ & - C_2 \cos(\mu + 1)\varphi - C_1 \sin(\mu - 1)\varphi + \\ & \left. + C_2 \cos(\mu - 1)\varphi \right) \frac{\mu - 1 + 2\lambda}{4\mu}. \end{aligned} \quad (13.22)$$

Получив u_r и u_φ , воспользуемся третьим уравнением (13.18) для нахождения зависимости $\mu = \mu(\lambda)$ при фиксированном φ_j . С этой целью, после подстановки в него явных формул для u_r и u_φ , приравняем к нулю коэффициенты при линейно независимых функциях $\cos(\mu + 1)\varphi$ и $\sin(\mu + 1)\varphi$ соответственно. В результате придем к системе уравнений

$$\begin{aligned} C_1 [(1 - 2\lambda) \sin \mu \varphi_j \cos \varphi_j - \mu \cos \mu \varphi_j \sin \varphi_j] + \\ + C_2 [(1 - 2\lambda) \sin \mu \varphi_j \sin \varphi_j + \mu \sin \mu \varphi_j \sin \varphi_j] = 0, \\ C_2 [(1 - 2\lambda) \sin \mu \varphi_j \cos \varphi_j + \mu \cos \mu \varphi_j \sin \varphi_j] + \\ + C_1 [\mu \sin \mu \varphi_j \sin \varphi_j - (1 - 2\lambda) \sin \mu \varphi_j \sin \varphi_j] = 0. \end{aligned}$$

Теперь для нахождения ее нетривиального решения C_1, C_2 приравняем к нулю определитель. Это и порождает искомое уравнение (13.15), хорошей иллюстрацией к которому являются расчеты из [167].

Рассмотрим некоторые следствия. Пусть в (13.15) собственное значение λ совпадает с минимальным, т. е. $\lambda = \lambda_{\min}(S_0) = \mathcal{L}^2$. Тогда

справедлива оценка

$$\lambda_{\min}(S_0) \leq \frac{1}{2} \left(1 - \frac{\sin \varphi}{\varphi} \right), \quad (13.23)$$

где $\varphi = \min\{\varphi_j, \pi\}$.

Действительно, учитывая необходимое и достаточное условие принадлежности собственной функции p к пространству P , т.е. неравенство $\operatorname{Re} \mu > 0$, из уравнения (13.15) имеем оценку

$$\sin^2 \varphi_j = (1 - 2\lambda)^2 \frac{\sin^2 \mu \varphi_j}{\mu^2} \leq (1 - 2\lambda)^2 \varphi_j^2,$$

которая при использовании обозначения $\varphi = \min_j\{\varphi_j, \pi\}$ приводит к неравенству

$$|1 - 2\lambda| \geq \frac{\sin \varphi}{\varphi}.$$

Из полученного выражения следуют две оценки:

$$\lambda \leq \frac{1}{2} \left(1 - \frac{\sin \varphi}{\varphi} \right) \quad \text{или} \quad \lambda \geq \frac{1}{2} \left(1 + \frac{\sin \varphi}{\varphi} \right).$$

Последняя из них малосодержательна, так как даже для областей с гладкими границами ($\varphi = \pi$) хорошо известно [69], что $\lambda_{\min}(S_0) \leq 1/2$. Следовательно, справедливо неравенство (13.23).

Отметим, что оценка (13.23) может быть получена и другими способами. В частности, она следует из анализа существенного спектра Коссера в случае многоугольных областей [145, 155]; уравнение (13.15) встречалось также в [143].

Если граница области Ω имеет хотя бы один малый угол φ , то это приводит к асимптотической оценке для константы Ладыженской

$$\mathcal{L} \leq \frac{\varphi}{2\sqrt{3}} \quad \text{при} \quad \varphi \rightarrow 0.$$

Практическим приложением этого результата является получение количественной информации о том, насколько трудоемко будет нахождение численного решения уравнений Стокса в таких областях.

Возможны также обобщения полученной оценки на случай трехмерных областей, когда φ_j являются плоскими углами трехгранных углов и т.п. [51].

В случае правильного n -угольника на основе неравенства Фридрикса в [164] была получена оценка снизу для константы Ладыженской. Комбинируя ее с (13.23), имеем

$$\frac{1 - \sin(\pi/n)}{2} \leq \mathcal{L}^2 \leq \frac{1}{2} \left(1 - \frac{n \sin(2\pi/n)}{\pi(n-2)} \right). \quad (13.24)$$

Проанализируем уравнение (13.15) и оценки (13.24) применительно к случаю прямоугольной области. Для $\varphi = \pi/2$ и $\lambda = \mathcal{L}^2$ имеем

$$0,146446 \dots = \sin^2 \frac{\pi}{8} \leq \lambda \leq \frac{\pi - 2}{2\pi} = 0,181690 \dots \quad (13.25)$$

и

$$(1 - 2\lambda)^2 \sin^2 \frac{\mu\pi}{2} - \mu^2 = 0. \quad (13.26)$$

Несложно убедиться в том, что вещественное решение (13.26), отличное от нуля, при условии (13.25) монотонно убывает и удовлетворяет ограничениям

$$0 \leq \mu(\lambda) \leq \frac{1}{2}.$$

Например, при $\lambda = 0,16$ решение (13.26) имеет вид $\mu(\lambda) = 0,3977 \dots$. При $(\pi - 2)/2\pi < \lambda \leq 1/2$ нетривиальные вещественные решения отсутствуют, а при $\lambda \leq \sin^2(\pi/8)$ — монотонно возрастают от $1/2$ и стремятся к единице. Это означает, что собственная функция, отвечающая $\lambda_{\min}(S_0) = \mathcal{L}^2$, имеет наихудшую гладкость для области квадратной формы (т. е. наименьший показатель $(\mu - 1) < 0$); затем, по мере увеличения одной из сторон, константа Ладыженской монотонно убывает (13.14), а соответствующая собственная функция постепенно выглаживается.

В завершение раздела отметим, что имеется естественное усложнение задачи Стокса, связанное с различными постоянными значениями коэффициента вязкости в подобластях (так называемая, задача с интерфейсом). В этом случае можно определить аналог inf-sup-неравенства [179], и исследование константы Ладыженской и LBB-констант становится весьма нетривиальной проблемой [99, 100].

13.6. РАЗНОЕ

Существует значительное число работ, использующих специальные свойства задач для уравнений в частных производных, которые служат прообразами матричных задач с седловыми операторами. Используемые в них идеи весьма оригинальны и разнообразны, поэтому стоит иметь их в виду.

13.6.1. Обобщенная задача Стокса

Рассмотрим в области $\Omega \subset \mathbb{R}^s$ ($s = 2, 3$) нестационарную первую краевую задачу для уравнений Стокса в переменных скорость–

давление [56]:

$$\begin{aligned}\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + \operatorname{grad} p &= \mathbf{f} && \text{в } \Omega, \\ \operatorname{div} \mathbf{u} &= 0 && \text{в } \Omega, \\ \mathbf{u} &= 0 && \text{на } \partial\Omega, \\ \mathbf{u}(x, 0) &= \mathbf{u}_0(x) && \text{в } \Omega.\end{aligned}\tag{13.27}$$

Уравнения (13.27) описывают движение вязкой ($\nu > 0$ — постоянный коэффициент кинематической вязкости), несжимаемой, однородной и изотропной жидкости при малых скоростях под воздействием внешних сил. Неизвестными здесь являются вектор-функция $\mathbf{u} = (u_1(x, t), \dots, u_s(x, t))$ (скорость жидкости) и скалярная функция $p = p(x, t)$ (давление), определенная с точностью до константы. Нулевые граничные условия означают, как правило, что движение происходит в некотором неподвижном объеме с твердыми стенками, целиком заполненным жидкостью.

Произвольная неявная по времени дискретизация после соответствующей перенормировки приводит к обобщенной задаче Стокса с параметром $\alpha \geq 0$:

$$\begin{aligned}-\Delta \mathbf{u} + \alpha \mathbf{u} + \operatorname{grad} p &= \mathbf{f}, \\ \operatorname{div} \mathbf{u} &= 0, \\ \mathbf{u}|_{\partial\Omega} &= 0.\end{aligned}\tag{13.28}$$

Случай $\alpha = 0$ соответствует классической задаче Стокса, нестационарные же уравнения порождают большой параметр $\alpha \sim (\nu \delta t)^{-1}$, где δt — шаг интегрирования по времени.

Построение многосеточных предобусловливателей для задачи Стокса проанализировано в работах [102, 121, 152, 196]. Многоуровневые алгоритмы для той же задачи изучались в [185, 197].

Сочетание модельных седловых операторов с итерациями в подпространстве и методом фиктивных областей предложено в работах [14, 112].

Граничные значения для гармонической составляющей скорости было предложено находить в работе [191].

Сходимость итерационного метода в подпространстве гармонических функций для давления в дифференциальном случае обоснована в [172].

Итерирование граничных условий для скорости, основанное на применении решений вспомогательных задач с ненулевыми следами на частях границы, предложено и обосновано в [136].

Строить методы решения задачи Стокса аналогично *обратным* задачам предлагалось в [105, 106].

13.6.2. Уравнения Ламе в теории упругости и слабосжимаемая жидкость

Рассмотрим в области $\Omega \subset \mathbb{R}^s$ ($s = 2, 3$) стационарную первую краевую задачу для линейных уравнений теории упругости с постоянными коэффициентами Ламе μ, λ , записанную относительно вектора перемещений \mathbf{u} [77]:

$$\begin{aligned} \mu \Delta \mathbf{u} + (\lambda + \mu) \operatorname{grad} \operatorname{div} \mathbf{u} &= \mathbf{f} & \text{в } \Omega, \\ \mathbf{u} &= 0 & \text{на } \partial\Omega. \end{aligned} \quad (13.29)$$

Неизвестными здесь являются компоненты вектор-функции $\mathbf{u} = (u_1(x), \dots, u_s(x))$. Нулевые граничные условия означают, как правило, что форма тела остается неизменной. Введем новую переменную

$$p = -\frac{\lambda + \mu}{\mu} \operatorname{div} \mathbf{u} \quad (13.30)$$

и обозначение параметра

$$\varepsilon = \frac{\mu}{\lambda + \mu}.$$

Тогда задачу (13.29) после соответствующей перенормировки можно записать как седловую

$$\begin{aligned} -\Delta \mathbf{u} + \operatorname{grad} p &= \mathbf{f}, \\ -\operatorname{div} \mathbf{u} - \varepsilon p &= 0, \\ \mathbf{u}|_{\partial\Omega} &= 0. \end{aligned} \quad (13.31)$$

К аналогичному виду приводятся линеаризованные уравнения слабосжимаемой жидкости [79]:

$$\begin{aligned} -\nu \Delta \mathbf{u}_\varepsilon - \varepsilon^{-1} \operatorname{grad} \operatorname{div} \mathbf{u}_\varepsilon &= \mathbf{f} & \text{в } \Omega, \\ \mathbf{u} &= 0 & \text{на } \partial\Omega. \end{aligned}$$

Здесь $\varepsilon > 0$ – малый параметр, характеризующий сжимаемость.

13.6.3. Смешанный подход для эллиптических уравнений

Рассмотрим в области $\Omega \subset \mathbb{R}^s$ ($s = 2, 3$) первую краевую задачу для эллиптического уравнения второго порядка

$$\begin{aligned} -\operatorname{div} K \operatorname{grad} p &= f & \text{в } \Omega, \\ p &= 0 & \text{на } \partial\Omega, \end{aligned} \quad (13.32)$$

где $K = \{k_{ij}\}_{i,j=1}^{s'}$ — симметричная положительно определенная матрица, элементы которой являются ограниченными функциями пространственных переменных (значения размерностей s и s' не обязательно совпадают). Такая постановка является традиционной модельной задачей в механике сплошных сред или гидродинамике пористых областей [63, 77].

Введем новую переменную \mathbf{u} следующим образом:

$$\mathbf{u} = K \operatorname{grad} p. \quad (13.33)$$

Тогда задачу (13.32) можно записать в виде

$$\begin{aligned} K^{-1} \mathbf{u} - \operatorname{grad} p &= 0 & \text{в } \Omega, \\ \operatorname{div} \mathbf{u} &= -f & \text{в } \Omega, \\ p &= 0 & \text{на } \partial\Omega. \end{aligned} \quad (13.34)$$

Типичными являются случаи, когда K определяет тензор упругости/проницаемости, \mathbf{u} представляет вектор напряжений/скорости, а p — функцию смещения/давления.

Пусть, с точностью до краевых условий, справедливо

$$P = L_2(\Omega), \quad U = \{v \mid v \in (L_2(\Omega))^s, \operatorname{div} v \in L_2(\Omega)\}.$$

Детали аккуратного учета краевых условий в определении пространств описываются в [127].

Определим обобщенное решение задачи (13.34) следующим образом: найти $\mathbf{u} \in U$ и $p \in P$, удовлетворяющие системе интегральных тождеств

$$\begin{aligned} (K^{-1} \mathbf{u}, v) + (p, \operatorname{div} v) &= 0 & \forall v \in U, \\ (\operatorname{div} \mathbf{u}, q) &= -(f, q) & \forall q \in P. \end{aligned} \quad (13.35)$$

Несколько другой способ сведения (13.35) к равносильной системе уравнений, а также алгоритмы для ее решения были рассмотрены в [17, 47, 170]. Сначала делается перенормировка входных данных. Пусть

$$k = \frac{1}{2} \min_{i,j} k_{ij}, \quad \bar{f} = \frac{f}{k}.$$

Тогда матрицу коэффициентов в (13.35) можно записать в виде

$$K = I + W, \quad W \geq 1.$$

Исходное уравнение при этом будет иметь форму

$$-\Delta p - \operatorname{div} W \operatorname{grad} p = \bar{f}.$$

Вводя новую переменную $\mathbf{u} = W \operatorname{grad} p$, получим эквивалентную формулировку

$$\begin{aligned} W^{-1}\mathbf{u} - \operatorname{grad} p &= 0 && \text{в } \Omega, \\ \operatorname{div} \mathbf{u} + \Delta p &= -\bar{f} && \text{в } \Omega, \\ p &= 0 && \text{на } \partial\Omega. \end{aligned}$$

При численном решении после перехода к конечномерным подпространствам относительно коэффициентов разложения неизвестных функций получим систему вида (1.2). Сравнение эффективности различных алгоритмов для такого подхода проводилось в [18].

13.6.4. Другие применения

Укажем еще некоторые приложения задач с седловыми операторами, их число непрерывно растет, и охватить все не представляется возможным.

При решении эллиптических задач в областях сложной формы используются методы композиции [60] и декомпозиции [132] в сочетании с нестыкующимися сетками. Сведение к подзадам на интерфейсах и в подобластях порождают системы уравнений с седловыми точками [28].

Алгоритмы решения алгебраических систем с сингулярно возмущенными матрицами, основанные на сведении к седловым задачам, анализировались в [173].

Разрешимость уравнений с седловой точкой для нахождения оптимальных коэффициентов кубатурных формул исследовалась в [78].

Вариационные постановки задач в усиленных пространствах Соболева и регуляризация систем типа Стокса также порождают системы указанного типа. Систематическое изложение этих вопросов содержится в [42].

ЧИСЛЕННЫЙ АНАЛИЗ LBB-УСЛОВИЯ

Цель настоящей главы — обратить внимание на различия между значениями константы Ладыженской \mathcal{L} и ее дискретными аналогами \mathcal{L}_h , возникающие вследствие анизотропии области и отсутствия необходимой гладкости у ее границы. Главным побудительным мотивом для этого является отсутствие подходящей теории сходимости решений сеточных спектральных задач для седловых операторов, что может порождать трудности при интерпретации результатов численного моделирования в гидродинамике.

Основными объектами изучения являются: младшее собственное значение $\lambda_{\min}(S_0^h)$ и соответствующая собственная функция задачи $S_0^h p_h = \lambda p_h$:

$$\begin{aligned} -\Delta \mathbf{u}_h + \operatorname{grad} p_h &= 0, \\ \operatorname{div} \mathbf{u}_h &= \lambda p_h. \end{aligned} \quad (14.1)$$

При этом в качестве метода дискретизации непрерывной задачи используется только конформный метод конечных элементов ($U_h \subset U$, $P_h \subset P$), порождающий LBB-устойчивую схему. Напомним, что величины $\lambda_{\min}(S_0^h)$ и \mathcal{L}_h связаны между собой соотношением

$$\lambda_{\min}(S_0^h) = \mathcal{L}_h^2,$$

а LBB-константа \mathcal{L}_0 является равномерной по h оценкой снизу для \mathcal{L}_h . Чтобы не отвлекаться на ненужные пересчеты, в таблицах будет приводиться именно то, что вычисляется, т. е. значения $\gamma_h = \lambda_{\min}(S_0^h)$.

14.1. ЗАДАЧА С ГЛАДКИМ РЕШЕНИЕМ

Прежде чем конструировать конкретные конечномерные пространства U_h, P_h и решать соответствующую спектральную матричную задачу, попытаемся определить, что именно является искомым результатом.

Пусть младшая собственная функция p_{\min} непрерывной задачи (13.5) обладает необходимой гладкостью, например,

$$p_{\min} \in P \cap W_2^1(\Omega)$$

и выполнено следующее условие аппроксимации:

$$\begin{aligned} & \|u - u_h\|_U + \|p - p_h\|_P \leq \\ & \leq c \left(\inf_{v_h \in U_h} \|u - v_h\|_U + \inf_{q_h \in P_h} \|p - q_h\|_P \right) = O(h) \end{aligned} \quad (14.2)$$

с постоянной c , не зависящей от h . Тогда, учитывая некомпактность непрерывного оператора S_0 , «наилучшей» [111] представляется оценка скорости сходимости вида

$$|\mathcal{L}^2 - \gamma_h| = O(h^2). \quad (14.3)$$

Поэтому в аккуратных численных экспериментах должно наблюдаться монотонное поведение значений γ_h и их сходимость к \mathcal{L}^2 со вторым (как следствие аппроксимации и устойчивости) порядком относительно параметра дискретизации области.

Выберем в качестве задачи с гладким решением $S_0 p = \lambda p$ задачу (13.5) в области Ω , имеющей форму концентрического кольца,

$$\Omega = \{(r, \varphi) \mid R_1 < r < R_2, 0 \leq \varphi \leq 2\pi\}.$$

Тогда из первой части приложения имеем (см. раздел 13.3), что минимальное собственное значение

$$\lambda_{\min}(S_0) = \frac{1}{2} \left(1 - \sqrt{\frac{x^2 - 1}{x^2 + 1} \frac{1}{\ln x}} \right), \quad x = \frac{R_2}{R_1} > 1$$

двукратно, а соответствующие ему (при $m = \pm 1$) собственные функции обладают требуемой гладкостью и ортогональны в метрике пространства P как друг другу, так и остальным собственным функциям. Отметим также, что нижняя граница существенного спектра оператора S_0 равна $1/2$ и является точкой сгущения конечнократных собственных значений $\lambda_m(S_0)$ при $|m| \rightarrow \infty$, поэтому распределение собственных значений не мешает вычислению $\lambda_{\min}(S_0) < 1/2$.

Понизим размерность задачи (13.5), воспользовавшись естественной периодичностью решения относительно φ . Кроме того, чтобы избежать вычислений с комплексными числами и избавиться от кратности минимальных собственных значений, рассмотрим систему уравнений, зависящую от параметра m ($m = 1, 2, \dots$)

$$\begin{aligned} & -\Delta_r v_m + \frac{1}{r^2} v_m - \frac{2m}{r^2} w_m + \frac{\partial p_m}{\partial r} = 0, \\ & -\Delta_r w_m + \frac{1}{r^2} w_m - \frac{2m}{r^2} v_m + \frac{m}{r} p_m = 0, \\ & \frac{1}{r} \left(\frac{\partial}{\partial r} (r v_m) - m w_m \right) = \lambda p_m. \end{aligned} \quad (14.4)$$

Здесь радиальный оператор Δ_r определен как

$$\Delta_r = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \right) - \frac{m^2}{r^2}.$$

Система (14.4) следует из (13.5), если решение искать в виде

$$\begin{aligned} [v, p]^T &= \sum_{m=1}^{\infty} [v_m(r), p_m(r)]^T \sin m\varphi, \\ w &= \sum_{m=1}^{\infty} w_m(r) \cos m\varphi, \quad \mathbf{u} = [v, w]^T, \end{aligned}$$

с учетом ортогональности тригонометрических множителей в метрике $L_2(0, 2\pi)$. Использование «альтернативного» разложения

$$[v, p]^T = \sum_{m=1}^{\infty} [v_m(r), p_m(r)]^T \cos m\varphi, \quad w = \sum_{m=1}^{\infty} w_m(r) \sin m\varphi,$$

приводит к замене в системе (14.4) значений параметра m на $-m$. Это преобразование сохраняет как радиальные части собственных функций, так и соответствующие собственные значения. Поэтому без ограничения общности можно в дальнейшем обсуждать численное решение только системы (14.4). Напомним, что из $\mathbf{u} \in U$ для (14.4) следуют граничные условия

$$v_m(R_2) = w_m(R_2) = v_m(R_1) = w_m(R_1) = 0. \quad (14.5)$$

Окончательно для численного решения имеем задачу (14.4) – (14.5), зависящую от параметра $|m| > 0$. Значение $m = 0$ не рассматривается, так как для $p \in P$, $\mathbf{u} \in U$ спектральная задача имеет только тривиальное решение $p \equiv 0$, $\mathbf{u} \equiv 0$.

Обратим внимание, что использование периодичности фактически ликвидирует возможную анизотропию исходной двухмерной области (при малых δ , когда $R_2/R_1 = 1 + \delta$) за счет перехода к одномерной расчетной области.

Введем на отрезке $[R_1, R_2]$ равномерную сетку для аппроксимации давления p с шагом h_p

$$r_j = R_1 + jh_p, \quad j = 0, 1, \dots, N_p, \quad h_p N_p = R_2 - R_1, \quad (14.6)$$

и определим набор кусочно-постоянных функций

$$\psi_j(r) = \begin{cases} 1 & \text{при } r \in [r_j, r_j + h_p], \\ 0 & \text{иначе} \end{cases}$$

при $j = 1, 2, \dots, N_p$. Целое N_p для удобства будем считать четным.

Для аппроксимации скорости \mathbf{u} потребуется более мелкая сетка ($h_u = h_p/2$)

$$r_i = R_1 + ih_u, \quad i = 0, 1, \dots, N_u, \quad h_u N_u = R_2 - R_1,$$

и набор кусочно-линейных функций

$$\theta_i(r) = \theta\left(\frac{r - R_1}{h_u} - i\right), \quad i = 1, 2, \dots, N_u - 1,$$

определенных через стандартную функцию -- «крышку»

$$\theta(t) = \begin{cases} 1 - |t| & \text{при } |t| \leq 1, \\ 0 & \text{при } |t| > 1. \end{cases}$$

Определим конечномерные пространства $P_h \subset P$, $U_h \subset U$ как линейные оболочки соответствующих наборов линейно независимых функций

$$P_h = \text{span}\{\psi_j(r)\}, \quad 1 \leq j \leq N_p, \quad U_h = \text{span}^2\{\theta_i(r)\}, \quad 1 \leq i \leq N_u - 1,$$

и будем искать приближенные решения (14.4) – (14.5) в виде

$$p_m^h = \sum_{j=1}^{N_p} p_j \psi_j(r), \quad [v_m^h, w_m^h]^T = \sum_{i=1}^{N_u-1} [v_i, w_i]^T \theta_i(r).$$

Это дает возможность определить матричную задачу

$$S_0^h \mathbf{y} \equiv B^T A^{-1} B \mathbf{y} = \lambda C \mathbf{y}, \quad \mathbf{x} = -A^{-1} B \mathbf{y}, \quad (14.7)$$

относительно коэффициентов разложения по базисам

$$\mathbf{y} = [p_1, p_2, \dots, p_{N_p}]^T, \quad \mathbf{x} = [v_1, w_1, v_2, w_2, \dots, v_{N_u-1}, w_{N_u-1}]^T.$$

Здесь C — матрица Грама, т. е. $c_{ij} = \int_{R_1}^{R_2} r \psi_i(r) \psi_j(r) dr$.

Элементы матриц A, B, C найдем аналитическим интегрированием, оператор S_0^h вычислим в явном виде (по столбцам) из равенства $S_0^h = S_0^h I$, где I — единичная квадратная матрица размерности N_p . Вспомогательные системы вида $A \mathbf{x} = \mathbf{b}$ решим с помощью матричной 2×2 прогонки [76]. Сохранение симметрии оператора S_0^h будем контролировать величиной

$$\left\| \frac{S_0^h - (S_0^h)^T}{2} \right\|_E, \quad \|A\|_E^2 = \sum_{i,j} |a_{i,j}|^2;$$

при расчетах с двойной точностью она не превосходила 10^{-13} для $N_p \leq 1024$. Полную задачу для обобщенной проблемы собственных значений

$$C^{-1/2} S_0^h C^{-1/2} \mathbf{z} = \lambda \mathbf{z}, \quad \mathbf{z} = C^{1/2} \mathbf{y},$$

будем решать подпрограммами из пакета EISPACK [2], реализующими QL — алгоритм.

Приведем результаты расчетов для иллюстрации качества приближенного метода. Зафиксируем параметры кольца: $R_1 = 1$, $R_2 = 3$, вычислим по аналитической формуле точное минимальное собственное значение

$$\mathcal{L}^2 = 0,0733293479446 \dots$$

и для нормированных собственных функций

$$\|p_{\min}\|_P = \|p_{\min}^h\|_P = 1$$

приведем в таблице 1 посчитанные величины:

Таблица 1

N_p	$\mathcal{L}^2 - \gamma_h$	$\ p_{\min} - p_{\min}^h\ _P$
4	0,84609 E-3	0,26074 E-1
8	0,20403 E-3	0,13606 E-1
16	0,50523 E-4	0,68805 E-2
32	0,12600 E-4	0,34501 E-2
64	0,31482 E-5	0,17263 E-2
128	0,78694 E-6	0,86331 E-3
256	0,19673 E-6	0,43168 E-3
512	0,49181 E-7	0,21584 E-3
1024	0,12295 E-7	0,10792 E-3

Легко заметить, что здесь имеет место экспериментальная (совпадающая с «наилучшей» при используемой аппроксимации) скорость сходимости

$$\mathcal{L}^2 - \gamma_h = O(h_p^2), \quad \|p_{\min} - p_{\min}^h\|_P = O(h_p),$$

сопровождающаяся монотонным возрастанием γ_h .

В таблице приведены величины, посчитанные при $m = 1$; для других значений параметра m асимптотические свойства решения сохраняются (при соответствующем увеличении постоянных множителей).

Отметим, что использование кусочно-линейных базисных функций для аппроксимации давления позволяют получить для гладких собственных функций более точные (по сравнению с кусочно-постоянными) приближения в метрике P ; асимптотика сходимости собственных значений при этом не меняется, так как определяется

не только от пространством P_h , но и пространством U_h . Соответствующие численные эксперименты, как и приведенные выше, опубликованы в [99, 100].

Рассмотренный здесь модельный позитивный пример позволяет определить ориентиры для теоретических исследований сходимости в случае, когда полностью отсутствуют неблагоприятные факторы.

14.2. ЗАДАЧА С НЕГЛАДКИМ РЕШЕНИЕМ

Простейшей областью, в которой решение задачи $S_0^h p_h = \lambda p_h$ имеет особенности в угловых точках, является область Ω прямоугольной формы

$$\Omega = \{(x, y) | 0 < x < l_1, 0 < y < l_2\}.$$

Причем наиболее сильно особенности выражены при $l_1 = l_2$, т. е. когда Ω представляет собой квадрат, как это следует из первой части приложения (см. раздел 13.5).

Для исследования младшего собственного значения $\lambda_{\min}(S_0^h)$ и соответствующей собственной функции задачи (14.1) методом конечных элементов на равномерных по каждому направлению сетках построены две схемы. Они используют кусочно-линейное или кусочно-билинейное поле скоростей по отношению к разбиению области Ω на треугольники или прямоугольники соответственно. В обоих случаях поле скоростей является непрерывным над $\bar{\Omega} = \Omega \cup \partial\Omega$, а давление — кусочно-постоянным над Ω . Каждая схема удовлетворяет LBB-условию [156, 158] и, если решение непрерывной задачи (13.5) обладает необходимой гладкостью, то имеет место условие аппроксимации (14.2).

14.2.1. Схема 1

Пусть $h_i N_i = l_i$, $i = 1, 2$. Будем считать целые числа N_i четными. Тогда множество прямых вида

$$\begin{aligned} x &= 2ih_1, & i &= 0, 1, \dots, N_1/2; \\ y &= 2jh_2, & j &= 0, 1, \dots, N_2/2, \end{aligned}$$

разбивает $\bar{\Omega}$ на элементарные прямоугольники. Разобьем каждый элементарный прямоугольник на два треугольника, проводя диагональ, параллельную прямой $y = h_2 x / h_1$. Таким образом мы получим регулярное («северо-восточное») разбиение T_h прямоугольной области $\bar{\Omega}$ на треугольники. Далее разобьем каждый треугольник из T_h на четыре треугольника средними линиями; это даст триангуляцию $T_{h/2}$.

Определим пространства

$$U_h = \{ \mathbf{u}_h \mid \mathbf{u}_h \in S_1(\Delta), \Delta \in T_{h/2}; \mathbf{u}_h \in C(\bar{\Omega}); \mathbf{u}_h = 0 \text{ на } \partial\Omega \},$$

$$P_h = \left\{ p_h \mid p_h \in S_0(\Delta), \Delta \in T_h; \int_{\Omega} p_h dx dy = 0 \right\}.$$

Здесь $C(\bar{\Omega})$ — пространство непрерывных на $\bar{\Omega}$ функций, а $S_r(\Delta)$ — пространство многочленов степени не выше r , определенных на множестве $\Delta \subset \mathbb{R}^2$, или векторное (размерности 2) пространство функций, каждая компонента которых принадлежит $S_r(\Delta)$.

В этом случае схема 1 будет определяться соотношениями

$$\begin{aligned} (\text{grad } \mathbf{u}_h, \text{grad } \mathbf{v}_h) - (p_h, \text{div } \mathbf{v}_h) &= 0 \quad \forall \mathbf{v}_h \in U_h, \\ (\text{div } \mathbf{u}_h, q_h) &= \lambda(p_h, q_h) \quad \forall q_h \in P_h. \end{aligned} \quad (14.8)$$

Здесь круглые скобки означают скалярное произведение в пространстве $L_2(\Omega)$, а выражение $(\text{grad } \mathbf{u}, \text{grad } \mathbf{v})$ для векторов $\mathbf{u} = (u_1, u_2)^T$, $\mathbf{v} = (v_1, v_2)^T$ определено как

$$(\text{grad } \mathbf{u}, \text{grad } \mathbf{v}) = \sum_{i=1}^2 \left(\frac{\partial u_i}{\partial x}, \frac{\partial v_i}{\partial x} \right) + \left(\frac{\partial u_i}{\partial y}, \frac{\partial v_i}{\partial y} \right).$$

Эта схема была предложена и обоснована в работе [39]; в настоящее время для нее применяют стандартное обозначение $P_{\text{iso}}P_2 - P_0$; выражения для матричных элементов схемы можно найти в [96, 107].

14.2.2. Схема 2

Разбиение $\bar{\Omega}$ на элементарные прямоугольники проведем как при построении схемы 1 и обозначим его через Q_h . Затем каждый полученный элемент дополнительно разобьем на четыре подобных, соединив середины противоположных сторон прямыми. Построенное таким образом разбиение области обозначим за $Q_{h/2}$.

Рассмотрим аппроксимацию поля скоростей кусочно-билинейными функциями по отношению к разбиению $Q_{h/2}$, непрерывными на $\bar{\Omega}$ и обращающимися в нуль на $\partial\Omega$, т. е.

$$U_h = \{ \mathbf{u}_h \mid \mathbf{u}_h \in \hat{S}_1(\square), \square \in Q_{h/2}; \mathbf{u}_h \in C(\bar{\Omega}); \mathbf{u}_h = 0 \text{ на } \partial\Omega \}.$$

Для аппроксимации давления будем использовать кусочно-постоянные функции, определенные на больших прямоугольниках разбиения Q_h и имеющие нулевые средние на Ω , т. е.

$$P_h = \left\{ p_h \mid p_h \in \hat{S}_0(\square), \square \in Q_h; \int_{\Omega} p_h dx dy = 0 \right\}.$$

В данном случае $\hat{S}_r(\square)$ имеет смысл пространства полиномов не выше r по каждой координате.

Теперь, с учетом определения пространств, схему 2 можно также получить из соотношений (14.8); выражения для матричных элементов схемы можно найти в [96, 107].

14.2.3. Вычислительные аспекты

Границы спектра оператора S_0^h будем определять с помощью метода итерирования подпространства [73], подробный анализ которого имеется в [80]. Выбор этого алгоритма обусловлен следующими причинами. Степенной метод [15] (простые векторные итерации [73]) в данном случае непригоден, так как вычисляемые собственные значения не обязательно однократны. С другой стороны, несмотря на то что существуют простые в использовании и надежные программы, реализующие алгоритм Ланцоша [81] (вообще говоря, более предпочтительный при такого рода расчетах), большая размерность задачи позволяет хранить в оперативной памяти одновременно небольшое количество итерируемых векторов. Поэтому не оказывается другого выхода, как отбрасывать предыдущие векторы последовательности Крылова и использовать итерирование подпространства.

При вычислениях использовалась программа EA12 из библиотеки численного анализа HARWELL [2], позволившая достичь довольно высокой точности. Процесс останавливался, когда нормированная на собственный вектор невязка (в сеточном аналоге пространства $L_2(\Omega)$) становилась менее чем 10^{-10} . В основе этого лежит следующее утверждение [43, 73]: если число μ и нормированный в евклидовой метрике $\|\cdot\|_2$ вектор x рассматриваются как приближенная собственная пара симметричной матрицы A и $r = Ax - \mu x$ — соответствующая им невязка, то найдется собственное значение λ матрицы A такое, что $|\lambda - \mu| \leq \|r\|_2$. Здесь важно, что в данном случае можно не разделять понятия невязки и ошибки, а значит, говорить о высокой достоверности полученных результатов.

Напомним, что максимальное собственное значение оператора S_0 равно единице, причем оно имеет бесконечную кратность. Проведенные расчеты показали, что максимальное собственное значение оператора S_0^h равно 1 с большой степенью точности для обеих схем, причем кратность единицы велика. Достоверно можно утверждать, что она заведомо больше 10 — дальнейшие эксперименты в этом направлении не проводились из-за ограниченности вычислительных ресурсов.

Остальные расчеты были связаны с нахождением минимального ненулевого собственного значения γ_h оператора S_0^h . Отметим, что при реальном счете нижняя граница спектра всегда равна нулю, так как для надежности результатов вычисления ведутся по всему пространству, а не только на подпространстве функций, ортогональных единице в сеточном аналоге метрики $L_2(\Omega)$. Конечно, это является одним из факторов, осложняющих счет, зато уменьшается вероятность пропустить подпространство, содержащее искомую младшую собственную функцию.

14.2.4. Расчеты для квадратной области

Зафиксируем параметры области $l_1 = l_2 = 1$ и для обеих схем приведем в таблице 2 зависимость младшего собственного значения γ_h оператора S_0^h от параметра дискретизации $h = h_1 = h_2$.

Таблица 2

h	Схема 1	Схема 2
1/8	0,225616	0,313859
1/16	0,212847	0,271451
1/32	0,205198	0,246103
1/64	0,200189	0,230040
1/128	0,196717	0,219313
1/256	0,194198	0,211805

Дополнительно на рис. 1 и 2 приведем изображения собственных функций, соответствующие значениям γ_h при $h = 1/128$.

Легко заметить, что результаты расчетов для квадратной области имеют качественные отличия от расчетов для концентрического кольца.

Во-первых, величины LBB-констант для представленных схем различаются между собой и весьма далеки от значения константы Ладыженской. Величины γ_h при $h \rightarrow 0$ монотонно убывают и ограничены снизу (в силу выполнения LBB-условия), значит сходятся; но не к значению \mathcal{L}^2 , более того, они не принадлежат интервалу (13.25).

Главные причины этого — ярко выраженная сингулярность собственной функции и потеря симметрии непрерывной задачи при переходе к дискретной. Исходная область (квадрат) имеет четыре оси симметрии, а разбиения $T_{h/2}$ и $Q_{h/2}$ соответственно одну и две, что естественным образом проявляется в форме младшей собственной функции.

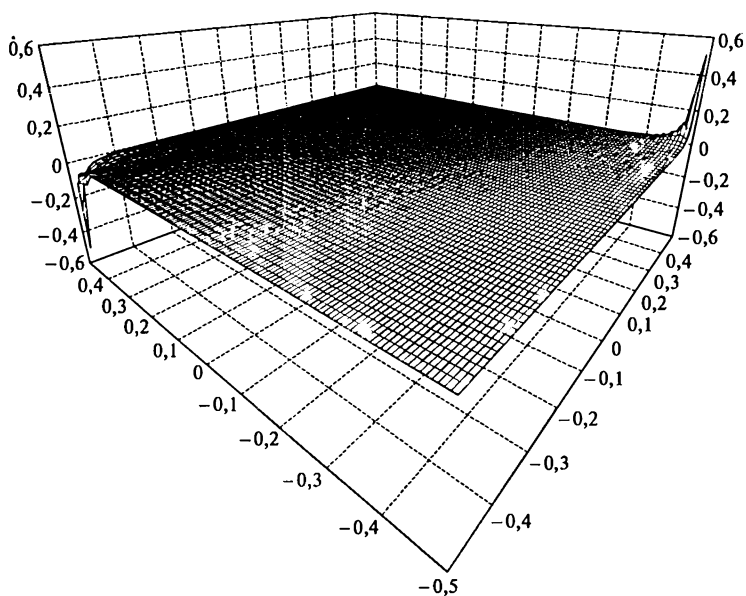


Рис. 1. Младшая собственная функция в схеме 1 при $h = 1/128$

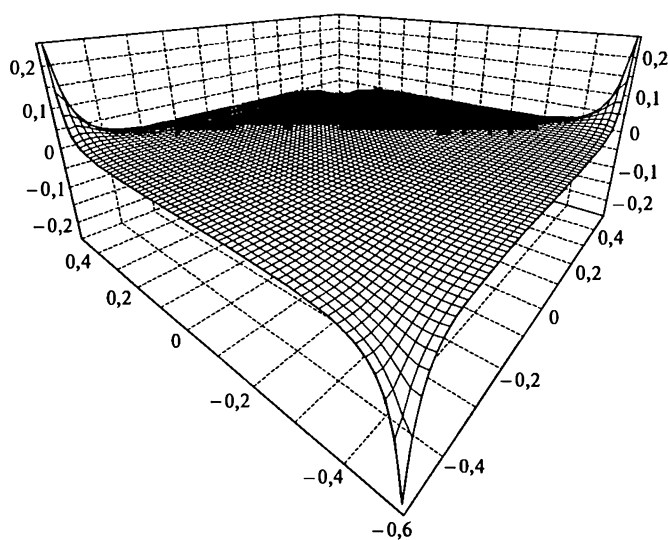


Рис. 2. Младшая собственная функция в схеме 2 при $h = 1/128$

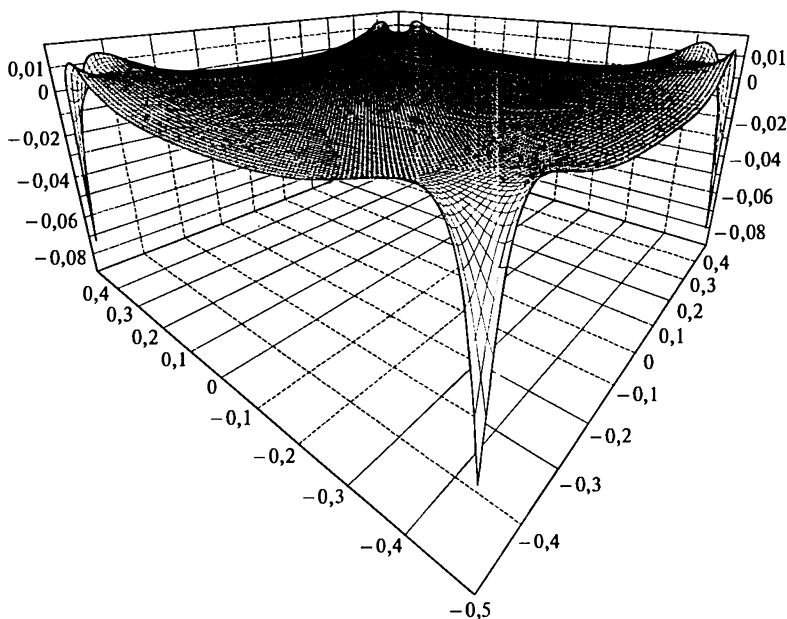


Рис. 3. Собственная функция, имеющая четыре оси симметрии

Для сравнения на рис.3 приведено изображение собственной функции, имеющей четыре оси симметрии, которое было получено на основе одной из схем метода конечных разностей [9, 96]. Ясно, что множества решений задачи (14.1) для схем 1 и 2 такого рода функций не содержат.

Во-вторых, экспериментальная скорость сходимости γ_h к своим предельным значениям (в данном случае – LBB-константам) существенно хуже, чем $O(h^2)$. Кроме того, она замедляется при уменьшении h , что следует из анализа асимптотики убывания разности значений γ_h из соседних строк табл. 2.

Скорее всего, для каждой из схем существует уравнение (аналогичное (13.15), но более сложное), связывающее параметры h , γ_h и «сеточный» показатель особенности собственной функции μ_h . Это резко усложняет задачу оценивания LBB-константы, так как процедура экстраполяции (вообще говоря, необходимая, в силу неконструктивной монотонности γ_h) становится непригодной.

Резюмируем вышесказанное: даже в случае области простейшей формы определение значения LBB-константы для конкретной схемы является непростой задачей. Для квадратной области и рассматри-

ваемых схем оценкой снизу, видимо, может служить величина константы Ладыженской, для которой, в свою очередь, точное значение пока не известно, но имеются двусторонние оценки (13.25).

14.2.5. Расчеты для прямоугольной области

Рассмотрим прямоугольную область

$$\Omega = \{(y_1, y_2) \mid 0 < y_i < \ell_i, i = 1, 2\}, \quad \ell_1 = \ell, \quad \ell_2 = 1$$

с параметром $\ell \geq 1$. Из первой части приложения (см. раздел 13.4) имеем, что при увеличении ℓ константа Ладыженской \mathcal{L} убывает как $O(\ell^{-1})$. При этом анализ уравнения (13.15) показывает, что соответствующая собственная функция становится более гладкой. Познакомимся с численными экспериментами, иллюстрирующими теоретические положения. Частично они представлены в [96, 107], целиком и с дополнениями — в [8].

Таблица 3 содержит значения γ_h , полученные для различных ℓ и $h = h_1 = h_2$.

Таблица 3

ℓ	h	Схема 1	Схема 2
2	1/8	0,151379	0,162477
	1/16	0,150740	0,155139
	1/32	0,150359	0,152273
	1/64	0,150164	0,151060
	1/128	0,150067	0,150507
4	1/8	0,048072	0,049321
	1/16	0,047848	0,048174
	1/32	0,047757	0,047843
	1/64	0,047729	0,047752
8	1/8	0,012730	0,012969
	1/16	0,012655	0,012714
	1/32	0,012628	0,012642
	1/64	0,012620	0,012624

Самое заметное — с ростом отношения сторон ℓ результаты расчетов для обеих схем приближаются друг к другу.

Если зафиксировать некоторое значение h , например, 1/64, то в расчетах обнаруживается быстрый выход LBB-константы на асимптотику $O(\ell^{-1})$.

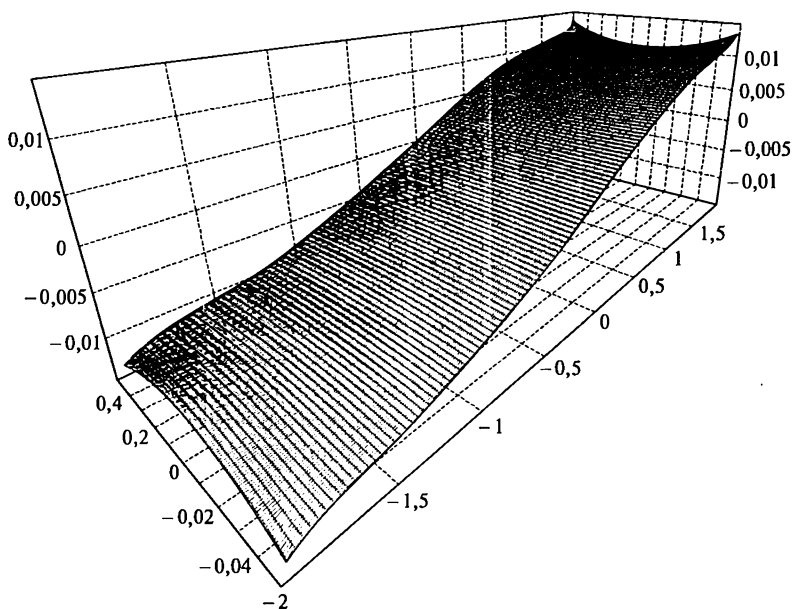


Рис. 4. Младшая собственная функция в схеме 2 при $\ell = 4$, $h_1 = h_2 = 1/64$

Если зафиксировать некоторое значение ℓ , например, 2, то экспериментальная скорость сходимости по параметру h существенно увеличивается по сравнению с $\ell = 1$: она становится похожей на $O(h)$ в то время, как из табл. 2 не следует даже асимптотика $O(\sqrt{h})$. Причем с ростом ℓ эта сходимость улучшается.

Приведем для полноты картины на рис. 4 изображение младшей собственной функции, посчитанной по схеме 2 при $\ell = 4$ и $h_1 = h_2 = 1/64$, и подытожим результаты численного анализа. Кажется правдоподобным, что сингулярность собственных функций является самым значимым фактором при исследовании константы Ладыженской и LBB-констант различных дискретных аналогов для задачи Стокса. При существенном ослаблении особенности (например, за счет увеличения параметра ℓ в прямоугольной области) происходит не только сближение LBB-констант для разных схем между собой, но и их уравнивание с константой Ладыженской.

Этот вывод хорошо согласуется с многочисленными расчетами из работы [167], проведенными как в двухмерном, так и в трехмерном случаях.

Отметим, что использование спектральной задачи (13.5) не является единственным подходом к вычислению константы Ладыженской; например, можно применять специальное интегральное уравнение [10, 96]. Хотя основные трудности и выводы из расчетов при этом, конечно, сохраняются.

ЧИСЛЕННЫЙ АНАЛИЗ РОЛИ ОПЕРАТОРА $\beta BC^{-1}B^T$ В СХОДИМОСТИ GMSOR

Основная цель этой главы — исследовать влияние слагаемого $\beta BC^{-1}B^T u^k$ на сходимость алгоритма GMSOR (12.3), применяемого для численного решения стационарной двумерной задачи Навье-Стокса:

$$\begin{aligned} -\Delta \mathbf{u} + \operatorname{Re}(\mathbf{u} \cdot \nabla) \mathbf{u} + \operatorname{grad} p &= 0 & \text{в } \Omega, \\ \operatorname{div} \mathbf{u} &= 0 & \text{в } \Omega, \\ \mathbf{u} &= \mathbf{g} & \text{на } \partial\Omega, \\ \int_{\Omega} p d\Omega &= 0, \end{aligned} \quad (15.1)$$

где Ω — ограниченная область в \mathbb{R}^2 , $\mathbf{g} = (g_1, g_2)^T : \partial\Omega \rightarrow \mathbb{R}^2$ — заданные краевые условия, $\operatorname{Re} \geq 0$ — число Рейнольдса, $\mathbf{u} = (u_1, u_2)^T : \Omega \rightarrow \mathbb{R}^2$ и $p : \Omega \rightarrow \mathbb{R}$ — искомые распределения скоростей и давления.

Наиболее распространенный подход к численному решению задачи (15.1) состоит в сведении непрерывной системы при помощи сеточных аппроксимаций, методов конечных элементов и др. (см., например, [79, 151, 158]) к конечномерной системе (в общем случае нелинейной) алгебраических уравнений с седловой точкой вида (12.1).

Алгоритм GMSOR в форме (12.3) для решения дискретной задачи содержит три независимых итерационных параметра $\alpha, \tau > 0$, $\beta \in \mathbb{R}$. Напомним, что для линейной симметричной регулярной задачи, соответствующей случаю $\operatorname{Re} = 0$, было показано, что оптимальным является выбор параметра $\beta = 0$ (теоремы 7.2.3, 7.3.3). При решении же нелинейной задачи подобный результат отсутствует, однако исторически сложились три подхода к выбору итерационных параметров:

- 1) $\alpha, \tau > 0$, $\beta = 0$ — двухпараметрический алгоритм Эрроу — Гурвица (1958, [108]),
- 2) $\alpha, \tau > 0$, $\beta = \alpha^{-1}$ — двухпараметрический алгоритм Кобелькова (1978, [46]),
- 3) $\alpha, \tau > 0$, $\beta \in \mathbb{R}$ — трехпараметрический алгоритм (1999, [129]).

Сравнение этих подходов проведем на двух модельных задачах при различном выборе предобусловливателя Q : задаче о течении в квадратной каверне ($Q = A$) и задаче обтекания тела в трубе ($Q \neq A$).

15.1. АЛГОРИТМ ЧИСЛЕННОЙ ОПТИМИЗАЦИИ

Для приближенного вычисления «оптимальных» параметров в алгоритме (12.3) воспользуемся следующей идеей, основанной на представлениях оптимального показателя q_K , полученных для линейных задач в теоремах 7.2.2, 7.3.2. Рассмотрим, например, представление

$$q_{K_1} = \min_{\alpha', \tau' > 0, \beta' \in \mathbb{R}} L'(\alpha', \beta', \tau'),$$

где

$$K_1 = K_1(\delta, \Delta, \gamma, \Gamma),$$

$$L'(\alpha', \beta', \tau') = \max_{\substack{s=\delta, \Delta \\ t=\gamma, \Gamma}} \left| 1 - (\tau's + \beta'ts) \pm \sqrt{(\tau's + \beta'ts)^2 - 2\alpha'ts} \right|,$$

$$\tau' = \frac{\tau}{2}, \quad \alpha' = \frac{\tau}{2\alpha}, \quad \beta' = \tau \frac{\beta + 1/\alpha}{2}.$$

Обратим внимание, что определения параметров α, β, τ алгоритмов (7.1) и (12.3) отличаются, однако это не влияет на сущность рассуждений.

Известно (см. доказательство теоремы 7.2.2), что функция $L'(\alpha', \beta', \tau')$ обладает единственным локальным минимумом, который также является строгим глобальным минимумом, причем для любых зависимостей вида

$$\alpha'(x) = \alpha'_0(1-x) + \alpha'_1 x, \quad \beta'(x) = \beta'_0(1-x) + \beta'_1 x, \quad \tau'(x) = \tau'_0(1-x) + \tau'_1 x$$

функция вида $L'(\alpha'(x), \beta'(x), \tau'(x))$ как функция $x \in \mathbb{R}$ принадлежит классу $YS(\mathbb{R})$, т. е. является унимодальной в \mathbb{R} . Таким образом, для численного нахождения q_{K_1} можно эффективно использовать метод оптимального покординатного спуска [6, 15], имеющий в данном случае глобальную сходимость, причем для нахождения минимума $L'(\alpha', \beta', \tau')$ при фиксированных двух переменных можно воспользоваться методом бисекции для унимодальных функций [6].

Для приближенного вычисления функции $L'(\alpha', \beta', \tau')$ воспользуемся свойством [49]

$$\rho(T(\alpha', \beta', \tau')) = \lim_{k \rightarrow +\infty} \|T^k\|^{1/k} = \lim_{k \rightarrow +\infty} \sqrt[k]{\frac{\|T^k z\|}{\|z\|}},$$

которое имеет место для почти всех $z \in Z$; здесь $T = T(\alpha', \beta', \tau')$ — линейный оператор перехода метода. Таким образом, можно предположить асимптотическое равенство

$$L'(\alpha', \beta', \tau') \approx \sqrt[k]{\|r^k\| / \|r^0\|},$$

где r^k — вектор невязки алгоритма на k -й итерации, $k \in \mathbb{N}$, $k \gg 1$.

Все приведенные рассуждения легко адаптируются к случаям $\beta = 0$ и $\beta = 1/\alpha$. Отметим, что все этапы (но не их обоснования!) в предложенной схеме вычисления «оптимальных» параметров могут быть без изменений применены в нелинейном случае. Практический анализ показывает, что данный алгоритм дает эффективный результат и позволяет существенно сократить время численной оптимизации.

15.2. ЗАДАЧА О КВАДРАТНОЙ КАВЕРНЕ

Рассмотрим задачу (15.1) в области

$$\Omega = \{(x, y) \mid 0 < x < 1, 0 < y < 1\}$$

с граничными условиями

$$g|_{x=0} = 0, \quad g|_{x=1} = 0, \quad g|_{y=0} = 0, \quad g_2|_{y=1} = 0, \quad g_1|_{y=1} = 1,$$

Эта задача называется «задачей о квадратной каверне» и является, пожалуй, самым популярным тестом для алгоритмов решения задачи Навье–Стокса. Для ее приближенного решения часто используют так называемые аппроксимации на смещенных сетках, впервые предложенные В. И. Лебедевым в работе [57] и широко применявшиеся в работах Г. М. Кобелькова [48]. Одно из несомненных удобств такого подхода состоит в том, что имеется возможность прямым методом обращать оператор A с помощью быстрого дискретного преобразования Фурье, если сетки равномерные и число узлов в одном из направлений является степенью двойки [76]. Это естественным образом мотивирует выбор $Q = A$; оператор C в данном случае положим равным I .

Для численного решения задачи будем использовать равномерную сетку с числом узлов равным $2^6 = 64$ по каждому направлению с общим числом независимых переменных порядка 12000, а в качестве начального приближения нелинейной задачи — предварительно вычисленное решение при $Re = 0$ с сеточной L_2 -нормой невязки, не превышающей 10^{-12} . В качестве критерия остановки алгоритма примем уменьшение сеточной L_2 -нормы невязки в 10^3 раз по сравнению

Таблица 4

Re	N_{2PA}	N_{2PB}	N_{3P}
0	10	11	8
1	12	13	10
5	13	13	10
10	16	14	11
20	26	15	15
50	80	22	22
100	217	31	31
200	1177	60	53
500	—	365	235
1000	—	1705	726

с начальной. Не останавливаясь на вопросах сходимости решения разностной задачи к дифференциальной отметим, что полученные картины течений хорошо согласуются с известными результатами про задачу о квадратной каверне [154].

Результаты расчетов представлены в табл. 4 и на рис. 5 и характеризуют зависимость числа итераций для каждого метода при «оптимальном» выборе параметров. Здесь приняты следующие обозначения: N_{2PA} — число итераций в методе Эрроу—Гурвица, N_{2PB} — число итераций в методе Кобелькова, N_{3P} — число итераций в трехпара-

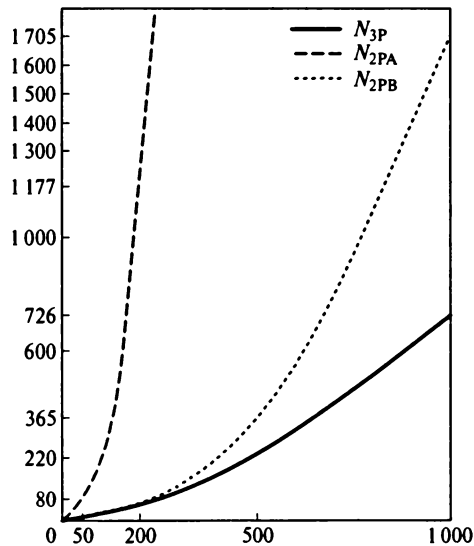


Рис. 5. Графики зависимости числа итераций от Re

метрическом методе. Прочерк в таблице означает, что сходимость вообще не была достигнута (при описанном подходе).

15.3. ОБТЕКАНИЕ ТЕЛА В ТРУБЕ

Рассмотрим задачу (15.1) в области Ω , изображенной на рис. 6: с граничными условиями

$$g|_{\Gamma_2, \Gamma_4, \Gamma_5} = 0, \quad g|_{\Gamma_1, \Gamma_3} = ((H - y)y/H^2, 0)^T.$$

Эта задача моделирует состояние воздушной среды, окружающей автомобиль, при движении по дороге с постоянной скоростью (см., например, [55]).

Для приближенного решения задачи воспользуемся аппроксимацией на основе параметрических неконформных четырехугольных конечных элементов типа $\tilde{Q}1/Q0$ [181], которые являются аналогом известных треугольных неконформных элементов [146], часто используемых при решении эллиптических задач. Основными преимуществами этого подхода являются LBB-устойчивость, простота и эффективность реализации [190], что позволило положить их в основу универсального пакета программ численной гидродинамики FEATFLOW [3].

При расчетах использовалась сетка (из примера «ASMO1» к свободно распространяемому пакету FEATFLOW), полученная равномерным измельчением в 4 раза (т. е. каждый четырехугольник делился на 4^2 подобных) базовой сетки, изображенной на рис. 7. При этом общее число независимых переменных составило 14328.

В качестве Q^{-1} — «сглаживающего» оператора для A , использован оператор перехода в ускоренном при помощи линейного стационарного метода Чебышева (50 итераций) попеременно-треугольном методе с итерационным параметром $\omega = 0,96$ [76]. В качестве C

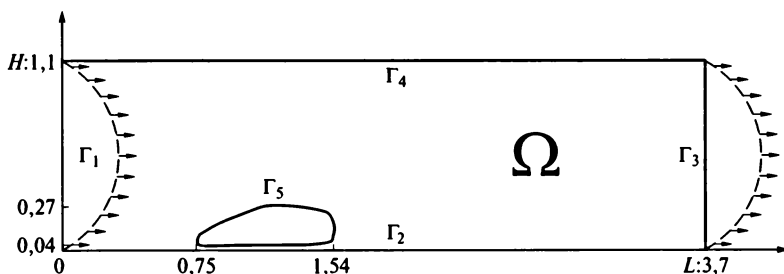


Рис. 6. Область Ω

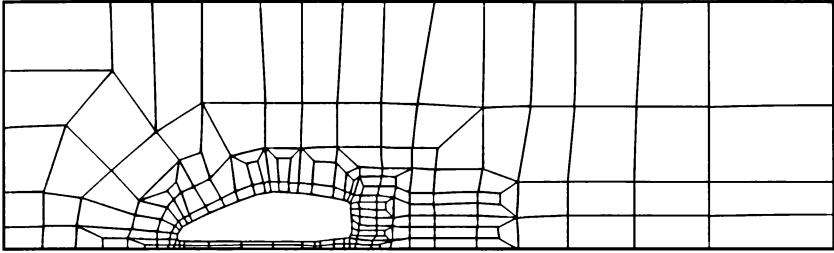


Рис. 7. Базовая сетка

выбрана матрица масс базисных элементов в P_h . Полученная численным образом оценка обусловленности имеет вид

$$(\text{cond}_2(Q^{-1}A))^{-1} \approx 0,14.$$

Начальное приближение выбрано нулевым. Критерий окончания работы — уменьшение сеточной L_2 -нормы невязки в 10^3 раз. Не останавливаясь на вопросах сходимости решения разностной задачи к дифференциальной отметим, что полученные картины течений хорошо согласуются с альбомом течений FEATFLOW [3].

Расчеты показывают, что скорость сходимости алгоритмов практически не зависит от величины равномерного измельчения сетки, что хорошо согласуется с теоретическими результатами.

График зависимости «оптимального» числа итераций от величины Re для каждого из случаев представлен на рис. 8.

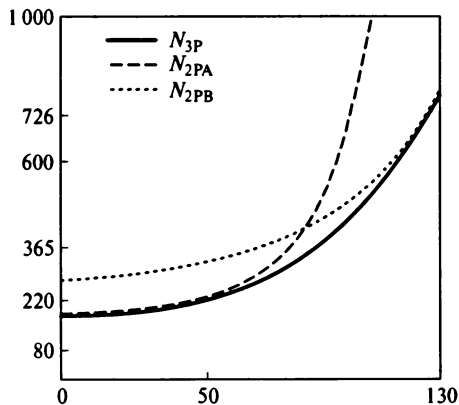
Рис. 8. Графики зависимости числа итераций от Re

Таблица 5

Re	N_{2PA}	N_{2PB}	N_{3P}
0	177	273	174
10	181	278	178
20	185	283	182
30	191	288	189
40	204	304	200
50	216	320	212
60	231	347	230
70	261	368	253
80	286	381	285
90	393	427	368
100	560	471	453
110	—	531	520
120	—	655	635
130	—	805	787

Сделаем выводы на основе проведенных численных экспериментов.

Для начала отметим, что вычислительные сложности одной итерации методов практически совпадают, что позволяет сконцентрироваться только на анализе числа итераций.

При небольших числах Re трехпараметрический метод и метод Эрроу—Гурвица обладают практически одинаковой эффективностью, независимо от качества предобуславливателя Q . Этого и следовало ожидать, так как этот случай «близок» к линейному, для которого оптимальное значение β равно 0 согласно теоремам 7.2.3, 7.3.3. Что касается метода Кобелькова, то в этой ситуации его не рекомендуется применять, так как оператор $Q^{-1}A$ «плохо» обусловлен.

Когда же число Re становится достаточно большим, метод Эрроу—Гурвица теряет свою конкурентноспособность на фоне двух других методов. Это свидетельствует о существенном влиянии слагаемого $\beta BC^{-1}B^T u^k$ на эффективность метода GMSOR в нелинейном случае.

Как показывает второй тест, при «плохой» обусловленности оператора $Q^{-1}A$ имеется тенденция к сближению оптимального числа итераций трехпараметрического алгоритма и алгоритма Кобелькова. Последнее дает основание предполагать, что стратегия выбора $\beta = 1/\alpha$ является достаточно близкой к оптимальной при больших значениях $\text{cond}_2(Q^{-1}A)$ и Re.

СПИСОК ЛИТЕРАТУРЫ

- [1] Математическая энциклопедия. Т.3. — М.: Советская энциклопедия, 1984.
- [2] Зарубежные библиотеки и пакеты программ по вычислительной математике / Ред. У.Кауэлл. — М.: Наука, 1993.
- [3] Finite Element Analysis Tool (FEATFLOW). <http://www.featflow.de>.
- [4] Агошков В. И., Дубовский П. Б., Шутяев В. П. Методы решения задач математической физики. — М.: Физматлит, 2002.
- [5] АЛЕКСЕЕВ В. М., ТИХОМИРОВ В. М., ФОМИН С. В. Оптимальное управление. — М.: Наука, 1979.
- [6] АОКИ М. Введение в методы оптимизации. — М.: Наука, 1977.
- [7] АРИСТОВ П. П. Об ускорении сходимости одного итерационного метода решения задачи Стокса // *Известия вузов. Математика*, 1994, № 9, с. 3–10.
- [8] АРИСТОВ П. П. Исследование семейства алгоритмов для решения линейных уравнений гидродинамики и теории упругости. Дисс. канд. физ.-мат. наук. — М.: МГУ им. Ломоносова, 1995.
- [9] АРИСТОВ П. П., ЧИЖОНКОВ Е. В. О некоторых конечно-разностных аппроксимациях задачи Стокса // *Фундаментальная и прикладная математика*, 1995, том 1, № 3, с. 573–580.
- [10] АРУШАНЯН И. О., ЧИЖОНКОВ Е. В. Численное исследование константы в $\inf\text{-}\sup$ неравенстве методом граничных интегральных уравнений // *Пакеты прикладных программ*. — М.: Изд-во МГУ, 1997, с. 49–59.
- [11] АСТРАХАНЦЕВ Г. П. Анализ алгоритмов типа Эрроу-Гурвица // *ЖВМ и МФ*, 2001, том 41, № 1, с. 17–28.
- [12] АХИЕЗЕР Н. И. Лекции по теории аппроксимации. — М.: Наука, 1965.
- [13] БАЙОККИ К., КАПЕЛО А. Вариационные и квазивариационные неравенства. Приложения к задачам со свободной границей. — М.: Наука, 1988.
- [14] БАХВАЛОВ Н. С. Эффективный итерационный метод решения уравнений Ламе для почти несжимаемых сред и уравнений Стокса // *Доклады Академии наук СССР*, 1991, том 319, № 1, с. 13–17.

- [15] Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы, 6-е изд. — М.: Бином. Лаборатория знаний, 2008.
- [16] Бахвалов Н. С., Кобельков Г. М., Чижонков Е. В. Эффективные методы решения уравнений Навье-Стокса // *Численное моделирование в аэрогидродинамике*. — М.: Наука, 1986, с. 37–45.
- [17] Бахвалов Н. С., Кобельков Г. М., Чижонков Е. В. Итерационный метод решения эллиптических задач со скоростью сходимости, не зависящей от разброса коэффициентов. — М.: Препринт ОВМ АН СССР № 190, 1988.
- [18] Богачев К. Ю. Эффективные алгоритмы решения жестких эллиптических задач с большими параметрами // *Тр. матем. центра им. Н. И. Лобачевского. Т. 2. Итерационные методы решения линейных и нелинейных сеточных задач*. — Казань: УНИПРЕСС, 1999, с. 3–44.
- [19] Боговский М. Е. Решение первой краевой задачи для уравнения неразрывности в несжимаемой среде // *Доклады Академии наук СССР*, 1979, том 248, № 5, с. 1037–1040.
- [20] Быченков Ю. В. Об одном трехпараметрическом методе решения уравнений Навье-Стокса // *ЖВМ и МФ*, 2002, том 42, № 9, с. 1420–1427.
- [21] Быченков Ю. В. Исследование и оптимизация многопараметрических алгоритмов для решения задач с седловыми операторами. Дисс. канд. физ.-мат. наук. — М.: МГУ им. Ломоносова, 2003.
- [22] Быченков Ю. В. Об оптимизации одного класса алгоритмов для решения несимметричных седловых задач // *ЖВМ и МФ*, 2005, том 45, № 7, с. 1157–1166.
- [23] Быченков Ю. В. Оптимизация обобщенного метода переменных симметричных и кососимметричных итераций для решения симметричных седловых задач // *ЖВМ и МФ*, 2006, том 46, № 6, с. 983–995.
- [24] Быченков Ю. В. О спектральных свойствах одного пучка операторов // *ЖВМ и МФ*, 2007, том 47, № 2, с. 197–205.
- [25] Быченков Ю. В. О преобуславливании седловых задач методом переменных симметричных и кососимметричных итераций // *ЖВМ и МФ*, 2009, том 49, № 3, с. 411–421.
- [26] Быченков Ю. В., Чижонков Е. В. Об одном подходе к регуляризации задач с седловой точкой // *Материалы Четвертого Всероссийского семинара «Сеточные методы для краевых задач и приложения»*. — Казань: Изд-во Казанского мат.общества, 2002, с. 40–42.

- [27] БЫЧЕНКОВ Ю. В., ЧИЖОНКОВ Е. В. К решению нерегулярных задач с седловой точкой // *Вестник МГУ, Сер. 1. Математика. Механика*, 2004, № 1, с. 17–20.
- [28] ВАСИЛЕВСКИЙ Ю. В. Методы решения краевых задач с использованием нестыкующихся сеток // *Тр. Матем. центра им. Н. И. Лобачевского. Т.2. Итерационные методы решения линейных и нелинейных сеточных задач.* – Казань: УНИ-ПРЕСС, 1999, с. 94–121.
- [29] ВАСИЛЕВСКИЙ Ю. В., ОЛЬШАНСКИЙ М. А. Краткий курс по многосеточным методам и методам декомпозиции области. – М.: МАКС Пресс, 2007.
- [30] ВОЕВОДИН В. В., КУЗНЕЦОВ Ю. А. Матрицы и вычисления. – М.: Наука, 1984.
- [31] ГЛАВАЧЕК И., ГАСЛИНГЕР Я., НЕЧАС И., ЛОВИШЕК Я. Решение вариационных неравенств в механике. – М.: Мир, 1986.
- [32] ГЛОВИНСКИ Р., ЛИОНС Ж.-Л., ТРЕМОЛЬЕР Р. Численное исследование вариационных неравенств. – М.: Мир, 1979.
- [33] ГОДУНОВ С. К. Лекции по современным аспектам линейной алгебры. – Новосибирск: Научная книга (ИДМИ), 2002.
- [34] ГОРЕЛОВА М. В., ЧИЖОНКОВ Е. В. Об одном итерационном методе с модельным седловым оператором на верхнем слое // *Материалы Третьего Всероссийского семинара «Теория сеточных методов для нелинейных краевых задач».* – Казань: Изд-во Казанского мат. общества, 2000, с. 39–41.
- [35] ГОРЕЛОВА М. В., ЧИЖОНКОВ Е. В. О решении седловых задач методами с модельными седловыми операторами на верхнем слое // *Известия Вузов. Математика*, 2003, № 8, с. 19–27.
- [36] ГОРЕЛОВА М. В., ЧИЖОНКОВ Е. В. К предобуславливанию седловых задач с помощью седловых операторов // *ЖВМ и МФ*, 2004, том 44, № 9, с. 1523–1533.
- [37] ДЕМЬЯНОВ В. Ф., МАЛОЗЁМОВ В. Н. Введение в минимакс. – М.: Наука, 1972.
- [38] ДЬЯКОНОВ Е. Г. О применении эквивалентных по спектру операторов для решения разностных аналогов сильно эллиптических систем // *Доклады Академии наук СССР*, 1965, том 63, № 6, с. 1314–1317.
- [39] ДЬЯКОНОВ Е. Г. Оценки вычислительной работы для краевых задач с оператором Стокса // *Известия вузов. Математика*, 1983, № 7, с. 46–58.
- [40] ДЬЯКОНОВ Е. Г. О некоторых классах седловых градиентных методов // *Вычислительные процессы и системы.* Вып. 5. М.: Наука, 1987, с. 101–115.

- [41] Дьяконов Е. Г. Минимизация вычислительной работы. Асимптотически оптимальные алгоритмы для эллиптических задач. — М.: Наука, 1989.
- [42] Дьяконов Е. Г. Энергетические пространства и их применения. — М.: Изд. отдел ф-та ВМиК МГУ им. М. В. Ломоносова, 2001.
- [43] Икрамов Х. Д. Несимметричная проблема собственных значений. — М.: Наука, 1991.
- [44] Ильин В. П. Методы неполной факторизации для решения алгебраических систем. — М.: Наука, Физматлит, 1995.
- [45] Кобельков Г. М. Об эквивалентных нормировках подпространств L_2 // *Analysis Mathematica*, 1977, № 3, с. 177–186.
- [46] Кобельков Г. М. О методах решения уравнений Навье–Стокса // *Доклады Академии СССР*, 1978, том 243, № 4, с. 843–846.
- [47] Кобельков Г. М. О решении эллиптических уравнений с сильно меняющимися коэффициентами. — М.: Препринт ОВМ АН СССР № 145, 1987.
- [48] Кобельков Г. М. О численных методах решения уравнений Навье–Стокса в переменных скорость–давление // *Вычислительные процессы и системы*. Вып. 8. — М.: Наука, 1991, с. 204–236.
- [49] Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. 7-е изд. — М.: Физматлит, 2006.
- [50] Кондратьев В. А. Краевые задачи для эллиптических уравнений в областях с коническими или угловыми точками // *Труды Моск. матем. общества*, 1967, том 16, с. 209–292.
- [51] Кондратьев В. А., Олейник О. А. Краевые задачи для уравнений с частными производными в негладких областях // *Успехи матем. наук*, 1983, том 38, № 2(230), с. 3–76.
- [52] Кондратьев В. А., Олейник О. А. О зависимости констант в неравенстве Корна от параметров, характеризующих геометрию области // *Успехи матем. наук*, 1989, том 44, № 6(270), с. 157–158.
- [53] Красногорский А. М. О разрешимости некоторых краевых задач в областях с негладкой границей. Дисс. канд. физ.-мат. наук. М.: МЭИ, 2006.
- [54] Красносельский М. А., Забрейко П. П. Геометрические методы нелинейного анализа. — М.: Наука, 1975.
- [55] Кульпина И. Э., Перминов С. М., Писковский В. О., Соколов А. Г. Численное моделирование процесса обтекания автомобиля // *Матем. моделирование*, 1994, том 6, № 1, с. 54–68.

- [56] ЛАДЫЖЕНСКАЯ О. А. Математические вопросы динамики вязкой несжимаемой жидкости. — М.: Наука, 1970.
- [57] ЛЕБЕДЕВ В. И. Метод сеток для уравнений типа Соболева // *Доклады Академии СССР*, 1956, том 114, № 6, с. 1166–1169.
- [58] ЛЕБЕДЕВ В. И. О КР-методе ускорения сходимости итераций при решении кинетического уравнения // *Численные методы решения задач математической физики*. — М.: Наука, 1966, с. 154–161.
- [59] ЛЕБЕДЕВ В. И. Теория игр и оптимальность итерационных методов // *Разностные и вариационно-разностные методы*. Вып. 4. — Новосибирск: Изд-во ВЦ СО АН СССР, 1979, с. 11–22.
- [60] ЛЕБЕДЕВ В. И. Метод композиции. — М.: ОВМ АН СССР, 1986.
- [61] ЛЕБЕДЕВ В. И. Функциональный анализ и вычислительная математика. — М.: Физматлит, 2005.
- [62] ЛЕБЕДЕВ В. И., ФИНОГЕНОВ С. А. О порядке выбора итерационных параметров в чебышевских циклических итерационных методах // *ЖВМ и МФ*, 1971, том 11, № 2, с. 425–438.
- [63] ЛОЙЦЯНСКИЙ Л. Г. Механика жидкости и газа. — М.: Наука, 1973.
- [64] МАРЧУК Г. И., ЛЕБЕДЕВ В. И. Численные методы в теории переноса нейтронов. — М.: Атомиздат, 1981.
- [65] МИЛЮТИН С. В. О методах решения одной седловой задачи // *Тр. матем. центра им. Н. И. Лобачевского: численные методы решения задач математической физики*. — Казань: Изд-во Казанского мат. общества, 2004, том 26, с. 203–209.
- [66] МИЛЮТИН С. В. Практическая оптимизация трехпараметрического итерационного метода для расчета течений бингамовской жидкости // *Вычислительные методы и программирование*, 2008, том 9, № 1, с. 38–43.
- [67] МИЛЮТИН С. В. Об одном алгоритме расчета течений бингамовской жидкости // *ЖВМ и МФ*, 2009, том 49, № 3, с. 567–577.
- [68] МИХЛИН С. Г. Дальнейшее исследование спектра Коссера // *Вестн. ЛГУ. Сер. Математика*, 1967, № 7 (Вып. 2), с. 96–102.
- [69] МИХЛИН С. Г. Спектр пучка операторов теории упругости // *Успехи матем. наук*, 1973, том 28, № 3 (171), с. 43–82.
- [70] ОЛЬШАНСКИЙ М. А. Об одной задаче типа Стокса с параметром // *ЖВМ и МФ*, 1996, том 36, № 2, с. 75–86.
- [71] ОЛЬШАНСКИЙ М. А. Лекции и упражнения по многосеточным методам. — М.: Физматлит, 2005.

- [72] ОЛЬШАНСКИЙ М. А., ЧИЖОНКОВ Е. В. О наилучшей константе в $\inf\text{--}\sup$ условии для вытянутых прямоугольных областей // *Мат. заметки*, 2000, том 67 (Вып. 3), с. 387–396.
- [73] ПАРЛЕТТ Б. Симметричная проблема собственных значений. — М.: Мир, 1983.
- [74] РЕПИН С. И. Оценки отклонения от точных решений некоторых краевых задач с условием несжимаемости // *Алгебра и анализ*, 2004, том 16, № 5, с. 121–164.
- [75] РОУЧ П. Вычислительная гидродинамика. — М.: Мир, 1980.
- [76] САМАРСКИЙ А. А., НИКОЛАЕВ Е. С. Методы решения сеточных уравнений. — М.: Наука, 1978.
- [77] СЕДОВ Л. И. Механика сплошной среды. Т. 1. — М.: Наука, 1973.
- [78] СОВОЛЕВ С. Л. Введение в теорию кубатурных формул. — М.: Наука, 1974.
- [79] ТЕМАМ Р. Уравнения Навье–Стокса. Теория и численный анализ. — М.: Мир, 1981.
- [80] ТЫРТЫШНИКОВ Е. Е. Методы численного анализа. — М.: Издательский центр «Академия», 2007.
- [81] УИЛКИНСОН ДЖ.Х. Алгебраическая проблема собственных значений. — М.: Наука, 1970.
- [82] ХОРН Р., ДЖОНСОН Ч. Матричный анализ. — М.: Мир, 1989.
- [83] ЧИЖОНКОВ Е. В. О сходимости одного алгоритма для решения задачи Стокса // *Вестник Моск. ун-та. Сер. 15. Вычисл. матем. и киберн.*, 1995, № 2, с. 12–17.
- [84] ЧИЖОНКОВ Е. В. К оптимизации алгоритмов решения задачи Стокса // *Вестник Моск. ун-та. Сер. 1. Математика. Механика*, 1995, № 6, с. 93–96.
- [85] ЧИЖОНКОВ Е. В. К сходимости метода искусственной сжимаемости // *Вестник Моск. ун-та. Сер. 1. Математика. Механика*, 1996, № 2, с. 13–20.
- [86] ЧИЖОНКОВ Е. В. К решению уравнений Стокса в областях с большим разбросом линейных размеров // *Материалы Всероссийского семинара «Теория сеточных методов для нелинейных краевых задач»*. — Казань: Казанский фонд «Математика», 1996, с. 105–107.
- [87] ЧИЖОНКОВ Е. В. О спектральных свойствах оператора давления, порожденного уравнениями Стокса // *Материалы международной конференции и Чебышевских чтений*. Том 2. — М.: Изд-во мех.-мат. фак-та МГУ, 1996, с. 363–366.

- [88] Чижонков Е. В. О сходимости алгоритма Эрроу–Гурвица для алгебраической системы типа Стокса // *Доклады Академии наук*, 1998, том 361, № 5, с. 600–602.
- [89] Чижонков Е. В. О сходимости модифицированного метода SSOR для алгебраической системы типа Стокса // *Численный анализ: методы и программы*. — М.: Изд-во МГУ, 1998, с. 83–91.
- [90] Чижонков Е. В. Некоторые результаты о сходимости алгоритма Эрроу–Гурвица для алгебраической системы типа Стокса // *ЖВМ и МФ*, 1999, том 39, № 3, с. 521–531.
- [91] Чижонков Е. В. Об алгоритме Эрроу–Гурвица с переменными итерационными параметрами // *Известия вузов. Математика*, 1999, № 5 (444), с. 65–72.
- [92] Чижонков Е. В. О сходимости метода MSOR с переменными итерационными параметрами. — М.: Препринт мех.-мат. ф-та МГУ им. Ломоносова № 1, 1999.
- [93] Чижонков Е. В. О сходимости модифицированного метода Якоби для алгебраической системы типа Стокса // *Оптимизация численных методов: Труды межд. научной конф. Часть I* / Отв. ред. М. Д. Рамазанов. — Уфа: ИМБИЦ УНЦ РАН, 2000, с. 169–177.
- [94] Чижонков Е. В. К решению алгебраической системы типа Стокса при блочно диагональном предобуславливании // *ЖВМ и МФ*, 2001, том 41, № 4, с. 449–557.
- [95] Чижонков Е. В. Об одном обобщенном релаксационном методе решения линейных задач с седловым оператором // *Матем. моделирование*, 2001, том 13, № 12, с. 107–114.
- [96] Чижонков Е. В. Релаксационные методы решения седловых задач. — М.: ИВМ РАН, 2002.
- [97] Чижонков Е. В. Об ускорении сходимости метода Ланцоша при решении алгебраических систем с седловой точкой // *ЖВМ и МФ*, 2002, том 42, № 4, с. 504–513.
- [98] Чижонков Е. В. К оптимизации методов с конструктивными седловыми операторами на верхнем слое // *Математические идеи П. Л. Чебышева и их приложение к современным проблемам естествознания: Тезисы докладов международной конференции*. — Обнинск: Изд-во ОИАЭ, 2002, с. 90–91.
- [99] Чижонков Е. В. К решению задачи Стокса с интерфейсом // *Материалы Седьмого Всероссийского семинара «Сеточные методы для краевых задач и приложения»*. — Казань: Издательство Казанского государственного университета, 2007, с. 301–306.

- [100] ЧИЖОНКОВ Е. В. О численном решении задачи Стокса с интерфейсом // *ЖВМ и МФ*, 2009, том 49, № 1, с. 111–122.
- [101] ШАБАТ Б. В. Введение в комплексный анализ. — М.: Наука, 1961.
- [102] ШАЙДУРОВ В. В. Многосеточные методы конечных элементов. — М.: Наука, 1989.
- [103] ЭРРОУ К., ГУРВИЦ Л., УДЗАВА Х. Исследования по линейному и нелинейному программированию. — М.: ИЛ, 1962.
- [104] ЯНКЕ Е., ЭМДЕ Ф., ЛЕШ Ф. Специальные функции. — М.: Наука, 1968.
- [105] AGOSHKOV V. Optimal control methods in inverse problems and computational processes // *J. Inverse Ill-Posed Problems*, 2001, vol. 9, № 3, p. 205–218.
- [106] AGOSHKOV V., BARDOS C., BULEEV S. Solution of the Stokes problem as an inverse problem // *Comp. Meth. in Applied Math.*, 2002, vol. 2, № 3, p. 213–232.
- [107] ARISTOV P. P., CHIZHONKOV E. V. On the Constant in the LBB condition for rectangular domains. Report № 9534 (September 1995). — Dept. of Math., Univ. of Nijmegen. The Netherlands, 1995.
- [108] ARROW K., HURWICZ L., UZAWA H. Studies in Linear and Nonlinear Programming. — Stanford, CA: Stanford University Press, 1958.
- [109] ASHBY S. F., MANTEUFFEL T. A., SAYLOR P. E. A taxonomy for conjugate gradient methods // *SIAM J. Numer. Anal.*, 1990, vol. 27, p. 1542–1568.
- [110] BABUŠKA I., AZIZ A. K. Survey lectures on the mathematical foundations of the finite element method. Chapter 5, p. 111–184. In : The mathematical foundations of the finite element method with applications to partial differential equations, A. K. Aziz ed. — New York: Academic Press, 1972.
- [111] BABUŠKA I., OSBORN J. Eigenvalue Problems. Handbook of Numerical Analysis, V. II. Ciarlet P. G. and Lions J. L., editors. North Holland, 1991.
- [112] BAKHVALOV N. S. Solution of the Stokes nonstationary problems by the fictitious domain method // *Russ. J. Numer. Anal. Math. Modelling*, 1995, vol. 10, № 3, p. 163–172.
- [113] BANK R. E., WELFERT B. D., YSERENTANT H. A class of iterative methods for solving saddle point problems // *Numer. Math.*, 1990, № 56, p. 645–666.
- [114] BENZI M., GANDER M. J., GOLUB G. H. Optimization of the Hermitian and Skew-Hermitian Splitting Iteration for Saddle-Point

- Problems // *BIT Numerical Mathematics*, 2003, vol. 43, p. 881–900.
- [115] BENZI M., GOLUB G.H. An iterative method for generalized saddle point problems // *Technical Report SCCM-02-12, Scientific Computing and Computational Mathematics Program. Department of Computer Science, Stanford University*, 2002.
 - [116] BENZI M., GOLUB G.H. A Preconditioner for Generalized Saddle Point Problems // *SIAM Journal on Matrix Analysis and Applications*, 2004, vol. 26, № 1, p. 20–41.
 - [117] BENZI M., GOLUB G., LIESEN J. Numerical solution of saddle point problems // *Acta Numerica*, 2005, vol. 14, p. 1–137.
 - [118] BENZI M., SIMONCINI V. Spectral Properties of the Hermitian and Skew-Hermitian Splitting Preconditioner for Saddle Point Problems // *SIAM Journal on Matrix Analysis and Applications*, 2004, vol. 26, № 2, p. 377–389.
 - [119] BOFFI D., BREZZI F., GASTALDI L. On the convergence of eigenvalues for mixed formulations // *Ann. Scuola Norm. Sup. Pisa, Cl. Sci.*, 1997, vol. XXV, p. 131–154.
 - [120] BOFFI D., BREZZI F., GASTALDI L. On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form // *Math. Comp.*, 1999, vol. 69, № 229, p. 121–140.
 - [121] BRAESS D., SARAZIN R. An efficient smoother for the Stokes problem // *Appl. Numer. Math.*, 1997, vol. 23, p. 3–19.
 - [122] BRAMBLE J. H. A proof of the inf – sup condition for the Stokes equations on Lipschitz domains // *Math. Models. Methods. Appl. Sci.*, 2003, vol. 13, № 3, p. 361–371.
 - [123] BRAMBLE J. H., PASCIAK J. E. A Preconditioning Technique for Indefinite Systems Resulting from Mixed Approximations of Elliptic Problems // *Math. Comp.*, 1988, vol. 50, № 181, p. 1–17.
 - [124] BRAMBLE J. H., PASCIAK J. E., VASSILEV A. T. Analysis of the inexact Uzawa algorithm for saddle point problems // *SIAM Journ. Numer. Anal.*, 1997, vol. 33, № 4, p. 1072–1092.
 - [125] BRAMBLE J. H., PASCIAK J. E., VASSILEV A. T. Inexact Uzawa algorithms for nonsymmetric saddle point problems // *Math. Comp.*, 2000, vol. 69, № 230, p. 667–689.
 - [126] BREZZI F. On the existence, uniqueness and approximation of saddle - point problems arising from Lagrange multipliers // *R. A. I. R. O. Anal. Numer.*, 1974, vol. 8, p. 129–151.
 - [127] BREZZI F., FORTIN M. Mixed and Hybrid Finite Element Methods. — New York: Springer-Verlag, 1991.
 - [128] BYCHENKOV YU.V. Optimization of one class of nonsymmetrizable algorithms for saddle point problems //

- Russ. J. Numer. Anal. Math. Modelling*, 2002, vol. 17, № 6, p. 521–546.
- [129] BYCHENKOV YU.V., CHIZHONKOV E. V. Optimization of one three-parameter method of solving an algebraic system of the Stokes type // *Russ. J. Numer. Anal. Math. Modelling*, 1999, vol. 14, № 5, p. 429–440.
- [130] BYCHENKOV YU.V., CHIZHONKOV E. V. On optimization of one symmetric algorithm for saddle point problem // *International Conference on Computational Mathematics* / Ed.: G. A. Mikhailov, V. P. Il'in, Y. M. Laevsky. Proceedings of ICCM-2002. — Novosibirsk: ICM and MG Publisher, 2002, p. 97–103.
- [131] CAHOUE J., CHABART J. P. Some fast 3D finite element solvers for the generalized Stokes problem // *Internat. J. Numer. Methods Fluids*, 1988, vol. 8, p. 869–895.
- [132] CHAN T., SMITH B. Domain decomposition and multigrid algorithms for elliptic problems on unstructured meshes // *Domain decomposition methods in scientific and engineering computing. Contemporary Mathematics*, 1994, vol. 180, p. 175–189.
- [133] CHIZHONKOV E. V. On Methods for Solving the Stokes Problem for a Weakly-Compressible and Incompressible Fluids // *Advanced Mathematics: Computations and Applications* / Ed.: A. S. Alekseev, N. S. Bakhvalov. Proceedings of AMCA-95. — Novosibirsk: NCC Publisher, 1995, p. 167–171.
- [134] CHIZHONKOV E. V. On the Constant in the LBB condition for ring domains. Report № 9537 (October 1995). — Dept. of Math., Univ. of Nijmegen. The Netherlands, 1995.
- [135] CHIZHONKOV E. V. The limit convergence factor of the preconditioned Arrow-Hurwicz algorithm for saddle point problems // *Book of abstracts of FEM3D (International Conference: Finite Element Methods for Three-dimensional Problems)*. — Jyväskylä, Finland, 2000, p. 17–18.
- [136] CHIZHONKOV E. V., KARGIN A. V. On solution of the Stokes problem by the iteration of boundary conditions // *Rus. J. Numer. Anal. Math. Modelling*, 2006, vol. 21, № 1, p. 21–38.
- [137] CHIZHONKOV E. V., LEBEDEV V. I. On acceleration of the convergence of one iterative method // *Russ. J. Numer. Anal. Math. Modelling*, 2000, vol. 15, № 5, p. 383–395.
- [138] CHIZHONKOV E. V., OL'SHANSKII M. A. On the domain geometry dependence of the LBB condition // *Mathem. Modelling and Numer. Anal.*, 2000, vol. 34, № 5, p. 935–951.
- [139] CLARKE F. H. Optimization and Nonsmooth Analysis. — New York: Wiley, 1983.

-
- [140] COSSERAT E., COSSERAT F. Sur les équations de la théorie de l'élasticité // *C.R. Acad. Sci. (Paris)*, 1898, vol. 126, p. 1089–1091.
- [141] COSSERAT E., COSSERAT F. Sur la déformation infiniment petite d'un ellipsoïde élastique // *C.R. Acad. Sci. (Paris)*, 1898, vol. 127, p. 315–318.
- [142] COSSERAT E., COSSERAT F. Sur la déformation infiniment petite d'une enveloppe sphérique élastique // *C.R. Acad. Sci. (Paris)*, 1901, vol. 133, p. 326–329.
- [143] COSTABEL M., DAUGE M. On the Cosserat spectrum in polygons and polyhedra. Slides EPFL (<http://www.maths.univ-rennes.fr/~dauge>). — Lausanne, Switzerland, 2000.
- [144] CROUZEIX M. Etude d'une methode de linearisation. Resolution des equations de Stokes stationaries. Application aux equations des Navier–Stokes stationares // *Cahiere de l'IRIA*, 1974, № 12, p. 139–244.
- [145] CROUZEIX M. On an operator related to the convergence of Uzawa's algorithm for the Stokes equation // *Computational Science for 21st Century*. M. O. Bristeau, G. Etgen, W. Fitzgibbon, J. L. Lions, J. Periaux and M. F. Wheeler, editors. Chichester: Wiley, 1997, p. 242–249.
- [146] CROUZEIX M., RAVIART P. A. Conforming and non-conforming finite element methods for solving the stationary Stokes equations // *R. A. I. R. O.*, 1973, № R-3, p. 77–104.
- [147] DAFERMOS M. Some remarks on Korn's inequality // *Z. Angew. Math. Phys.*, 1968, vol. 19, p. 913–920.
- [148] DAUGE M. Stationary Stokes and Navier-Stokes Systems on Two- or Three-Dimensional Domains with Corners. Part I: Linearized Equations // *SIAM J. Math. Anal.*, 1989, vol. 20, № 1, p. 74–97.
- [149] DOBROWOLSKI M. On the LBB-constant on stretched domains // *Math. Nachr.*, 2003, vol. 254–255, p. 64–67.
- [150] DOBROWOLSKI M. On the LBB condition in the numerical analysis of the Stokes equations // *Appl. Numer. Math.*, 2005, vol. 54, № 3–4, p. 314–323.
- [151] D'YAKONOV E. G. Optimization in Solving Elliptic Problems. — Boca Raton: CRC Press, 1996.
- [152] ELMAN H. C. Multigrid and Krylov Subspace Methods for the Discrete Stokes Equations // *J. Numer. Methods Fluids*, 1996, vol. 22, p. 755–770.
- [153] ELMAN H. C., GOLUB G. H. Inexact and preconditioned Uzawa algorithms for saddle point problems // *SIAM J. Numer. Anal.*, 1994, vol. 31, № 6, p. 1645–1661.

- [154] ERTURK E., CORKE T. C., GOKCOL C. Numerical Solutions of 2-D Steady Incompressible Driven Cavity Flow at High Reynolds Numbers // *Int. J. Numer. Meth. Fluids*, 2005, vol. 48, p. 747–774.
- [155] FRIEDRICHS K. O. On certain inequalities and characteristic value problem for analytic functions and for functions of two variables // *Trans. Amer. Math. Soc.*, 1937, vol. 41, p. 321–364.
- [156] GIRAULT V., RAVIART P. A. Finite element methods for Navier–Stokes equations. — Berlin: Springer, 1986.
- [157] GOUBET O. Study of the Uzawa operator for the channel flow problem // *Applicable Analysis*, 1994, vol. 55, p. 235–258.
- [158] GUNZBURGER M. Finite element methods for viscous incompressible flow: a guide to theory, practice and algorithms. — Boston: Academic Press, 1989.
- [159] HACKBUSCH W. Multi-Grid Methods and Applications. — Berlin: Springer, 1985.
- [160] HENRICI P. Applied and computational complex analysis. Vol. 1. — New York: John Wiley & Sons Inc., 1988.
- [161] HESTENES M. R., STIEFEL E. Methods of conjugate gradients for solving linear systems // *J. Res. Nat. Bur. Standarts Sect. B*, 1952, vol. 49, p. 409–436.
- [162] HORGAN C. O. Inequalities of Korn and Friedrichs in elasticity and potential theory // *Z. Angew. Math. Phys.*, 1975, vol. 26, p. 155–164.
- [163] HORGAN C. O. Korn's inequalities and their applications in continuum mechanics // *SIAM Review*, 1995, vol. 37, № 4, p. 491–511.
- [164] HORGAN C. O., PAYNE L. E. On inequalities of Korn, Friedrichs and Babuška–Aziz // *Arch. Ration. Mech. Anal.*, 1983, vol. 82, p. 165–179.
- [165] ILIASH JU., ROSSI T., TORVANEN J. Two iterative methods for solving the Stokes problem. Tech. Report 2. — University of Jyväskylä, Department of Mathematics, Laboratory of Scientific Computing, 1993.
- [166] KELLER C., GOULD N. I. M., WATHEN A. J. Constraint preconditioning for indefinite linear systems // *SIAM J. Matrix Anal. Appl.*, 2000, vol. 21, p. 1300–1317.
- [167] KEßLER M. Die Ladyzhenskaya-Konstante in der numerischen Behandlung von Strömungsproblemen. Dissertation, Institut für Angewandte Mathematik und Statistik der Universität Würzburg. — Würzburg, 2000.

-
- [168] KLAOWONN A. Block-triangular preconditioners for saddle point problems with penalty term // *SIAM J. Sci. Comput.*, 1998, vol. 19, № 1, p. 172–184.
- [169] KLAOWONN A. An optimal preconditioner for a class of saddle point problems with a penalty term // *SIAM J. Sci. Comput.*, 1998, vol. 19, № 2, p. 540–552.
- [170] KOBEL'KOV G. M. Fictitious domain method and the solution of elliptic equations with highly varying coefficients // *Russ. J. Numer. Anal. Math. Modelling*, 1987, vol. 2, № 6, p. 407–488.
- [171] KOBEL'KOV G. M., OL'SHANSKII M. A. Effective preconditioning of Uzawa type schemes for a generalized Stokes problem // *Numer. Math.*, 2000, vol. 86, p. 443–470.
- [172] KOZHEVNIKOV A. The basic boundary value problems of static elasticity theory and their Cosserat spectrum // *Mathem. Zeitschrift*, 1993, vol. 213, p. 241–274.
- [173] KUZNETSOV YU. A. New iterative methods for singular perturbed positive definite matrices // *Russ. J. Numer. Anal. Math. Modelling*, 2000, vol. 15, № 1, p. 65–71.
- [174] LANGER U., QUECK W. On the convergence factor of Uzawa's algorithm // *J. Comp. Appl. Math.*, 1986, vol. 15, p. 191–202.
- [175] NITSCHKE J. A. On Korn's second inequality // *R. A. I. R. O. Numer. Anal.*, 1981, vol. 15, p. 237–248.
- [176] OLEINIK O. A. Korn's type inequalities and applications to elasticity. In: *Convegno Internazionale in memoria di Vito Volterra*. — Seiten, 1991.
- [177] OL'SHANSKII M. A. On numerical solution of nonstationary Stokes equations // *Russ. J. Numer. Anal. Math. Modelling*, 1995, vol. 10, № 1, p. 81–92.
- [178] OL'SHANSKII M. A., REUSKEN A. Grad-Div stabilization for the Stokes equations // *Math. Comp.*, 2004, vol. 73, № 248, p. 1699–1718.
- [179] OL'SHANSKII M. A., REUSKEN A. Analysis of a Stokes interface problem // *Numer. Math.*, 2006, vol. 103, p. 129–149.
- [180] QUECK W. The convergence factor of preconditioned algorithms of the Arrow-Hurwicz type // *SIAM J. Numer. Anal.*, 1989, vol. 26, № 4, p. 1016–1030.
- [181] RANNACHER R., TUREK S. A simple nonconforming quadrilateral Stokes element // *Numer. Meth. Part. Diff. Eq.*, 1992, № 8, p. 97–111.
- [182] REPIN S. A posteriori estimates for the Stokes problem // *J. Math. Sci. (New York)*, 2002, vol. 109, № 5, p. 1950–1964.

- [183] SAAD Y. Iterative methods for sparse linear systems. — Boston: WPS, 1996.
- [184] SAAD Y., SCHULTZ M. H. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems // *SIAM J. Sci. Comput.*, 1986, vol. 7, p. 856–869.
- [185] SARIN V., SAMEN A. An efficient iterative method for the generalized Stokes problem // *SIAM J. Sci. Comput.*, 1998, vol. 19, № 1, p. 206–226.
- [186] SHELDON J. On the numerical solution of elliptic difference equations // *Math. Tables Aids Comput.*, 1955, vol. 9, p. 101–112.
- [187] SILVESTER D., WATHEN A. Fast iterative solution of stabilized Stokes systems. Part II: using general block preconditioners // *SIAM J. Numer. Anal.*, 1994, vol. 31, № 5, p. 1352–1367.
- [188] STOYAN G. Iterative Stokes solvers in the harmonic Velte subspace // *Computing*, 2001, vol. 67, p. 13–33.
- [189] TONG Z., SAMEH A. On an iterative method for saddle point problem // *Numer. Math. Electronic Edition*, 1998, vol. 79, p. 643–646.
- [190] TUREK S. Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approach. — Springer-Verlag, 1999.
- [191] VALEDINSKY V. D., CHIZHONKOV E. V. Structure of Solution to Stokes Problem and Efficient Numerical Method // *Sov. J. Numer. Anal. Math. Modelling*, 1990, vol. 5, № 4/5, p. 419–423.
- [192] VAN DER VORST H. A. BI-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems // *SIAM J. Sci. Comput.*, 1992, vol. 13, p. 631–644.
- [193] VANKA S. P. Block-implicit multigrid solution of Navier–Stokes equations in primitive variables // *J. Comput. Physics*, 1986, vol. 65, p. 138–158.
- [194] VASSILEVSKI P. S., LAZAROV R. D. Preconditioning Mixed Finite Element Saddle-Point Elliptic Problems // *Numer. Lin. Alg. with Appl.*, 1996, vol. 3, p. 1–20.
- [195] VELTE W. On inequalities of Friedrichs and Babuška–Aziz in dimension three // *Z. Anal. Anwend.*, 1998, vol. 17, № 4, p. 843–857.
- [196] VERFURTH R. A combined conjugate gradient–multigrid algorithm for the numerical solution of the Stokes problem // *IMA J. Numer. Anal.*, 1984, vol. 4, p. 441–455.
- [197] VERFURTH R. A multilevel algorithm for the mixed problem // *SIAM J. Numer. Anal.*, 1984, vol. 21, p. 264–271.

-
- [198] WATHEN A., SILVESTER D. Fast iterative solution of stabilized Stokes systems. Part I: Using simple diagonal preconditioners // *SIAM J. Numer. Anal.*, 1993, vol. 30, № 3, p. 630–649.
 - [199] WITTUM G. Multi-grid methods for the Stokes and Navier–Stokes equations // *Numer. Math.*, 1989, vol. 54, p. 543–564.
 - [200] YOUNG D. M. Iterative Solution of Large Linear Systems. — New York: Academic Press, 1971.
 - [201] ZHONG-ZHI BAI, GOLUB G. H. Generalized Preconditioned Hermitian and Skew-Hermitian Splitting Iteration Methods for Saddle-Point Problems // *Tech. Rep. TR-2005-007-A, Department of Mathematics and Computer Science, Emory University, Atlanta, GA*, SCCM Reports, № 7.
 - [202] ZULEHNER W. Analysis of iterative methods for saddle point problems: a unified approach // *Math. Comp.*, 2002, vol. 71, № 238, p. 479–505.

ОГЛАВЛЕНИЕ

Предисловие	3
Часть I. Релаксационные методы	7
Глава 1. Вводные сведения	8
1.1. Основные обозначения и постановка задачи	8
1.2. Метод Узавы — сопряженных градиентов	10
1.3. Вспомогательные утверждения	13
1.3.1. Две задачи на собственные значения	14
1.3.2. Базис специального вида из собственных векторов	15
1.3.3. Полезное начальное приближение	17
1.4. Библиография и комментарии	19
Глава 2. Модифицированные методы релаксации.	
Общий анализ	22
2.1. Сведения о методах релаксации	22
2.1.1. Общие понятия	22
2.1.2. Метод Якоби	23
2.1.3. Метод SOR	25
2.1.4. Метод SSOR	26
2.2. Модифицированный метод Якоби (MJOR)	28
2.2.1. Построение метода	28
2.2.2. Спектр оператора перехода	29
2.2.3. Условие сходимости	31
2.2.4. Задача асимптотической оптимизации	33
2.2.5. Оптимизация в подпространстве	38
2.3. Модифицированный метод SOR (MSOR)	40
2.3.1. Спектр оператора перехода	40
2.3.2. Условие сходимости	41
2.3.3. Задача асимптотической оптимизации	43
2.4. Модифицированный метод SSOR (MSSOR)	45
2.4.1. Спектр оператора перехода	46

2.4.2. Условие сходимости	49
2.4.3. Задача асимптотической оптимизации	49
2.5. Трехпараметрический метод (3MSOR)	50
2.5.1. Спектр оператора перехода	50
2.5.2. Задача асимптотической оптимизации	51
2.5.3. Частный случай: (β, τ) -метод	52
2.5.4. Задача асимптотической оптимизации (β, τ) — метода	54
2.6. Библиография и комментарии	55
Глава 3. Оценки погрешности методов MJOR и MSOR	58
3.1. Оценки из общей теории	59
3.1.1. Оптимальный одношаговый метод	59
3.1.2. Циклический метод с чебышевскими парамет- рами	59
3.1.3. Полуитерационный метод Чебышева	60
3.1.4. Стационарный трехслойный метод	61
3.1.5. Методы сопряженных направлений	61
3.2. Погрешность метода MJOR в случае постоянных параметров	63
3.2.1. Преобразование формул	63
3.2.2. Начальное приближение	64
3.2.3. Оценка погрешности с постоянными парамет- рами	65
3.3. Погрешность метода MJOR в случае переменных параметров	66
3.4. Погрешность метода MSOR в случае постоянных параметров	71
3.4.1. Преобразование формул	71
3.4.2. Начальное приближение	73
3.4.3. Полином ошибки	73
3.4.4. Оценка погрешности	74
3.5. Погрешность метода MSOR в случае переменных параметров	76
3.5.1. Преобразование формул	76
3.5.2. Выбор параметров для p , как в циклическом методе	78
3.5.3. Выбор параметров для p , как в трехслойных методах	78
3.5.4. Выбор параметров для u , как в трехслойных методах	80
3.6. Библиография и комментарии	84

Глава 4. Релаксационные методы для системы с параметром	86
4.1. Явный метод типа MSOR (MSORe)	86
4.1.1. Построение метода	86
4.1.2. Спектр оператора перехода	87
4.1.3. Условие сходимости	88
4.1.4. Задача асимптотической оптимизации	90
4.2. Неявный метод типа MSOR (IMSORe)	91
4.2.1. Построение метода	91
4.2.2. Спектр оператора перехода	91
4.2.3. Условие сходимости	93
4.2.4. Задача асимптотической оптимизации	95
4.3. Погрешность метода MSORe в случае постоянных параметров	96
4.3.1. Преобразование формул	96
4.3.2. Начальное приближение	98
4.3.3. Полином ошибки	98
4.3.4. Оценка погрешности	99
4.4. Погрешность метода MSORe в случае переменных параметров	100
4.4.1. Преобразование формул	100
4.4.2. Выбор параметров для p , как в циклическом методе	102
4.4.3. Выбор параметров для p , как в трехслойных методах	103
4.4.4. Выбор параметров для u , как в трехслойных методах	105
4.5. Библиография и комментарии	108
Глава 5. Методы для нормальных уравнений	109
5.1. Оптимизация метода для базовой системы	110
5.1.1. Спектр оператора равносильной задачи	110
5.1.2. Минимизация числа обусловленности	111
5.1.3. Наилучшая оценка погрешности	112
5.2. Оптимизация метода для системы с параметром	113
5.2.1. Спектр оператора равносильной задачи	113
5.2.2. Минимизация числа обусловленности	115
5.2.3. Наилучшая оценка погрешности	118
5.3. Оптимизация метода для случая равносильной системы	118
5.3.1. Спектр оператора равносильной задачи	119
5.3.2. Минимизация числа обусловленности	121

5.3.3. Наилучшая оценка погрешности	124
5.4. Библиография и комментарии	124
Часть II. Обобщенные методы	125
Глава 6. Предварительные результаты	126
6.1. Классы оптимизации	126
6.2. Что происходит с алгоритмом Узавы?	128
6.3. Нерегулярные задачи с седловой точкой.	131
6.4. Вспомогательные утверждения	134
6.4.1. Сведения из анализа	134
6.4.2. Спектральные свойства одного пучка операторов	140
6.4.3. Свойства некоторых классов функций	149
6.5. Ключевые этапы построения и анализа алгоритмов .	156
6.6. Библиография и комментарии.	161
Глава 7. Блочнo треугольное предобусловливание (GMSOR)	163
7.1. Формулировка метода и его свойства.	163
7.2. Симметричная регулярная задача: оптимизация в классе K_1	165
7.3. Симметричная регулярная задача: оптимизация в классе K_2	171
7.4. Симметричная нерегулярная задача: оценка в классе K_{2s}	178
7.5. Несимметричная регулярная задача: оценка в классе K_3	180
7.6. Библиография и комментарии.	184
Глава 8. Блочнo диагональное предобусловливание	187
8.1. Обобщенный метод MJOR (GMJOR)	187
8.1.1. Формулировка метода	187
8.1.2. Связь спектров операторов перехода GMJOR и GMSOR	188
8.1.3. Оптимизация метода в классах K_1, K_2, K_3, K_{2s}	189
8.1.4. Случай $\beta = 0$	190
8.2. Обобщенный метод Ланцоша (GMLan)	194
8.2.1. Построение метода	194
8.2.2. Оптимизация в классе K_1	196
8.2.3. Оптимизация в классе K_2	202
8.2.4. Оценка в классе K_{2s}	210
8.3. Библиография и комментарии.	211

Глава 9. Симметризация специального вида	214
9.1. Обобщенный метод Bramble–Pasciak (GMBP)	215
9.1.1. Построение метода	215
9.1.2. Оптимизация метода в классе K_1	218
9.1.3. Оптимизация метода в классе K_2	221
9.1.4. Оценка в классе K_{2s}	224
9.2. Библиография и комментарии	225
Глава 10. Модельные седловые операторы	228
10.1. Неконструктивный подход	228
10.1.1. Построение методов	228
10.1.2. Оценка в классе K_1	230
10.1.3. Оценка в классе K_2	232
10.1.4. Оценка в классе K_{2s}	235
10.2. Конструктивное предобусловливание	236
10.2.1. Построение методов	236
10.2.2. Оценка в классе K_2	236
10.2.3. Оценка в классе K_{2s}	241
10.3. Библиография и комментарии	242
Глава 11. Методы попеременных симметричных и кососимметричных итераций	244
11.1. Стационарный метод (GPHSSI)	245
11.1.1. Формулировка метода	245
11.1.2. Безусловная сходимость метода	245
11.1.3. Оптимизация в классе K_2	247
11.2. HSS-предобусловливание	252
11.2.1. Построение методов	252
11.2.2. Оценка спектра в классе K_2	254
11.2.3. Анализ сходимости методов чебышевского типа и GMRES	256
11.3. Библиография и комментарии	263
Глава 12. Нелинейные задачи и блочно треугольное предобусловливание	265
12.1. Уравнения с кососимметричным возмущением	265
12.1.1. Постановка задачи	265
12.1.2. Оценка скорости сходимости	268
12.2. Уравнения с сильно монотонным оператором	277
12.2.1. Постановка задачи	277
12.2.2. Вспомогательные факты и утверждения	279
12.2.3. Оценка скорости сходимости	281
12.3. Библиография и комментарии	284

Часть III. Приложение к гидродинамике	286
Глава 13. Inf-sup неравенство и смежные вопросы	289
13.1. О задаче Стокса и спектре Коссера	290
13.2. Неравенства Фридрихса и Корна в двухмерном случае	292
13.3. Точные значения и оценки снизу константы Ладыженской	293
13.4. Анизотропия области	297
13.5. Угловые точки на границе	299
13.6. Разное	303
13.6.1. Обобщенная задача Стокса	303
13.6.2. Уравнения Ламе в теории упругости и слабо-сжимаемая жидкость	305
13.6.3. Смешанный подход для эллиптических уравнений	305
13.6.4. Другие применения	307
Глава 14. Численный анализ LBB-условия	308
14.1. Задача с гладким решением	308
14.2. Задача с негладким решением	313
14.2.1. Схема 1	313
14.2.2. Схема 2	314
14.2.3. Вычислительные аспекты	315
14.2.4. Расчеты для квадратной области	316
14.2.5. Расчеты для прямоугольной области	319
Глава 15. Численный анализ роли оператора $\beta BC^{-1}B^T$ в сходимости GMSOR	322
15.1. Алгоритм численной оптимизации	323
15.2. Задача о квадратной каверне	324
15.3. Обтекание тела в трубе	326
Список литературы	329

ИМЕЕТСЯ В ПРОДАЖЕ:



Галкин В. А. Анализ математических моделей: системы законов сохранения, уравнения Больцмана и Смолуховского / В. А. Галкин. — 2009. — 408 с. : ил. — (Математическое моделирование).

Монография посвящена вопросам обоснования корректности задач для систем нелинейных уравнений, имеющих прикладное значение в математической физике. Содержание книги направлено на выявление и анализ основных математических структур, связанных с вопросами обоснования методов математического моделирования, приводящих к нелинейным системам законов сохранения, включающих в себя систему Навье—Стокса газовой динамики, уравнения Больцмана, Смолуховского, Власова в физической кинетике. Сюда же примыкают задача Стефана и модели тепломассопереноса, связанные с выращиванием кристаллов.

Для специалистов в области прикладной математики, физической кинетики и газовой динамики, а также для студентов и аспирантов соответствующих специальностей.



ИЗДАТЕЛЬСТВО

«БИНОМ
Лаборатория знаний»

125167, Москва, проезд Аэропорта, д. 3

Телефон: (499) 157-5272

e-mail: binom@lbz.ru, <http://www.lbz.ru>

Оптовые поставки:

(499) 174-7616, 171-1954, 170-6674

НАУЧНАЯ И УЧЕБНАЯ ЛИТЕРАТУРА

■ ВЫСШАЯ МАТЕМАТИКА

ИМЕЕТСЯ В ПРОДАЖЕ:



Колесниченко А. В. Турбулентность и самоорганизация. Проблемы моделирования космических и природных сред / А. В. Колесниченко, М. Я. Маров. — 2009. — 632 с. : ил., [16] с. цв. вкл. — (Математическое моделирование).

Монография посвящена разработке континуальных моделей турбулизированных природных сред — моделей, лежащих в основе постановок и численных расчетов задач, связанных с образованием, структурой и эволюцией различных астро- и геофизических объектов. Стохастические модельные подходы к соответствующим задачам рассмотрены как отражение процессов самоорганизации в диссипативных открытых системах. Приведены примеры возникновения упорядоченностей в различных космических объектах и природных средах в процессе их эволюции.

Для научных сотрудников, работающих в областях астрофизики, геофизики, планетологии, аэрономии и космических исследований, а также для студентов старших курсов и аспирантов соответствующих специальностей.



ИЗДАТЕЛЬСТВО
«БИНОМ
Лаборатория знаний»

125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272
e-mail: binom@lbz.ru, <http://www.lbz.ru>
Оптовые поставки:
(499) 174-7616, 171-1954, 170-6674

НАУЧНАЯ И УЧЕБНАЯ ЛИТЕРАТУРА

■ ВЫСШАЯ МАТЕМАТИКА

ИМЕЕТСЯ В ПРОДАЖЕ:



Брушлинский К. В. Математические и вычислительные задачи магнитной газодинамики / К. В. Брушлинский. — 2009. — 200 с. : ил. — (Математическое моделирование).

Монография относится к актуальной области математического моделирования в современных задачах физики плотной плазмы. Изложены математические вопросы магнитной газодинамики, представлены численные модели соответствующих физических процессов. При исследовании двумерных МГД-течений специальное внимание уделено роли и моделированию эффекта Холла. Обсуждаются особенности численного решения МГД-задач. Приведены примеры расчетов магнитных ловушек для удержания плазмы и дан подробный обзор моделей ускорения плазмы магнитным полем в каналах.

Для научных работников, аспирантов и студентов старших курсов, интересующихся МГД-моделированием плазмы, в том числе начинающих работать в этой области и не имеющих узкоспециальной подготовки.



ИЗДАТЕЛЬСТВО
«БИНОМ»
Лаборатория знаний»

125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272
e-mail: binom@lbz.ru, <http://www.lbz.ru>
Оптовые поставки:
(499) 174-7616, 171-1954, 170-6674



Быченков Юрий Владимирович – кандидат физико-математических наук, сотрудник кафедры вычислительной математики механико-математического факультета МГУ им. М.В. Ломоносова. Основные научные интересы: конструирование и оптимизация итерационных методов решения задач с седловой точкой.



Чижонков Евгений Владимирович – доктор физико-математических наук, профессор кафедры вычислительной математики механико-математического факультета МГУ им. М.В. Ломоносова. Основные научные интересы: конструирование и оптимизация итерационных методов решения задач с седловой точкой, математическое моделирование

в гидродинамике и физике лазерно-плазменных взаимодействий, численная стабилизация неустойчивых решений нелинейных уравнений математической физики. Автор более 100 научных работ.

Монография относится к актуальной области численного анализа – решению больших систем линейных уравнений, имеющих в качестве решения седловую точку. Впервые в одной книге рассматриваются идеи построения и подробно анализируются вопросы сходимости и оптимизации всех известных в настоящее время итерационных методов для таких задач. Результаты представлены в виде удобных для использования формул и их асимптотических характеристик. Приложение ориентировано на численное моделирование в гидродинамике и смежных областях.

Для научных работников, аспирантов и студентов старших курсов, интересующихся вычислительной математикой, а также инженеров и исследователей, которые используют в своей деятельности итерационные методы.

ISBN 978-5-9963-0118-8

