ОГЛАВЛЕНИЕ

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 предобусловливанием), 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 удается 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-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

Глава 1

ВВОДНЫЕ СВЕДЕНИЯ

В главе приводятся основные обозначения и постановка задачи, рассматривается базовый для решения блочных седловых задач метод Узавы 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} $(1 \leqslant i, j \leqslant 2)$:

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

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} невырожденна, 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:

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

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 определение 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 $\varepsilon \geqslant 0$ наиболее распространенного вида:

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

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 невырожденна при любом $\varepsilon \geqslant 0$. 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. $(S_0 p, p) \geqslant 0$ для произвольного $p \in P$. Поэтому при $\varepsilon > 0$ матрица S_ε (и следовательно, L_ε) невырожденна 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 $N_u \geqslant N_p$. 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

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

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

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

Отсюда имеем, что если матрица L_0 невырожденна, 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 $k = 1, 2, \ldots$ формулы имеют вид:

$$r^k = r^{k-1} - a_k S_0 s^{k-1}, \qquad w^k = C^{-1} r^k,$$
$$d_k = \frac{\left(w^k, r^k\right)}{\left(w^{k-1}, r^{k-1}\right)}, \qquad s^k = w^k + d_k s^{k-1},$$
$$a_{k+1} = \frac{\left(w^k, r^k\right)}{\left(S_0 s^k, s^k\right)}, \qquad p^{k+1} = p^k - a_{k+1} s^k.$$

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)

$$\gamma C \leqslant S_0 \leqslant \Gamma C, \quad 0 < \gamma, \tag{1.5}$$

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

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

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)

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

При фиксированных γ и Γ величина

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

является неулучшаемой, 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 C 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).