Российская академия наук
Институт вычислительной математики

На правах рукописи

Оселедец Иван Валерьевич

УДК 519.6

Нелинейные аппроксимации матриц

01.01.07 — Вычислительная математика

ДИССЕРТАЦИЯ
На соискание учёной степени
кандидата физико-математических наук

Научный руководитель
чл.-корр. РАН, проф. Тыртышников Е. Е.

Москва 2007

Содержание

Введение

i.1. Нелинейные аппроксимации матриц: зачем и как

К решению линейных систем уравнений — основной задаче линейной алгебры и матричного анализа — сводится подавляющее большинство практических вычислительных задач. Однако, несмотря на наличие универсальных методов, многие приложения приводят к «большим» системам, которые не могу быть решены даже на современных суперкомпьютерах.

В данной диссертации развиваются эффективные методы вычислений с плотными матрицами, в которых сами матрицы и результаты матричных операций аппроксимируются матрицами специальной структуры, определённой относительно малым числом параметров. Зависимость от параметров носит нелинейный характер, поэтому речь идёт о методах *нелинейной матричной аппроксимации*.

Плотные матрицы описываются N^2 параметрами. Если мы хотим ускорить работу с ними, то необходимо построить «сжатое» представление матрицы с помощью меньшего числа параметров. Матрицы, описываемые малым числом параметров, будем называть *структурированными*.

Часто структура матрицы видна сразу или следует из физических свойств задачи. Например, в задачах с оператором, инвариантным относительно сдвига, получающиеся матрицы имеют *тёплицеву* (или блочно-тёплицеву в многомерных задачах) структуру, т.е. элемент матрицы зависит лишь от разности индексов: $a_{ij} = b_{i-j}$. Для тёплицевых матриц существуют быстрые алгоритмы, основанные на БПФ. Тёплицевы матрицы — классический пример матриц с *линейной структурой*. Можно привести другие примеры: ганкелевы матрицы ($a_{ij} = b_{i+j}$), ленточные матрицы, разреженные матрицы.

Ещё один важнейший класс матриц — *матрицы малого ранга*, т.е. матрицы вида

$$A = UV^\top,$$

$U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{m\times r}$, где ранг $r \ll m, n$. Это — пример матрицы с *нелинейной структурой*: её элементы зависят от параметров (элементов матриц U и V) нелинейно.

Таким образом, эффективные алгоритмы могут быть основаны на *нелинейных малопараметрических аппроксимациях матриц*. Однако далеко не всегда очевидно, как получить эффективное малопараметрическое представление матрицы. Более того, чтобы быстро работать с такими структурами, мы должны уметь выполнять матричные

операции (сложение, умножение, обращение) именно в терминах малопараметрического представления. В общем случае возможность сохранения структуры при операциях зависит от выбранного типа структуры. Например матрица, обратная к тёплицевой матрице, уже не будет тёплицевой. В то же время тёплицевы матрицы можно вложить в более широкий класс матриц *малого ранга смещения*, который уже замкнут относительно операции обращения. К сожалению, даже этот класс не замкнут относительно операции умножения. Поэтому выполнение матричных операций с сохранением малопараметического формата может быть только приближённым.

Тёплицевы матрицы соответствуют одномерным интегральным уравнениям, где использование сеток большой размерности не является необходимым. На практике значительно более интересным представляется решение многомерных уравнений. Для ядер, инвариантных относительно сдвига и дискретизации на равномерной сетке, получаются *многоуровневые тёплицевы матрицы*. Такие матрицы тоже можно умножать на вектор за квазилинейное время, однако до сих пор универсальных прямых методов решения таких систем за то же время неизвестно. Существующие формулы (формулы Гохберга-Хайнига) содержат не $\mathcal{O}(N)$ параметров, а $\mathcal{O}(N^{3/2})$, и, видимо, удобных формул с меньшим числом параметров не существует. Что же делать? Ответ прост. Вместо *точных формул* мы предлагаем использовать некоторые *приближённые формулы*. Из каких соображений можно исходить при получении приближённых формул? По существу, изучению этого вопроса (или, точнее, методов поиска ответа на данный вопрос) и посвящена данная диссертация.

i.2. Основные результаты работы

При решении любой задачи всегда хочется получить сначала некоторый общий подход. В данной диссертации предложены два таких общих подхода для двух классических задач *матричного анализа* — обращение матриц и построение *предобуславливателей*. Напомним, о чём идёт речь. Пусть нам нужно решить линейную систему

$$Ax = b,$$

с помощью какого-нибудь стандартного итерационного метода. Однако часто бывает так, что требуется большое количество итераций для сходимости, поэтому решают эквивалентную систему вида

$$AP^{-1}x = b,$$

где матрица P *легко обратима*, (т.е. $P^{-1}f$ можно вычислить достаточно быстро). Матрица P называется *предобуславливателем*. Как проверить, что матрица P *хорошая*? Обычно хотят добиться того, чтобы матрица AP^{-1} была хорошо обусловлена. Однако задача оптимизации числа обусловленности является довольно сложной, и поэтому её заменяют гораздо более простой задачей аппроксимации вида

$$\|A - P\| \to \min, \tag{1}$$

где P принадлежит некоторому классу быстрообратимых матриц, а $\|\ \|$ — некоторая (обычно фробениусова) матричная норма. Такой подход, например, применяется для построения циркулянтных предобуславливателей для тёплицевых матриц[1]. Однако, как известно из теории итерационных методов, хорошая обусловленность достаточна, но отнюдь не необходима для быстрой сходимости итерационных методов. Большую роль играет наличие *кластеров* собственных значений предобусловленной системы около 1. Это означает, что подавляющее большинство собственных значений, за исключением может быть конечного числа, находится в окрестности 1. А как проверить наличие кластеров? Оказывается, что все теоремы о существовании кластеров явно или неявно опираются на представление матрицы A в виде

$$A = P + R + E, \tag{2}$$

где P — предобуславливатель, R — матрица малого ранга и E — матрица малой нормы, т.е. выполняется приближённое равенство

$$A \approx P + R, \tag{3}$$

где R — поправка малого ранга. Наше предложение (и общий подход!) состоит в том, чтобы использовать (2) , а не (1) в качестве отправной точки для построения предобуславливателя P. Если мы зафиксируем класс предобуславливателей (например, циркулянтные матрицы), то мы получим некоторую задачу *нелинейной матричной аппроксимации*, в которой необходимо находить и P, и R. Обычно находят P, а R оценивают; однако, как увидим позднее, гораздо более эффективно находить P и R одновременно. Выбирая различные классы матриц P, мы получаем целую россыпь новых задач нелинейной матричной аппроксимации, для которых можно пытаться придумать эффективные алгоритмы решения. В данной диссертации рассмотрены следующие

[1] Такие предобуславливатели называются предобуславливателями Т. Чэна (T. Chan)

классы матриц, в которых ищутся предобуславливатели: циркулянтные матрицы, разреженные матрицы с известным шаблоном, матрицы вида

$$TST^{\top},$$

где T — матрица какого-либо *быстрого преобразования* (Фурье, синус-преобразование, косинус-преобразование, вейвлет-преобразование), а S — разреженная матрица. Идея построения основана на *методе чёрных точек*, который является обобщением *крестового метода* [?, 40, 13] для приближения матрицами малого ранга. Фактически, с помощью метода *чёрных точек* можно для заданной матрицы A построить за число операций порядка $\mathcal{O}(N)$ приближение (если оно, конечно, существует) вида

$$A \approx S + R, \tag{4}$$

где S — разреженная матрица с известным шаблоном, R — матрица малого ранга. Область применимости метода не ограничивается только предобуславливанием. Нетрудно увидеть в (4) задачу приближения матрицы A матрицей малого ранга *за исключением* малого числа элементов. Такая задача возникает при *заполнении пропусков в больших массивах данных*, например данных наблюдений или данных, связанных с исследованием ДНК. Также в диссертации предлагается обобщение метода на случай, когда положение «испорченных» элементов неизвестно и требует определения.

Другая задача, подробно рассматриваемая в диссертации — задача приближённого обращения структурированных матриц. Пусть есть некоторый класс матриц, который допускает быстрое умножение. Предлагается использовать следующие *итерации Ньютона*:

$$X_{k+1} = 2X_k - 2X_kAX_k, k = 0, \cdots, \tag{5}$$

Нетрудно показать, что $X_k \to A^{-1}$ с квадратичной скоростью (для всех начальных приближений X_0, достаточно близких к A^{-1}. Для обычных матриц такой алгоритм, конечно, непрактичен — сложность его составляет $\mathcal{O}(N^3)$ операций на одну итерацию.

Однако для структурированных матриц, допускающих быстрое умножение, метод оказывается очень эффективен. При этом возникают дополнительные сложности, связанные с *восстановлением структуры* матрицы после каждой итерации. В итоге, итерационный процесс можно записать в виде

$$X_{k+1} = R(2X_k - 2X_kAX_k), k = 0, \cdots, \tag{6}$$

где R — некоторый (нелинейный) *оператор проектирования*. В диссертации показано, что при достаточно общих предположениях на оператор R, метод сохраняет квадратичную скорость сходимости. Оказывается, что для структурированных матриц можно построить *модификацию* метода Ньютона, которая оказывается гораздо более быстрой; в «точной» арифметике два итерационных процесса эквивалентны, а в приближённой модифицированный метод требует умножения матриц гораздо более простой структуры, что приводит к уменьшению вычислительных затрат.

Используя различные форматы данных, теперь можно получать различные методы. Для многоуровневых матриц удобным форматом оказываются матрицы *малого тензорного ранга*:

$$A \approx A_r = \sum_{k=1}^{r} U_k \otimes V_k, \tag{7}$$

где $U \otimes V$ — блочная матрица вида

$$[u_{ij}V].$$

Выгода от такого представления очевидна — вместо $N^2 = n^4$ параметров требуется всего лишь $rn^2 = N$ параметров, а r обычно порядка 10-20. Как же вычислять такое представление? Оказывается, что после простой перестановки индексов задача сводится к задаче аппроксимации переставленной матрицы матрицей малого ранга. Для решения этой задачи особенно эффективен метод неполной крестовой аппроксимации, который и применяется в дальнейшем. После этого, можно поставить вопрос о дальнейшем сжатии матриц-факторов U_i V_i. Это необходимо, так как умножение на вектор в тензорном формате всё ещё требует более чем линейное по N число операций $\mathcal{O}(N^{3/2})$. Это уже не очень приятно при $N = 10000$. Возможно несколько подходов. Один из наиболее успешных и простых в реализации — использование вейвлет-преобразований для каждого фактора, после чего каждый фактор становится уже разреженным:

$$\widehat{U}_i = WU_iW^\top$$

В качестве вейвлет-преобразований можно использовать, например, классические преобразования Добеши. Однако эти преобразования приспособлены к равномерным сеткам. Если же дискретизация уравнения происходит на неравномерных сетках, то возникает необходимость построения специальных преобразований, приспособленных

к неравномерным сеткам. В диссертации предложен быстрый алгоритм построения таких преобразований и построены явные формулы, которые требуют лишь знания расположения узлов неравномерной сетки. В главе 4 естественным образом объединяются результаты о тензорных аппроксимациях и итерационном обращении матриц. Общие теоремы и общий подход применяются для обращения *дважды тёплицевых матриц*. Эти матрицы описываются $\mathcal{O}(N)$ параметрами, однако неизвестно ни одного алгоритма (вида «чёрный ящик») для решения систем с такими матрицами с почти линейной сложностью. В данной диссертации предлагается метод, имеющий сложность $\mathcal{O}(\sqrt{N})$ для достаточного широкого подкласса дважды тёплицевых матриц(в частности тех, которые получаются при дискретизации интегральных уравнений). Для этого используются тензорные аппроксимации, причём факторы имеют дополнительную «внутреннюю» структуру малого ранга смещения. Быстрые арифметические операции в таком формате, необходимые «ингредиенты» для метода Ньютона, получаются почти автоматически (единственным нетривиальным местом является вычисление оператора проектирования).

Важно отметить, что для того, чтобы получить *теоретически обоснованный* алгоритм, необходимо доказать, что *обратная матрица* тоже представима в таком формате. В заключительной части диссертации получена оценка на тензорный ранг обратной матрицы. Однако оценок рангов смещения для факторов получить не удалось.

До данного момента были известны только тривиальные случаи, когда матрицы малого тензорного ранга имеют обратные тоже малого тензорного ранга, такие как

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

Для случая двух и большего числа слагаемых тензорный ранг обратной матрицы уже может быть практически произвольным. Однако всё же существуют классы матриц малого тензорного ранга с обратными матрицами тоже малого тензорного ранга.

При исследовании вопроса о структуре матриц, обратных к дважды тёплицевым был (сначала экспериментально) обнаружен интересный факт. Пусть матрица A имеет вид

$$A = I + D \otimes R + R \otimes D,$$

где R– матрица ранга 1, а D — диагональная матрица. Тогда тензорный ранг A^{-1} *не превышает* 5. Это утверждение удаётся обобщить

на матрицы вида

$$A = I + \sum_i D_i \otimes R_i + R_i \otimes D_i. \tag{8}$$

Доказательство является конструктивным и даёт способ представления обратной матрицы. То есть, обнаружен *класс матриц малого тензорного ранга, замкнутый относительно обращения.*

Как это связано с дважды тёплицевыми матрицами? Обращение такой матрицы мы начинаем с аппроксимации матрицей малого тензорного ранга с факторами вида $C+R$. Из вышеозначенных результатов вытекает, что соответствующая обратная матрица имеет относительно малый тензорный ранг. Таким образом, получен теоретически обоснованный алгоритм для обращения дважды тёплицевых матриц линейной сложности. Если бы у нас получилась оценка на ранг смещения для факторов, то мы получили бы полностью обоснованный алгоритм сублинейной сложности (по порядку зависимости от размера матрицы). Заметим, что такая сублинейная зависимость наблюдалась нами на модельных задачах.

i.3. Содержание работы по главам

Первая глава посвящена классической задаче теории структрированных матриц — построение циркулянтных предобуславливателей для тёплицевых (и не только!) матриц. И тёплицевы, и циркулянтные матрицы — матрицы линейной структуры и в многочисленных предыдующих работах использовались лишь методы на основе наилучших (в некоторой норме) линейных приближений циркулянтами заданной тёплицевой матрицы. Мы же сформулировали задачу, как задачу нелинейной аппроксимации заданной матрицы суммой циркулянта и матрицы малого ранга. Раздел 1.1 содержит краткое описание исходной задачи и состояния дел на данный момент. Там же сразу формулируется основная задача, которую мы будем решать — задача $C+R$ аппроксимации. В разделе 1.2 даются различные формулировки задачи $C+R$ аппроксимации и показывается, как эта задача связана с восстановлением неизвестных элементов в малоранговой матрице (матрице с «чёрными точками»). В разделе 1.3 формулируется алгоритм чёрных точек для решения задачи $D+R$ аппроксимации, к которой сводится задача $C+R$ аппроксимации. Этот алгоритм не является итерационным и доказана теорема о том, что алгоритм восстанавливает подавляющее большинство «пропусков» за конечное число итераций, на практике порядка 10-20. В разделе 1.4 формулируется

практическая, адаптивная версия метода чёрных точек, позволяющая строить $C+R$ и $D+R$ аппроксимации без какой-либо дополнительной информации — на вход надо лишь подать исходную матрицу и нужный параметр точности аппроксимации ε. Для матриц общего вида, описываемых n^2 параметрами, получается алгоритм сложности $\mathcal{O}(n^2)$. Для тёплицевых матриц это, конечно, непозволительная роскошь. Решению этого вопроса посвящён раздел 1.5. Основная задача, которую потребовалось решить — дать удобное, легко и быстро вычисляемое описание для Фурье-образа тёплицевой матрицы.

В разделе 1.6 приведены теоретические результаты, касающиеся существования $C+R$ аппроксимаций для класса тёплицевых матриц. Оказывается, такие аппроксимации существуют для практически всех матриц, встречающихся в литературе по супер-линейному предобуславливанию. В разделе 1.7 приведены численные эксперименты по построению $C+R$ аппроксимаций для различных тёплицевых матриц. В разделе 1.8 идея метода чёрных точек получает своё естественное продолжение. Приведён вариант метода, позволяющий восстанавливать неизвестные элементы в малоранговой матрице с произвольным расположением этих самых неизвестных элементов. Однако, в отличие от диагонального шаблона, надо внимательно следить за тем, какие элементы удалось восстановить, а какие нет. Для решения этой проблемы предложены два способа. В разделе 1.9 сделано самое общее возможное обобщение метода чёрных точек на случай, когда положение неизвестных элементов неизвестно. Для этого предложено максимизировать разреженность матрицы $A-R$ с помощью минимизации специального функционала.

Вторая глава посвящена построению специальных преобразований (вейвлет-преобразований) для сжатия матриц, построенных по неравномерным сеткам. Построение происходит на основе использования лифтинговой схемы. В разделе 2.1 описывается история вопроса и мотивируется необходимость построения таких новых преобразований. В разделе 2.2 даются основные понятия и определения, необходимые для дальнейшего изложения. В разделе 2.3 вводится самый важный в главе объект — вейвлет-пространство и описывается так называемая лифтинговая схема построения вейвлет-преобразований с требуемыми свойствами. В разделе 2.4 формулируются основные требования на вейвлет-преобразование: наличие заданного количества нулевых моментов и выписывается система линейных уравнений специального вида на коэффициенты, определяющие искомое преобразование. В разделе 2.5 формулируется основной результат главы и выписывается явная формула для решения основной системы. Раздел 2.6 посвя-

щён нахождению масштабирующих коэффициентов — показано, что их можно находить с помощью уже описанного алгоритма по аналогичным формулам. В разделе 2.8 описан конкретный, пошаговый способ реализации вейвлет-преобразования, требующий $\mathcal{O}(n)$ операций. Также описаны алгоритмы вычисления обратного и обратного транспонированного преобразования — они активно используются в численных расчётах для восстановления исходных данных по преобразованным. В разделе 2.8 приведены численные эксперименты, сравнивающие новые преобразования с преобразованиями Добеши. Показано, что выигрыш по степени сжатия составляет в различных примерах от 30% до 50% процентов.

Глава 3 посвящена общему подходу для построения алгоритмов обращения больших структурированных матриц и решению систем с такими матрицами. В разделе 3.1 дано краткое описание истории вопроса и дана формулировка задачи, описаны основные этапы построения тензорной аппроксимации со структурированными факторами. В разделе 3.3 описан первый способ построения предобуславливателя — масштабированный циркулянтный предобуславливатель. В разделе 3.3 начинается изложение одного из основных результатов диссертации — метода Ньютона для обращения матриц. В разделе 3.4 описан метод Ньютона с аппроксимациями, позволяющий строить быстро приближённые обратные к большим структурированным матрицам. Также в этом разделе предложен модифицированный метод Ньютона, который работает существенно быстрее для структурированных матриц. В разделе 3.5 представлены численные результаты алгоритма по решению уравнения Прандтля.

Глава 4 посвящена теоретическому и практическому изучению вопроса обращения двухуровневых тёплицевых матриц. В рамках развиваемого подхода построен построен метод Ньютона с аппроксимациями для обращения двухуровневых тёплицевых матриц с использованием введённого TDS-формата (Tensor-displacement-structure). Сам новый формат вводится в разделе 4.2. В разделе 4.3 описываются основные арифметические операции над матрицами в TDS формате — сложение, умножение, и показывается, что их можно выполнить за $\mathcal{O}(\sqrt{n}) \log^{\alpha} n)$ операций. Важнейший элемент одного шага метода Ньютона — оператор обрезания — описан в том же разделе. Показано, что задача сводится к вычислению фробениусова скалярного произведения двух TDS-матриц, и это вычисление может быть проведено очень быстро. В разделе 4.5 приводится напоминание о работе метода Ньютона и описывается способ выбора начального приближения. В разделе 4.6 приведены численные эксперименты. И, наконец, в разде-

ле 4.7 впервые получены теоретические результаты о структуре обратных матриц к матрицам малого тензорного ранга специального вида. Построен класс матриц, замкнутый относительно обращения. Объяснено, как свести задачу обращения двухуровневой тёплицевой матрицы, возникающей при дискретизации интегрального уравнения к задаче обращения структурированной матрицы из этого класса. Таким образом, в этом разделе получено теоретическое обоснование предложенных в этой и предыдущей главах быстрых алгоритмов приближённого обращения матриц.

Глава 1.
Метод чёрных точек и наилучшие циркулянтные предобуславливатели

1.1. Введение

Начнём мы с рассмотрения одной классической задачи — построения циркулянтных преобуславливателей и покажем, как она сводится к задаче нелинейной матричной аппроксимации.

Идея использования циркулянтов в качестве предобуславливателей к тёплицевым матрицам была впервые предложена Гильбертом Стрэнгом в 1986 году [31]. Его идея состояла в том, чтобы строить циркулянт, используя половину элементов из первого столбца и первой строчки тёплицевой матрицы. Другой популярный подход — так называемый оптимальный предобуславливатель Т. Чэна (T. Chan) — ближайший во фробениусовой норме циркулянт к заданной (тёплицевой) матрице [6]. Эти предобуславливатели легко построить, но в некоторых случаях они не работают (число итераций может сильно расти с увеличением размера матрицы n). Для «плохих» случаев было построено несколько методов и алгоритмов (см. обзор [8]). Однако, наиболее эффективные подходы явно используют информацию о так называемом символе (производящей функции) тёплицевой матрице (чуть ниже мы дадим определение, что такое символ). По этой причине они могут быть названы «функциональными», а не «матричными» (см. [25]) подходами. Более того, метод, подходящий для симметричных положительно определённых матриц, может не работать для незнакоопределённых или несимметричных матриц — существующие подходы построения циркулянтных преодобуславливателей не являются универсальными.

Мы предлагаем матричный подход для построения новых циркулянтных предобуславливателей, которые, по-видимому, являются наилучшими из всех известных циркулянтных предобуславливателей. В отличие от «функциональных» циркулянтных предобуславливателей для «плохих» символов, они строятся только по элементам матриц и работают не хуже лучших из известных предобуславливателей для одних и тех же символов. Но для реализации нашего матричного под-

хода нам потребовалось решить одну новую, нестандартную задачу нелинейной матричной аппроксимации. Основным результатом главы является «метод чёрных точек» и его быстрая версия для построения специальных аппроксимаций для тёплицевых матриц. Вкратце, если хороший циркулянтный предобуславливатель существует, то он может быть легко найден с помощью нашего алгоритма.

Начнём с самого начала. Если линейная система

$$Ax = b$$

решается с помощью какого-либо итерационного метода (такого, как CG или GMRES) и наблюдается медленная сходимость (что случается часто), тогда известное «лекарство» состоит в переходе к предобусловленной системе

$$AP^{-1}x = b,$$

где P называется предобуславливателем. Анализ качества предобуславливателя обычно начинается с погружения выбранной системы в последовательность систем (матриц коэффициентов, правых частей, предобуславливателей), параметризованных размером матрицы n. Далее, для того, чтобы предобуславливатель был «хорошим», добиваются выполнения следующих свойств

(a) Для AP^{-1} существует равномерная оценка на число обусловленности по n;

(b) Собственные значения AP^{-1} имеют *кластер* в 1. Это, значит, что подавляющее большинство собственных значений матрицы AP^{-1} находятся в ε–окрестности 1.

По крайней мере, для эрмитовых положительно определённых матриц и при некоторых дополнительных предположениях в общем случае, свойство (a) означает *линейную сходимость*, а свойство (b) даёт так называемую *суперлинейную сходимость* (см. [42]). Поэтому свойство (b) особенно интересно. Когда же существует кластер и как доказывается его существование? Существование кластера тесно связано с разложением вида [39]

$$A = P + R + E, \tag{1.1}$$

где $\operatorname{rank} R = r \ll n$ и $\|E\| \leq \varepsilon$. Отметим, что матрицы в правой части (1.1) зависят от n и ε.

Но если представление (1.1) так привлекательно, то почему бы не исходить прямо из него? Мы предлагаем строить предобуславливатели

P основываясь прямо на (1.1). Если мы будем выбирать P из некоторого матричного класса, то (1.1) становиться задачей аппроксимации со следующей (пока нестрогой) формулировкой:

C+R аппроксимация *Аппроксимировать матрицу* A*, суммой вида*

$$A \approx C + R,$$

где C *— циркулянт, (в* (1.1) *соответствует* P*) и* R *— матрица «малого» ранга.*

Рассмотрим, например, тёплицевы матрицы $A = [a_{i-j}]$ размера $n = 128, 256, 512$, порождённые символом $f = x^4$ (это означает, что a_k являются коэффициентами Фурье для f). Пусть $P = C$ в (1.1) является либо предобуславливателем Стрэнга, либо предобуславливателем Т. Чэна. Тогда, установив точность $\varepsilon = 10^{-2}$, найдём R с помощью отбрасывания сингулярных чисел матрицы

$$A - C = R + E$$

по порогу ε так, что $\|E\|_2 \leq \varepsilon$. В этом случае получаются следующие ранги для R:

Таблица 1.1. Зависимость $\operatorname{rank} R$ от n ($\varepsilon = 10^{-2}$)

s n	Стрэнг	Т. Чэн
128	8	20
256	8	24
512	8	24

Для этого фиксированного ε, rank R практически не зависит от размера матрицы n. Однако исследуем, как этот ранг зависит от ε для фиксированного n.

Таблица 1.2. Зависимость $\operatorname{rank} R$ от ε ($n = 256$).

ε	Стрэнг	Т. Чэн
10^{-3}	10	244
10^{-4}	18	254
10^{-5}	50	256

Как мы видим, оба предобуславливателя сработали неудовлетворительно: кластера собственных значений нет — собственные значения не групируются около единицы. Ситуация с предобуславливателем

Стрэнга выглядит получше, однако видно, что ранги растут как $\varepsilon^{-\alpha}$, $\alpha \sim 1$. Матрица A плохо обусловлена, поэтому мы должны аппроксимировать её с высокой точностью и ни один из двух рассмотренных предобуславливателей не даёт собственный кластер. Однако данная матрица может быть *очень точно аппроксимирована* суммой циркулянта и матрицы *достаточно малого* ранга. Но соответствующий циркулянт не имеет ничего общего ни с преобуславливателем Стрэнга, ни с предобуславливателем Т. Чэна. Более того, будет доказано, что для довольно общего класса тёплицевых матриц (включая все примеры в статьях по суперлинейным предобуславливателям) существуют аппроксимации суммой циркулянта и матрицы малого ранга[63] с оценкой вида

$$r = \operatorname{rank} R = \mathcal{O}(\log \varepsilon^{-1}(\log \varepsilon^{-1} + \log n)).$$

Поэтому мы можем быть уверены, что «хороший» циркулянт существует и заинтересованы в том, как его вычислить.

1.2. Задачи C+R и D+R аппроксимации

Так как каждый циркулянт диагонализуется с помощью дискретного преобразования Фурье (Discrete Fourier Transform—DFT)

$$C = \frac{1}{n} F^* D F,$$

где F это матрица дискретного преобразования Фурье, а D — диагональная матрица из собственных значений, то задача $C + R$ аппроксимации может быть переформулирована следующим образом:

$$\widehat{A} = \frac{1}{n} F A F^* \approx D + R. \tag{1.2}$$

Поэтому, общая задача $C + R$ аппроксимации легко сводится к задаче $D + R$ аппроксимации, где D — диагональная матрица. В силу унитарности матрицы Фурье F задачи $D + R$ и $C + R$ аппроксимации эквивалентны.

Теперь давайте строго определим, что же такое «приближённо» и «малый ранг». Если мы зафиксируем ранг малоранговой части разложения, то получится следующая задача аппроксимации:

D+R I: *Для заданной матрицы* A *и целого числа* $r > 0$ *найти матрицу* B *вида* $B = D + R$, *где* $\operatorname{rank} R \leqslant r$, *и диагональную матрицу* D, *которые минимизируют* $\|A - B\|_F$.

Мы можем исключить из формулировки либо R либо D. Если мы исключим R, то получится оптимизационная задача в терминах сингулярных чисел.

D+R II: *Для заданной матрицы* A *и целого числа* $r > 0$ *найти диагональную матрицу* D*, которая минимизирует*

$$\sigma_{r+1}(A - D).$$

Важно отметить, что это негладкая, невыпуклая задача оптимизации, которая, по-видимому, имеет много локальных минимумов (нам не известно о каких либо эффективных методах её решения).

Если же мы исключим D, то получится следующая формулировка:

D+R III: *Для заданной матрицы* $A = [A_{ij}]$ *и целого числа* $r > 0$ *найти матрицу* $R = [R_{ij}]$ *ранга не выше* r*, которая минимизирует*

$$\sum_{i,j=1,\, i\neq j}^{n} (A_{ij} - R_{ij})^2.$$

Задача $D + R$ аппроксимации была впервые рассмотрена в [3], где был предложен итерационный метод её решения. Это был вариант метода переменных направлений, названный ADR (Alternating Diagonal Rank) с двухшаговой итерацией следующего вида:

Пусть заданы некоторые начальные приближения для D *и* R*, тогда для нахождения новых приближений к рещению* $\widehat{D}$ *и* $\widehat{R}$ *нужно действовать следующим образом:*

1. $\widehat{D} = \arg\min\limits_{D} ||A - D - R||_F$;

2. $\widehat{R} = \arg\min\limits_{R,\, \mathrm{rank}R \leqslant r} ||A - \widehat{D} - R||_F$.

Видно, что на каждом шаге невязка $||A - D - R||_F$ невозрастает. К сожалению, по-видимому, это единственное достоинство метода ADR. Часто для его сходимости требуется большое количество итераций. Иногда он «застревает» в локальном минимуме. Он имеет большую вычислительную сложность — $\mathcal{O}(n^3)$ операций на каждом шаге, что делает его (в вышеприведённом виде) неприемлемым для практического использования. Существует способ видоизменить этот метод (совсем нетривиально) так, чтобы он сходился к глобальному минимуму, однако вычислительная сложность всё равно остаётся очень высокой. Тем не менее, если имеется хорошая аппроксимация полученная с помощью какого-нибудь другого метода (например, с помощью

алгоритма, предлагаемого в данной диссертации), можно попытаться построить некоторую быструю версию ADR для улучшения заданной аппроксимации к решению.

Давайте посмотрим более внимательно на формулировку D + R III. Вспомним, что нам интересен случай, кого матрица A хорошо аппроксимируется суммой диагональной матрицы и матрицы ранга r. Поэтому начнём с предположения, что матрица A является *точной* суммой диагональной матрицы и матрицы ранга r. Как можно восстановить D и R, зная только их сумму? Ответ мы узнаем совсем скоро.

1.3. Чёрные точки, малые ранги и скелетоны

Задача формулируется следующим образом. Пусть матрица A является *точной* суммой диагональной матрицы и матрицы ранга r. Зная, что $A = D + R$, как восстановить D и R по A?

Очевидно, что в матрице R мы знаем все внедиагональные элементы. Поэтому, всё что нужно определить — это диагональные элементы матрицы R. Перед описанием общего метода, рассмотрим следующий простой пример матрицы ранга 2 и размера 6×6:

$$A = \begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 6 & 7 & 8 & 9 & 10 \\ 6 & 7 & 8 & 9 & 10 & 11 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{pmatrix}.$$

(Это действительно матрица ранга 2, так как $a_{ij} = i + j$).

Предположим теперь, что мы не знаем диагональных элементов матрицы A:

$$A = \begin{pmatrix} \bullet & 3 & 4 & 5 & 6 & 7 \\ 3 & \bullet & 5 & 6 & 7 & 8 \\ 4 & 5 & \bullet & 7 & 8 & 9 \\ 5 & 6 & 7 & \bullet & 9 & 10 \\ 6 & 7 & 8 & 9 & \bullet & 11 \\ 7 & 8 & 9 & 10 & 11 & \bullet \end{pmatrix}.$$

Диагональные элементы обозначены чёрными точками. Вопрос такой: как дополнить внедиагональную часть, заменив чёрные точки числами так, чтобы получившаяся матрица имела ранг 2? Можно применить следующую простую идею. Возьмём подматрицу, образован-

ную столбцами 4,5,6 и строками 2,3,4:

$$\widehat{A} = \begin{pmatrix} 6 & 7 & 8 \\ 7 & 8 & 9 \\ \bullet & 9 & 10 \end{pmatrix}.$$

Мы хотим получить матрицу ранга 2, так что эти 3 столбца должны быть линейно зависимы; поэтому, первый столбец должен быть линейной комбинацией второго и третьего столбца (которые, как легко видеть, линейно независимы). Коэффициенты этой линейной комбинации легко определяются из решения следующей системы:

$$\begin{pmatrix} 7 & 8 \\ 8 & 9 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \end{pmatrix}.$$

Как нетрудно видеть, выбор этих строк и столбцов неединственнен. В данном примере, различные способы выбора «опорных» строк и столбцов приведут к одному и тому же результату. На практике же, выбор «правильных» строк и столбцов, используемых для восстановления очень важен, и неправильный выбор может привести (и часто приводит) к неустойчивости.

Опишем теперь приведённую процедуру в общем случае и докажем, что она действительно восстанавливает чёрные точки.

Рассмотрим произвольную матрицу B ранга r, возьмём r линейно независимых строк и r линейно независимых столбцов из B и образуем матрицы $L \in \mathbb{R}^{n\times r}$ (из столбцов) и $U \in \mathbb{R}^{r\times n}$ (из строк). Пусть $\widehat{B}$ обозначает подматрицу размера $r \times r$, находящуюся на пересечении этих выбранных строк и столбцов. Тогда подматрица $\widehat{B}$ невырождена и матрица B может быть представлена в виде

$$B = L\widehat{B}^{-1}U,$$

который называется *скелетным разложением*.

Основное утверждение состоит в том, что матрица ранга r однозначно определяется по своим линейно независимым r столбцам и строкам. Построим теперь скелетное разложение для матрицы с чёрными точками на диагонали. Для нашего примера, строки 3,4 и столбцы 1,2 дают нам невырожденную подматрицу на пересечении, и поэтому мы можем написать

$$A = \begin{pmatrix} \bullet & 3 \\ 3 & \bullet \\ 4 & 5 \\ 5 & 6 \\ 6 & 7 \\ 7 & 8 \end{pmatrix} \begin{pmatrix} 4 & 5 \\ 5 & 6 \end{pmatrix}^{-1} \begin{pmatrix} 4 & 5 & \bullet & 7 & 8 & 9 \\ 5 & 6 & 7 & \bullet & 9 & 10 \end{pmatrix}. \tag{1.3}$$

Для того, чтобы определить, как работать с чёрными точками, введём следующую «арифметику чёрных точек»:

$$\bullet\bullet = \bullet,$$

$$\bullet x = x\bullet = \bullet,$$

$$\bullet + x = x + \bullet = \bullet,$$

где x — обычное число. Поэтому, умножение матриц в (1.3) даёт

$$A = \begin{pmatrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ 4 & 5 & \bullet & \bullet & 8 & 9 \\ 5 & 6 & \bullet & \bullet & 9 & 10 \\ 6 & 7 & \bullet & \bullet & \underline{10} & 11 \\ 7 & 8 & \bullet & \bullet & 11 & \underline{12} \end{pmatrix}.$$

Это означает, что мы нашли (подчёркнутые) диагональные элементы (5,5) и (6,6). Так как все внедиагональные элементы в A заданы, теперь нам известны два полных столбца с номерами $5, 6$ и две полных строчки с номерами $5, 6$ из исходной A и можно снова использовать скелетное разложение для восстановления оставшихся неизвестных элементов полной матрицы.

В общем случае мы делаем то же самое.

Метод чёрных точек. *Пусть задана матрица A, допускающая представление в виде $A = D + R$ с диагональной матрицей D и матрицей R ранга r. Тогда для нахождения по крайней мере $n - 2r$ столбцов и строк неизвестной матрицы R нужно действовать следующим образом:*

1. *Взять в A невырожденную подматрицу $\hat{A}$ размера $r \times r$, которая не содержит диагональных элементов A. Пусть строчки и столбцы этой подматрицы имеют в A индексы $i_1, \ldots, i_r$ и $j_1, \ldots, j_r$ соответственно, и пусть матрицы L и U размерами $n \times r$ и $r \times n$ составлены из этих строк и столбцов.*

2. *Образовать матрицу* $Q = L\hat{A}^{-1}U$ *(всё ещё с чёрными точками) и обнаружить, что элементы*

$$Q_{ij} = R_{ij}, \qquad i \neq j_1, ..., j_r, \quad j \neq i_1, ..., i_r, \tag{1.4}$$

больше не являются чёрными точками. Следовательно, к этому моменту нам известны по крайней мере $n - 2r$ *диагональных элементов матрицы* R.

Алгоритм основан на следующей простой теореме.

Теорема 1 *Элементы определённой выше матрицы* Q *удовлетворяют (1.4).*

Доказательство. Используя определение L и U, получаем, что

$$Q_{ij} = \sum_{k=1}^{r} \sum_{l=1}^{r} R_{ij_k} (\hat{A}^{-1})_{kl} R_{i_l,j}.$$

Если $i \neq j_1, ..., j_r$ и $j \neq i_1, ..., i_r$, тогда ни один из элементов вида R_{i,j_k}, R_{j,i_l} не находится на главной диагонали. Поэтому, все эти элементы известны и соотвествущие элементы Q должны совпадать с соответствующими элементами R. □

В наших приложениях $r \ll n$, поэтому первые два шага метода чёрных точек позволяют восстановить большинство диагональных элементов. Для нахождения оставшихся элементов R нужен ещё один шаг:

(3) *Если* n *достаточно велико (*$n - 2r \geqslant r$*, или, что эквивалентно,* $n \geqslant 3r$*), тогда нам известно по крайней мере* r *столбцов и* r *строчек матрицы* R*. Предположим, что эти* r *столбцов и строчек линейно независимы. Тогда, используя их для построения скелетного разложения, восстановим оставшиеся чёрные точки в* R.

Отметим, что это шаг основан на предположении, что первые два шага дали нам r линейно независимых столбцов и строчек с известными элементами. Для этого достаточно потребовать, что в A есть две невырожденные подматрицы размера $r \times r$, не содержащие общих столбцов и строчек и не содержащие также диагональных элементов матрицы A.

1.4. Адаптивная версия метода чёрных точек

Осталось несколько необсуждавшихся проблем. Во-первых, матрица A может не являться точной суммой диагональной матрицы и матрицы малого ранга. Вместо этого, пусть

$$A = D + R + E,$$

где E такое, что $\|E\| \leq \varepsilon$ может рассматриваться как некий «шум». Более того, значение r ранга R (зависящее от ε) может быть неизвестно заранее. Поэтому мы получаем задачу, где ранг r нужно определить, исходя из заданного порога точности ε.

В случае наличия шума, выбор столбцов и строчек (или, что одно и то же, выбор опорной подматрицы), по которым строится скелетное разложение, играет основную роль. Вопрос в том, какая подматрица является наилучшей. Если бы чёрных точек не было, ответом является так называемый *принцип максимальноо объёма* [13]: если подматрица $\widehat{A}$ имеет максимальный объём (т.е. максимальный по модулю определитель) среди всех подматриц размера $r \times r$, тогда оценка на ошибку скелетного разложения выглядит следующим образом:

$$|(A - L\widehat{A}^{-1}U)_{ij}| \leqslant (r+1)\,\sigma_{r+1}(A), \tag{1.5}$$

где $\sigma_{r+1}(A)$ — $(r+1)$ сингулярное число матрицы A. Мы предполагаем, что при наличии чёрных точек такой же «хорошей» подматрицей будет подматрица максимального объёма среди всех подматриц без чёрных точек.

Так как нахождение подматрицы максимального объёма составляет нетривиальную задачу, мы можем попытаться заменить её на подматрицу с достаточно большим объёмом. Такая подматрица может быть получена с помощью различных вариантов *метода неполной крестовой аппроксимации* (см. [14, 38, 1]).

Приведём описание крестового метода. Он работает следующим образом (здесь R это ненулевая матрица, которую нужно аппроксимировать малоранговой матрицей):

1. Инициализация: $k = 0$, $R^k = R$.

2. Выбор ведущего элемента: $(i_0, j_0) = \arg\max_{i,j} |R^k_{ij}|$.

3. Вычислить крест-скелетон:

$$C^k = \frac{u_k v_k^\top}{R^k_{i_0 j_0}},$$

где u_k это строка с номером i_0 матрицы R^k и v_k — столбец с номером j_0 в матрице R^k.

4. Вычислить новую невязку $R^{k+1} = R^k - C^k$.

5. Если норма невязки $\|R^{k+1}\|$ достаточно мала, тогда остановиться и вернуть $\sum_{i=0}^{k} C^i$ как аппроксимацию ранга r к матрице R. Иначе, увеличить k на 1 и перейти к шагу 2.

На каждом шаге мы вычитаем из матрицы одну матрицу ранга 1, называемую *скелетоном*. Она определяется по выбранным столбцу и строчке так, чтобы получившийся скелетон совпадал с матрицей R^k по одному столбцу и одной строчке. Это некоторый дискретный аналог интерполяции.

Так как мы «интерполируем» на каждом шаге одну строчку и один столбец, ведущие элементы элементы (т.е. элементы $R^k_{i_0 j_0}$) нужно выбирать так, чтобы исключить «большие» элементы в матрице невязки. Когда алгоритм завершается, он на самом деле возвращает вариант скелетного разложения с предположительно хорошей подматрицей (выбор ведущего элемента обычно приводит к подматрице достаточно большого обьёма).

Суммарная стоимость приведённого варианта составляет $\mathcal{O}(n^2 r)$ из-за того, что на каждом шаге ищется максимальный элемент по всей матрице (полное пивотирование). Изначально, однако, метод неполной крестовой аппроксимации был создан в надежде на то, что он может аппроксимировать матрицу малого ε-ранга, используя только лишь малое число её элементов [14]. Если матрица имеет точный ранг r, тогда гауссово исключение с выбором ведущего элемента даёт нулевой ведущий элемент точно после r шагов. На самом деле, точно такой же подход может быть применён и при наличии шума. Однако, можно реализовать различные стратегии методов выбора ведущих элементов (например, выбор по строке или столбцу). Особую вычислительную выгоду дают методы, использующие лишь небольшое число элементов матрицы. С частичным выбором ведущего элемента мы можем, конечно, получить больший коэффициент перед $\sigma_{r+1}(A)$ в оценке (1.5); он

зависит от того, как близок объём получающейся подматрицы к подматрице максимального объёма [13]. Для некоторого класса матриц, этот коэффициент может составлять 2^r вместо $r+1$. В любом случае, даже не смотря на потерю точности (которая компенсируется увеличением r), существенное снижение вычислительной сложности делает частичный выбор ведущего элемента единственным практически эффективным алгоритмом для метода неполной крестовой аппроксимации.

Метод неполной крестовой аппроксимации может быть легко приспособлен для случая матриц с неизвестными элементами (чёрными точками). Мы только должны следить за тем, как чёрные точки распространяются на каждом шаге. Строки и столбцы, содержащие чёрные точки, составят наши «чёрные списки» (обновляемые на каждом шаге).

Адаптивный метод чёрных точек.

1. *Инициализация:*

$$k = 0, \quad R_{ii} = 0, \quad R_{ij} = A_{ij}, \quad i \neq j, \quad R^k = R,$$

$$\mathcal{L}_r = \emptyset, \quad \mathcal{L}_c = \emptyset \quad (\text{«чёрные списки»}).$$

2. *Найти ведущий элемент:*

$$(i_0, j_0) = \arg \max_{i \neq j,\, i \notin \mathcal{L}_c,\, j \notin \mathcal{L}_r} |R^k_{ij}|.$$

3. *Вычислить крест-скелетон:*

$$C^k = \frac{u_k v_k^\top}{R^k_{i_0 j_0}},$$

где u_k — i_0*-ая строка а* v_k — j_0*-ый столбец матрицы* R^k.

4. *Вычислить новую невязку:* $R^{k+1} = R^k - C^k$.

5. *Добавить элемент* i_0 *к* $\mathcal{L}_c$ *и элемент* j_0 *к* $\mathcal{L}_r$.

6. *Вычислить ошибку:*

$$\delta^k = \Big(\sum_{i,j \in S} (R^k_{ij})^2\Big)^{1/2}, \quad S = \{i, j : 1 \leqslant i, j \leqslant n,\ i \neq j,\ i \notin \mathcal{L}_r,\ j \notin \mathcal{L}_c\}.$$

7. *Если* δ_k *достаточно мало, выйти и возвратить*

$$d_i = (\sum_{m=0}^{k} C^m)_{ii}, \quad i \notin \mathcal{L}_r, \quad i \notin \mathcal{L}_c,$$

как аппроксимации к соответствующим диагональным элементам R_{ii}. *Иначе, увеличить* k *на* 1 *и перейти к шагу 2.*

Приведённый алгоритм на самом деле возвращает вариант скелетного разложения с некоторой адаптивно выбранной подматрицей $\hat{R}$, размер которой не был задан заранее и ожидается, что это «хорошая» подматрица для аппроксимации. Ошибка измеряется по элементам в известной части матрицы.

Если n достаточно велико, тогда метод чёрных точек возвращает приближения ко всем неизвестным элементам, за исключением $2r$ диагональных элементов. После этого мы должны запустить наш метод во второй раз, выбирая ведущие элементы только в полностью известных столбцах и строках. В конце концов, мы получаем скелетную аппроксимацию для матрицы R:

$$R \approx \sum_{k=1}^{r} x_k y_k^\top = XY^\top.$$

Выражение для диагональной части $D + R$ аппроксимации имеет вид

$$D \approx \operatorname{diag}(A - XY^\top).$$

Как уже отмечалось выше, описанный адаптивный метод чёрных точек предназначен для задачи $D + R$ аппроксимации. Однако, он может быть легко приспособлен для многих других проблем с другими шаблонами расположения чёрных точек.

Представленные выше алгоритмы требуют $\mathcal{O}(n^2(\log n + r))$ операций для построения $C{+}R$ аппроксимации к неструктурированной матрице A. Эта «цена» складывается из двух частей: использование БПФ для вычисления FAF^* и из-за стратегии полного выбора ведущего элемента. В случае тёплицевых матриц это неприемлемая сложность. В дальнейшем мы покажем, как наши методы могут быть адаптированы для тёплицевого случая. В результате будет предложен алгоритм сложности $\mathcal{O}(n(\log n + r^2))$.

1.5. Тёплицев случай

1.5.1. Быстрое вычисление образа Фурье для тёплицевой матрицы Пусть дана тёплицева матрица T и A — её образ Фурье:

$$T = [t_{i-j}], \qquad A = \frac{1}{n} F T F^*.$$

Внедиагональные элементы матрицы $A = [A_{kl}]$ можно параметризовать следующим простым и эффективным образом.

Лемма 1

$$A_{kl} = \frac{\nu_k - \nu_l}{n(w^{k-l} - 1)}, \quad 0 \le k, l \le n-1, \quad k \ne l, \tag{1.6}$$

$$\nu_k = w^{-k}\widetilde{\nu}_k, \qquad w = e^{\frac{2\pi i}{n}},$$

$$\widetilde{\nu} = F\widetilde{t}, \qquad \widetilde{t}_k = t_{k-n} - t_k, \quad 0 \le k \le n-2, \qquad \widetilde{t}_{n-1} = 0.$$

Доказательство. По определению образа Фурье A,

$$nA_{kl} = \sum_{\alpha=0}^{n-1}\sum_{\beta=0}^{n-1} w^{-\alpha k} t_{\alpha-\beta} w^{\beta l} = \sum_{\beta=0}^{n-1} w^{\beta(l-k)} \sum_{\alpha=-\beta}^{n-\beta-1} w^{-\alpha k} t_\alpha.$$

Изменяя порядок суммирования, получаем:

$$\begin{aligned}
nA_{kl} &= \sum_{\alpha=-n+1}^{-1} w^{-\alpha k} t_\alpha \sum_{\beta=-\alpha}^{n-1} w^{\beta(l-k)} + \sum_{\alpha=1}^{n-1} w^{-\alpha k} t_\alpha \sum_{\beta=0}^{n-\alpha-1} w^{\beta(l-k)} \\
&= \sum_{\alpha=1}^{n-1} w^{\alpha l} t_{-\alpha} \frac{w^{(n-\alpha)(l-k)} - 1}{w^{l-k} - 1} + \sum_{\alpha=1}^{n-1} w^{-\alpha k} t_\alpha \frac{w^{(n-\alpha)(l-k)} - 1}{w^{l-k} - 1} \\
&= \frac{1}{w^{k-l} - 1}\Big(\sum_{\alpha=1}^{n-1}(w^{\alpha k} t_{-\alpha} - w^{\alpha l} t_{-\alpha}) + \sum_{\alpha=1}^{n-1}(w^{-\alpha l} t_\alpha - w^{-\alpha k} t_\alpha)\Big) \\
&= \frac{\nu_k - \nu_l}{w^{k-l} - 1},
\end{aligned}$$

где

$$\nu_k = \sum_{\alpha=1}^{n-1}(w^{\alpha k} t_{-\alpha} - w^{-\alpha k} t_\alpha) = \sum_{\alpha=1}^{n-1} w^{-\alpha k}(t_{\alpha-n} - t_\alpha).$$

Видно, что для вычисления v требуется вычислить одно дискретное преобразование Фурье и умножить на диагональную матрицу. Главная диагональ A также вычисляется за одно дискретное преобразование Фурье. Как только предварительный шаг завершён, любой элемент A может быть вычислен с помощью(1.6) очень быстро.

Вместо полного выбора ведущего элемента, мы должны использовать какую-нибудь стратегию частичного выбора. Мы предлагаем использовать *ладейную схему*:

1. *На каждом шаге вычислить наддиагональ матрицы-невязки* R^k*:*
$$S = [(R^k)_{12}, \ldots, (R^k)_{n-1,n}].$$
2. *Найти максимальный по модулю элемент в* S *и его индексы* $(i_0, i_0 + 1)$.
3. *Найти максимальный по модулю элемент в* i_0*-ой строчке матрицы* R^k *и использовать его для вычисления следующего креста.*

После этих модификаций суммарная сложность метода чёрных точек для тёплицевых матриц снижается до

$$\mathcal{O}(n(\log n + r^2)).$$

Множитель r^2 появляется из-за того, что для вычисления какого-либо столбца или строчки матрицы R^k мы должны вычислить соответствующие элементы в предыдущих $k - 1$ крестах, поэтому сложность ведёт себя как

$$n(0 + 1 + 2 + \ldots + (r - 1)) = \mathcal{O}(nr^2).$$

В итоге, для тёплицевого случая, $C+R$ аппроксимация может быть вычислена быстро, если только требуемый ранг $r \ll n$. Верхние оценки на ранг r, который мы называем *циркулянтным рангом*, будут представлены в следующем параграфе.

1.6. Существование C+R аппроксимации для некоторых классов тёплицевых матриц

Как было заявлено ранее, некоторые широкие и практически важные классы символов приводят к тёплицевым матрицам, которые могут быть очень точно аппроксимированы суммой циркулянта и матрицы малого ранга.

Лемма 2 Пусть T — нижнетреугольная тёплицева матрица с первым столбцом $t_k = \alpha\rho^k$, *и при этом* $\rho^n \neq 1$. Тогда $T = C + R$, где C — циркулянтная матрица, а R — матрица ранга 1. Если же $\rho^n = 1$, то ранг малоранговой добавки равен 2.

Доказательство. В качестве R достаточно взять одноранговую тёплицеву матрицу

$$R = [r_{i-j}], \quad r_k = \frac{\alpha\rho^k}{1 - \frac{1}{\rho^n}}, \quad k = -n+1, \ldots, n-1.$$

Напрямую проверяется, что матрица $C = T - R$ циркулянтная. Если же $\rho^n = 1$, то в качестве r_k возьмём функцию вида:

$$r_k = \beta k \rho^k,$$

Тогда

$$c_k - ck - n = t_k - r_k + r_{k-n} = \rho^k(\alpha - \beta n).$$

(мы использовали, что $\rho^{-n} = 1$). Нам нужно, чтобы $c_k - ck - n = 0$. Для этого выражение справа должно тождественно равняться нулю, т.е. $\beta = \alpha/n$. Ранг же матрицы R будет равным 2, т.к. $r_{i-j} = \beta\rho^i\rho^{-j}(i-j)$.

Следствие 1 Если тёплицева матрица T порождена символом

$$f(z) = \frac{1}{1 - \rho z}, \quad z = e^{it}, \quad |\rho| \neq 1,$$

то $T = C + R$, где C — циркулянтная матрица, а R — матрица ранга 1.

Доказательство. Возможны два случая: $|\rho| < 1$ и $|\rho| > 1$. В первом случае $f(z)$ раскладывается в ряд Тейлора

$$f(z) = \sum_{k=0}^{\infty} \rho^k z^k,$$

поэтому мы можем применить лемму 2. В случае $|\rho| > 1$ мы можем разложить $f(z)$ в ряд Лорана. Единственное отличие от первого случая состоит в том, что мы получаем не нижнетреугольную, а верхнетреугольную матрицу.

Это был случай функции с простым полюсом. А что будет, если символ f будет иметь полюс более высокого порядка? Для этого нам потребуется доказать следующую лемму.

Лемма 3 Пусть T — нижнетреугольная тёплицева матрица с первым столбцом $t_k = s(k)\rho^k$, q — *натуральное число, а* $s(x)$ — *многочлен степени не выше* q, ρ — *некоторое число,* Тогда $T = C + R$, где C — циркулянтная матрица, и при этом

$$\operatorname{rank} R = \begin{cases} q+1, & \rho^n \neq 1; \\ q+2, & \rho^n = 1. \end{cases}$$

Доказательство. В качестве R возьмём тёплицеву матрицу специального вида:

$$R = [r_{i-j}], \quad r_k = p(k)\rho^k,$$

где p — многочлен , который ещё предстоит определить.

Нам нужно, чтобы матрица $C = T - R$ была циркулянтом. Это означает, что

$$c_k - c_{k-n} = t_k - r_k + r_{k-n} = 0, \quad k = 0, \ldots n-1.$$

Отсюда получается следующее уравнение на неизвестный многочлен p:

$$p(k) - p(k-n)\rho^{-n} = s(k).$$

Возможны два случая. Пусть $\rho^n = 1$. Тогда полином p определяется с точностью до постоянной, и мы можем положить $p(0) = 0$.

$$\begin{aligned} p(n) &= s(n), \\ p(2n) &= p(n) + s(2n) = s(n) + s(2n), \\ \ldots &= \ldots \\ p(kn) &= s(kn) + s(k-1)n + \ldots + s(n). \end{aligned}$$

т.е. значения многочлена p задаются в точках Построим теперь в точках $0, n, \ldots, (q+1)n$ многочлен $(q+1)$ степени с этими значениями. Тогда многочлен $p(x) - p(x-n)$ будет многочленом q-ой степени и будеть совпадать с $s(x)$ (в силу выполнения интерполяционных условий).

Ранг матрицы R не больше $q+2$. Действительно,

$$r_{i-j} = p(i-j) = \rho^{i-j} \sum_{k=0}^{q+1} p_k (i-j)^k.$$

В каждом слагаемом раскроем скобки:

$$(i-j)^k = \sum_{m=0}^{k} C_k^m (-1)^{k-m} i^m j^{k-m},$$

подставим в выражение для r_{i-j} и поменяем порядок суммирования:

$$r_{i-j} = \rho^{i-j} \sum_{k=0}^{q+1} \sum_{m=0}^{k} p_k C_k^m (-1)^{k-m} i^m j^{k-m} = \rho^{i-j} \sum_{m=0}^{q+1} i^m (-1)^m j^m \sum_{k=m}^{q+1} C_k^m p_k j^{-k}.$$

Видно, что r_{i-j} представимо в виде

$$r_{i-j} = \sum_{m=0}^{q+1} u_{im} v_{jm},$$

т.е. является матрицей ранга не выше $q+2$.

Случай $\rho^n \neq 1$ разбирается аналогично. Единственное отличие состоит в том, что многочлен p можно выбирать не $q+1$, а q-ой степени, так как добавляется одно уравнение на старший коэффициент p (который сокращался в выражении $p(x) - p(x-n)$).

Матрица $C \equiv T - R$ будет циркулянтной в силу того, что

$$c_k - c_{k-n} = t_k - t_{k-n} - r_k + r_{k-n} = t_k - r_k + r_{k-n}.$$

Теорема 2 Пусть тёплицева матрица T порождена рациональным тригонометрическим символом

$$f(z) = P(z) + \frac{Q(z)}{L(z)}, \quad z = e^{it},$$

где P, Q, L — многочлены, L не имеет корней на единичной окружности, степень Q меньше степени L и они не имеют общих корней. Тогда

$$T = C + R,$$

где C–циркулянтная матрица, и при этом $\operatorname{rank} R \leqslant \deg P + \deg L + 1$.

Доказательство. Представляем $\frac{Q(z)}{L(z)}$ в виде суммы простых дробей:

$$\frac{Q}{L} = \frac{c_0}{(z-z_0)^{\alpha_0}} + \ldots + \frac{c_d}{(z-z_d)^{\alpha_d}},$$

где d — степень многочлена L. Рассмотрим теперь отдельно каждую простую дробь вида $\frac{c}{(z-a)^\alpha}$. Элементы соответствующих такому символу тёплицевых матриц можно получить, разлагая $\frac{c}{(z-a)^\alpha}$ либо в ряд Тейлора, либо в ряд Лорана, в зависимости от того, где находится полюс a относительно единичной окружности. В одном случае

это приведёт к нижнетреугольным тёплицевым матрицам, в другом — к верхнетреугольным, поэтому можно ограничиться лишь изучением одного из двух случаев, второй рассматривается аналогично. Коэффициенты ряда Тейлора (или ряда Лорана) функции $\frac{c}{(z-a)^\alpha}$ легко находятся. А именно,

$$\frac{c}{(z-a)^\alpha} = \sum_{k=0}^{\infty} f_k z^k, \text{ где}$$

$$f_k = \frac{(-1)^\alpha}{a^\alpha} \frac{(\alpha+k-1)!}{(\alpha-1)!k!a^k}.$$

Видно, что эти коэффициенты удовлетворяют условиям леммы 3, применяя которую мы и получаем требуемый результат.

Перейдём теперь от рациональных символов к символам с логарифмическими особенностями.

Лемма 4 Пусть T — нижнетреугольная тёплицева матрица с первым столбцом

$$t_k = \begin{cases} 0, & k = 0; \\ \rho^k k^{-\alpha}, & k = 1, \ldots, n-1, \quad \alpha > 0. \end{cases}$$

Тогда для любого ε существуют циркулянтная матрица C и матрица R ранга r такие, что

$$|(T - C - R)_{ij}| \leqslant |T_{ij}|\varepsilon,$$

причём

$$r \leqslant \log \varepsilon^{-1}[c_0 + c_1 \log \varepsilon^{-1} + c_2 \log n],$$

где c_0, c_1, c_2 зависят только от α.

Доказательство. Для любого ε существуют f_m, w_m такие, что [44]

$$|k^{-\alpha} - \sum_{m=1}^{r} w_m e^{-f_m k}| \leqslant k^{-\alpha}\varepsilon, \qquad r \leq \log \varepsilon^{-1}[c_0 + c_1 \log \varepsilon^{-1} + c_2 \log n].$$

Остается применить лемму 2.

Следствие 2 Пусть тёплицева матрица T порождена символом

$$f(z) = \log(z - \zeta), \quad z = e^{it}, \quad |\zeta| = 1.$$

Тогда для любого ε существуют циркулянтная матрица C и матрица R ранга r такие, что

$$|(T - C - R)_{ij}| \leqslant |T_{ij}|\varepsilon,$$

причём

$$r \leqslant \log \varepsilon^{-1}[c_0 + c_1 \log \varepsilon^{-1} + c_2 \log n].$$

Для доказательства следствия достаточно заметить, что ряд Лорана рассматриваемой функции $f(z)$ имеет коэффициенты вида $f_k = k^{-\alpha}\rho^k$. После этого мы можем применить лемму 3.

Аналогичным способом доказывается следующее утверждение:

Следствие 3 Пусть тёплицева матрица T порождена символом

$$f = (z - \zeta)^\alpha \log(z - \zeta), \quad z = e^{ix}, \quad |\zeta| = 1, \quad \alpha \in \mathbb{N}.$$

Тогда для любого ε существуют циркулянтная матрица C и матрица R ранга r такие, что

$$|(T - C - R)_{ij}| \leqslant |T_{ij}|\varepsilon,$$

причём

$$r \leqslant \log \varepsilon^{-1}[c_0 + c_1 \log \varepsilon^{-1} + c_2 \log n] + 2\alpha.$$

Результаты данного раздела объединяет следующая

Теорема 3 Пусть тёплицева матрица T порождена кусочно-аналитическим символом вида

$$f = g + \sum_{\alpha=0}^{l} \sum_{k=0}^{m} A_{k\alpha}(z - \zeta_k)^\alpha \log(z - \zeta_k), \quad z = e^{it}, |\zeta_k| = 1,$$

где функция g является аналитической в кольце, содержащем $|z| = 1$. Тогда для любого ε существуют циркулянтная матрица C и матрица R такие, что

$$|(T - C - R)_{ij}| \leqslant |T_{ij}|\varepsilon,$$

причём

$$\operatorname{rank} R \leqslant \log \varepsilon^{-1}[c_0 + c_1 \log \varepsilon^{-1} + c_2 \log n] + c_3, \tag{1.7}$$

а c_0, c_1, c_2, c_3 не зависят от n и ε.

Здесь, кроме слагаемых с логарифмическими особенностями, добавлена ещё и функция, аналитическая в кольце, содержащем $|z| = 1$. Для такой функции коэффициенты ряда Фурье убывают экспоненциально, и соответствующие тёплицевы матрицы могут быть аппроксимированы ленточными тёплицевыми матрицами. Ленточные тёплицевы матрицы могут быть приближены суммой циркулянта и матрицей малого ранга очень просто: достаточно добавить ненулевые элементы в двух противоположных углах матрицы.

Теорема 2 утверждает, что тёплицевы матрицы, порожденные произвольным рациональным тригонометрическим символом являются точной суммой циркулянта и матрицы фиксированного (не зависящего от n) ранга. Теорема 3 относится к случаю, когда символ является суммой аналитической функции и функции с логарифмическими особенностями. Соответствующие тёплицевы матрицы могут быть аппроксимированы матрицей вида $C + R$ с очень высокой точностью. На первый взгляд может показаться, что класс таких символов достаточно узок. Однако, этот «шаблон» для символов *содержит все примеры в литературе, посвящённые построению суперлинейных предобуславливателей.* Действительно, функции вида $(z-\zeta_k)^\alpha \log(z-\zeta_k)$ имеют скачок в производной с номером α.

Для примера, рассмотрим символ $f = x^4$, опредённый на интервале $-\pi < x < \pi$ и продолженный как 2π–периодическая функция на все остальные x. Он имеет скачки в первой и третьей производной. Вычитая из f функции вида

$$A(z-\zeta)\log(z-\zeta) + B(z-\zeta)^3\log(z-\zeta),$$

где ζ — точка, в которой происходит скачок, а A и B пропорциональны величине скачков, мы получаем аналитическую функцию, которая может быть аппроксимирована тригонометрическими полиномами (которые приводят к ленточным тёплицевым матрицам) с экспоненциально убывающей (по степени полинома) ошибкой.

Суммируя всё вышеизложенное, можно сказать, что все функции с конечным числом скачков конечного порядка могут быть аппроксимированы суммой циркулянта и матрицы малого ранга с оценкой вида (1.7).

1.7. Численные эксперименты

Циркулянты, получающиеся в результате $C + R$ аппроксимации тёплицевых матриц, естественно использовать в качестве предобуславливателей в методе сопряжённых градиентов (когда это возможно)

или GMRES (во всех других случаях). Мы использовали эти циркулянты для тёплицевых матриц, порождённых следующими типичными символами (определёнными на интервале $-\pi < x < \pi$ и продолженными как 2π-периодические функции на все вещественные x):

(A) Положительно определённые эрмитовы тёплицевы матрицы:

(1) $f_1 = |x|$,
(2) $f_2 = x^2$,
(3) $f_3 = |x|^3$,
(4) $f_4 = x^4$,
(5) $f_5 = x^2(x-\pi)^2$,
(6) $f_6 = (x+\pi)^2$.

(B) Знаконеопределённые эрмитовы тёплицевы матрицы, взятые из [7]:

(7) $f_7 = x^2(x^2+1)\mathrm{sgn}(x)$,
(8) $f_8 = \mathrm{sgn}(x-\pi+2)\mathrm{sgn}(x+\pi-2)(\cos(x+2)+1)(\cos(x-2)+1)$,
(9) $f_9 = ((\frac{x}{\pi})^2-1)^2 - 0.9$.

(C) Неэрмитовы тёплицевы матрицы ($z = e^{ix}$):

(10) $f_{10}(z) = \frac{z^4-1}{(z-\frac{3}{2})(z-\frac{1}{2})}$,
(11) $f_{11}(z) = \frac{(z+1)^2(z-1)^2}{(z-\frac{3}{2})(z-\frac{1}{2})}$.

В Таблице 1.3 приведены соответствующие времена (в единицах времени, равных времени вычисления одного умножения тёплицевой матрицы на вектор) для ладейной схемы. Эти матрицы аппроксимированы с относительной точностью $\varepsilon = 10^{-7}$. Отметим, что на практике для различных символов могут требоваться различные точности. Если матрица не очень плохо обусловлена, мы можем использовать большее ε. Но для того, чтобы решать системы с плохообусловленной матрицей, мы должны взять ε достаточно малым.

n	128	256	512	1024
f_1	24	23	22	20
f_2	15	18	18	17
f_3	23	23	22	20
f_4	16	17	19	18
f_5	19	21	20	15
f_6	8	9	10	10
f_7	19	23	22	20
f_8	21	24	22	20
f_9	12	12	11	11
f_{10}	4	3	3	3
f_{11}	5	4	4	4

Таблица 1.3. Время (в matvecs) для построения $C + R$ аппроксимации ($\varepsilon = 10^{-7}$).

Тёплицевы матрицы, порождённые рассматриваемыми символами, удовлетворяют условиям Теорем 2 и 3, и поэтому могут быть хорошо аппроксимированы матрицами вида $C + R$. Для символов f_{10} и f_{11} оказалось, что невязка порядка машинной точности. На самом деле, это наблюдение привело нас к формулировке Теоремы 2, о которой нам не было известно до начала экспериментов. В Таблице 1.4 приведены ранги малоранговых поправок (матриц R в $C + R$ представлении), вычисленные с помощью «ладейной» схемы. Для краткости будем называть эти ранги «циркулянтными рангами».

n	128	256	512	1024
f_1	36	37	38	41
f_2	19	23	23	26
f_3	28	29	32	32
f_4	20	21	23	24
f_5	27	25	22	18
f_6	17	20	22	22
f_7	27	32	38	35
f_8	28	33	34	30
f_9	15	15	12	10
f_{10}	4	4	4	4
f_{11}	6	6	6	6

Таблица 1.4. «Циркулянтные ранги», ($\varepsilon = 10^{-7}$).

Определённый интерес представляет зависимость циркулянтных рангов от ε. Их типичное поведение приведено в Таблице 1.5.

Когда $C+R$ аппроксимация построена, мы используем C^{-1} в качестве явного предобуславливателя. Для случая симметричной и поло-

ε	10^{-4}	10^{-5}	10^{-6}	10^{-7}
Ранг	13	19	21	24

Таблица 1.5. Зависимость циркулянтного ранга от ε, $n = 512$, символ f_4.

жительно определённой матрицы важно, чтобы C тоже была симметричной и положительно определённой матрицей. Численные эксперименты показывают, что симметрия нашим алгоритмом поддерживается.

Однако, наши циркулянты иногда имеют нулевые или отрицательные собственные значения, что мешает их использованию в качестве предобуславливателя. В этом случае, мы исправляем их, меняя «неправильные» (нулевые или отрицательные) собственные значения на 1. Это малоранговая коррекция, которая делает циркулянты положительно определёнными. В Таблице 1.6 показано число отрицательных и нулевых собственных значений циркулянтов C в $C + R$ аппроксимации для семейства символов вида $|x|^k$, $k = 1, 2, 3, 4$. Оказалось, что это число не зависит от n — для других n результаты получились абсолютно такими же.

$\lvert x\rvert$	x^2	$\lvert x\rvert^3$	x^4
0	1	1	1

Таблица 1.6. Число отрицательных/нулевых собственных значений для построенных циркулянтов

Наконец, в Таблице 1.7 показано число итераций, требуемых для решения предобусловленной системы. Порог остановки итерационного метода был установлен на 10^{-6}.

1.8. Метод чёрных точек для произвольного шаблона

Область применения метода чёрных точек отнюдь не ограничивается лишь диагональными матрицами. Вместо диагональной матрицы можно рассматривать *произвольный шаблон разреженности* S. Тогда получается следующая задача:

$$A \approx S + R, \tag{1.8}$$

где S — разреженная матрица, а R — матрица малого ранга. Такие задачи часто встречаются в приложениях. Например, пусть A — неко-

n	128	256	512	1024
f_1	8	8	9	8
f_2	6	6	6	6
f_3	13	16	17	20
f_4	15	16	16	20
f_5	3	3	3	3
f_6	5	5	5	5
f_7	12	12	13	14
f_8	10	10	11	11
f_9	3	4	4	4
f_{10}	9	9	9	9
f_{11}	8	9	9	9

Таблица 1.7. Число итераций в методе сопряжённых градиентов

торый большой массив данных с *пропусками* — такая ситуация часто возникает в математической статистике при анализе рядов наблюдений. Требуется заполнить пропуски, исходя из того, что исходный массив был малого ранга. Для решения задач вида (1.8) предлагались различные методы, однако все они являются итерационными и имеют достаточно высокую вычислительную сложность. Мы же покажем, что метод чёрных точек позволяет получить приближение быстро и за конечное число операций (т.е. метод не является итерационным). Также очень интересена ситуация, когда нам *неизвестен шаблон матрицы* S. В этом случае не удаётся получить какие-либо теоретические результаты. Однако возможно построить эвристический алгоритм, который позволяет определить положение ненулевых элементов в матрице S (т.е. определить, какие элементы малоранговой матрицы были «испорчены»). На практике это может, например, давать возможность обнаруживать *ошибки* экспериментаторов, которые измерили элементы матрицы A. Однако вначале подробно рассмотрим случай, когда нам *известно* положение ненулей в S.

Схема метода чёрных точек для случая произвольного шаблона S практически не отличается от случая диагонального шаблона. Если ранг матрицы R равен r, то мы выбираем $r \times r$ подматрицу $\widehat{A}$ и вычисляем скелетное разложение вида

$$B = L\widehat{A}^{-1}U, \tag{1.9}$$

где L и U — матрицы размеров $n \times r$ и $r \times m$ из строк и столбцов исходной матрицы. Матрицы C и R содержат пропуски («чёрные точки»). Нетрудно понять, как будут расположены чёрные точки в матрице B. А именно, элемент (i, j) матрицы B будет известным, если в i-ой строке матрицы C нет чёрных точек и в j-ом столбце матрицы R

тоже нет чёрных точек. Для выбора подматрицы опять воспользуемся методом неполной крестовой аппроксимации. На каждом шаге к шаблону разреженности добавляются новые элементы. Однако, как было объяснено выше, на каждом шаге фактически «вырезаются» некоторые строки и столбцы. Поэтому удобно завести два булевых массива длины n и m для того, чтобы помечать строки и столбцы, которые нельзя использовать при исключении. Полностью описание алгоритма выглядит так.

Алгоритм 1 Пусть дана матрица A, шаблон разреженности $\mathcal{S}$ и допустимая точность аппроксимации ε. Требуется построить матрицу R как можно меньшего ранга r и разреженную матрицу S с шаблоном $\mathcal{S}$ так, чтобы

$$\|A - S - R\| \leqslant \|A\|\varepsilon.$$

1. *Инициализация:* $\mathcal{I} = \emptyset, \mathcal{J} = \emptyset, r = 0$.

2. *Найти ведущий элемент: Если* $r = 0$, *то*

$$(i^*, j^*) = \arg\max_{(i,j)\notin\mathcal{S}} |A|_{ij},$$

иначе

$$(i^*, j^*) = \arg\max_{(i,j)\notin\mathcal{S}, i\notin\mathcal{I}, j\notin\mathcal{J}} |A - UV^\top|_{ij}.$$

3. *Если* $r = 0$, *то*

$$B = A,$$

иначе

$$B = A - UV^\top.$$

4. *Посчитать крест:*

$$C_r = u_r v_r^\top, (u_r)_i = B_{ij^*}, (v_r)_j = \frac{B_{i^*j}}{B_{i^*j^*}}.$$

Если $r = 0$ *то*

$$U = [u_r], V = [v_r].$$

Иначе

$$U := [U, u_r], V := [V, v_r].$$

5. *Пересчитать ранг:*

$$r = r + 1.$$

6. *Пересчитать «чёрные списки»:*

$$\mathcal{I} = \mathcal{I} \cup \{i : (i, j^*) \in \mathcal{S}\}, \mathcal{J} = \mathcal{J} \cup \{j : (i^*, j) \in \mathcal{S}\}.$$

7. *Проверить критерий остановки, если он не выполнен, вернуться к шагу 1.*

Что же получается в результате работы алгоритма? А в результате работы алгоритма получается матрица

$$\widehat{A} = UV^\top,$$

ранга $(r + 1)$ в которой известна целая (как мы надеемся, большая!) прямоугольная подматрица размера

$$(n - \mathrm{card}(\mathcal{I})) \times (m - \mathrm{card}(\mathcal{J})),$$

($\mathrm{card}(\cdot)$ — число элементов в множестве).

Как же теперь найти оставшиеся «чёрные точки»? Тут есть два пути. Один, более сложный в программировании, и, возможно, менее устойчивый, состоит в следующем. Мы запускаем алгоритм ещё раз, но накладываем на ведущие элементы дополнительные ограничения:

- Если есть целиком известный столбец или строчка, выбирать ведущие элементы оттуда.
- Чтобы алгоритм после очередного «прогона» приводил к появлению новых известных элементов, требуем, чтобы на каждом новом шаге выбирались столбцы и строки, отличные от предыдущих.

Отметим важное отличие от случая диагонального шаблона. Там, после первого шага, мы находили $(n - 2r)$ полных столбцов и строчек. В случае произвольного шаблона $\mathcal{S}$ такой гарантии нет — скорее всего, даже мы не найдём ни одной полной строчки или столбца. Просто неизвестных элементов станет меньше. Однако от нескольких проходов алгоритма и выбора не совсем оптимальных ведущих элементов может, в принципе, накопиться погрешность.

Другой, гораздо более простой в реализации способ состоит в следующем. Вспомним, что выполняются равенства

$$A_{ij} = \sum_{\alpha=1}^{r+1} U_{i\alpha} V_{j\alpha}, i \notin \mathcal{I}, j \notin \mathcal{J}.$$

Тогда можно посчитать V с помощью алгоритма, основанного на простой идее. Для каждого $j \notin \mathcal{J}$ построим вектор

$$a_i = \begin{cases} 0, & i \in \mathcal{I}; \\ A_{ij}, & i \notin \mathcal{I}, \end{cases}$$

и матрицу

$$\widehat{U}_{i\alpha} = \begin{cases} 0, & i \in \mathcal{I}; \\ U_{i\alpha}, & i \notin \mathcal{I}. \end{cases}$$

Фактически мы просто «вырезали» неудобные нам строки из j-ого столбца матрицы A. Теперь для нахождения j-ой строчки матрицы V необходимо решить линейную переопределённую задачу

$$\widehat{U}v = a.$$

После этого нужно проделать аналогичную процедуру для нахождения матрицы U.

Какова цена этого «упрощения»? Если применять обычные методы типа QR-разложения, то потребуется nr^2 операций для нахождения одного столбца V. Несложно построить модификации, учитывающие происхождение матрицы $\widehat{U}$. Мы не будем обращать на это большого внимания, так как скорость вычислений метода в экспериментах оказалась достаточно высокой при $n \approx 1000-50000$. Так как мы работаем с полной матрицей, содержащей n^2 элементов, то главным ограничением становится не скорость работы алгоритма, а память, требуемая для хранения элементов матрицы. На данный момент размеры порядка десятков тысяч являются предельными. Достоинством нового метода является то, что он всегда сходится и является более устойчивым, то есть восстанавливает элементы матриц с меньшей погрешностью.

1.9. Неизвестный шаблон

Рассмотрим теперь более сложный случай, когда нам *неизвестен шаблон разреженности* $\mathcal{S}$. Что же делать? На самом деле понятно, что нам нужно. Необходимо найти такую матрицу малого ранга R, чтобы матрица $A-R$ была *максимально разреженной*. Однако у читателя может возникнуть естественный вопрос, а что же такое «разреженность»? Как её померить? В литературе имеется большое количество вариантов количественного определения разреженности. Мы остановимся на одном из них. Пусть дан некоторый вектор $x \in \mathbb{R}^n$. Тогда определим «разреженность» $S(x)$ по формуле

$$S(x) = \sum_{i=1}^{n} \log(|x_i| + \delta), \tag{1.10}$$

где δ — некоторый небольшой параметр. Теперь, когда у нас есть определение разреженности, мы можем сформулировать задачу:

$$R^* = \arg\min_R S(A - R). \tag{1.11}$$

Применим для решения задачи (1.11) метод Гаусса-Ньютона. Нетрудно показать, что на каждом итерационном шаге придётся решать минимизационную задачу вида

$$R^{(k+1)} = \arg\min_R \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} |a_{ij} - R_{ij}| \tag{1.12}$$

где веса w_{ij} вычисляются по матрице с предыдущей итерации $R^{(k)}$:

$$w_{ij} = \frac{1}{|a_{ij} - R_{ij}^{(k)}| + \delta}. \tag{1.13}$$

Фактически, мы получили *взвешенную задачу приближения матрицы матрицей малого ранга*:

$$||W \circ (A - R)|| \to \min, \tag{1.14}$$

с некоторым положительными весами $W = [w_{ij}]$. ($\circ$ — поэлементное произведение матриц). Очевидно, что задача приближения суммой разреженной матрицы и матрицы малого ранга может быть представлена в виде (1.14). Достаточно лишь положить

$$w_{ij} = \begin{cases} 1, & (i,j) \in \mathcal{S}; \\ 0, & \text{иначе.} \end{cases}$$

Сделаем теперь предположение, что матрица является суммой разреженной матрицы и матрицы малого ранга:

$$A = S + R,$$

и текущее приближение к решению достаточно близко к точному:

$$R^{(k)} \approx R.$$

Тогда заменить использованную выше норму любой удобной нормой, например, фробениусовой:

$$||W \circ (A - R)||_F \to \min.$$

Здесь опять развилка: возможно использование двух различных подходов. Первый состоит в том, что весовая матрица W фильтруется по некоторому порогу η:

$$w^*_{ij} = \begin{cases} 1, & w_{ij} > \eta * w_{max}; \\ 0, & \text{иначе.} \end{cases}$$

Здесь w_{max} — максимальный по модулю элемент матрицы W. После этого решается задача $S + R$ аппроксимации с известным шаблоном, задаваемым матрицей W^* с помощью метода чёрных точек. Задача состоит в том, чтобы не ошибиться с положением чёрных точек — это контролируется параметром η.

Другой подход состоит в использовании метода переменных направлений для минимизации функционала $||W \circ (A - R)||_F$. Вспоминим, что $R = UV^\top$, $U \in \mathbb{R}^{n\times r}, V \in \mathbb{R}^{m\times r}$. Метод переменных направлений состоит в следующем.

- Фиксируем U, находим V из минимизации квадратичного функционала $||W \circ (A - UV^\top)||$.
- Фиксируем V, находим U из минимизации квадратичного функционала $||W \circ (A - UV^\top)||$.

Теперь определимся с тем, как мы будем пересчитывать матрицу весов W. Оказывается, что при использовании фробениусовой нормы формула (1.13) уже не является столь эффективной и требует замены. Смысл этой формулы очевиден. Элементам с маленькими значениям невязки $(A - R)_{ij}$ соответствуют большие веса, а элементам с большими значениями невязки — маленькие. Оказывается, что функция

$$\frac{1}{|x| + \delta}$$

является более эффективной.

1.10. Выводы

В этой главе мы рассмотрели классическую задачу теории структрированных матриц — построение циркулянтных предобуславливателей для общих и тёплицевых матриц. Эта задача была переформулирована как задача аппроксимации матрицы суммой циркулянта и матрицы малого ранга. Последняя задача была полностью решена с помощью построенного метода чёрных точек. Для тёплицева случая

построен вариант метода линейной по размеру матрицы сложности, доказаны теоремы о существовании $C + R$ аппроксимации для тёплицевых матриц широкого класса. Метод чёрных точек оказался применим не только для построения циркулянтных предобуславливателей, но и для задачи $S + R$ аппроксимации (аппроксимации матрицы суммой разреженной матрицы плюс матрицы малого ранга). Такой метод был также построен в данной главе.

Глава 2. Нестандартные вейвлет-преобразования

2.1. Введение

В предыдущей главе мы рассмотрели задачу аппроксимации матрицы (тёплицевой или матрицы общего вида) суммой циркулянта и матрицы малого ранга. Циркулянтные матрицы диагонализуются с помощью преобразования Фурье. Однако в вычислениях оказывается полезным использовать другие быстрые преобразования для сжатия данных помимо преобразования Фурье, например так называемые вейвлет-преобразования (иногда вместо слова «вейвлеты» употребляют слова «локальные волны», «всплески»).

Классические вейвлеты и их применение в иерархическом анализе данных обычно связаны с равномерными сетками и использованием преобразования Фурье [51, 58]. Практический численный анализ приводит, как правило, к неравномерным сеткам. Построение функций и преобразований со свойствами классических вейвлетов в этом случае также возможно, но требует совершенно иной техники. В данной работе для построения нестандартных вейвлетов («нестандартность» связана с использованием неравномерных сеток) используются B-сплайны, построенные по неравномерной сетке на интервале. Нестандартные вейвлеты используются затем для построения быстрых дискретных преобразований вейвлетовского типа, служащих для «сжатия» данных. Последнее обеспечивается тем, что нестандартные вейвлеты строятся с заданным количеством нулевых моментов. Основным результатом главы является запись линейной системы для параметров, определяющих искомое преобразование(лифтинговых коэффициентов) через разделённые разности и явное её решение. Получены явные, удобные формулы для вычисления лифтинговых коэффициентов. При использовании полученных преобразований для сжатия матриц оказалось, что при заданной погрешности аппроксимации они обеспечивают более высокое сжатие данных, чем соответствующее вейвлет-преобразование Добеши.

2.2. Основные понятия и определения

Напомним определение B-сплайна (см.[47]).

Определение

Пусть заданы некоторые точки, среди которых по крайней мере 2 различных:

$y_0 \leqslant .... \leqslant y_{k+1}$, $y_0 < y_{k+1}$.

Тогда B-сплайном k-ого порядка, построенным по этим точкам, называется функция

$$B(x) = [y_0; ...; y_{k+1}](y - x)_+^k,$$

где через $[y_0; ...; y_{k+1}]$ обозначен оператор разделённой разности, взятой по точкам $y_0, ..., y_{k+1}$. Функция $(y - x)_+^k$ определяется как

$$(y - x)_+^k = \begin{cases} 0, y < x, \\ (y - x)^k, y \geqslant x. \end{cases}$$

Носителем B-сплайна является множество $supp(B) = [y_0, y_{k+1}]$. Теперь рассмотрим произвольную сетку

$$a = x_1 \leqslant x_2 \leqslant ... \leqslant x_{n+k+1} = b,$$

удовлетворяющую условию

$$x_i < x_{i+k+1}, i = 1, ..., n. \tag{2.1}$$

На этой сетке возможны совпадающие точки. Построим по этим точкам B-сплайны k-ого порядка.

$$B_i(x) = [x_i; ...; x_{i+k+1}](y - x)_+^k, \tag{2.2}$$

$$i = 1, ..., n.$$

Ввиду (2.1), носители этих сплайнов являются отрезками. Определим пространство V как линейную оболочку этих функций: $V = Span(B_1, ..., B_n)$. Размерность этого пространства равна n. Теперь введём «более грубые» B-сплайны. Для этого рассмотрим дополнительную сетку $\widetilde{x}_i$, $i = 1, ..., N + k + 1$, которая содержится в сетке $\{x_i\}$ (т.е. для любого номера i существует номер s такой, что $\widetilde{x}_i = x_s$). От этой сетки мы потребуем, чтобы её граничные точки совпадали с граничными точками исходной сетки:

$$x_1 = \widetilde{x}_1,$$

$$x_{n+k+1} = \widetilde{x}_{N+k+1}.$$

Пример

Если общее количество точек $n+k+1$ нечётно, в качестве $\widetilde{x}_i$ можно взять все нечётные точки:

$$\widetilde{x}_i = x_{2i-1}, 1 \leqslant i \leqslant (n+k+2)/2. \tag{2.3}$$

В этом случае $N = \frac{n-k}{2}$. Построим по сетке $\{\widetilde{x}_i\}$ неравномерные B-сплайны:

$$\widetilde{B}_i = [\widetilde{x}_i; ...; \widetilde{x}_{(i+k+1)}](y-x)^k_+,$$
$$i = 1, ..., N.$$

Рассмотрим пространство $\widetilde{V} = Span(\widetilde{B}_1, ..., \widetilde{B}_N)$. Размерность этого пространства равна N. Пространства V, $\widetilde{V}$ называются *масштабирующими пространствами*. Рассмотрим два примера масштабирующих пространств.

Пример 1

Для любой функции f из V $supp(f) \subset \bigcup_{i=1}^{n} supp(B_i) = [a, b]$. Если все точки x_i различны, то f непрерывна, и поэтому

$$f|_{x=a} = f|_{x=b} = 0.$$

Пример 2

Если мы хотим, чтобы функции из V не обращались в 0 на границе(например, в b), нужно выбрать $\{x_i\}$, $i = 1, ..., n+1$ различными, а $x_i = x_{n+1}$ при $i = n+2, ..., n+k+1$.

2.3. Вейвлет-пространство. Масштабирующие и лифтинговые коэффициенты.

Рассмотрим функцию $\widetilde{B}_i$. Её носителем является множество $supp(\widetilde{B}_i) = [\widetilde{x}_i, \widetilde{x}_{i+k+1}]$. Для простоты будем считать все эти точки различными. В силу определения дополнительной сетки существуют номера s_0, s_1 такие, что $\widetilde{x}_i = x_{s_0}$, $\widetilde{x}_{i+k+1} = x_{s_1}$. В случае, когда точки x_i различны, известно,что B-сплайны $B_{s_0}, ..., B_{s_1-k-1}$ образуют базис в пространстве $(k-1)$ раз дифференцируемых функций, которые на каждом отрезке $[x_i, x_{i+1}]$, $i = s_0, ..., s_1 - 1$ являются полиномами k-ой степени. $\widetilde{B}_i$ принадлежит этому пространству, поэтому

$$\widetilde{B}_i = \sum_{s=s_0}^{s_1-k-1} r_{is} B_s.$$

Коэффициенты r_{is} называются *масштабирующими коэффициентами (refinement coefficients)*. Это свойство неравномерных B-сплайнов является необходимым для того, чтобы использовать их при построении вейвлетов. Вейвлет-пространством W называется пространство, которое является дополнением (необязательно ортогональным) к пространству $\widetilde{V}$ в V.

$$\widetilde{V} \dotplus W = V$$

(прямая сумма). Его размерность равна $\dim W = \dim V - \dim \widetilde{V} = n - N$. Базис в W мы будем обозначать через $\{\widetilde{\psi}_i, i = 1, ..., n - N\}$. Рассмотрим какое-нибудь пространство W с известным базисом.

$$\widetilde{\psi}_i = \sum_s \beta_{is} B_s,$$

$$i = 1, ..., n - N.$$

Например, в примере (2.3) при $k = 1$ в качестве $\widetilde{\psi}_i$ можно выбрать функции $\widetilde{\psi}_i = B_{2i-1}, i = 1, ..., (n+1)/2$. Теперь построим новое вейвлет-пространство с базисом

$$\widetilde{\psi}_i = \sum_s \beta_{is} B_s - \sum_{j=j_{min}(i)}^{j_{max}(i)} \alpha_{ij} \widetilde{B}_j. \qquad (2.4)$$

Преобразование (2) является частным случаем общего преобразования, называемого лифтинговой схемой. Коэффициенты α_{ij} называются *лифтинговыми коэффициентами*. В данной работе эти коэффициенты выбираются так, чтобы $\{\widetilde{\psi}_i\}$ имели заданное количество нулевых моментов (p-ый момент функции f определяется как (f, x^p), где $(,)$ - скалярное произведение в $L_2(a, b)$). Подробно лифтинговая схема и различные способы выбора лифтинговых коэффициентов рассматриваются в [34].

2.4. Основная система

Потребуем теперь, чтобы функции $\widetilde{\psi}_i$ имели m нулевых моментов. Это означает, что

$$(\widetilde{\psi}_i, x^p) = 0, \quad p = 0, 1, \ldots, m,$$

где $(,) = (,)_{L_2(a,b)}$ - скалярное произведение в $L_2(a, b)$. Подставляя выражение для $\widetilde{\psi}_i$, получим следующую систему на лифтинговые ко-

эффициенты

$$\sum_s \beta_{is}(B_s, x^p) = \sum_{j=j_{min}}^{j_{max}} \alpha_{ij}(\widetilde{B}_j, x^p). \tag{2.5}$$

Для вычисления моментов B-сплайнов нам понадобится следующая

Лемма 5 Пусть $B(x)$ - B-сплайн k-ого порядка, построенный по произвольным (совпадающим или нет) точкам $a = y_0 \leqslant y_1 \leqslant \ldots \leqslant y_k \leqslant y_{k+1} = b$. Тогда p-ый момент сплайна вычисляется по формуле

$$\int_a^b B(x)x^p dx = \frac{k!p!}{(k+p+1)!}[y_0; \ldots; y_{k+1}]x^{k+p+1}. \tag{2.6}$$

Доказательство. Воспользуемся интегральным выражением для разделённых разностей(см. [45]): для любой $(k+1)$ раз дифференцируемой на $[a, b]$ функции f

$$[y_0; \ldots; y_{k+1}]f = \frac{1}{k!}\int_a^b B(x)f^{(k+1)}(x)dx. \tag{2.7}$$

Полагая в (2.7) $f(x) = x^{k+p+1}$ получим, что

$$\int_a^b B(x)x^p dx = \frac{k!p!}{(k+p+1)!}[y_0; \ldots; y_{k+1}]x^{k+p+1}.$$

Лемма доказана.

Используя лемму 5 и (2.5), запишем систему, которой удовлетворяют лифтинговые коэффициенты:

$$L_i x^{k+p+1} = \sum_j \alpha_{ij}[\widetilde{x}_j; \widetilde{x}_{(j+1)}; \ldots; \widetilde{x}_{(j+k+1)}]x^{k+p+1}, \tag{2.8}$$

$$0 \leqslant p \leqslant m.$$

Через L_i обозначен следующий оператор:

$$L_i = \sum_s \beta_{is}[x_s; \ldots; x_{s+k+1}].$$

В этой системе i - фиксировано. Индекс j меняется от j_{min} до j_{max}, причём для того,чтобы число неизвестных совпадало с числом уравнений необходимо условие $j_{max} = j_{min} + m$.

2.5. Решение основной системы

Займёмся теперь решением системы (2.8). Так как (2.8) должно выполняться при $0 \leqslant p \leqslant m$, то оно равносильно тому, что

$$L_i P(x) = \sum_j \alpha_{ij} [\widetilde{x}_j; \widetilde{x}_{(j+1)}; ...; \widetilde{x}_{(j+k+1)}] P(x) \tag{2.9}$$

для любого многочлена $P(x) = \sum_{s=0}^{m} a_s x^{s+k+1}$. Но так как разделённая разность $(k+1)$-го порядка, взятая от многочлена степени не выше k равна 0, то основная система эквивалентна уравнению(2.9). При этом $P(x)$ – уже произвольный многочлен степени $(m+k+1)$.

Найдём теперь такие многочлены $P_j(x)$, $j_{min} \leqslant j \leqslant j_{max}$, что

$$[\widetilde{x}_r; \widetilde{x}_{(r+1)}; ...; \widetilde{x}_{(r+k+1)}] P_j(x) = \delta_{rj}, \tag{2.10}$$

$$j_{min} \leqslant r \leqslant j_{max}.$$

Тогда, подставляя эти многочлены в (2.9), получим, что

$$\alpha_{ij} = L_i P_j(x)$$

Для решения системы (2.8) достаточно указать какие-нибудь многочлены $P_j(x)$. Пусть сначала все $\widetilde{x}_i, i = j_{min}, ..., j_{max} + k + 1$ различны. Тогда многочлен P_j однозначно определяется своими значениями в этих точках. Докажем следующую теорему.

Теорема 4 1. Многочлен $P_j(x)$ такой, что

$$P_j(\widetilde{x}_r) = \begin{cases} \widetilde{x}_j - \widetilde{x}_{j+1}, j_{min} \leqslant r \leqslant j \\ 0, j < r \leqslant j_{max} + 1, \end{cases} \tag{2.11}$$

удовлетворяет (2.10) при $k = 0$.

2. Многочлен $P_j(x)$ такой, что

$$P_j(\widetilde{x}_r) = \begin{cases} q(\widetilde{x}_r), j_{min} \leqslant r \leqslant j \\ 0, j < r \leqslant j_{max} + k + 1 \end{cases} \tag{2.12}$$

$$q(x) = (\widetilde{x}_j - \widetilde{x}_{(j+k+1)}) \prod_{l=j+1}^{j+k} (x - \widetilde{x}_l),$$

удовлетворяет (2.10) при $k \geqslant 1$.

Доказательство

1. То, что (2.11) удовлетворяет (2.10), проверяется непосредственной подстановкой.

2. Проверим, что (2.12) действительно даёт решение системы (2.10). Возможны 3 случая:

а) Пусть $j < r \leqslant j_{max}$. Тогда $P_j(\widetilde{x}_s) = 0$ для всех $s = r, ..., r + k + 1$. Поэтому

$$[\widetilde{x}_r; ...; \widetilde{x}_{(r+k+1)}]P_j(x) = 0.$$

б) Пусть $j_{min} \leqslant r \leqslant j - 1$. Заметим, что в силу построения $P_j(x)$, $P_j(\widetilde{x}_s) = q(\widetilde{x}_s)$ при всех $j_{min} \leqslant s \leqslant j + k$ (так как при $j < s \leqslant j + k$ $P_j(\widetilde{x}_s) = 0 = q(\widetilde{x}_s)$). Поэтому

$$[\widetilde{x}_r; ...; \widetilde{x}_{(r+k+1)}]P_j(x) = [\widetilde{x}_r; ...; \widetilde{x}_{(r+k+1)}]q(x) = 0,$$

так как разделённая разность порядка $k + 1$ от многочлена степени не выше k равна 0.

в) Пусть $r = j$. Тогда

$$[\widetilde{x}_r; ...; \widetilde{x}_{(r+k+1)}]P_j(x) = \frac{q(\widetilde{x}_j)}{\prod_{l=j+1}^{j+k+1}(\widetilde{x}_j - \widetilde{x}_l)} = 1.$$

Теорема доказана.

Запишем теперь многочлен $P_j(x)$ в виде, подходящим и для случая совпадающих узлов. $P_j(x)$ представим в виде

$$P_j(x) = \prod_{s=j+1}^{j_{max}+k+1} (x - \widetilde{x}_s)\widetilde{P}_j(x),$$

где многочлен $\widetilde{P}_j(x)$ задаётся интерполяционными условиями

$$\widetilde{P}_j(\widetilde{x}_r) = \frac{q(\widetilde{x}_r)}{\prod_{s=j+1}^{j_{max}+k+1}(\widetilde{x}_r - \widetilde{x}_s)} = f(\widetilde{x}_r), r = j_{min}, ..., j,$$

где

$$f(x) = \frac{q(x)}{\prod_{s=j+1}^{j_{max}+k+1}(x - \widetilde{x}_s)}.$$

При $k = 0$

$$f(x) = \frac{\widetilde{x}_j - \widetilde{x}_{j+1}}{\prod_{s=j+1}^{j_{max}+k+1}(x - \widetilde{x}_s)},$$

а при $k \geqslant 1$

$$f(x) = \frac{\widetilde{x}_j - \widetilde{x}_{j+k+1}}{\prod_{s=j+k+1}^{j_{max}+k+1}(x - \widetilde{x}_s)}.$$

Поэтому, если мы запишем $\widetilde{P}_j$ как интерполяционный многочлен Ньютона:

$$\widetilde{P}_j = \sum_{r=j_{min}}^{j} \prod_{s=j_{min}}^{r} \frac{x - \widetilde{x}_s}{x - \widetilde{x}_{j_{min}}} [\widetilde{x}_{j_{min}}; \ldots; \widetilde{x}_r] f(x), \tag{2.13}$$

мы получим формулу, верную и для совпадающих узлов. Действительно, так как $\widetilde{x}_{j+k+1} > \widetilde{x}_j$, все разделённые разности от функции f, входящие в выражение (2.13), определены и в случае совпадающих узлов. Поэтому выражение (2.13) для совпадающих узлов получается предельным переходом.

2.6. Нахождение масштабирующих коэффициентов

Покажем, как находить масштабирующие коэффициенты. Напомним, что масштабирующие коэффициенты определяются как коэффициенты разложения сплайнов $\widetilde{B}_i$ по исходным сплайнам :

$$\widetilde{B}_i = \sum_{s=s_0}^{s_1-k-1} r_{is} B_s$$

. Умножим это равенство на $x^p, 0 \leqslant p \leqslant s_1 - s_0 - k - 1$ и проинтегрируем от a до b. Ввиду (2.6) мы получим следующую систему на коэффициенты r_{is}:

$$[\widetilde{x}_i; \ldots; \widetilde{x}_{i+k+1}] x^{k+p+1} = \sum_{s=s_0}^{s_1-k-1} r_{is} [x_s; \ldots; x_{s+k+1}] x^{k+p+1},$$

$$0 \leqslant p \leqslant s_1 - s_0 - k - 1.$$

Эта система – система в точности того же вида, как и система (2.8). Здесь $m = s_1 - s_0 - k - 1$. Поэтому, используя полученные в предыдущем параграфе результаты, можно выписать формулы для масштабирующих коэффициентов.

2.7. Алгоритм вычисления вейвлет-преобразования

Доказанная теорема даёт нам следующий алгоритм вычисления вейвлет-преобразования.

Алгоритм 2 *Имея:*

- *сетку* $x_i, i = 1, \ldots, n + k + 1$,
- *подсетку* $\widetilde{x}_i, i = 1, \ldots, N + k + 1$,
- *массивы* $j_{min}(i),\ i = 1, \ldots, n - N$, *и* $\beta_{is},\ i = 1, \ldots, n - N,\ s = s_0, \ldots, s_1$,

вычислить лифтинговые коэффициент при помощи следующего псевдо-кода :

```
do i = 1, n − N
  do j = j_min(i), j_min(i) + m
    do r = j_min(i), j_min(i) + m + k + 1
      Вычислить q_j(x̃_r) .
    end do
    Вычислить P_j(x), используя формулы интерполяции
Ньютона.
    Вычислить α_ij = Σ_{s=s_0}^{s_1} β_is [x_i; ...; x_{i+k+1}] P_j(x) .
  end do
end do
```

Для построения дискретного преобразования мы берём вектор

$$a = [a_1, \ldots, a_n]^\top$$

и рассматриваем его как разложение некоторой функции $f \in V$:

$$f = \sum_{i=1}^{n} a_i \varphi_i,$$

где $\varphi_i,\ i = 1, \ldots, n$, представляют собой *дуальный* базис к базису $B_i \in V$. Обозначим через φ_i и ψ_i дуальные базисы к $\widetilde{B}_i$ и $\widetilde{\psi}_i$, соответственно. Тогда

$$f = \sum_{i=1}^{n} a_i \varphi_i = \sum_{i=1}^{N} c_i \varphi_i + \sum_{i=1}^{n-N} d_i \psi_i, \tag{2.14}$$

и требуемое преобразование выглядит как

$$a \to (c^T, d^T)^T,$$

$$c = [c_1, \ldots, c_N]^\top, \quad d = [d_1, \ldots, d_{n-N}]^\top.$$

Алгоритм 3 (Один уровень вейвлет-преобразования.)
Имея

- *Компоненты вектора* $a_i,\ i = 1, \ldots, n$,
- *Число нулевых моментов* m,
- *Порядок сплайна* k,
- *Массивы коэффициентов*
  $r_{is},\ i = 1 \leqslant i \leqslant N, \quad \alpha_{ij},\ 1 \leqslant i \leqslant N, \quad \beta_{is}, 1 \leqslant i \leqslant n - N$,
- *Массивы индексов* $j_{min}(i),\ 1 \leqslant i \leqslant n - N$,

вычислить компоненты преобразованного вектора $z_i,\ i = 1, \ldots, n$, *с помощью следующего кода* :

```
do i = 1, N
   z_i = Σ_s r_is a_s .
end do
do i = 1, n-N
   z_{i+N} = Σ_s β_is a_s − Σ_{j=j_min(i)}^{j_min(i)+m} α_ij z_j .
end do
```

Алгоритмы 2 и 3 реализуют один уровень вейвлет-преобразования. Для вычисления l-уровневого преобразования нужно применить их рекурсивно l раз. Для этого необходимо лишь задать последовательность вложенных сеток.

Также нам потребуются в дальнейшем *обратное вейвлет-преобразование* и *обратное транспонированное преобразование*. В общем случае, преобразование задаваемое алгоритмом 3 нельзя обратить «явно». Рассмотрим пока более подробно случай $k = 1$. Предположим для простоты, что общее число точек нечётно и $\widetilde{x}_i = x_{2i-1}$, $\beta_{is} = \delta_{(2i-1)s}$, $N = (n-1)/2$. В этом случае,

$$\widetilde{B}_i = r_{i1}B_{2i-1} + r_{i2}B_{2i} + r_{i3}B_{2i+1}.$$

Матрица преобразования может быть представлена в виде

$$W = LRP,$$

где P — матрица перестановки такая, что

$$P[a_1, \ldots, a_n]^\top = [a_2, a_4, \ldots, a_{n-1}, a_1, a_3, \ldots, a_n]^\top,$$

R — блочная матрица вида

$$R = \begin{bmatrix} D & B \\ 0 & I \end{bmatrix}, \quad D = \mathrm{diag}\,(r_{12}, ..., r_{(n-1)2}),$$

B — бидиагональная матрица размера $N \times (n - N)$ с элементами

$$B_{ii} = r_{i1}, \quad B_{i,i+1} = r_{i3}, \qquad i = 1, \ldots, N.$$

Матрица L — блочная матрица вида

$$L = \begin{bmatrix} I & 0 \\ -A & I \end{bmatrix},$$

где ненулевые элементы A являются лифтинговыми коэффициентами:

$$A_{ij} = \alpha_{ij}, \quad i = 1, \ldots, n - N, \quad j = j_{min}(i), ..., j_{min}(i) + m.$$

Обратное преобразование имеет следующий вид:

$$W^{-1} = P^{\top} R^{-1} L^{-1},$$

$$L^{-1} = \begin{bmatrix} I & 0 \\ A & I \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}B \\ 0 & I \end{bmatrix}.$$

Обратное транспонированное преобразование имеет вид

$$W^{-\top} = L^{-\top} R^{-\top} P, L^{-\top} = \begin{bmatrix} I & A^{\top} \\ 0 & I \end{bmatrix}, \quad R^{-\top} = \begin{bmatrix} D^{-1} & 0 \\ -B^{\top} D^{-1} & I \end{bmatrix}. \qquad (2.15)$$

Отметим, что матрицы D, A, B являются ленточными. Поэтому, W, W^{-1}, $W^{-\top}$ могут быть умножены на вектор за $O(n)$ операций.

Нетрудно также отказаться от ограничения $k = 1$. В этом случае, для вычисления обратного преобразования, придётся решать линейную систему с ленточной матрицей.

В заключение определим матричное вейвлет-преобразование матрицы Z размера $p \times p$ по формуле

$$\widetilde{Z} = WZW^{\top}. \qquad (2.16)$$

2.8. Численные эксперименты

В этом параграфе мы покажем, что нестандартные вейвлет-преобразования, адаптированные к заданной сетке превосходят преобразования Добеши. Будем сравнивать вейвлет-преобразование Добеши и нестандартное вейвлет-преобразование с парметрами $k = 1$ и $m = 4$. Для этого мы зануляем все элементы, которые меньше порога 10^{-6} и сравниваем число ненулевых элементов в каждой матрице. Вычислительная сложность нестандартных вейвлет-преобразований совпадает со сложностью вейвлет преобразований Добеши порядка 3.

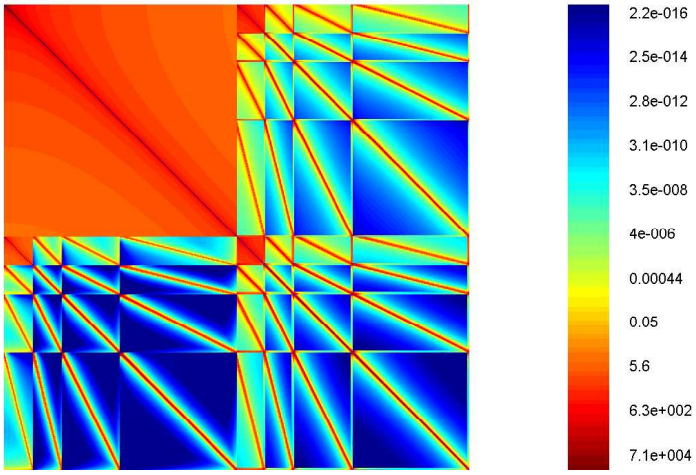

Рис. 2.1. Матрица из примера 1 с $p = 1$. Слева вверху: исходная матрица; Справа вверху: D-3; слева снизу: NS-4 справа: D-4.

2.8.1. Пример 1 На рисунке 2.1 показана $n \times n$ матрица

$$a_{ij} = \begin{cases} 0.0 & \text{если } i = j \\ 1/|x_i - x_j|^p & \text{иначе} \end{cases} \tag{2.17}$$

(где $p = 1$) определённая на сетке $x_i = 1 - \cos(i\pi/2n)$, вместе с её преобразованными версиями. Гладкость исходной матрицы (левый верхний угол рисунка) приводит к большому количеству маленьких элементов в преобразованной матрице. Нестандартные преобразования(левый нижний угол) приводят к гораздо большему количеству очень маленьких элементов чем и Добеши-3 (правый верхний угол) и Добеши 4 (правый нижний угол).

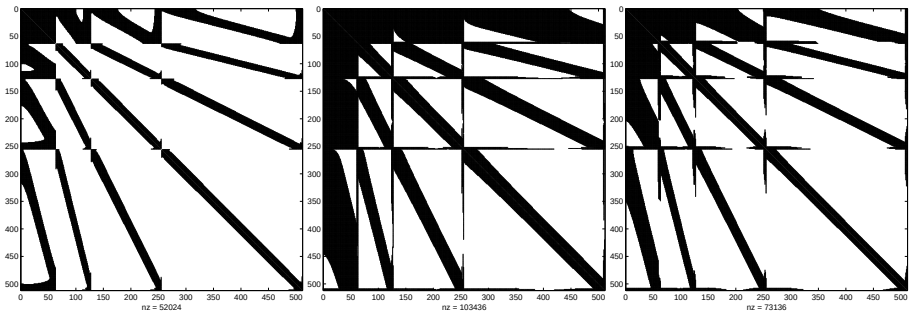

Рис. 2.2. Приближения к вейвлет-преобразованным матрицам из примера 1, с $p = 1$. Слева: новые, нестандартные преобразования с 4 нулевыми моментами; центр: D-3; справа: D-4.

Картина становится более наглядной, если мы сравним портреты разреженности при использовании порога 10^{-6}. На рисунке 2.2 показа-

ны матрицы, аппроксимированные с помощью преобразования Добеши и с помощью нестандартных вейвлетов. Число ненулевых элементов при использовании нестандартных вейвлетов в $1,5$ раза меньше, чем при использовании преобразования Добеши порядка 4, и в два раза меньше числа ненулевых элементов в матрице, преобразованной с помощью преобразования Добеши порядка 3. Это типично для других функциональных матриц на этой сетке. В частности, мы протестировали матрицы вида (2.17) для $p = 1/2, 1, 3/2, 2, 5/2$. В каждом случае, нестандартные вейвлет-преобразования давали существенное уменьшение числа ненулевых элементов при заданной точности. Это продемонстрировано на рисунке 2.3. На нём показана фробениусова норма ошибки в зависимости от степени сжатия (доли ненулевых элементов в разреженной матрице). Как мы можем видеть, нестандартные вейвлеты требует на 40%. меньше памяти для хранения преобразованной матрицы.

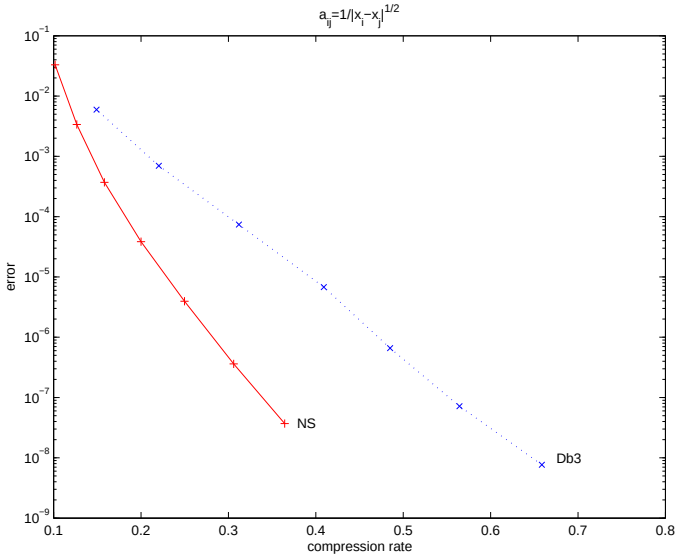

Рис. 2.3. Фробениусова норма ошибки в зависимости от степени сжатия для нестандартных вейвлетов и для преобразования Добеши-3 для матрицы из примера 1 с $p = 1/2$

2.8.2. Пример 2 Рассмотрим $n \times n$ матрицу

$$a_{ij} = \begin{cases} 0.0 & \text{если } i = j \\ -\ln|x_i - x_j| & \text{иначе.} \end{cases}$$

определённую на сетке $x_i = \ln(i)/\ln(n)$.

На рисунке 2.2 показаны результаты (с порогом 10^{-6}) для Добеши-3 (слева) и для нестандартных вейвлетов. Число ненулевых элементов

при использовании нестандартных вейвлетов в два раза меньше, чем при использовании преобразований Добеши.

Эксперименты с другими функциями на этой сетке дают похожие результаты с выигрышем 50% в числе ненулевых элементов.

Даже при использовании «фактора неортогональности», равного 10, (т.е. матрицы приближаются с порогом 10^{-7}), выигрыш от использования нестандартных вейвлетов составляет от 30% до 40%.

2.9. Выводы

В данной главе были построены новые, нестандартные вейвлет-преобразования для сжатия данных и матриц, связанных с известными неравномерными сетками. Получены явные, удобные формулы для вычисления параметров, определяющих искомое преобразование, причём это сделано для произвольной гладкости базисных функций и произвольного количества нулевых моментов. Проведённые численные эксперименты показали, что построенные преобразования дают существенный выигрыш в сжатии данных по сравнению с классическими преобразованиями Добеши.

Глава 3.
Тензорные аппроксимации матриц со структурированными факторами

3.1. Введение

В предыдущей главе подробно рассматривались задачи нелинейной аппроксимации, возникающие при построении матричных предобуславливателей вида $C + R, S + R$ и т.п. Перейдём теперь к абсолютно другой, на первый взгляд, задаче аппроксимации матриц — задаче тензорной аппроксимации. Прежде, скажем несколько слов о тех практических задачах, к которым мы собираемся применять методы тензорной аппроксимации. Это, в первую очередь, многомерные интегральные уравнения. В диссертации речь пойдёт в основном о двумерных интегральных уравнениях, однако будет обозначено, как обобщить полученные результаты на случай большего количества измерений.

В качестве приложения рассматривается задача потенциального обтекания прямоугольного крыла (квадрат $\Pi = [0,\ 1] \times [0,\ 1]$ в плоскости Oxy) установившимся потоком $V(x, y, z)$. Для потенциала возмущённой скорости Φ получаем краевую задачу (с условием непротекания на границе)

$$\triangle\Phi = 0; \qquad \frac{\partial\Phi}{\partial n} = -V_z, \quad x \in \Pi; \qquad \Phi,\ \nabla\Phi \to 0, \quad x \to \infty. \tag{3.1}$$

При поиске решения в виде потенциала двойного слоя получается следующее уравнение Прандтля:

$$-\int_0^1\int_0^1 \frac{U(x,y)}{[(x-x_0)^2+(y-y_0)^2]^{3/2}}dxdy = F(x_0,y_0) \quad (\ = V_z(x_0,y_0,0)\ ). \tag{3.2}$$

Интеграл в (3.2) понимается в смысле конечной части по Адамару.

Для численного решения уравнения (3.2) применяется метод дискретных вихрей [46],[52]. Введём сетки $0 \le x_0 \le \ldots \le x_p = 1, \quad 0 \le y_0 \le \ldots \le y_p = 1,$ и будем искать приближённое решение $u_p(x, y) \approx u(x, y)$ в виде кусочно-постоянной функции:

$$U_p(x,y) = \sum_{i=1}^{p}\sum_{j=1}^{p} u_{ij}\ \varphi_{ij}(x,y),$$

$$\varphi_{ij}(x,y)=\begin{cases}1, & \text{при} \quad (x,y)\in(x_{i-1},\ x_i)\times(y_{j-1},\ y_j),\\ 0 & \text{иначе.}\end{cases}$$

Выбрав точки коллокации $(x_{0i},\ y_{0j})\in(x_{i-1},\ x_i)\times(y_{j-1},\ y_j)$, получаем систему линейных алгебраических уравнений вида

$$\sum_{i=1}^{p}\sum_{j=1}^{p}u_{ij}a_{ij}^{km}=f_{km}\equiv F(x_{0k},y_{0m}),\quad 1\le k,m\le p, \tag{3.3}$$

где

$$a_{ij}^{km}=-\int_{x_{i-1}}^{x_i}\int_{y_{j-1}}^{y_j}\frac{dxdy}{[(x-x_{0k})^2+(y-y_{0m})^2]^{3/2}}=$$

$$=-\frac{\sqrt{(x_{0k}-x_{i-1})^2+(y_{0m}-y_j)^2}}{(x_{0k}-x_{i-1})(y_{0m}-y_j)}+\frac{\sqrt{(x_{0k}-x_i)^2+(y_{0m}-y_j)^2}}{(x_{0k}-x_i)(y_{0m}-y_j)}-$$

$$-\frac{\sqrt{(x_{0k}-x_i)^2+(y_{0m}-y_{j-1})^2}}{(x_{0k}-x_i)(y_{0m}-y_{j-1})}+\frac{\sqrt{(x_{0k}-x_{i-1})^2+(y_{0m}-y_{j-1})^2}}{(x_{0k}-x_{i-1})(y_{0m}-y_{j-1})}.$$

В матрично-векторной форме система (3.3) принимает вид

$$Au=f,$$

где $A=[a_{ij}^{km}]$ - двухуровневая матрица (согласно терминологии [48]), в которой элемент a_{ij}^{km} находится на пересечении строки $i+(j-1)p$ и столбца $k+(m-1)p$; в векторах $u=[u_{ij}]$ и $f=[f_{km}]$ элементы u_{ij} и f_{km} занимают позиции $i+(i-1)p$ и $k+(m-1)p$ соответственно.

Сходимость приближённого решения к единственному решению уравнения (3.2) была исследована в [53], но только для случая равномерных сеток. Было показано, что имеет место интегральная сходимость $\|U-U_p\|_{L_1(\Pi)}\to 0,\quad p\to\infty$, и равномерная сходимость на любом множестве точек на расстоянии не меньше δ от границы квадрата Π:

$$|U(x_{0k},y_{0m})-U_p(x_{0k},y_{0m})|<A_\delta|\ln h|^{9/4}h^{1/4-\varepsilon}, \tag{3.4}$$

где $0<\varepsilon<1/4$, а $h=1/p$ – шаг сетки. Оценка (3.4) указывает на то, что для достижения даже умеренной точности требуется решать линейные системы достаточно больших размеров.

Будем рассматривать два типа сеток:

(1) Равномерная сетка:

$$x_i=y_i=i/p,\quad i=0,1,\dots,p,$$

$$x_{0i}=y_{0i}=(i-0.5)/p,\quad i=1,2,\dots,p.$$

(2) Неравномерная чебышёвская сетка:

$$x_i = y_i = (1 - \cos\frac{\pi i}{p})/2, \quad i = 0, 1, \ldots, p,$$

$$x_{0i} = y_{0i} = (1 - \cos\frac{\pi(i - 0.5)}{p})/2, \quad i = 1, 2, \ldots, p.$$

Что можно сказать о точности оценки (3.4)? Для изучения этого вопроса путем численного эксперимента необходимо уметь решать системы (3.3) для достаточно больших p. Соответствующие результаты и выводы представлены в разд. 4. Для случая неравномерных сеток результатов о сходимости вообще нет, поэтому результаты численного эксперимента особенно интересны.

В нестационарных задачах аэрогидродинамики представляет интерес решение систем вида (3.3) для большого числа различных правых частей. Именно для таких задач важно получать достаточно точные структурированные аппроксимации к A^{-1} с малой памятью для хранения и быстрой процедурой умножения на вектор. Подобные аппроксимации могут использоваться и как явные предобусловливатели. Заметим, что в случае равномерной сетки матрица A оказывается дважды теплицевой [48] (блочно теплицевой с теплицевыми блоками). В этом случае для A^{-1} известны формулы Гохберга-Хайнига [12], но они содержат $O(n^{3/2})$ параметров и поэтому малополезны. В случае неравномерных сеток вообще неизвестно какое-либо явное описание строения матриц A и A^{-1}. Поэтому поиск «хороших» аппроксимаций и развитие соответствующих вычислительных технологий представляются очень перспективным направлением с точки зрения практических вычислений. Аппроксимации плотных матриц, возникающих при решении интегральных уравнений [15, 16, 18, 27, 33, 40, 38] основываются на разбиении исходной матрицы на блоки и аппроксимации отдельных блоков матрицами малого ранга. Однако в случае областей, являющихся тензорными произведениями одномерных удаётся избежать использования сложных многоуровневых блочных структур с помощью *тензорных аппроксимаций*. Объяснить это можно следующим образом. Такие методы, как мультипольный метод Рохлина, мозаично-скелетонный метод и другие основаны на разбиении матрицы по принципу *источник-приёмник*. При построении тензорных аппроксимаций используется разделение по геометрическим координатам. А именно, для матрицы A мы будем искать аппроксимацию в виде

$$A \approx A_r = \sum_{k=1}^{r} U_k \otimes V_k, \tag{3.5}$$

где $U \otimes V$ — тензорное (кронекерово) произведение матриц, определяемое как блочная матрица вида $[u_{ij}V]$. Покажем связь между (3.5) и разделением переменных. Запишем приближённое равенство (3.5) в индексной форме (нумеруя, как и договаривались, элементы матрицы A 4 индексами):

$$a_{i_1 i_2 j_1 j_2} \approx \sum_{k=1}^{r} U^{(k)}_{i_1 j_1} V^{(k)}_{i_2 j_2}.$$

Объединяя теперь пару $(i_1 j_1)$ в один общий индекс i и пару $(i_1 j_1)$ в общий индекс j получаем

$$\widehat{A}_{ij} = a_{i_1 i_2 j_1 j_2} \approx \sum_{k=1}^{r} U^{(k)}_{i} V^{(k)}_{j}. \tag{3.6}$$

Индексы, объединённые в пары, соотвествуют номерам по направлению x и y соответственно для «источников» и «приёмников». Нетрудно теперь увидеть в (3.6) задачу малоранговой аппроксимации матрицы $\widetilde{A}$. Для этой задачи у нас уже есть алгоритм неполной крестовой аппроксимации, позволяющий находить малоранговую аппроксимацию за $\mathcal{O}(Nr)$ вычислений элементов матрицы и $\mathcal{O}(Nr^2)$ дополнительных операций (здесь $N = n_1 n_2$).

После применения метода неполной крестовой аппроксимации мы можем сохранить матрицу в оперативной памяти. Однако этого недостаточно — необходимо ещё уметь умножать матрицу на вектор. Использование одного лишь тензорного формата потребует $\mathcal{O}(N^{3/2})$ операций, что уже довольно существенно при $N \sim 1000$. Поэтому нужно сжимать факторы! Один из возможных подходов состоит в построении $C + R$ аппроксимации каждого фактора или структурированных представлений, основанных на так называемом малом ранге смещения (этот подход будет подробно описан в Главе 4). Но это не единственный вариант. Второй предлагаемый подход основан на использовании *вейвлетов*. К каждому фактору применяется *вейвлет-преобразование* W, после чего фактор становится псевдоразреженным. Чуть ниже мы подробнее опишем этот шаг, а пока обсудим другой вопрос. Если мы умеем быстро умножать матрицу на вектор, то можно запустить любой удобный итерационный процесс. Однако, как известно, без использования *предобуславливателя* обычно требуется большое коли-

чество итераций. В этой главе мы предложим несколько методов предобуславливания: тензорный предобуславливатель (тензорного ранга 1), предобуславливатель основанный на неполном LU-разложении и двухуровневый циркулянтный преобуславливатель. Сразу скажем, что при использовании неравномерных сеток циркулянтный преобуславливатель работает плохо, и нужно использовать масштабирование.

Образованная таким образом тройка «тензоры + вейвлеты + преобуславливатель» позволяет очень эффективно решать двумерные интегральные уравнения. Мы рассматриваем два варианта предобуславливателей: масштабированный циркулянтный предобуславливатель и предобуславливатель, основанный на использовании метода Ньютона.

Предлагаемый алгоритм состоит из следующих этапов:

(A) Приближаем A суммой тензорных произведений

$$B = \sum_{k=1}^{r} U_k \otimes V_k, \tag{3.7}$$

$$\|B - A\|_F \leqslant \varepsilon \|A\|_F, \tag{3.8}$$

где U_k и V_k размеров $n_1 \times n_1$ and $n_2 \times n_2$ соответственно, и $N = n_1 n_2$ размер матрицы A. Для простоты изложения будем предполагать, что $n_1 = n_2 = n = N^{1/2}$.

(B) Применяем вейвлет преобразования с заданным числом нулевых моментов m к каждому фактору:

$$P_k = WU_kW^T, Q_k = WV_kW^T, 1 \leqslant k \leqslant r. \tag{3.9}$$

Здесь W — матрица вейвлет-преобразования, например преобразования Добеши или нестандартного преобразования, построенного нами. P_k и Q_k — *псевдоразреженные матрицы*. Выбирая подходящий порог $\tau = \tau(\varepsilon, P_k, Q_k)$ и занулив все элементы в P_k и Q_k меньшие по модулю, чем τ, мы приближаем матрицу B матрицей вида

$$C = W^{-T} \otimes W^{-T} D W \otimes W \approx B, D = \sum_{k=1}^{r} P_k^\tau \otimes Q_k^\tau. \tag{3.10}$$

Приближение строится с заданной точностью ε :

$$\|B - C\|_F \leqslant \varepsilon \|B\|_F. \tag{3.11}$$

(C) Построение предобуславливателя F^{-1} для матрицы A. Есть несколько вариантов.

(D) Применяем GMRES для решения системы

$$CF^{-1}y = b. \tag{3.12}$$

Возвращаем $F^{-1}y$ как приближение к точному решению x.

Пусть теперь $B = \sum_{k=1}^{r} U_k \otimes V_k$. Применим вейвлет-преобразование к каждому тензорному фактору:

$$P_k = WU_kW^\top, \quad Q_k = WV_kW^\top.$$

Затем мы выбираем порог τ и заменяем матрицы P_k и Q_k разреженными матрицами P_k^τ и Q_k^τ соответственно, получая аппроксимацию матрицы B:

$$C = \left(W^{-1} \otimes W^{-1}\right) \left(\sum_{k=1}^{r} P_k^\tau \otimes Q_k^\tau\right) \left(W^{-\top} \otimes W^{-\top}\right). \tag{3.13}$$

Легко проверить, что

$$\| \sum_{k=1}^{r} P_k \otimes Q_k - \sum_{k=1}^{r} P_k^\tau \otimes Q_k^\tau \|_F \leqslant \varepsilon_W \| \sum_{k=1}^{r} P_k \otimes Q_k \|_F, \tag{3.14}$$

где

$$\varepsilon_W = \frac{\sum_{k=1}^{r} (\|P_k - P_k^\tau\|_F \|Q_k\|_F + \|P_k\|_F \|Q_k - Q_k^\tau\|_F)}{\| \sum_{k=1}^{r} P_k \otimes Q_k \|_F}$$

легко вычисляется. Нестандартные преобразования *неортогональны*, поэтому мы должны писать

$$\|C - B\|_F \leqslant \gamma \varepsilon_W \|B\|_F, \tag{3.15}$$

где γ — «фактор неортогональности». Численные эксперименты показывают, что он порядка 10. Когда используются ортогональные преобразования, такие, как преобразования Добеши, $\gamma = 1$. Но ε_W оказывается гораздо меньше в случае нестандартных преобразований, и финальная ошибка аппроксимации меньше.

Если B — аппроксимация A такая, что

$$\|B - A\|_F \leqslant \varepsilon_K \|A\|_F,$$

то ошибка аппроксимации матрицы A матрицей C оценивается как

$$\|C - A\|_F \leqslant (\varepsilon_K + \gamma\varepsilon_W + \gamma\varepsilon_K\varepsilon_W). \tag{3.16}$$

Выбирая r и τ правильным образом, мы сможем поддерживать требуемую точность аппроксимации матрицы A. Важным результатом шага (B) является лучшее сжатие данных, но для нас это не самое главное. Основная цель этого шага — уменьшение сложности матрично-векторного произведения. Если ν обозначает число ненулей во всех P_k^τ и Q_k^τ, то C может быть умножено на вектор за $O(\nu\sqrt{n})$ операций.

3.2. Масштабированные циркулянтные предобуславливатели

С помощью вышеописанных преобразований матрица A может быть умножена на вектор быстро и с необходимой точностью. Будем использовать итерационный метод типа GMRES или PCG для решения линейной системы с матрицей A. Каждая итерация выполняется быстро, но для плохообусловленной матрицы таких итераций может быть очень много, особенно если мы переходим на неравномерные сетки. Нужен хороший предобуславливатель.

В работе [11] было предложено два варианта предобуславливателей (названные ILUT и IKT). При построении обоих предобуславливателей используется представление (3.7).Для построения ILUT использовалась неполная факторизация с динамическим выбором порога в духе [28]. Для построения IKP — разреженная обратная к первому (наибольшему по норме) слагаемому тензорного ранга 1 в сумме. Однако для неравномерных сеток IKP оказался неэффективным, а для ILUT потребовалось слишком много памяти.

Другая идея состоит в том, чтобы строить предобуславливатель прямо по матрице A, но используя лишь $O(n)$ её элементов. Мы предлагаем «воскресить» хорошо известные конструкции многоуровневых циркулянтных предобуславливателей (см [37, 32, 41]). Однако применение именно циркулянтов эффективно лишь на равномерных сетках. На неравномерных сетках мы предлагаем новый подход — использовать масштабирование. Сначала мы масштабируем матрицу

$$\hat{A} = D_1 A D_2 \tag{3.17}$$

с подходящими диагональными матрицами D_1 и D_2. Если диагональные элементы A положительны, мы можем выбрать $D_1 = D_2$ так, чтобы все диагональные элементы $\hat{A}$ были равны 1 (другие возможности связаны с выравниванием норм столбцов и строчек $\hat{A}$).

Оптимальный двухуровневый циркулянтный предобуславливатель Q для $\hat{A}$ это блочный циркулянт с циркулянтными блоками, удовлетворяющий

$$\|\hat{A} - Q\|_F = \min_{\hat{Q} \in \mathcal{T}} \|\hat{A} - \hat{Q}\|_F, \tag{3.18}$$

где $\mathcal{T}$ — множество всех двухуровневых циркулянтов с такими же размерами, как и матрица $\hat{A}$. Так как $\hat{A}$ является двухуровневой матрицей, элементы $\hat{a}_{ij}$ могут быть проиндексированы мультииндексами (i_1, j_1) и (i_2, j_2), где (i_1, j_1) определяет блок содержащий a_{ij}, а (i_2, j_2) определяет место этого элемента внутри блока. Элементы первого столбца $q_{(i_1,i_2)}$ оптимального циркулянта Q (он имеет длину p^2 и его элементы можно естественным образом пронумеровать парой индексов) выражаются по формуле [37]

$$q_{(i_1,i_2)} = \frac{1}{n} \left(\sum_{l=0}^{p-1} \sum_{k=0}^{p-1} \hat{a}_{(l,i_1+l),(k,i_2+k)} \right), \tag{3.19}$$

где элементы $\hat{A}$ считаются периодическими во всех 4 индексах.

Основной трудностью при вычислении Q по формуле (3.19) является то, что получается алгоритм сложности $O(n^2)$, что неприемлемо. Мы предлагаем строить *приближённую обратную матрицу*. В (3.19) мы вычисляем средние значения элементов, находящихся на i_2-ой периодический диагонали каждого блока вдоль i_1-ой периодической блочной диагонали. Естественно заменить среднее значение по всей последовательности на среднее значение по некоторой подвыборке с заданным шагом.

Когда Q построен, мы получаем предобуславливатель вида

$$M = D_1^{-1} Q D_2^{-1}. \tag{3.20}$$

Двухуровневые циркулянты диагонализуются двумерным преобразованием Фурье:

$$Q = \frac{1}{n} (F^* \otimes F^*) \Lambda (F \otimes F), \tag{3.21}$$

где F — матрица БПФ и Λ — диагональная матрица с собственными значениями. Следовательно,

$$M^{-1} = \frac{1}{n} D_2 (F^* \otimes F^*) \Lambda^{-1} (F \otimes F) D_1 \tag{3.22}$$

является *явным* предобуславливателем для матрицы A. Используя БПФ, мы можем умножать F на вектор за $O(n \log n)$ операций, и поэтому M^{-1} может быть умножена на вектор $O(n \log(n))$ операций.

3.3. Приближённое обращение структурированных матриц

Перейдём теперь к изложению одного из основых результатов диссертации — построению быстрых алгоритмов приближённого обращения структурированных матриц.

В вычислительной алгебре известен метод уточнения обратной матрицы (см. [57]), описанный Хотеллингом [20] и Шульцем [30]. Это не что иное, как метод Ньютона с k-й итерацией вида $\Phi'(X_{k-1})(X_k - X_{k-1}) = -\Phi(X_{k-1})$ для решения нелинейного уравнения

$$\Phi(X) \equiv A - X^{-1} = 0.$$

Очевидно, $\Phi(X + \delta X) - \Phi(X) = X^{-1} - (X + \delta X)^{-1} = X^{-1}\ \delta X\ (X + \delta X)^{-1}$, и поэтому $\Phi'(X_{k-1})\delta X = X_{k-1}^{-1}\ \delta X\ X_{k-1}^{-1}$. Следовательно, $X_k - X_{k-1} = -X_{k-1}(A - X_{k-1}^{-1})X_{k-1}$ и, окончательно, k-е приближение к A^{-1} имеет вид

$$X_k = 2X_{k-1} - X_{k-1}AX_{k-1} \tag{3.23}$$

и сводится к двум операциям умножения матриц.

Будем применять метод (3.23) для обращения матриц малого тензорного ранга:

$$A = \sum_{i=1}^{r} U_i \otimes V_i, \tag{3.24}$$

При перемножении матриц их тензорные ранги перемножаются. Поэтому на каждой итерации метода Ньютона мы ищем аппроксимацию к X_k с меньшим тензорным рангом. В итоге метод оказывается эффективным в тех случаях, когда A^{-1} допускает аппроксимацию малого тензорного ранга. Важно отметить, что в работе предлагается простая модификация метода Ньютона, намного более эффективная, чем (3.23), для таких случаев. Кроме того, для уменьшения вычислительной сложности кронекеровские сомножители аппроксимируются матрицами вида $WSW^\top$, где W - вейвлет-преобразование, а S - разреженная матрица.

3.4. Методы построения приближённой обратной матрицы

Начальное приближение X_0 к A^{-1} может быть быстро улучшено с помощью итераций вида (3.23). При этом легко проверить, что для невязок $R_k = I - AX_k$ выполняется соотношение $R_{k+1} = R_k^2$, доказывающее квадратичную сходимость метода Ньютона при условии $\rho(R_0) < 1$, где ρ – спектральный радиус. В качестве начального приближения

всегда можно взять $X_0 = \alpha A^*$ при некотором $\alpha > 0$. При этом оценка числа итераций для достижения точности $||A^{-1} - X_k||_2/||A^{-1}||_2 \leq \varepsilon$ имеет вид

$$\log_2(c^2 + 1) + \log_2 \ln(\frac{1}{\varepsilon}),$$

где c - спектральное число обусловленности матрицы A.

Применять метод Ньютона для обращения матриц общего вида (не обладающих какой-то структурой) не очень целесообразно: во-первых, каждый шаг требует двух дорогостоящих матричных умножений, во-вторых, метод Ньютона для плохо обусловленной матрицы может сходиться очень медленно. Если мы умеем быстро умножать матрицы с какой-либо структурой, то возникает вопрос о поддержании этой структуры в течение всех итераций.

3.4.1. Метод Ньютона с аппроксимациями Пусть $R(X)$ — нелинейный оператор в пространстве $(n \times n)$-матриц и M-константа Липшица для оператора $I - R$:

$$||[X - R(X)] - [Y - R(Y)]|| \leqslant M||X - Y||.$$

Будем говорить, что матрица $R(X)$ является аппроксимацией матрицы X.

На k-й итерации метода Ньютона с аппроксимациями осуществляется переход от $X_{k-1} = Z_{k-1}$ к новому приближению $X_k = R(X_k)$ (для каких-то матриц Z_{k-1} и Z_k):

$$Z_k = 2X_{k-1} - X_{k-1}AX_{k-1}, \qquad X_k = R(Z_k). \tag{3.25}$$

Теорема 5 *Предположим, что*

$$R(A^{-1}) = A^{-1}. \tag{3.26}$$

Тогда для всех достаточно близких к A^{-1} начальных приближений $X_0 = R(X_0)$ метод (3.25) порождает последовательность матриц X_k, сходящуюся к A^{-1} квадратично:

$$||A^{-1} - X_k|| \leq (1 + M)\,||A||\;||A^{-1} - X_{k-1}||^2, \qquad k = 1, 2, \ldots. \tag{3.27}$$

Доказательство. Справедливы соотношения

$$||X_k - Z_k|| = ||R(Z_k) - Z_k|| = ||R(Z_k) - R(A^{-1}) + A^{-1} - Z_k|| \;\leq\; M||Z_k - A^{-1}||.$$

Используя неравенство треугольника, получаем

$$||A^{-1} - X_k|| \;\leq\; ||A^{-1} - Z_k|| + ||Z_k - X_k|| \;\leq\; (1 + M)||A^{-1} - Z_k||. \tag{3.28}$$

Из уравнения (3.25) легко вывести, что $A^{-1}-Z_k = (A^{-1}-X_{k-1})\,A\,(A^{-1}-X_{k-1})$. Следовательно,

$$||A^{-1} - Z_k|| \;\le\; ||A||\,||A^{-1} - X_{k-1}||^2. \tag{3.29}$$

Неравенство (3.27) очевидным образом вытекает из (3.28) и (3.29). Теорема доказана.

Теорема 5 сводит исследование сходимости метода Ньютона к исследованию структуры обратной матрицы (которая выражается условием (3.26).

Рассмотрим различные примеры оператора $R(A)$. Пусть $\|\cdot\|$ обозначает унитарно инвариантную норму (например, спектральную или фробениусову норму) и $\Pi_r(A)$ – наилучшее для данной нормы приближение к A среди всех матриц ранга не выше r:

$$\rho_{r+1}(A) \equiv ||A - \Pi_r(A)|| = \min_{\mathrm{rank}B \le r} ||A - B||.$$

В случае спектральной нормы $\rho_{r+1}(A)$ равно $(r+1)$–ому (в порядке неубывания) сингулярному числу матрицы A, а для фробениусовой нормы – квадратному корню из суммы квадратов сингулярных чисел с номерами от $r+1$ до n.

Пусть теперь L – линейный обратимый оператор в пространстве $(n\times n)$-матриц. Теперь фиксируем некоторое r и будем рассматривать аппроксимации вида:

$$R(A) = L^{-1}(\Pi_r(L(A))). \tag{3.30}$$

Такие операторы можно использовать в методе Ньютона с аппроксимации.

Аппроксимации, понижающие тензорный ранг, являются частным случаем аппроксимаций вида (3.30). В самом деле, для двухуровневой матрицы

$$A = [a_{ij}^{km}], \quad 1 \le i,j,k,m \le p,$$

определим новую двухуровневую матрицу $L(A)$ следующим образом:

$$L(A) = [b_{ij}^{km}], \quad b_{ij}^{km} = a_{ik}^{jm}. \tag{3.31}$$

Тогда, как замечено в [43], тензорный ранг матрицы A будет совпадать с рангом матрицы $L(A)$, при этом задача оптимальной в норме Фробениуса аппроксимации A матрицей тензорного ранга не выше r сводится к аналогичной задаче аппроксимации $L(A)$ матрицей ранга не выше r (см. [54, 36]).

В качестве другого примера оператора $R(A)$ можно рассмотреть применение вейвлетовских преобразований с последующей спарсификацией:

$$R(A) = W^{-1} S_\tau(WAW^T) W^{-T}, \tag{3.32}$$

где S_τ-оператор спарсификации, все элементы по модулю меньшие τ заменяющий нулём.

Отметим еще один результат о сходимости метода Ньютона с аппроксимациями в случае, когда условие (3.26) нарушено. Теперь оценка для $\|Z_k - X_k\|$ принимаем вид

$$\|Z_k - X_k\| \;\le\; M\|Z_k - A^{-1}\| + \|L^{-1}\|\|A^{-1} - R(A^{-1})\|$$

Положим

$$c = (1+M)\,\|A\|, \quad \varepsilon = M\|A^{-1} - R(A^{-1})\|$$

Для величин $\delta_k \equiv \|A^{-1} - X_k\|$ находим

$$\delta_k \;\le\; c\,\delta_{k-1}^2 + \varepsilon. \tag{3.33}$$

Теорема 6 *Пусть m – наибольший номер такой, что $\sqrt{\varepsilon/c} \le \delta_{k-1}$ при всех $1 \le k \le m$. Если $2c\delta_0 < 1$, то метод (3.25) дает квадратичное уменьшение погрешности на итерациях $1 \le k \le m$:*

$$2c\delta_k \le (2c\delta_{k-1})^2, \tag{3.34}$$

и при этом $\delta_m \le 2\varepsilon$.

Доказательство. Неравенство $\sqrt{\varepsilon/c} \le \delta_{k-1}$ влечет за собой $c\varepsilon \le (c\delta_{k-1})^2$. Значит, $c\delta_k \le (c\delta_{k-1})^2 + c\varepsilon \le 2(c\delta_{k-1})^2$, откуда и получаем (3.34).

3.4.2. Модифицированный метод Ньютона Рассмотрим две матрицы в тензорном формате:

$$M^1 = \sum_{i=1}^{r_1} A_i^1 \otimes B_i^1, \qquad M^2 = \sum_{i=1}^{r_2} A_i^2 \otimes B_i^2.$$

Используя известные свойства тензорного произведения, получим

$$M^1 M^2 = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} (A_i^1 A_j^2) \otimes (B_i^1 B_j^2).$$

Поэтому сложность умножения матриц имеет вид $O(r_1 r_2 n^{3/2})$, что уже намного лучше, чем $O(n^3)$ при стандартном правиле умножения матриц. Для дальнейшего уменьшения вычислительных затрат

применим к A и начальному приближению X_0 двумерное вейвлет-преобразование:

$$\widetilde{A} = (W \otimes W)A(W^\top \otimes W^\top), \qquad \widetilde{X}_0 = (W \otimes W)X_0(W^\top \otimes W^\top).$$

(W - матрица одномерного вейвлет-преобразования) и начнём выполнять метод Ньютона с матрицами $\widetilde{A}$ и $\widetilde{X}_0$. Ожидается, что тензорные сомножители матриц X_k будут разреженными (это подтверждается численными экспериментами), вследствие чего сложность умножения матриц сильно понижается.

Тензорный формат сохраняется на каждой итерации метода Ньютона. Но тензорный ранг возводится в квадрат на каждой итерации, поэтому нужно найти способ уменьшить число слагаемых в тензорном представлении X_k. Как отмечалось выше, проблема аппроксимации X_k матрицей X_k^ε более низкого тензорного ранга с оценкой погрешности $\|X_k - X_k^\varepsilon\|_F \leqslant \varepsilon\|X_k\|_F$ сводится к нахождению малоранговой аппроксимации к матрице $L(Z_k)$. Поскольку ранг матрицы $L(Z_k)$ также мал, речь идет об аппроксимации малоранговой матрицы матрицей еще более малого ранга. Подобная задача эффективно решается с помощью процедуры, называемой *рекомпрессией* [38],[17]. Будем писать $X_k^\varepsilon = R_\varepsilon(X_k)$.

Однако прямое применение метода Ньютона (даже с рекомпрессией) требует все же большого объема вычислений. Действительно, пусть характерный тензорный ε-ранг матриц A и A^{-1} равен 15-20. Тогда на каждом шаге (когда X_k близко к A^{-1}) производится ~ 200 умножений разреженных матриц порядка p. Для уменьшения вычислительной сложности предлагается следующая модификация метода Ньютона:

$$X_k = X_{k-1}(2I - X_{k-1}), \quad Y_k = Y_{k-1}(2I - X_{k-1}), \qquad k = 1, 2, \ldots, \quad (3.35)$$

где $Y_0 = I$ и X_0 – невырожденная матрица такая, что спектральный радиус матрицы $I - X_0$ меньше 1. Последнее условие сразу же дает

$$\lim_{k\to\infty} X_k = I,$$

а поскольку из (3.35) легко вывести, что $X_{k+1} = X_0 Y_k$, получаем

$$\lim_{k\to\infty} Y_k = X_0^{-1}.$$

Матрица A системы (3.3) положительно определена, поэтому можно взять $X_0 = \alpha A$, где $\alpha < 2/\|A\|_2$. Тогда $\|I - X_0\|_2 < 1$. В данном

случае

$$\lim_{k\to\infty} Y_k = \frac{A^{-1}}{\alpha}.$$

Модифицированный метод Ньютона с аппроксимациями будет иметь такой вид:

$$X_k = R_\varepsilon(X_{k-1}(2I - X_{k-1})), \quad Y_k = R_\varepsilon(Y_{k-1}(2I - X_{k-1})), \qquad k = 1, 2, \ldots. \tag{3.36}$$

Почему предложенный вариант более эффективен, чем (3.25)? Поскольку матрицы X_k сходятся к единичной матрице, их тензорный ранг становится небольшим. Также возрастает разреженность сомножителей, поэтому матричных умножений становится меньше, а сложность каждого из них уменьшается – с точки зрения вычислительной сложности метод (3.36) является «ускоряющимся».

Результаты вычислений обратной матриц для различных p и для различных сеток приведены ниже в Таблицах 1 и 2.

$n = p^2$	4096	16384	65536	262144
ε	10^{-5}	10^{-5}	10^{-5}	10^{-5}
Тензорный ранг A	8	8	9	10
Время обращения	2.68 сек.	15.39 сек.	1.47 мин.	7.29 мин.
Число итераций	7	8	9	10
Тензорный ранг A^{-1}	14	15	15	15
Невязка	$4 \cdot 10^{-6}$	$3 \cdot 10^{-6}$	$1.5 \cdot 10^{-4}$	$1.1 \cdot 10^{-4}$

Таблица 3.1. Обращение матриц A в случае равномерной сетки

$n = p^2$	4096	16384	65536
ε	10^{-5}	10^{-5}	10^{-3}
Тензорный ранг A	15	16	18
Время обращения	18.1 сек.	1.4 мин	10 мин.
Число итераций	12	15	16
Тензорный ранг A^{-1}	21	21	21
Невязка	$8 \cdot 10^{-5}$	$4 \cdot 10^{-5}$	$8 \cdot 10^{-2}$

Таблица 3.2. Обращение матриц A в случае чебышёвской сетки.

Для сравнения обычного и модифицированного метода Ньютона с аппроксимациями заметим: при $p = 256$ в случае равномерной сетки для нахождения приближённой обратной по обычному методу Ньютона потребовалось 12 итераций, чтобы достичь невязки

$$\|AX_k - I\|_F = 4.88 \cdot 10^{-3}.$$

Вычисления заняли 4 минуты 35 секунд, что примерно в 3 раза дольше, чем в случае модифицированного метода Ньютона.

3.5. Численные результаты

3.5.1. Масштабированный циркулянтный преобуславливатель На данный момент неизвестно никаких аналитических решений уравнения Прандтля и отсутствуют теоремы о сходимости численных схем. Поэтому численный эксперимент для малых h (и больших n) может пролить свет на свойства этого уравнения и этой численной схемы.

n	16129	65025	261121	1046529
r	20	22	25	20
ε	$9.7 \cdot 10^{-8}$	$8.4 \cdot 10^{-8}$	$9.8 \cdot 10^{-8}$	$8.1 \cdot 10^{-6}$
Матрица-на-вектор	0.3 sec	1.6 sec	7.7 sec	15.6 sec
Число итераций	28	30	33	38
Построение предобуславливатель	4.0 sec	16.4 sec	1.1 min	4.4 min
Время решения	11.2 sec	59.4 sec	5.7 min	14.5 min
Относительная ошибка	$5.8 \cdot 10^{-7}$	$1.1 \cdot 10^{-6}$	$9.9 \cdot 10^{-7}$	$2.8 \cdot 10^{-5}$

Таблица 3.3. Результаты для неравномерной сетки.

Таблица 3.3 содержит некоторые результаты для чебышевской (неравномерной) сетки. «Относительная ошибка» это ошибка решения линейной системы. Для того, чтобы её померить мы? сделали правую часть равной сумме первого, пятого и десятого столбца матрицы A, так что мы знаем точное решение и можем вычислить точность.

В Таблице 3.4 мы сравниваем поведение GMRES с масштабированным циркулянтным предобуславливателем и без него.

n	16129	65025	261121	1046529
Число итераций	137	336	> 600	> 600
Число итераций(с предобусл.)	28	30	33	38

Таблица 3.4. Масштабированный циркулянтный предобуславливатель.

Сравним теперь сходимость приближённого решения к точному решению интегрального уравнения. Для этого возьмём специальную правую часть как результат применения интегрального оператора к функции $u(x, y) = x$. Интеграл вычисляется аналитически, однако результат крайне громоздок и занимает несколько страниц, поэтому мы его здесь не приводим. Поточечные ошибки решения на равномерной и чебышёвской сетках ($p = 255$) приведены на рисунке 3.1.

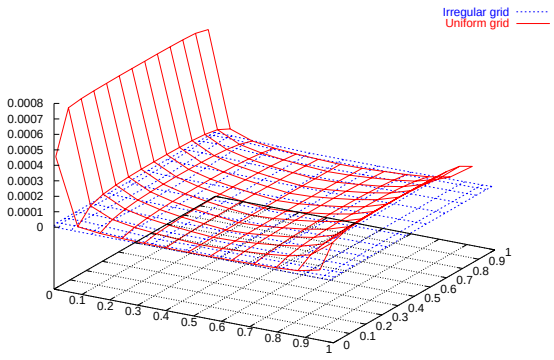

Рис. 3.1. Поточечная ошибка для различных сеток

Ошибка L_2 и в равномерной нормах представлены на рисунке 3.2. Очевидно, что неравномерная сетка даёт существенно лучшую сходимость.

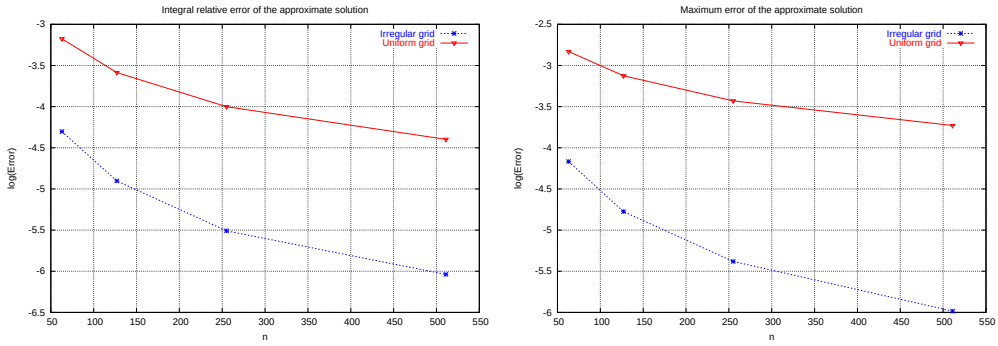

Рис. 3.2. L_2 и максимальная ошибки в зависимости от p

Проведённые эксперименты показывают, что неравномерная сетка существенно лучше, чем равномерная. Однако мы использовали очень специальную правую часть (даже неограниченную) так что полученные результаты не совсем подходят к физической постановке задачи (хотя они и очень обнадёживающие). Интересно, что из физических свойст задачи следует, что решение u должно равняться нулю на границе области для любой ограниченной f. Все результаты ниже получены для $f = 1$. Мы строим $-u$ вместо u. Рисунок 3.3 содержит результаты для неравномерной сетки и разного количества точек сетки. На рисунке 3.4 аналогичные результаты представлены для равномерной сетки.

В обоих случаях мы видим, что численное решение стабилизируется. Для получения численных оценок скорости сходимости, мы построили L_2 норму решения на рисунке 3.5 для обоих сеток. Явно видно, что неравномерная сетка даёт существенно более быструю сходимость.

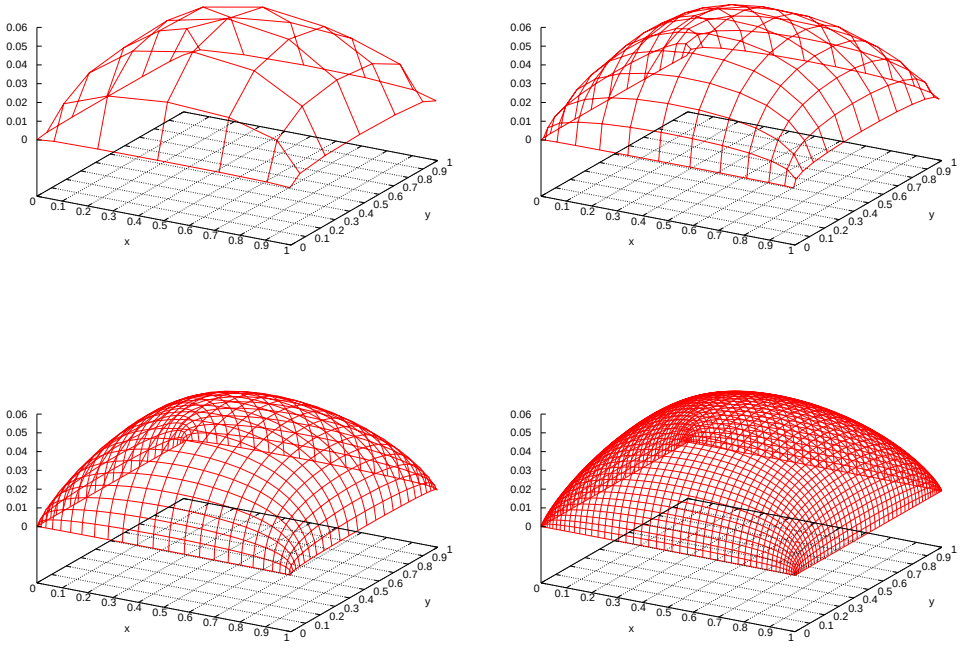

Рис. 3.3. Решения на неравномерной сетке ($p = 63, 127, 255, 511$).

3.5.2. Предобуславливатели на основе метода Ньютона Приведём результаты численных экспериментов с использованием приближённых обратных, построенных по методу Ньютона. Матрицы A аппроксимировались с точностью 10^{-7}, метод Ньютона проводился с точностью 10^{-5} для равномерной сетки, и с точностью 10^{-3} - для чебышёвской.

n	16129	65025	261121
Время построения предобусловливателя	18 сек	99 сек	342.3 сек
Число итераций	3	3	4
Время решения	0.4 сек	2 сек	13.5 сек

Таблица 3.5. Равномерная сетка

n	3969	16129	65025
Время построения предобусловливателя	18 сек	84 сек	737 сек
Число итераций	3	3	5
Время решения	0.2 сек	0.9 сек	6.6 сек

Таблица 3.6. Чебышёвская сетка

В качестве предобусловливателей можно использовать приближённые обратные не к матрице A, а к её аппроксимациям тензорного

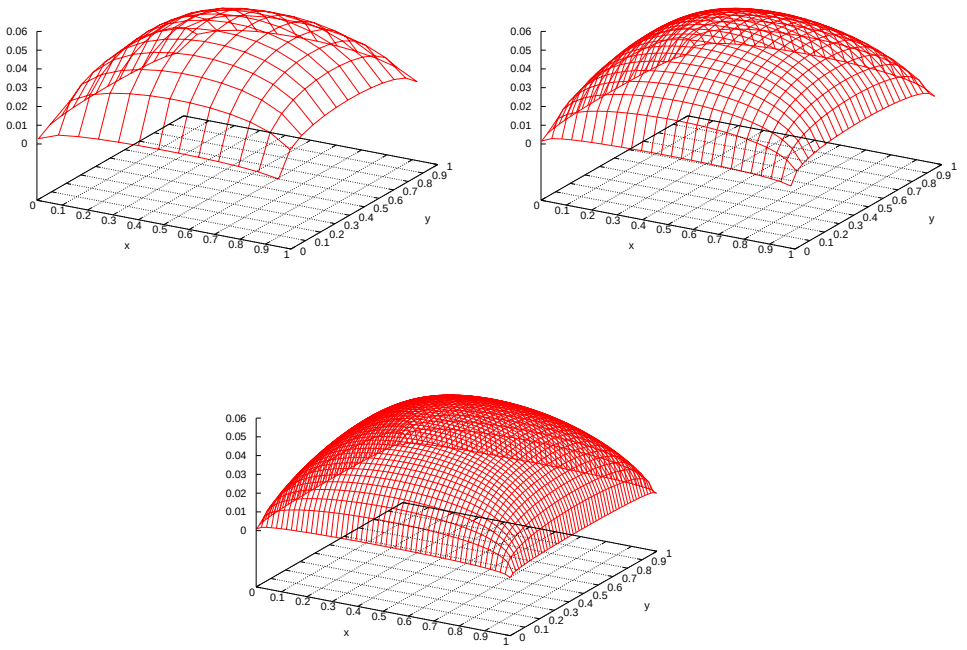

Рис. 3.4. Решение на равномерной сетке ($p = 127, 255, 511$).

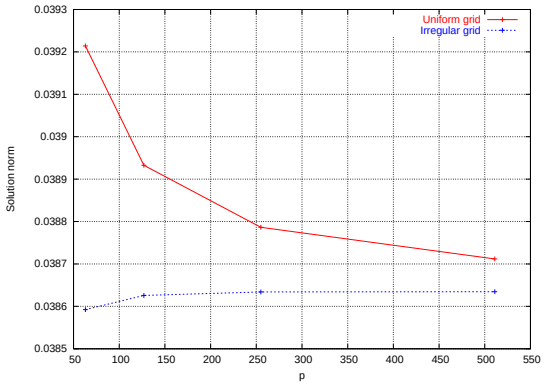

Рис. 3.5. L_2-норма решения

ранга r. В работе [11] был рассмотрен случай с $r = 1$ и соответствующий предобусловливатель был назван IKP-предобусловливателем (Incomplete Kronecker Product). Однако метод Ньютона позволяет находить приближённые обратные и при $r > 1$. Соответствующие численные результаты приведены в Таблице 3.7 для равномерной сетки при $p = 511$. Метод Ньютона проводился с точностью $\varepsilon = 10^{-3}$.

r	1	2	3	5
Время построения предобусловливателя	17 сек	30 сек	41 сек	58 сек
Число итераций	33	15	10	8
Время решения	82.1 сек	39.8 сек	27.7 сек	22.2 сек

Таблица 3.7. IKP-предобусловливатель с разными r

3.6. Выводы

В данной главе был предложен новый общий подход к построению алгоритмов приближённого обращения структурированных матриц большой размерности, возникающих при дискретизации интегральных уравнений. Основная идея — использование метода Ньютона обращения матриц. Доказана теорема о том, что метод Ньютона с аппроксимациями сохраняет квадратичную сходимость. Построена модификация метода Ньютона, которая оказалась существенно более быстрой при обращении структурированных матриц. Вся схема метода реализована для обращения матриц, представленных в виде суммы небольшого числа тензорных произведений с факторами, сохранёнными в вейвлетовском базисе. При этом в качестве вейвлет-преобразований существенно более эффективными оказались преобразования, построенные в предыдущей главе. В качестве приложения было решено с высокой точностью на очень мелких сетках интегральное уравнение Прандтля. При этом исследовались различные предобуславливатели: масштабированный циркулянтный преобуславливатели и предобуславливатель, основанный на метод Ньютона.

Глава 4.
Супер-быстрое обращение двухуровневых тёплицевых матриц

4.1. Введение

Применим теперь разработанный общий подход для обращения очень важного класса матриц — двухуровневых тёплицевых матриц. Эти матрицы имеют простую и запоминающуюся структуру и естественно ожидать, что существуют быстрые алгоритмы для работы с ними. Однако на данный момент не существует каких-либо удобных представлений для обратных к двухуровневым тёплицевым матрицам, таких как классическая формула Гохберга-Семенцула для обратной к тёплицевой матрице. Более того, существуют серьёзные основания полагать, что такой формулы может и не быть. Но никто не мешает нам использовать аппроксимации — как мы уже успели убедится, во многих практически важных случаях приближённые формулы успешно заменяют точные. Надо лишь указать достаточно широкий и удобный формат для приближённой обратной матрицы и реализовать стандартные матричные операции.

Мы будем рассматривать двухуровневые тёплицевы матрицы, которые являются блочно тёплицевыми с тёплицевыми блоками. Пусть p одновременно и блочный размер и размер блоков, тогда $n = p^2$ — размер матрицы и двухуровневая тёплицева матрица матрица опреляется $O(n)$ параметрами. Известная формула Гохберга-Хайнига [12] содержит $O(p^3) = O(n^{3/2})$ параметров, что много по сравнению с $O(n)$. Более эффективный подход основан, конечно, на использовании тензорных аппроксимаций (что мы и пытаемся продемонстировать в данной Главе). Он применим к двухуровневым тёплицевым матрицам малого тензорного ранга. Наилучшая аппроксимация матрицей малого тензорного ранга будет иметь, кроме этого, дополнительную структуру — факторы тензорного представления будут тёплицевыми. Как следствие, такие матрицы представляются не $O(n)$, а $O(\sqrt{n})$ параметрами. Поэтому следует ожидать, что приближённые обратные матрицы тоже описываются таким же числом параметров и могут быть, как будет показано далее, вычислены за $o(n)$ операций. Удивительно, но факт: этот специальный подкласс двухуровневых тёплицевых матриц включает в себя подавляющее количество практически интересных

матриц.

Будем применять метод Ньютона:

$$X_i = 2X_{i-1} - X_{i-1}AX_{i-1}, \quad i = 0, 1, \ldots, \tag{4.1}$$

где X_0 — некоторое начальное приближение к A^{-1}.

Для вычислений нам, как и ранее, потребуется два основных средства.

- Быстрый метод умножения матриц заданной структуры
- Метод поддержания структуры.

И, кроме того, мы ещё не определили тот тип структурированных матриц, которые мы будем использовать на промежуточных итерациях.

Первый пункт означает, что должен существовать некоторый быстрый алгоритм для умножения матриц X_k и A.

Однако если X_k не принадлежат к коммутативным алгебрам (циркулянты, диагональные матрицы и т.п.) то следующее приближение может быть «менее структурированным». Как следствие, вычислительная сложность матрично-матричного умножения растёт с номером итерации. Для того чтобы замедлить этот рост (а он может быть очень существенным), нам нужно сохранять эту структуру используя «грубую силу» — некий метод замены данной матрицы X_{k+1} на близкую матрицу с «лучшей структурой». Введём оператор нелинейного проектирования $R(X)$, действующий на пространстве $n \times n$ матриц, который и будет осуществлять требуемое приближение. Тогда получаются итерации следующего вида:

$$X_i = R(2X_{i-1} - X_{i-1}AX_{i-1}). \quad i = 0, 1, \ldots. \tag{4.2}$$

Как мы уже видели, такие итерации успешно применяются для обращения матриц различной структуры: матрицы малого ранга смещения, [4, 26] матриц малого тензорного ранга. Как было показано в главе 3, если обратная матрица структурирована, то итерации сохраняют квадратичную скорость сходимости.

Этот результат вдохновляет, так давайте же предложим на основе этого результата алгоритм вычисления приближённой обратной матрицы к заданной двухуровневой тёплицевой матрице. Основная идея состоит в том, чтобы совместить два эффективных представления матриц: матрицы малого тензорного ранга и матрицы малого ранга смещения. Для этого мы введём новый матричный формат — TDS формат (tensor displacement structure), и будем предполагать, что A

и A^{-1} должны быть в TDS формате, по крайней мере приближённо. Строгой теории, обосновывающей данной предположение с неулучшаемыми оценками у нас пока нет; однако некоторые теоретические соображения и результаты численных экспериментов подтверждают это. В любом случае, мы всегда можем проверить результат, полученный нашим алгоритмом — если невязка $\|AX - I\|$ получается небольшой, то всё хорошо. Однако полное доказательство и точные формулировки условий на тёплицевы матрицы пока скрыты от нас. Численные эксперименты на модельных задачах показывают, что сложность построенного алгоритма — $\mathcal{O}(\sqrt{n}\log^{\alpha} n)$.

Дальнейшее изложение в данной главе построено по следующей схеме. Сначала мы определяем TDS формат и описываем, как представлять двухуровневую тёплицевую матрицу в таком формате. Потом мы описываем все основные матричные операции в TDS формате (сложение, умножение) и, что самое важное, представляем быструю процедуру *рекомпрессии* (другими словами, определяем оператор R).

Затем мы ещё раз обсудим итерации Ньютона с обрезанием и модификацию, которая серьёзно ускоряет вычисления.

А в конце мы представим некоторые численные эксперименты.

4.2. TDS формат

Сначало вспомним общие обозначения, применяемые в теории многоуровневых матриц. Мы будем пользоваться обозначениям, введёнными в [37]. Возможны различные конструкции понятия «ранга смещения»; мы будем пользоваться подходом [19]. В этой статье содержится далеко идущее обобщение определения, впервые введённого в [21].

Определение 1 *Матрица* T *называется двухуровневой с вектором размеров* (n_1, n_2) *если в ней* $n_1 \times n_1$ *блоков и каждый блок имеет размер* $n_2 \times n_2$. *Такая матрица называется двухуровневой тёплицевой матрицей, если*

$$T = [a(\mathbf{i} - \mathbf{j})], \tag{4.3}$$

где $\mathbf{i} = (i_1, i_2)$ *и* $\mathbf{j} = (j_1, j_2)$ *мультииндексы, определяющие положение элемента в двухуровневой матрице:* (i_1, j_1) *определяет положение блока и* (i_2, j_2) *определяет положение элемента внутри блока.*

Определение 2 *Оператор* L *назовём оператором типа Сильвестра, если*

$$L(M) = \nabla_{A,B}(M) = AM - MB \tag{4.4}$$

и типа Штейна, если

$$L(M) = \triangle_{A,B}(M) = M - AMB. \tag{4.5}$$

Значение $\alpha \equiv \operatorname{rank}(L(M))$ *называется рангом смещения матрицы* M. *Любая из* $n \times \alpha$ *матриц* G *и* H *скелетного разложения*

$$L(M) = GH^{\top}$$

называется генератором M. *Матрица, заданная своими генераторами, называется матрицей (малого) ранга смещения. По самому своему определению, ранги смещения и генераторы матрицы зависят от выбора оператора смещения* L.

Мы будем использовать операторы типа Штейна. На самом деле большой разницы нет, но несколько более удобным (по разным причинам) нам представляется использования операторов именно такого типа. Разным классам структурированных матриц можно поставить в соответствие разные операторы смещения. Например, для тёплицевых матриц можно рассмотреть матрицы $Z_a, Z_b^{\top}$, где

$$Z_a = Z + ae_0e_{n-1}^{\top}, \qquad Z_b = Z + be_0e_{n-1}^{\top},$$

Z — матрица «сдвига вниз» и a, b — какие-то скаляры. Пусть $\triangle_{Z_a,Z_b^{\top}}(M) = GH^{\top}$ и $G = [g_1, ..., g_\alpha]$, $H = [h_1, ..., h_\alpha]$. Тогда

$$(1 - ab)M = \sum_{j=1}^{\alpha} Z_a(g_j)Z_b^{\top}(h_j). \tag{4.6}$$

Здесь $Z_a(g)$ и $Z_b(h)$ определяются следующим образом. Пусть c — скаляр и v — вектор; тогда $Z_c(v)$ это тёплицева матрица с элементами

$$(Z_c(v))_{ij} = \begin{cases} v_{i-j}, & i - j \geq 0, \\ c\, v_{n+i-j}, & i - j < 0. \end{cases}$$

Если M невырождена, то M^{-1} можно выразить по формуле такого же вида, как (4.6), которая в этом случае может быть проинтерпретирована как одно из возможных обобщений формулы Гохберга-Семенцула на матрицы типа тёплицевых. И в последней формуле, и

в (4.6), матрица суть сумма произведений тёплицевых матриц из различных алгебр; однако, в формуле Гохберга-Семенцула и в формуле (4.6) используются различные алгебры. Отметим также, что если M является тёплицевой матрицей, то $\alpha \leq 2$.

Определение 3 *Говорят, что матрица A находится в тензорном формате с тензорным рангом r, если*

$$A = \sum_{k=1}^{r} A_k^1 \otimes A_k^2. \tag{4.7}$$

Для заданной матрицы A, мы можем попытаться аппроксимировать её матрицей малого тензорного ранга. Напомним, как это делается. Пусть

$$\mathcal{V}_n(A) = [b_{(i_1,j_1)(i_2,j_2)}]$$

— двухуровневая матрица с вектором размеров (n_1, n_1) и (n_2, n_2), и мы определим её по правилу

$$b_{(i_1,j_1)(i_2,j_2)} = a_{(i_1,i_2)(j_1,j_2)}.$$

Как нетрудно видеть (и что уже, на самом деле, подмечалось в предыдущих главах) тензорный ранг A равен рангу $\mathcal{V}_n(A)$. Более того,

$$\|A - A_r\|_F = \|\mathcal{V}_n(A) - \mathcal{V}_n(A_r)\|_F,$$

что сводит задачу об оптимальном приближении матрицы матрицей малого тензорного ранга к задаче об оптимальной малоранговой аппроксимации. Последняя может быть решена, например, с помощью SVD или алгоритма бидиагонализации Ланцоша, или, что гораздо быстрее и эффективнее — методом неполной крестовой аппроксимации. Однако в случае двухуровневых тёплицевых матриц эту задачу можно решить гораздо проще [22] («проще» в смысле вычислительной сложности, а не алгоритмической реализации)

Пусть $T = [a(\mathbf{i} - \mathbf{j})]$ — двухуровневая тёплицева матрица, $\mathbf{i}, \mathbf{j}$ — двумерные мультииндексы. a зависит от только от разности $\mathbf{i} - \mathbf{j} = (i_1 - j_1, i_2 - j_2)$, т.е. её можно рассматривать как двумерный массив размера $(2n_1 - 1) \times (2n_2 - 1)$. Составим матрицу «свободных параметров»

$$W(A) = [a_{\mu\nu}], \quad 1 - n_1 \leq \mu \leq n_1 - 1,\ 1 - n_2 \leq \nu \leq n_2 - 1, \tag{4.8}$$

построим для неё наилучшую аппроксмацию ранга r

$$W(A) \approx \sum_{k=1}^{r} u_k v_k^\top,$$

$$U^k = [u^k_{i_1-j_1}], \quad 0 \le i_1, j_1 \le n_1 - 1,$$
$$V^k = [v^k_{i_2-j_2}], \quad 0 \le i_2, j_2 \le n_2 - 1,$$

и построим тензорную аппроксимацию вида

$$T \approx T_r = \sum_{k=1}^{r} U^k \otimes V^k. \tag{4.9}$$

Можно показать, что это — оптимальное приближение матрицей тензорного ранга r во фробениусовой норме. Вычислительная сложность алгоритма — вычисление малоранговой аппроксимации к матрице размера $(2n_1 - 1) \times (2n_2 - 1)$. С помощью метода неполной крестовой аппроксимации это можно сделать за $\mathcal{O}(n)$ операций. Что замечательно и приятно, так это то, что тензорные факторы тоже являются (одноуровневыми) тёплицевыми матрицами. Важнейшим параметром, определяющим эффективность алгоритма, является тензорный ранг r. От чего он зависит? Оказывается, что он зависит от двух параметров: от требуемой точности аппроксимации и размера матриц n. Сама зависимость полностью определяется свойствами символа (порождающей функции) матрицы T. Можно показать, что если символ матрицы T обладает определёнными свойствами гладкости (так называемая асимптотическая гладкость), то соответствующие двухуровневые тёплицевы матрицы могут быть хорошо аппроксимированы суммой тензорных произведений тёплицевых матриц. Легко видеть, что такой формат разрушится сразу после первой же итерации метода Ньютона — поэтому мы погрузим этот формат в другой, более общий, который уже будет в некотором смысле «инвариантен» относительно итераций метода Ньютона (если на секундочку отвлечься от проблем с ростом рангов и т.д. и т.п.)

Определение 4 *Будем говорить, что двухуровневая матрица A находится в TDS (tensor-displacement structure) формате, если она одновременно в тензорном формате (4.7) и при этом каждый фактор тензорного представления является матрицей малого ранга смещения, заданной своими генераторами.*

Пусть r — тензорный ранг, s — максимальный ранг смещения в факторах. Нетрудно видеть, что для хранения матрицы в TDS формате требуется $O(\sqrt{n}rs)$ ячеек памяти.

4.3. Арифметика TDS формата

4.3.1. Основные арифметические операции Пусть матрицы A и B имеют тёплицевы ранги смещения α и β соответственно. Хорошо известно, что

- Матрично-векторное произведение Ax может быть вычислено за $O(\alpha n \log n)$ операций;
- Матрично-матричное произведение AB может быть вычислено за $O(\alpha\beta n \log n)$ операций, причём ранг смещения матрицы AB будет не больше, чем $\alpha + \beta$.

4.4. Основные арифметические операции в тензорном формате

Пусть две матрицы M_1 и M_2 находятся в тензорном формате

$$M^1 = \sum_{i=1}^{r_1} A_i^1 \otimes B_i^1, \quad M^2 = \sum_{i=1}^{r_2} A_i^2 \otimes B_i^2,$$

тогда произведение

$$M^1 M^2 = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (A_i^1 A_j^2) \otimes (B_i^1 B_j^2)$$

также находится в тензорном формате. Однако, для хранения результата требуется больше памяти, так как тензорные ранги перемножаются. Но для хранения результата требуется больше памяти, так как тензорные ранги перемножаются. Сумма двух тензорно-структурированных матриц тоже является матрицей в тензорном формате — тензорные ранги здесь складываются (даже ничего и делать не надо, лишь объединить два массива).

Как видно, тензорные ранги растут, и растут существенно. Поэтому нужно предложить какой-то разумный алгоритм «обрезания», т.е. аппроксимацией заданной матрицы малого тензорного ранга матрицей меньшего тензорного ранга с некоторой заданной наперёд точностью. Это задача может быть решена очень эффективно с помощью процедуры *à la* SVD, называемой рекомпрессией. Так как проблема вычисления аппроксимации малого тензорного ранга к матрице A эквивалентна проблеме вычисления малоранговой аппроксимации к матрицы $\mathcal{V}_n(A)$, мы можем использовать известную процедуру рекомпрессии для аппроксимации малого ранга, которая выглядит следующим образом.

Пусть дана матрица малого ранга $B = UV^\top$, $U, V \in \mathbb{R}^{n\times r}$, мы можем найти $q \leqslant r$ и матрицы $\widetilde{U}, \widetilde{V}^\top \in \mathbb{R}^{n\times q}$, приближающие A с требуемой точностью ε:

$$\|B - \widetilde{U}\widetilde{V}^\top\|_F \leqslant \varepsilon\|B\|_F. \tag{4.10}$$

Всё, что необходимо сделать, это найти SVD матрицы B. Так как B уже в малоранговом формате, мы действуем следующим образом:

(1) Найдём QR-разложение матриц U и V: $U = Q_u R_u, V = Q_v R_v$;

(2) Найдём SVD $r \times r$ матрицы $R_u R_v^\top$: $R_u R_v^\top = U_1 \Sigma V_1^\top$.

После этих операций $(Q_u U_1)\Sigma(Q_v V_1)$ — сингулярное разложение матрицы B. Теперь, возьмём наименьшее возможное q такое, что

$$\sigma_{q+1}^2 + \ldots + \sigma_r^2 \leqslant \varepsilon\|B\|_F.$$

Когда r мало, самая дорогая часть метода — это вычисление QR-разложений, сложность которых $O(nr^2)$, т.е. линейна по размеру матрицы. Но напомним, что столбцы U и V получаются из переставленных тензорных факторов, имеющие малый ранг смещения. Помогает ли это проводить рекомпрессию быстрее? Ответ на этот вопрос положительный, и и в следующем параграфе мы опишем требуемый алгоритм.

4.4.1. TDS-рекомпрессия Посмотрим более внимательно на шаги рекомпрессии. QR-разложение можно реализовать через процесс ортогонализации Грама-Шмидта, применённый к векторам $u_1, \ldots, u_r$. Ортогональность определяется с помощью обычного скалярного произведения $(x, y) = \sum_{k=1}^{n} x_i \overline{y}_i$. Теперь, вместо того, чтобы работать с векторами, будем работать напрямую с их матричными прототипами. В качестве скалярного произведения для матриц нужно использовать *фробениусово скалярное произведение*:

$$(A, B)_F = \mathrm{tr}(AB^*).$$

Другие операции, требуемые в алгоритме Грама–Шмидта — сложения векторов и умножения на числа — могут быть осуществлены прямо над соответствующими матрицами, без перехода к их векторному представлению. Более того, использование представления матриц как матриц с малым рангом смещения приводит к сложности $O(\sqrt{n}\log n)$.

Нам осталось решить лишь один вопрос. Для данных $p \times p$ матриц A и B с рангами смещения α и β нам нужно найти $\mathrm{tr}(AB^*)$. Для этого сначала вычислим произведение AB^*. Это — важное отличие от случая общих матриц, так как для них вычисление произведения напрямую совсем нецелесообразно, оно имеет сложность $O(n^3)$,

а сложность вычисления фробениусова скалярного произведения, как нетрудно видеть, $O(n^2)$. Для матриц с малым рангом смещения всё не так. Произведение AB^* может быть вычислено за $O((\alpha+\beta)p\log p)$ операций и ранг смещения произведения не превышает $\alpha+\beta$. И остался один последний шаг — научиться вычислять след матрица типа тёплицевой быстро. Получим простую формулу для этого следа.

Лемма 6 *Пусть* C — $p\times p$ *матрица* $\triangle_{Z_a,Z_b^\top}(C) = GH^T$, $G = [g^1,...,g^\alpha]$, $H = [h^1,...,h^\alpha]$, *где* $h^i, g^i \in \mathbb{R}^p$. *Тогда*

$$\mathrm{tr}(C) = \frac{1}{1-ab}\sum_{r=1}^{\alpha}\sum_{k=0}^{p-1} h_k^r g_k^r (p-k+abk). \tag{4.11}$$

Доказательство. Из (4.6) следует, что матрица C может быть представлена как

$$C = \frac{1}{1-ab}\sum_{j=1}^{\alpha} Z_a(g_j)Z_b^\top(h_j).$$

Поэтому,

$$\mathrm{tr}(C) = \frac{1}{1-ab}\sum_{j=1}^{\alpha} \mathrm{tr}(Z_a(g_j)Z_b^\top(h_j)). \tag{4.12}$$

Каждый член в (4.12) имеет вид

$$\mathrm{tr}(Z_a(g)Z_b^\top(h)) = \sum_{i=0}^{p-1}(Z_a(g)Z_b^\top(h))_{ii} = \sum_{i=0}^{p-1}\sum_{k=0}^{p-1} Z_a(g)_{ik}Z_b(h)_{ik} =$$

$$= \sum_{i=0}^{p-1}\sum_{k=0}^{i} g_{i-k}h_{i-k} + ab\sum_{i=0}^{p-1}\sum_{k=i+1}^{p-1} g_{p+i-k}h_{p+i-k}.$$

Первое слагаемое преобразуем как

$$\sum_{i=0}^{p-1}\sum_{k=0}^{i} g_{i-k}h_{i-k} = \sum_{i=0}^{p-1}\sum_{k=0}^{i} g_k h_k = \sum_{k=0}^{p-1} h_k g_k (p-k),$$

второе слагаемое преобразуем аналогично:

$$\sum_{i=0}^{p-1}\sum_{k=i+1}^{p-1} g_{p+i-k}h_{p+i-k} = \sum_{k=0}^{p-1} k h_k g_k.$$

4.4.2. Оператор обрезания Оператор обрезания $R(X)$ может быть определён двояким образом. Мы можем либо наложить ограничения на требуемую точность, либо на получаемые ранги. Конкретный выбор оператора сложно обосновать строго и приходится пользоваться различными эвристическими подходами. Если мы фиксируем ранги, то вычисление $R_{\rho,s}(X)$ происходит по следующей схеме:

(1) Находим наилучшую аппроксимацию X_ρ тензорного ранга ρ к матрице X, используя быстрый алгоритм рекомпрессии.

(2) Приближаем тензорные факторы некоторыми матрицами с рангом смещения s.

Можно проверить, что такой оператор удовлетворяет условиям нашей основной теоремы о сходимости. Поэтому метод Ньютона с оператором обрезания $R_{\rho,s}(X)$ сохраняет локальную квадратичную сходимость.

На практике, однако, бывает полезнее разрешить рангу меняться, но при этом зафиксировав требуемую точность. Соответствующий оператор обозначим через R_ε. Формально, шаги метода такие же, но ранги больше не являются постоянными. Первый шаг заканчивается построением малоранговой аппроксимации X_r, удовлетворяющей условию $||X - X_r|| \leq \varepsilon||X||$, и при этом имеющей минимально возможный тензорный ранг r. Второй шаг строит аппроксимации к факторам с наименьшим возможным суммарным рангом смещения, но при этом сохраняя требуемую точность.

4.5. Метод Ньютона и выбор начального приближения

Далее будем применять модифицированный метод Ньютона с обрезанием. Обсудим ещё один важный вопрос — как выбирать начальное приближение X_0 для A^{-1}. Особенно этот вопрос актуален для плохообусловленных матриц. Есть универсальный рецепт: взять $X_0 = \alpha A^*$ с подходящим $\alpha > 0$. Однако это достаточно плохой выбор. Но мы можем варьировать по своему усмотрению параметр точности ε. Для структурированных матриц он контролирует и окончательную точность результата, и ошибку обрезания, вносимую на каждой итерации. Поэтому он «регулирует» и ранги после и скорость вычислений. Когда итерационный процесс далёк от области быстрой сходимости, мы можем провести обрезание с гораздо большим значением параметра ε. Поэтому матричные операции становятся очень быстрыми на начальных итерациях. Потом ε должен уменьшаться и в конце стать равным требуемой финальной точности. Весь процесс можно описать следующим образом.

(1) Положим $X_0 = \alpha A^*$ и будем проводить вычисления с грубым порогом $\delta \gg \varepsilon$. В результате, мы получим грубое приближение M к обратной матрице, но преимущество состоит в том, что обрезанные по порогу δ Ньютоновские итерации имеют низкую вычислительную стоимость.

(2) Используем M как начальное приближение для новой последовательности Ньютоновских итераций с более высокой точностью ε.

Естественно, эта схема может быть распространена на три и более шагов с относительными ошибками δ_1, δ_2 и так далее. Скорее всего, наиболее эффективным было бы некоторое правило «непрерывного» изменения параметра δ для достижения оптимального времени счёта. Однако на данный момент это не реализовано не в последнюю очередь благодаря тому, что простая схема с двумя порогами работает вполне удовлетворительно.

4.6. Численные результаты

Решим теперь две модельных численных задачи. Для простоты будем полагать, что $n_1 = n_2 = \sqrt{n}$.

Первая — стандартный 5-точечный Лаплас. Это двухуровневая тёплицева матрица $[a_{i-j}]$ со свободными параметрами a_{ij} определяемыми как $a_{ij} = 0$, для $-n_1 + 1 \leqslant i \leqslant n_1 - 1$, $j = -n_2 + 1 \leqslant j \leqslant n_2 - 1$ кроме

$$a_{00}, \quad a_{0,\pm1} = -1, \quad a_{\pm1,0} = -1.$$

Второй — плотная двухуровневая тёплицева матрица, для которой a_{ij} задаются формулами

$$a_{ij} = -f(i+0.5, j-0.5)+f(i-0.5, j-0.5)-f(i-0.5, j+0.5)+f(i+0.5, j+0.5),$$

где

$$f(x, y) = \frac{\sqrt{x^2 + y^2}}{xy}.$$

Это — уже хорошо знакомая нам матрица, возникающая при решении уравнения Прандтля.

Результаты представлены в Таблицах 4.1 и 4.2. Мы вычислили тензорные ранги для приближённой обратной с $\varepsilon = 10^{-5}$ (это значит, что «тензорный ранг» и «средний ранг смещения» в этих таблицах означают на самом деле ε-ранг).

n	64^2	128^2	256^2	512^2
Время счёта	154 сек	333 сек	966 сек	2555 сек
Тензорный ранг A^{-1}	9	10	11	12
Средний ранг смещения A^{-1}	13.5	13.5	16.8	18.6

Таблица 4.1. Численные результаты для случая 1

n	64^2	128^2	256^2	512^2
Время счёта	270 сек	433 сек	817 сек	1710 сек
Тензорный ранг A^{-1}	13	13	12	11
Средний ранг смещения A^{-1}	8.5	9.3	9.5	9.7

Таблица 4.2. Численные результаты для случая 2

4.7. Структура обратных к двухуровневым матрицам специального вида

Представленые предыдущем пункте алгоритмы обращения двухуровневых тёплицевых матриц на основе тензорных аппроксимаций обнадёживают. Обнадёживают и результаты численных экспериментов. Однако хотелось бы найти им некоторое теоретическое обоснование. Например, почему обратная к матрице малого тензорного ранга тоже может быть приближена матрицей малого тензорного ранга? Возможно ли какое-нибудь матричное объяснение этому факту? На данный момент известны лишь тривиальные формулы вида

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

Для случая же двух факторов можно показать, что тензорный ранг в общем случае не превосходит n (а не n^2), а для трёх и более факторов тензорный ранг обратной может быть полным (т.е. n^2). Поэтому необходимо наложить дополнительные ограничения на структуру факторов. Рассмотрим матрицу вида

$$A = I + D \otimes R + R \otimes D, \tag{4.13}$$

где D — диагональная, а $R = uu^\top$ — матрица ранга 1. Тогда можно сформулировать следующую теорему.

Теорема 7 *Пусть матрица A обратима. Тогда тензорный ранг A^{-1} не превосходит 5.*

Отметим, что этот факт был сначала обнаружен экспериментально. Доказательству этого факта, естественным обобщениям и будет посвящено дальнейшее изложение.

4.7.1. Так почему же 5? Попытаемся получить явные формулы для обратной матрицы. Для этого нам необходимо уметь решать системы вида

$$Ax = e_i \otimes e_j.$$

В таком случае x будет соответствующим столбцом обратной матрицы. Стандартным образом заменим вектор x длины n^2 на матрицу размера $n \times n$. Тогда, используя свойства тензорного произведения, запишем линейную систему с матрицей A в виде

$$X + DXuu^\top + uu^\top D = e_i e_j^\top. \tag{4.14}$$

Отсюда легко видеть, что ранг матрицы X не превосходит 3. Однако этого ещё совершенно недостаточно для того, чтобы получить оценку именно на *тензорный ранг обратной матрицы к матрице* A. Поэтому продолжим. Введём векторы

$$x = DXu, y = DX^\top u.$$

Тогда уравнение (4.14) перепишется в виде

$$X = e_i e_j^\top - xu^\top - uy^\top.$$

Отсюда получаем уравнения на x, y:

$$x = DXu = D(e_i e_j^\top - xu^\top - uy^\top)u,$$

$$y = DX^\top u = D(e_j e_i^\top - yu^\top - ux^\top)u.$$

Перепишем их в матричной форме. Для этого воспользуемся равенствами

$$y^\top u = u^\top y, x^\top u = u^\top x$$

и получим, что

$$\begin{pmatrix} I + (u,u)D & Duu^\top \\ Duu^\top & I + (u,u)D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} d_i e_i u_j \\ d_j e_j u_i \end{pmatrix}.$$

Фактически, мы уже получили систему уравнений с достаточно простой матрицей — диагональная матрица плюс матрица ранга 2. Однако, для упрощения дальнейших вычислений, воспользуемся тем, что

матрица является блочно-циркулянтной. Поэтому её можно «диагонализовать», введя систему обозначений $\Lambda = I + (u,u)D$, $a = d_i e_i u_j$, $b = d_j e_j u_i$, и

$$\begin{pmatrix} \Lambda + Duu^\top & 0 \\ 0 & \Lambda - Duu^\top \end{pmatrix} \begin{pmatrix} I & I \\ I & -I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} I & I \\ I & -I \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix},$$

откуда

$$(\Lambda + Duu^\top)(x+y) = (a+b), \quad (\Lambda - Duu^\top)(x-y) = (a-b).$$

Воспользуемся теперь формулой Шермана-Вудбери-Моррисона:

$$x + y = (\Lambda + Duu^\top)^{-1}(a+b) = (\Lambda^{-1} - \Lambda^{-1}Du(1+\gamma)^{-1}u^\top\Lambda^{-1})(a+b),$$

$$x - y = (\Lambda - Duu^\top)^{-1}(a-b) = (\Lambda^{-1} + \Lambda^{-1}Du(1-\gamma)^{-1}u^\top\Lambda^{-1})(a-b),$$

где $\gamma = u^\top\Lambda^{-1}Du$. Введём дополнительно диагональную матрицу $\widehat{\Lambda} = \Lambda^{-1}D = \operatorname{diag}(\widehat{\lambda}_i)$ и векторы

$$\widehat{a} = \Lambda^{-1}a = \frac{d_i}{\lambda_i}u_j e_i, \widehat{b} = \Lambda^{-1}b = \frac{d_j}{\lambda_j}u_i e_j.$$

Используя эти обозначения получим, что

$$x + y = (I - \widehat{\Lambda}uu^\top\frac{1}{1+\gamma})(\widehat{a}+\widehat{b}) = \widehat{a}+\widehat{b} - \widehat{\Lambda}u\frac{1}{1+\gamma}\left(\widehat{\lambda}_i u_i u_j + \widehat{\lambda}_j u_i u_j\right),$$

$$x - y = (I + \widehat{\Lambda}uu^\top\frac{1}{1-\gamma})(\widehat{a}-\widehat{b}) = \widehat{a}-\widehat{b} + \widehat{\Lambda}u\frac{1}{1-\gamma}\left(\widehat{\lambda}_i u_i u_j - \widehat{\lambda}_j u_i u_j\right).$$

Осталось только выразить x и y:

$$x = \widehat{a} - \widehat{\Lambda}u\frac{u_i u_j}{1-\gamma^2}(\widehat{\lambda}_j - \gamma\widehat{\lambda}_i),$$

$$y = \widehat{b} - \widehat{\Lambda}u\frac{u_i u_j}{1-\gamma^2}(\widehat{\lambda}_i - \gamma\widehat{\lambda}_j).$$

Подставим теперь полученные x, y в выражение для X (4.14)

$$X = e_i e_j^\top - \widehat{a}u^\top - u\widehat{b}^\top + \widehat{\Lambda}uu^\top\frac{u_i u_j}{1-\gamma^2}(\widehat{\lambda}_i + \widehat{\lambda}_j)(1-\gamma),$$

или, проводя сокращения,

$$X = e_i e_j^\top - \widehat{a}u^\top - u\widehat{b}^\top + \widehat{\Lambda}uu^\top\frac{u_i u_j}{1+\gamma}(\widehat{\lambda}_i + \widehat{\lambda}_j).$$

Сделаем последний шаг, перейдя от X к формуле для обратной матрицы. Опуская детали, выпишем окончательную формулу:

$$A^{-1} = I - \widehat{\Lambda} \otimes uu^\top - uu^\top \otimes \widehat{\Lambda} + \frac{1}{1+\gamma}\widehat{\Lambda}uu^\top \otimes uu^\top \widehat{\Lambda} + \frac{1}{1+\gamma}uu^\top \widehat{\Lambda} \otimes \widehat{\Lambda}uu^\top. \tag{4.15}$$

Видно, что тензорный ранг обратной матрицы не превосходит 5. Более того, мы получили эффективный алгоритм нахождения обратной матрицы, так как единственным «параметром», требующим нахождения, является диагональная матрица $\widehat{\Lambda}$.

Интересно получить другой вывод формулы (4.15), или, хотя бы, первых её слагаемых. Позднее этот вывод пригодится нам в более сложном случае.

Будем искать приближение (вскоре поясним, в каком смысле мы понимаем здесь «приближение») к обратной матрице в виде

$$B = I + \Lambda \otimes uu^\top + uu^\top \otimes \Lambda.$$

(Здесь Λ не имеет никакого отношения к Λ, использовавшимися ранее). Умножим A на B:

$$AB = I + D \otimes uu^\top + \Lambda \otimes uu^\top + uu^\top \otimes \Lambda + (u,u)D\Lambda \otimes uu^\top + (u,u)uu^\top \otimes D\Lambda + (\ldots)),$$

где в $(\ldots))$ собраны слагаемые вида $Duu^\top \otimes uu^\top \Lambda$, ранг которых не зависит от n. Теперь поясним, из каких соображений мы будем подбирать Λ. Потребуем, чтобы произведение AB отличалось от единичной матрицы лишь на матрицу малого ранга, который при этом не зависит от n. Нетрудно заметить, что для этого достаточно потребовать, чтобы сократились «главные члены», т.е.

$$D \otimes uu^\top + (u,u)D\Lambda \otimes uu^\top + \Lambda \otimes uu^\top = (D + (u,u)D\Lambda + \Lambda) \otimes uu^\top) = 0.$$

Для этого достаточно положить

$$\Lambda = -(I + (u,u)D)^{-1}D.$$

Заметим, что эта матрица Λ в точности совпадает с матрицей $\widehat{\Lambda}$ из формулы (4.15). Мы постоили матрицу B вида

$$B = I + \Lambda \otimes uu^\top + uu^\top \otimes \Lambda,$$

для которой

$$AB = I + R,$$

где матрица R имеет ранг, ограниченный константой, не зависящей от n. Следовательно, матрица B — суперлинейный преобуславливатель для матрицы A и итерационный метод, такой как метод GMRES или сопряжённых градиентов, будет сходиться быстро.

4.7.2. Обобщение на случай большего числа слагаемых Естественным обобщением матрицы вида ((4.13)) будут матрицы состоящие из большего числа слагаемых такого же вида:

$$A = I + \sum_{i=1}^{r} D_i \otimes u_i u_i^\top + u_i u_i^\top \otimes D_i. \tag{4.16}$$

Построим для матрицы такого вида суперлинейный предобуславливатель, т.е. такую матрицу B, что

$$AB = I + R,$$

где R — матрица малого ранга, причём этот ранг ограничен константой, не зависящей от n. Воспользуемся приёмом, который мы использовали в конце предыдущего пункта, а именно, попытаемся угадать вид матрицы B. Будем искать матрицу B в виде

$$B = I + \sum_{i,j=1}^{r} \Lambda_{ij} \otimes R_{ij} + R_{ij} \otimes \Lambda_{ij}. \tag{4.17}$$

Здесь Λ_{ij} — диагональные матрицы, которые будут определены позднее, а матрицы R_{ij} определяются по формулам

$$R_{ij} = u_i u_j^\top.$$

В таких обозначениях матрица A может быть записана как

$$A = I + \sum_{i=1}^{r} D_i \otimes R_{ii} + R_{ii} \otimes D_i.$$

Легко получить основное свойство матриц R_{ij}, которое нам потребуется в дальнейшем:

$$R_{ij} R_{i'j'} = R_{ij'} \gamma_{ji'},$$

где $\Gamma = [\gamma_{ij}], \gamma_{ij} = (u_i, u_j)$ — матрица Грама векторов u_i. Эту матрицу при вычислениях можно вычислить заранее и очень быстро.

Выпишем теперь произведение AB:

$$AB = I + \sum_{i=1}^{r} D_i \otimes R_{ii} + R_{ii} \otimes D_i + \sum_{ij=1}^{r} \Lambda_{ij} \otimes R_{ij} + R_{ij} \otimes \Lambda_{ij} +$$

$$+ \sum_{iji'=1}^{r} (D_{i'} \otimes R_{i'i'} + R_{i'i'} \otimes D_{i'})(\Lambda_{ij} \otimes R_{ij} + R_{ij} \otimes \Lambda_{ij}) + (\ldots),$$

где в скобках собраны слагаемые с рангом, ограниченным константой, не зависящей от n.

Приведём теперь «подобные члены» так чтобы AB имело вид единичная матрица плюс матрица малого ранга. Нетрудно увидеть, что для этого нужно удовлетворить следующим тождествам:

$$\Lambda_{ij} + D_i \delta_{ij} + \left(\sum_{k=1}^{r} \Lambda_{kj} \gamma_{ik} \right) D_i = 0.$$

Это — блочная система $r \times r$, где каждый блок — диагональная матрица. На решение такой системы (и нахождение Λ_{ij}) требуется всего $\mathcal{O}(nr^3)$ операций. Поэтому мы показали существование суперлинейного предобуславливателя тензорного ранга r^2. Ранг малоранговый добавки тоже можно оценить, но это требует некоторых технических усилий. Мы приведём лишь финальный ответ. Ранг малоранговой добавки порядка $\mathcal{O}(r^3)$.

Заметим, что эти теоремы могут быть применены для установления структуры при приближённом обращении двухуровневых тёплицевых матриц, таких как матрица из уравнения Прандтля. Обозначим основные шаги доказательства.

1. Проверяем символ $f(x, y)$, порождающего двухуровневую тёплицеву матрицу, на асимптотическую гладкость. В случае положительного ответа, сразу получаем оценку на тензорный ранг матрицы.

2. Каждый фактор оптимальной аппроксимации малого тензорного ранга является тёплицевой матрицей. Проверяем, удовлетворяют ли «срезки» символа f (т.е. функции вида $f(x, y_i)$, при фиксированном y_i) условиям теоремы 3. Если удовлетворяют, то мы получаем аппроксимацию к исходной матрице вида

$$T = \sum_{i=1}^{r} (C_i + R_i) \otimes (C_i + R_i),$$

где C_i — циркулянты, а R_i — матрицы малого ранга (мы предположили случай симметрии). После применения преобразования Фурье, мы получаем матрицу вида

$$\widehat{T} = (F \otimes F)T(F^* \otimes F^*) = \sum_{i=1}^{r} (D_i + R_i) \otimes (D_i + R_i),$$

где D_i — диагональные матрицы.

3. Обращаем «главный член»:

$$T' = \Lambda^{-1/2}\widehat{T}\Lambda^{-1/2} = I + \sum_{i=1}^{r'} (\widehat{D}_i \otimes \widehat{R}_i + \widehat{R}_i \otimes \widehat{D}_i) + \widetilde{R},$$

где $\Lambda = \sum_{i=1}^{r} D_i \otimes D_i$, $\widehat{D}_i$ — диагональные матрицы, $\widehat{R}_i$ — матрицы малого ранга. При этом мы предполагаем, что $\Lambda^{-1/2}$ имеет малый тензорный ранг. Это нужно доказывать, однако иногда это можно сделать, так как Λ — диагональная матрица, и элементы $\Lambda^{-1/2}$ легко выписываются. Матрица $\widetilde{R}$ имеет малый ранг, т.е. ранг, ограниченный константой, не зависящей от n.

4. Матрица T' имеет требуемый вид с точностью до поправки малого ранга.

4.8. Выводы

В этой главе, на основе метода Ньютона с аппроксимациями, впервые построен метод обращения двухуровневых тёплицевых матриц сублинейной сложности. Это достигнуто с помощью использования специальной структуры для тензорных факторов. Оказывается, что эти факторы имеют малый приближённый ранг смещения. Возникает несколько сложных вопросов, которые были успешно решены. Быстрая арифметика в новом формате, быстрая процедура рекомпрессии — всё это оказалось возможно делать за сублинейное время. В последнем разделе впервые приведены теоретические результаты о структуре матрицы, обратной к матрицам малого тензорного ранга. Выделен специальный подкласс в классе таких матриц, который замкнут относительно обращения. Показано, как свести задачу обращения произвольной двухуровневой тёплицевой матрицы, получающейся из дискретизации интегрального уравнения к обращению матрицы из выделенного специального класса.

Заключение

В заключение диссертации сформулируем её основные результаты.

1. Решена задача построения оптимальных циркулянтных предобуславливателей на основе эффективных алгоритмов построения $C + R$ и $D + R$ аппроксимаций. Предложен и теоретически обоснован метод чёрных точек для решения задачи $C + R$ и $D + R$ аппроксимации. Для всех практически важных случаев тёплицевых матриц доказаны теоремы о существовании $C + R$ аппроксимации.

2. Построены явные формулы для построения вейвлет-преобразований на неравномерных сетках. Показано, что адаптированные к заданной сетке вейвлет-преобразования дают существенный выигрыш по сравнению с классическими преобразованиями Добеши, от 30% до 50%.

3. Предложен общий подход обращения структурированных матриц на основе метода Ньютона с аппроксимациями. Доказана теорема о сохранении квадратичной сходимости метода. Предложен модифицированный метод Ньютона, дающий существенный выигрыш при применении для обращения структурированных матриц. Предложен алгоритм обращения больших матриц на основе тензорных аппроксимаций со структурированными факторами. Использование тензорных аппроксимаций вместе с нестандартными вейвлет-преобразованиями позволило решать плотные системы с миллионом неизвестных в течении минут на обычной персональной станции.

4. Построен алгоритм супер-быстрого обращения двухуровневых тёплицевых матриц основанный на разработанном методе Ньютона с аппроксимациями. Для тёплицевых матриц подходящей структурой оказались матрицы малого тензорного ранга, причём каждый фактор имеет малый ранг смещения. На модельных примерах показано, что сложность алгоритма для обращения двухуровневых тёплицевых матриц составляет $O(\sqrt{n}\log^{\alpha} n)$, т.е. алгоритм имеет *сублинейную сложность*. Впервые получены результаты о структуре матриц, обратным к матрицам малого тензорного ранга специального вида, и показано, что обратные к матрицам, получающимся при дискретизации двумерного

интегрального уравнения, могут быть аппроксимированы матрицами малого тензорного ранга со структурированными факторами. Число параметров в таком структурированном представлении ведёт себя как $\mathcal{O}(\sqrt{n})$ — меньше, чем линейный размер матрицы.

Литература

[1] M. Bebendorf, Approximation of boundary element matrices, *Numer. Math.*, 2000, **86**, 565-589.

[2] M. Bebendorf and S. Rjasanow, Adaptive low-rank approximation of collocation matrices, *Computing*, 2003, **70**, 1–24.

[3] M. Bebendorf, S. Rjasanow and E. E. Tyrtyshnikov, Approximation using diagonal-plus-skeleton matrices, *Math. asp. of bound. elem. meth.* (Palaiseau, 1998), Chapman Hall/CRC, Boca Raton, FL, 2000, 45–52.

[4] D. A. Bini, B. Meini, Solving block banded block Toeplitz systems with structured blocks: algorithms and applications, *Structured Matrices: Recent Developments in Theory and Computation. Advances in Computation* (Edited by D.A.Bini, E.Tyrtyshnikov, P.Yalamov), Nova Science Publishers, Inc., Huntington, New York (2001).

[5] R. Chan, Circulant preconditioners for Hermitian Toeplitz systems, *Linear Algebra Appl.*, 1989, **10**, 542–550.

[6] T. F. Chan, An optimal circulant preconditioner for Toeplitz systems, *SIAM J. Sci. Statist. Comput.*, 1988, **9**, 766–771.

[7] R. H. Chan, D. Potts, G. Steidl, Preconditioners for Nondefinite Hermitian Toeplitz Systems, *SIAM Matrix Anal. Appl.*, 2001, **22**:3, 647–665.

[8] R. H. Chan, M. K. Ng and A. M. Yip, A Survey of Preconditioners for Ill-Conditioned Toeplitz Systems, Structured Matrices in Mathematics, Computer Science, and Engineering II, *Contemporary Mathematics*, 2001, **281**, 175–191.

[9] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications of Pure and Applied Mathematics*, October 1988, **41**:7, 909-996.

[10] B. W. Dickinson, Solution of linear equations with rational Toeplitz matrices, *Math. Comput.*, 1980, **34**:149, 227–233.

[11] Ford J. M., Tyrtyshnikov E. E., Combining Kronecker product approximation with discrete wavelet transforms to solve dense, function-related systems, *SIAM J. Sci. Comp.*, 2003, **25**:3, 961–981.

[12] Gohberg I., Heinig G., Inversion of finite-section Toeplitz matrices consisting of elements of a non-commutative algebra, *Rev. Roum. Math. Pures et Appl.*, 1974, **19**:5, 623-663.

[13] S. A. Goreinov, E. E. Tyrtyshnikov, The maximal-volume concept in approximation by low-rank matrices, *Contemporary Mathematics*, 2001, **208**, 47–51.

[14] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin, A theory of pseudo-skeleton approximations, *Linear Algebra Appl*, 1997, **261**, 1–21.

[15] W. Hackbusch, A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I. Introduction to $\mathcal{H}$-matrices, *Computing*, 1999, **62**:89–108.

[16] W. Hackbusch, B. N. Khoromskii, A sparse $\mathcal{H}$-matrix arithmetic. II. Application to multi-dimensional problems, *Computing*, 2000, **64**, 21–47.

[17] Hackbusch W., Khoromskii B. N., Tyrtyshnikov E. E., *Hierarchical Kronecker tensor-product approximations*, Max-Panck-Institut für Mathematik in den Naturwissenschaften, Leipzig, Preprint No. 35, 2003.

[18] W. Hackbusch, Z. P. Nowak, On the fast matrix multiplication in the boundary elements method by panel clustering, *Numer. Math.*, 1989, **54**:4, 463–491.

[19] G. Heinig, K. Rost, *Algebraic methods for Toeplitz-like matrices and operators*, Berlin, Academie-Verlag, 1984.

[20] Hotelling H., Analysis of a complex of statistical variables into principal components, *J. Educ. Psych.*, 1933, **24**, P I: 417-441, P II: 498-520.

[21] T. Kailath, S. Kung, M. Morf, Displacement ranks of matrcies and linear equations, *J. Math. Anal.and Appl.*, 1979, **68**, 395-407.

[22] J. Kamm, J. G. Nagy, Optimal Kronecker Product Approximations of Block Toeplitz Matrices, *SIAM J. Matrix Anal. Appl.*, 2000, **22**:1, 155-172.

[23] T.-K. Ku, C.-C. J. Kuo, Spectral properties of preconditioned rational Toeplitz matrices: the nonsymmetric case, *SIAM J. Matrix Anal. Appl.*, 1993, **14**:2, 512–544.

[24] I. K. Lifanov, *Singular integral equations and discrete vortices*, VSP, 1996.

[25] D. Noutsos, S. Serra Capizzano, P. Vassalos, A preconditioning proposal for ill-conditioned Hermitian two-level Toeplitz systems, *Numer. Linear Algebra Appl.*, 2005, **12**, 231–239.

[26] V. Y. Pan, Y. Rami, Newton's iteration for the inversion of structured matrices, *Structured Matrices: Recent Developments in Theory and Computation* (Eds. Bini D.A., Tyrtyshnikov E.E., Yalamov P.), Nova Science Publishers, Huntington, New York, 2001, 79-90.

[27] V. Rokhlin, Rapid solution of integral equations of classical potential theory, *J. Comput. Physics*, 1985, **60**, 187–207.

[28] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, An International Thomson Publishing Company, Boston, 1996.

[29] L. Schumaker, *Spline functions : basic theory*, Wiley, New York, 1981.

[30] Schulz G., Iterative Berechnung der reziproken Matrix, *Z. angew. Math. und Mech.*, 1933, **13**:1, 57-59.

[31] G. Strang, A proposal for Toeplitz matrix calculations, *Studies in Applied Mathematics*, 1989, **84**, 171–176

[32] V.V.Strela and E.E.Tyrtyshnikov, Which circulant preconditioner is better? *Math. Comput.*, 1996, **65**:213, 137–150.

[33] X. Sun, N. P. Pitsianis, A matrix version of the fast multipole method, *SIAM Review*, 2001, **43**:2, 289–300.

[34] W. Sweldens, The lifting scheme: A custom design construction of biorthogonal wavelets, *Appl. Comput. Harmon. Anal.*, 1996, **3**, 186-200.

[35] W. F. Trench, An algorithm for the inversion of finite Toeplitz matrices, *SIAM J.Appl. Math.*, 1964, **12**, 515-521.

[36] Tyrtyshnikov E. E., Kronecker-product approximations for some function-related matrices, *Linear Algebra Appl.*, 2004, **379**, 423–437.

[37] E. Tyrtyshnikov, Optimal and superoptimal circulant preconditioners, *SIAM J. Matrix Anal. Appl.*, 1992, **13**:2, 459-473.

[38] E. E. Tyrtyshnikov, Incomplete cross approximation in the mosaic-skeleton method, *Computing*, 2000, **64**:4, 367–380.

[39] E. E. Tyrtyshnikov, A unifying approach to some old and new theorems on distribution and clustering, *Linear Algebra Appl.*, 1996, **232**, 1–43.

[40] Tyrtyshnikov E. E., Mosaic–skeleton approximations. *Calcolo*, 1996, **33**:(1-2), 47–57.

[41] E. Tyrtyshnikov, R.Chan, Spectral Equivalence and Proper Clusters for Boundary Element Method Matrices, *Int. J. Numer. Meth. Engnr.*, 2000, **49**, 1211–1224.

[42] E. E. Tyrtyshnikov, N. L. Zamarashkin, and A. Yu. Yeremin, Clusters, preconditioners, convergence, *Linear Algebra Appl.*, 1997, **263**, 25–48.

[43] Van Loan C. F., Pitsianis N. P., Approximation with Kronecker products, *NATO Adv. Sci. Ser. E Appl. Sci.*, 1993, **232**, 293-314.

[44] N. Yarvin, V. Rokhlin, Generalized Gaussian quadratures and singular value decompositions of integral operators, *SIAM J. Sci. Comput.*, 1999 **20**:2, 699–718.

[45] А.А. Акопян, А.А. Саакян. Многомерные сплайны и полиномиальная интреполяция. *Успехи математических наук*, 1993, **48**:5.

[46] Белоцерковский С. М., Лифанов И. К., *Численные методы в сингулярных интегральных уравнениях*, М., Наука, 1985.

[47] Василенко В.А. *Сплайн-функции: теория, алгоритмы, программы*, Новосибирск: Наука, 1983. 216 с.

[48] Воеводин В. В., Тыртышников Е. Е., *Вычислительные процессы с теплицевыми матрицами*, М., Наука, 1987.

[49] Горейнов С. А., Замарашкин Н. Л., Тыртышников Е. Е., Псевдоскелетные аппроксимации матриц, *ДАН*, 1995, **343**:2, 151–152, 1995.

[50] И. Ц. Гохберг, А. А. Семенцул, Об обращении конечных тёплицевых матриц и их непрерывных аналогов, *Матем. исслед.*, 1972, **7**:2, 201-224.

[51] И. Добеши. *Десять лекций по вейвлетам*, Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001

[52] Лифанов И. К., Тыртышников Е. Е., Теплицевы матрицы и сингулярные интегральные уравнения, *Вычислительные процессы и системы*, Вып. 7, 94–278. - М., Наука, 1990.

[53] Лифанов И. К., Полтавский Л. Н., Обобщенные операторы Фурье и их применение к обоснованию некоторых численных методов в аэродинамике, *Матем. сб.*, 1992, **5**, 79-114.

[54] Тыртышников Е. Е., Тензорные аппроксимации матриц, порожденных асимптотически гладкими функциями, *Матем. сб.*, 2003, **194**:6, 147–160.

[55] Тыртышников Е. Е., Методы быстрого умножения и решение уравнений, *Матричные методы и вычисления,* ИВМ РАН, Москва, 1999, 4–41.

[56] Е. Е. Тыртышников, Тензорные аппроксимации матриц, порожденных асимптотически гладкими функциями, *Матем. сб.*, 2003, **194**:6, 147–160

[57] Фаддеев Д. К., Фаддеева В. Н., *Вычислительные методы линейной алгебры*, М.-Л., Физматгиз, 1963.

[58] К. Чуи. *Введение в вейвлеты*, Москва, «Мир», 2001

Публикации по теме диссертации

[59] V. Olshevsky, I. Oseledets, E. Tyrtyshnikov, Tensor properties of multilevel Toeplitz and related matrices, *Linear Algebra Appl.*, 2006, **412**, 1–21.

[60] Oseledets I.V., Tyrtyshnikov E.E., A unifying approach to the construction of circulant preconditioners, *Linear Algebra Appl.*, 2007, **418**, 435–449.

[61] Оселедец И.В., Тыртышников Е.Е., Приближённое обращение матриц при численном решении гиперсингулярного интегрального уравнения *ЖВМ и МФ* , 2005, **45**:2, 315–326.

[62] Оселедец И.В., Применение разделённых разностей и B-сплайнов для построени быстрых дискретных преобразований вейвлетовского типа на неравномерных сетках, *Мат. заметки*, 2005, **75**:5, 743-752

[63] Замарашкин Н.Л., Оселедец И.В., Тыртышников Е.Е., О приближении тёплицевых матриц суммой циркулянта и матрицы малого ранга, *ДАН*, 2006, **73**, 100-101.

[64] Ford J. M., Oseledets I. V., Tyrtyshnikov E. E., Matrix approximations and solvers using tensor products and non-standard wavelet transforms related to irregular grids, *Rus. J. Numer. Anal. and Math. Modelling*, 2004, **19**:2, 185-204.

[65] Оселедец И.В., Оценки снизу для сепарабельных аппроксимаций ядра Гильберта, *Матем. сб.*, 2007, **198**:3, 137-144.

[66] Оселедец И.В., Савостьянов Д.В., Ставцев С.Л., Применение нелинейных методов аппроксимации для быстрого решения задачи о распространении звука в мелком море. *Методы и технологии решения больших задач, ИВМ РАН, 2004, 171-192.*