Федеральное государственное бюджетное учреждение науки Институт вычислительной математики им. Г. И. Марчука Российской академии наук «ИВМ РАН»

На правах рукописи

Морозов Станислав Викторович

Построение чебышевских приближений для матриц и тензоров и их применения

Специальность 1.1.6 —
«Вычислительная математика»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
академик РАН, доктор физико-математических наук, профессор
Тыртышников Евгений Евгеньевич

Москва — 2024

Оглавление

Стр.

Введение

На сегодняшний день задача построения малоранговых приближений матриц является важным компонентом во многих областях науки, таких как вычислительная математика [1], вычислительная гидродинамика [2], рекомендательные системы [3], машинное обучение [4] и других. Эта задача может быть легко решена с помощью сингулярного разложения (SVD) в унитарно инвариантных нормах, таких как спектральная норма или норма Фробениуса. В то же время, в некоторых приложениях может потребоваться приближать матрицу поэлементно, то есть в чебышевской норме. Недавние результаты [5; 6] показывают, что эффективные приближения в норме Чебышева могут иметь большие перспективы. Так, например, некоторые функционально порожденные матрицы или матрицы, возникающие в моделях с латентными переменными, могут не иметь хороших приближений в унитарно-инвариантных нормах, но допускать эффективные поэлементные приближения. Другим интересным примером, демонстрирующим различие между приближениями в унитарно-инвариантных нормах и норме Чебышева, является единичная матрица. С точки зрения унитарно-инвариантых норм, единичная матрица является классическим примером матрицы полного ранга, которая не допускает малоранговых приближений. С другой стороны, можно показать [5], что в чебышевской норме ранг, требуемый для достижения заданной точности приближения ε, растет как $O(\log n)$ с размером матрицы n при фиксированной точности ε. Кроме того, стоит отметить, что вопрос о точности приближения единичной матрицы в норме Чебышева также связан с важной задачей функционального анализа об оценке поперечников по Колмогорову [7—10].

В современной науке также имеется множество примеров задач, в которых данные или решения представляются в виде тензоров. Примеры могут включать в себя теорию аппроксимации [11], механику сплошных сред [12], дифференциальные уравнения [13] и анализ данных [14]. Однако число элементов, требуемых для хранения d–мерного тензора $T \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ равно $n_1 \dots n_d$, что делает явное хранение и обработку тензоров неприемлемым даже для небольших значений d. Для решения этой проблемы используются малопараметрические представления тензоров. Наиболее популярными среди них являются каноническое разложение (canonical polyadic decomposition, CP) [15; 16], HOSVD [17; 18] и разложение в формате тензорного поезда (tensor-train decomposition, TT) [19]. На сегодняш-

ний день большинство алгоритмов строят аппроксимации в норме Фробениуса, однако в некоторых приложениях требуется приближать тензоры таким образом, что ошибка приближения каждого элемента ограничена и мала.

Данная диссертация посвящена вопросам построения малоранговых приближений матриц и тензоров в чебышевской норме, а также исследованию свойств построенных алгоритмов. Стоит отметить, что эти задачи являются трудными. На момент начала исследования было известно только о методах построения приближений ранга 1 (неоптимальных) для матриц [20]. Однако, даже для ранга 1 можно показать, что задача проверки существует ли приближение, гарантирующее точность ε, является NP-полной [21].

Целью данной работы является создание эффективных алгоритмов для построения малоранговых приближений матриц и тензоров в норме Чебышева. Кроме того, целью работы является теоретический анализ предложенных методов: изучение гарантий и свойств алгоритмов. Также работа ставит своей целью программную реализацию алгоритмов и их эмпирическое исследование.

Для достижения поставленных целей необходимо было решить следующие **задачи**. Важным компонентом для решения задачи о построении малоранговых приближений матриц и тензоров в чебышевской норме является *задача наилучшего равномерного приближения*, которая ставится следующим образом:

$$\|a - Vu\|_\infty \to \min_{u \in \mathbb{R}^r},$$

где $a \in \mathbb{R}^n$ и $V \in \mathbb{R}^{n \times r}$. Таким образом, необходимо было найти условия, при которых задача о построении наилучшего равномерного приближения корректна, а также предложить эффективные алгоритмы для ее решения. Автору неизвестно, чтобы задача о наилучшем равномерном приближении ранее изучалась в литературе. Кроме того, необходимо было изучить вопросы существования и единственности решения задач о построении малоранговых чебышевских приближений для матриц и тензоров, предложить алгоритмы их решения и исследовать вопросы корректности и свойства алгоритмов. Также необходимо было программно реализовать полученные алгоритмы и численно изучить их эффективность.

Научная новизна:

1. Впервые были исследованы свойства задачи о наилучшем равномерном приближении, предложен алгоритм решения задачи, доказана скорость его сходимости.

2. Впервые предложен метод, позволяющий строить чебышевские приближения произвольного ранга для матриц.
3. Впервые предложен метод, позволяющий строить оптимальные чебышевские приближения ранга 1 для матриц.
4. Впервые предложен метод, позволяющий строить чебышевские приближения тензоров в каноническом формате.

Теоретическая и практическая значимость. Диссертация имеет преимущественно теоретический характер. Представленные результаты позволяют строить малоранговые приближения матриц и тензоров в чебышевской норме. Практическая значимость работы состоит в реализации предложенных алгоритмов на языке C++ с интерфейсами для языка Python, в том числе с использованием технологий параллельного программирования OpenMP. Разработанные программы способны строить за разумное время приближения к матрицам с размерами до нескольких десятков тысяч строк и столбцов.

Методология и методы исследования. Результаты, полученные в диссертации, основаны на применении теоретических методов вычислительной математики и верифицированы при помощи большого количества численных экспериментов. В теоретических исследованиях были использованы методы линейной алгебры, анализа, общей топологии, оптимизации и дискретной математики.

Основные положения, выносимые на защиту: Основным результатом работы являются алгоритмы для решения задачи наилучшего равномерного приближения, а также задач малоранговой аппроксимации матриц и тензоров и теоретический анализ алгоритмов.

1. Предложен эффективный алгоритм решения задачи наилучшего равномерного приближения, доказаны гарантии его работы, оценена скорость сходимости.
2. Предложен метод переменных направлений для построения малоранговых чебышевских приближений матриц, теоретически изучены его свойства.
3. Предложен алгоритм, позволяющий находить оптимальные приближения ранга 1 для матриц в чебышевской норме.
4. Предложен метод переменных направлений, позволяющий строить эффективные малоранговые приближения тензоров в каноническом формате в чебышевской норме.

Достоверность полученных результатов обеспечивается большим количеством дополняющих друг друга теоретических результатов и численных экспериментов.

Апробация работы. Основные результаты работы докладывались на:

1. Matrix Equations and Tensor Techniques IX, Perugia, September 9-10, 2021.
2. Numerical Methods and Scientific Computing (NMSC21), CIRM Luminy, November 8-12, 2021.
3. Random Matrix Theory and Beyond, НТУ Сириус, 8-9 августа 2022.
4. Материалы и технологии XXI века, Казань, 30 ноября - 2 декабря 2022.
5. The 6th International Conference on Matrix Methods and Machine Learning in Mathematics and Applications, Москва, 15-18 августа 2023.
6. Матричные методы и интегральные уравнения, НТУ Сириус, 25-31 августа 2023.
7. Матричные методы и интегральные уравнения, НТУ Сириус, 12-15 августа 2024.

Личный вклад. Автор принимал активное участие в постановке задачи и получении всех основных результатов. В работе [22] автором были самостоятельно проанализированы условия корректности задачи наилучшего равномерного приближения, изучены ее свойства и предложен алгоритм решения. Доказательство теоремы о сходимости алгоритма было получено совместно с Е. Е. Тыртышниковым. В работе [23] автором был самостоятельно предложен метод построения оптимальных приближений ранга 1. Результаты о поведении знаков в методе переменных направлений были получены совместно с М. С. Смирновым. Метод переменных направлений для тензоров и все результаты о свойствах метода в [24] получены автором полностью самостоятельно. В [25] автором самостоятельно были получены результаты о сходимости метода. Быстрый алгоритм получен в результате обсуждений с А. И. Осинским и Д. А. Желтковым. Создание программных реализаций алгоритмов и проведение всех численных экспериментов было выполнено автором полностью самостоятельно.

Диссертационное исследование является самостоятельным и законченным трудом автора.

Публикации. Основные результаты по теме диссертации изложены в 4 печатных изданиях, 4 из которых изданы в журналах, рекомендованных ВАК, 3 — в периодических научных журналах, индексируемых Web of Science и Scopus.

Объем и структура работы. Диссертация состоит из введения, 3 глав и заключения. Полный объём диссертации составляет 138 страниц, включая 15 рисунков и 1 таблицу. Список литературы содержит 40 наименований.

В первой главе изучается задача наилучшего равномерного приближения. В Разделе 1.1 приводится формальная постановка задачи наилучшего равномерного приближения. В Разделе 1.2 вводятся используемые обозначения. В Разделе 1.3 изучается корректность задачи наилучшего равномерного приближения. Затем в Разделе 1.4 анализируются характеристические множества задачи, а в Разделе 1.5 доказываются критерии оптимальности решения. Раздел 1.6 содержит явные формулы для построения решения задачи размера $(r+1) \times r$, а в Разделе 1.7 доказывается комбинаторная формула решения задачи в вещественном случае. В Разделе 1.8 доказывается Теорема об альтернансе, являющаяся аналогом известной Теоремы Чебышева об альтернансе в случае задачи наилучшего равномерного приближения функций. Наконец, в Разделе 1.9 предлагается обобщенный алгоритм Ремеза для решения задачи наилучшего равномерного приближения и доказываются оценки на скорость его сходимости. В Разделе 1.10 приводится ускоренная версия алгоритма, которая в точной арифметике совпадает с обобщенным алгоритмом Ремеза, но имеет меньшую сложность. Разделы 1.7-1.10 посвящены вещественной задаче равномерного приближения. В Разделе 1.11 приводятся замечания о комплексной задаче.

Во второй главе диссертации изучается задача построения малоранговых приближений в чебышевской норме для матриц. В Разделе 2.1 приводится формальная постановка задачи малорангового приближения матриц в чебышевской норме, а в Разделе 2.2 описываются ее базовые свойства. В Разделе 2.3 предлагается метод переменных направлений для построения чебышевских приближений произвольного ранга, а также приводятся базовые свойства метода. В Разделе 2.4 вводится понятие двумерного альтернанса ранга r и доказывается, что наличие введенной структуры является необходимым условием оптимальности решения задачи, а также что все предельные точки метода переменных направлений обладают введенной структурой. Разделы 2.5-2.7 содержат подробный анализ метода переменных направлений для построения приближений ранга 1, а именно, в Разделе 2.5 обосновывается корректность метода переменных направлений, в Разделе 2.6 анализируется поведение знаков компонент векторов, возникающих в методе переменных направлений, а в Разделе 2.7 на основе проведенного анализа предлагается метод, позволяющий строить оптимальные чебышевские

приближения ранга 1. Наконец, в Разделе 2.8 приводятся результаты интенсивного численного исследования предложенного алгоритма, в том числе для матриц размера порядка нескольких десятков тысяч.

В третьей главе диссертации изучается задача построения малоранговых приближений в чебышевской норме для тензоров в каноническом формате. В Разделе 3.1 приводится формальная постановка задачи. В Разделе 3.2 предлагается метод переменных направлений для построения малоранговых чебышевских приближений тензоров в каноническом формате для произвольного ранга и базовые свойства алгоритма. Разделы 3.3-3.6 содержат детальный анализ случая приближений ранга 1, а именно, Раздел 3.3 обосновывает корректность предложенной процедуры, а в Разделе 3.4 анализируются знаки компонент векторов, возникающих в результате работы метода. В Разделе 3.5 вводится понятие трехмерного альтернанса и доказывается, что наличие введенной структуры является необходимым условием оптимальности решения задачи, а также, что все предельные точки метода переменных направлений удовлетворяют этому условию. В Разделе 3.6 анализируются вопросы сходимости алгоритма к точкам локального минимума и предлагается модификация метода переменных направлений, во многих случаях позволяющая строить оптимальные приближения ранга 1. Наконец, Раздел 3.7 содержит численное исследование построенного алгоритма для приближения различных тензоров.

Благодарности. Автор выражает благодарность академику РАН д.ф.-м.н. Евгению Евгеньевичу Тыртышникову за научное руководство и поддержку, старшему научному сотруднику ИВМ РАН к.ф.-м.н. Николаю Леонидовичу Замарашкину, научному сотруднику ИВМ РАН к.ф.-м.н. Дмитрию Александровичу Желткову, сотруднику Сколковского института науки и технологий к.ф.-м.н. Александру Игоревичу Осинскому, младшему научному сотруднику ИВМ РАН Матвею Станиславовичу Смирнову за плодотворные обсуждения по теме диссертации, а также доценту факультета ВМК МГУ к.ф.-м.н. Сергею Александровичу Матвееву и Сукманюк Софье Владимировне за помощь в работе над текстами статей и диссертации, многочисленные советы и поддержку.

Глава 1. О задаче наилучшего равномерного приближения по системе векторов

1.1 Постановка задачи

Пусть заданы матрица $V \in \mathbb{C}^{n\times r}$, где $n > r$ и вектор $a \in \mathbb{C}^n$. Рассмотрим задачу

$$\|a - Vu\|_\infty \to \min_{u\in\mathbb{C}^r}. \tag{1.1}$$

Будем называть задачу (1.1) задачей *наилучшего равномерного приближения* вектора a по системе векторов V. В данной главе рассматриваются вопросы корректности задачи (1.1), а также предлагается эффективный метод ее решения.

1.2 Обозначения

В данном разделе вводятся обозначения, которые будут использованы в тексте работы. Пусть $V \in \mathbb{C}^{n\times r}$. Будем обозначать нижним индексом $v_j \in \mathbb{C}^n$ j–ый столбец матрицы V, а верхним индексом $v^j \in \mathbb{C}^r$ — j–ую строка матрицы V.

Пусть J является упорядоченным множеством из k натуральных чисел $1 \leqslant j_1, j_2, \ldots, j_k \leqslant n$. Круглыми скобками будем обозначать упорядоченные множества, например, $J = (j_1, j_2, \ldots, j_k)$. Обозначим через $V(J)$ подматрицу матрицы V, содержащую строки с номерами из множества J. Если $n = r + 1$, будем также обозначать через $V^{\setminus j}$ подматрицу матрицы V, содержащую все строки, кроме строки с номером j, то есть $V^{\setminus j} = V((1, \ldots, j-1, j+1, \ldots, r+1))$. Для матриц размера $(r+1)\times r$ будем также использовать обозначение $D_j(V) = \det V^{\setminus j}$. Аналогично, если $a \in \mathbb{C}^n$, будем обозначать через $a(J)$ подвектор вектора a, содержащий элементы с номерами из множества J и через $a_{\setminus j}$ вектор $a((1, \ldots, j-1, j+1, \ldots, n))$.

Под $\operatorname{sign} x$, где $x \in \mathbb{C}$ будем понимать число такое, что

$$\operatorname{sign} x = \begin{cases} x/|x|, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

1.3 Существование, единственность, непрерывность

Доказательства, приведенные в данном разделе, получены по аналогии с результатами, описанными в [26] и [27] для задачи равномерного приближения непрерывной функции по системе заданных функций. Однако в явном виде результаты данного раздела не следуют из приведенных в [26; 27] и автору не известно, чтобы они публиковались ранее.

Введем полезные для дальнейшего обозначения. Пусть $V \in \mathbb{C}^{n\times r}$, $a \in \mathbb{C}^n$. Обозначим функцию ошибки для вектора u в j–ой компоненте через

$$F(j,u) = |a_j - u^T v^j|.$$

Обозначим также

$$K(j,\lambda) = \{u \in \mathbb{C}^r | F(j,u) \leqslant \lambda\}.$$

$$K(J',\lambda) = \bigcap_{j\in J'} K(j,\lambda) = \{u \in \mathbb{C}^r | F(j,u) \leqslant \lambda, \forall j \in J'\}.$$

Здесь через J' обозначено некоторое подмножество индексов $J = (1,2,\ldots,n)$. Легко понять, что верны следующие вложения:

$$K(j,\lambda') \subset K(j,\lambda''), \quad K(J',\lambda') \subset K(J',\lambda''), \quad 0 \leqslant \lambda' < \lambda''.$$

$$K(J'',\lambda) \subset K(J',\lambda), \quad J' \subset J''.$$

Лемма 1.1. *Пусть столбцы матрицы V линейно независимы. Тогда множество $K(j,\lambda)$ выпукло и замкнуто, а множество $K(J,\lambda)$ ограничено для любых j и $\lambda \geqslant 0$.*

Доказательство. Докажем замкнутость множеств $K(j,\lambda)$. Пусть $\hat{u}$ — предельная точка множества $K(j,\lambda)$. Тогда в любой ее окрестности существуют точки $u \in K(j,\lambda)$.

$$F(j,\hat{u}) \leqslant |F(j,\hat{u}) - F(j,u)| + F(j,u).$$

Так как $u \in K(j,\lambda)$, то $F(j,u) \leqslant \lambda$. Функция $F(j,u) = \left|a_j - u^T v^j\right|$ непрерывна по u при любом фиксированном j. Тогда для любого $\varepsilon > 0$ существует $\delta > 0$ такое, что если $\|u - \hat{u}\| < \delta$, то $|F(j,\hat{u}) - F(j,u)| < \varepsilon$. Таким образом, имеем, что

$$F(j,\hat{u}) < \varepsilon + \lambda$$

для любого $\varepsilon > 0$, откуда $\hat{u} \in K(j, \lambda)$ и замкнутость доказана.

Докажем ограниченность $K(J, \lambda)$. Рассмотрим вектор Vu при $\|u\|_1 = 1$. Величина $\|Vu\|_\infty$ является непрерывной по u функцией на компактном множестве, поэтому достигает в некоторой точке $\hat{u}$ своего минимального значения

$$M = \|V\hat{u}\|_\infty \leqslant \|Vu\|_\infty.$$

Поскольку столбцы V линейно независимы, любая их нетривиальная линейная комбинация не равна нулю и $M > 0$. Пусть $\|u\|_1 \geqslant \dfrac{C+1}{M}$, где $C > 0$ — некоторая константа. Тогда

$$\|a - Vu\|_\infty \geqslant \|Vu\|_\infty - \|a\|_\infty \geqslant \|u\|_1 M - \|a\|_\infty \geqslant C + 1 - \|a\|_\infty.$$

Тогда при $\|u\|_1 \geqslant \dfrac{C+1}{M}$ не может быть выполнено условие $F(J, u) \leqslant C - \|a\|_\infty$, то есть $u \notin K(J, \lambda)$ при $\lambda \leqslant C - \|a\|_\infty$. В силу произвольности C ограниченность доказана.

Докажем выпуклость множества $K(j, \lambda)$. Пусть $u_1, u_2 \in K(j, \lambda)$, то есть $F(j, u_1) \leqslant \lambda$ и $F(j, u_2) \leqslant \lambda$. Пусть $\tau \in (0, 1)$. Тогда

$$F(j, \tau u_1 + (1-\tau)u_2) = |a_j - (\tau u_1 + (1-\tau)u_2)^T v^j| = \\ |\tau(a_j - u_1^T v^j) + (1-\tau)(a_j - u_2^T v^j)| \leqslant \tau F(j, u_1) + (1-\tau)F(j, u_2) \leqslant \lambda.$$

□

Обозначим $\mu = \inf\limits_{u \in \mathbb{C}^r} \|a - Vu\|_\infty$.

Лемма 1.2. *Пусть существует λ_0 такое, что при любом λ, где $\mu < \lambda < \lambda_0$ и любом j множество $K(j, \lambda)$ замкнуто и $K(J, \lambda_0)$ ограничено. Тогда существует такой вектор $\hat{u} \in \mathbb{C}^r$, что*

$$\|a - V\hat{u}\|_\infty = \mu.$$

Доказательство. Пусть $\lambda > \mu$. Поскольку $\mu = \inf\limits_{u \in \mathbb{C}^r} \max\limits_j F(j, u)$, существует такая точка u_λ, что $\max\limits_j F(j, u_\lambda) \leqslant \lambda$, следовательно $u_\lambda \in K(J, \lambda)$.

Рассмотрим убывающую последовательность $\{\lambda_k\}$ таких, что $\mu < \lambda_k < \lambda_0$ и сходящуюся к μ. Пересечение любой совокупности замкнутых множеств замкнуто, поэтому $K(J, \lambda_k)$ также замкнуты. В силу убывания λ_k имеем

$$K(J, \lambda_0) \supset K(J, \lambda_1) \supset K(J, \lambda_2) \supset \ldots$$

Из рассуждений выше, множества $K(J, \lambda_k)$ не пусты. Если дана система замкнутых, вложенных друг в друга, непустых множеств из $\mathbb{C}^r$, хотя бы одно из которых ограничено, то их пересечение является непустым, замкнутым, ограниченным множеством. Следовательно,

$$K = \bigcap_{k=1}^{\infty} K(J, \lambda_k)$$

не пусто. Пусть $\hat{u} \in K$. Тогда $\hat{u} \in K(j, \lambda_k)$ для любых j и любого k. Следовательно $F(j, \hat{u}) \leqslant \lambda_k$ для любых j и k, откуда имеем, что $\max\limits_j F(j, \hat{u}) \leqslant \mu$. Однако в силу определения μ, $\max\limits_j F(j, \hat{u}) \geqslant \mu$. □

Докажем результат о существовании решения задачи (1.1).

Теорема 1.3. *Решение задачи*

$$\|a - Vu\|_\infty \to \min_{u \in \mathbb{C}^r}$$

существует для любых вектора $a \in \mathbb{C}^n$ и матрицы $V \in \mathbb{C}^{n \times r}$.

Доказательство. Если столбцы матрицы V линейно независимы, то результат сразу следует из Леммы 1.1 и Леммы 1.2. Если столбцы матрицы V линейно зависимы, то выберем в матрице V максимальную систему линейно независимых столбцов и обозначим получившуюся матрицу через $\tilde{V} \in \mathbb{C}^{n \times \tilde{r}}$. Поскольку образы матриц V и $\tilde{V}$ совпадают,

$$\inf_{u \in \mathbb{C}^r} \|a - Vu\|_\infty = \inf_{\tilde{u} \in \mathbb{C}^{\tilde{r}}} \|a - \tilde{V}\tilde{u}\|_\infty,$$

однако последняя задача имеет решение. Дополнив его нулями, получим требуемое утверждение. □

Перейдем к исследованию вопроса единственности. Для этого нам понадобится следующее

Определение 1.4. *Будем говорить, что матрица $V \in \mathbb{C}^{n \times r}$ при $n \geqslant r$ является чебышевской, если любые r строк матрицы V линейно независимы.*

Теорема 1.5. *Пусть матрица $V \in \mathbb{C}^{n \times r}$ и $n > r$. Тогда для того, чтобы для любого вектора $a \in \mathbb{C}^n$ существовало единственное решение задачи*

$$\|a - Vu\|_\infty \to \min_{u \in \mathbb{C}^r},$$

необходимо и достаточно, чтобы матрица V была чебышевской.

Доказательство. Пусть решение задачи единственно. Тогда столбцы матрицы V линейно независимы (в противном случае решение не единственно). Пусть столбцы матрицы V не образуют чебышевскую систему. Тогда существует набор номеров строк $J = (j_1, \ldots, j_r)$ такой, что строки матрицы $V(J)$ линейно зависимы. Тогда существуют векторы $b \in \mathbb{C}^r$ и $d \in \mathbb{C}^r$ такие, что $V(J)^T b = 0$ и $V(J)d = 0$. Покажем, что существует вектор $a \in \mathbb{C}^n$, для которого решение задачи о наилучшем равномерном приближении не единственно.

Пусть вектор $g \in \mathbb{C}^n$ удовлетворяет следующим свойствам:

1. $|g_j| \leqslant 1$ при $j = 1, \ldots, n$;
2. $g_{j_k} = \operatorname{sign} \overline{b_k}$ при $k = 1, \ldots, r$.

Выберем $\lambda > 0$ так, что $\lambda \|Vd\|_\infty = 1$ ($\|Vd\|_\infty > 0$ поскольку матрица V имеет ранг r, а вектор d ненулевой). Определим вектор $a \in \mathbb{C}^n$ как

$$a = g \odot (e - \lambda |Vd|),$$

где e обозначает вектор из всех единиц, $|Vd|$ вектор, содержащий модули вектора Vd, а $\odot$ обозначает адамарово (поэлементное) произведение. По определению g имеем, что $|g_j| \leqslant 1$, а $\lambda |(Vd)_j| \in [0, 1]$, откуда $|a_j| \leqslant 1$ для любого j.

Предположим, что

$$\inf_{u \in \mathbb{C}^r} \|a - Vu\|_\infty < 1. \tag{1.2}$$

Так как $V(J)d = 0$, то $a_{j_k} = g_{j_k}$ при $k = 1, \ldots, r$. Пусть $b_k \neq 0$, тогда $|a_{j_k}| = |\operatorname{sign} \overline{b_k}| = 1$. Тогда если $(Vu)_{j_k} = 0$, то $|a_{j_k} - (Vu)_{j_k}| = 1$ и мы приходим к противоречию с предположением (1.2). Следовательно, если $b_k \neq 0$, то и $(Vu)_{j_k} \neq 0$. Поскольку все b_k не могут быть нулевыми, существуют такие k, для которых выполнено $\overline{b_k}(Vu)_{j_k} \neq 0$. Пусть для такого j

$$|\arg(Vu)_{j_k} - \arg a_{j_k}| = |\arg(Vu)_{j_k} - \arg \overline{b_k}| \geqslant \pi/2.$$

Тогда

$$|(Vu)_{j_k} - a_{j_k}| = |(Vu)_{j_k} - \operatorname{sign} \overline{b_k}| \geqslant 1,$$

где вновь приходим к противоречию с (1.2). Значит для всех k, для которых $b_k \neq 0$, выполнено

$$|\arg(Vu)_{j_k} b_k)| = |\arg(Vu)_{j_k} - \arg \overline{b_k}| < \pi/2.$$

Но тогда для всех k от 1 до r выполнено $|\arg((Vu)_{j_k} b_k)| < \pi/2$, откуда $\operatorname{Re}((Vu)_{j_k} b_k) \geqslant 0$, причем не все из них равны нулю, следовательно

$$\operatorname{Re} \sum_{k=1}^{r} b_k (Vu)_{j_k} > 0,$$

однако $V(J)^T b = 0$. Приходим к противоречию и получаем, что $\inf\limits_{u\in\mathbb{C}^r} \|a - Vu\|_\infty = 1$.

Возьмем произвольное $\varepsilon \in (0, 1]$ и рассмотрим

$$|a_k - \varepsilon\lambda d^T v^k| \leqslant |a_k| + \varepsilon\lambda|d^T v^k| = |g_k|(1 - \lambda|d^T v^k|) + \varepsilon\lambda|d^T v^k| \leqslant$$
$$1 - \lambda|d^T v^k| + \varepsilon\lambda|d^T v^k| = 1 - \lambda|d^T v^k|(1 - \varepsilon) \leqslant 1.$$

Следовательно, для любого $\varepsilon \in (0, 1]$, вектор $\varepsilon\lambda d$ является решением задачи $\inf\limits_{u\in\mathbb{C}^r} \|a - Vu\|_\infty$, то есть существует бесконечно много решений.

Докажем обратное, пусть матрица V является чебышевской и $\hat{u} \in \mathbb{C}^r$ является вектором наилучшего равномерного приближения. Покажем, что тогда по крайней мере в $r + 1$ точке в векторе $w = a - V\hat{u}$ достигается максимальное по модулю значение, то есть существуют $j_1, \ldots, j_{r+1}$, в которых выполняется равенство

$$|w_{j_k}| = \|w\|_\infty, \quad k = 1, \ldots, r + 1.$$

Пусть таких точек $r_1 < r + 1$. Тогда, решая систему с r_1 уравнениями и r неизвестными, строки которой линейно независимы, получим вектор $p \in \mathbb{C}^r$ такой, что

$$(Vp)_{j_k} = w_{j_k}, \quad k = 1, \ldots, r_1.$$

Но тогда вектор $\hat{u} + \delta p$ при достаточно малом $\delta > 0$ дает решение лучше, чем $\hat{u}$. Пришли к противоречию.

Предположим, что существует по крайней мере два решения $\hat{u}_1$ и $\hat{u}_2$. Пусть $\mu = \inf\limits_{u\in\mathbb{C}^r} \|a - Vu\|_\infty$. Тогда

$$\mu \leqslant \left\|a - V\frac{\hat{u}_1 + \hat{u}_2}{2}\right\|_\infty \leqslant \frac{1}{2}\|a - V\hat{u}_1\|_\infty + \frac{1}{2}\|a - V\hat{u}_2\|_\infty = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu,$$

откуда следует, что $\tilde{u} = \dfrac{\hat{u}_1 + \hat{u}_2}{2}$ также является решением и по крайней мере в $r+1$ точке $j_1, \ldots, j_{r+1}$ в векторе $\tilde{w} = a - V\tilde{u}$ достигается максимальное по модулю значение μ. Однако в этих точках также выполнено

$$\mu = \left|a_{j_k} - \left(\frac{\hat{u}_1 + \hat{u}_2}{2}\right)^T v^{j_k}\right| \leqslant \frac{1}{2}|a_{j_k} - (V\hat{u}_1)_{j_k}| + \frac{1}{2}|a_{j_k} - (V\hat{u}_2)_{j_k}| \leqslant \frac{1}{2}\mu + \frac{1}{2}\mu = \mu,$$

откуда

$$\left|a_{j_k} - \left(\frac{\hat{u}_1 + \hat{u}_2}{2}\right)^T v^{j_k}\right| = \frac{1}{2}|a_{j_k} - (V\hat{u}_1)_{j_k}| + \frac{1}{2}|a_{j_k} - (V\hat{u}_2)_{j_k}| \quad k = 1, \ldots, r + 1.$$

Но последнее равенство возможно только если совпадают модули и аргументы, следовательно

$$a_{j_k} - (V\hat{u}_1)_{j_k} = a_{j_k} - (V\hat{u}_2)_{j_k},$$

а значит $(V(\hat{u}_1 - \hat{u}_2))_{j_k} = 0$ в $r+1$ позиции, но поскольку система столбцов V является чебышевской, отсюда следует, что $\hat{u}_1 = \hat{u}_2$. □

Заметим, что нами было доказано следующее

Утверждение 1.6. *Пусть матрица $V \in \mathbb{C}^{n\times r}$, где $n > r$, является чебышевской, и $\hat{u} \in \mathbb{C}^r$ является решением задачи*

$$\|a - Vu\|_\infty \to \min_{u\in\mathbb{C}^r}.$$

Тогда в векторе $w = a - V\hat{u}$ по крайней мере в $r+1$ точке достигается максимальное по модулю значение, то есть существуют $j_1, \ldots, j_{r+1}$, в которых выполняется равенство

$$|w_{j_k}| = \|w\|_\infty, \quad k = 1, \ldots, r+1.$$

Рассмотрим вопрос о непрерывности решения.

Теорема 1.7. *Пусть матрица $V \in \mathbb{C}^{n\times r}$, где $n > r$, является чебышевской. Обозначим через $\hat{u}(y, X)$ решение задачи о наилучшем равномерном приближении для матрицы $X \in \mathbb{C}^{n\times r}$ и вектора $y \in \mathbb{C}^n$. Тогда функция $\hat{u}(y, X)$ является непрерывной в точке (a, V) для любого $a \in \mathbb{C}^n$, то есть $\forall\varepsilon > 0\ \exists\delta = \delta(a, V, \varepsilon) > 0$ такое, что если $\|a - b\|_\infty + \|V - W\|_C < \delta$, то $\|\hat{u}(a, V) - \hat{u}(b, W)\|_\infty < \varepsilon$.*

Доказательство. Пусть последовательность пар (a_k, V_k) сходится к (a, V). Заметим, что если матрица V является чебышевской, то существует такое $\delta_1 > 0$, что если $\|V - W\|_C < \delta_1$, то W также является чебышевской. Не ограничивая общности будем считать, что матрицы V_k таковы, что $\|V - V_k\|_C \leqslant \dfrac{\delta_1}{2}$. Покажем, что последовательность векторов $\{\hat{u}(a_k, V_k)\}_k$ ограничена.

Рассмотрим вектор Wu при $\|u\|_1 = 1$ и $\|V - W\|_C \leqslant \dfrac{\delta_1}{2}$. Величина $\|Wu\|_\infty$ является непрерывной по u функцией на компактном множестве, поэтому достигает в некоторой точке своего минимального значения

$$M = \|\hat{W}\hat{u}\|_\infty \leqslant \|Wu\|_\infty$$

для любых u и W таких, что $\|u\|_1 = 1$ и $\|V - W\|_C \leqslant \dfrac{\delta_1}{2}$. Поскольку W чебышевская, любая нетривиальная линейная комбинация ее столбцов не равна 0 и $M > 0$. Пусть $\|u\|_1 \geqslant \dfrac{C+1}{M}$, где $C > 0$ — некоторая константа. Тогда

$$\|a_k - V_k u\|_\infty \geqslant \|V_k u\|_\infty - \|a_k\|_\infty \geqslant \|u\|_1 M - \|a_k\|_\infty \geqslant C + 1 - \|a_k\|_\infty.$$

Выберем C таким образом, что $C + 1 - \|a_k\|_\infty > \|a_k\|_\infty$ для любого k (это возможно, поскольку нормы a_k ограничены). Тогда при $\|u\|_1 \geqslant \dfrac{C+1}{M}$ вектор u для решения задачи о наилучшем равномерном приближении с матрицей V_k и правой частью a_k будет хуже, чем нулевой вектор. Отсюда следует, что существует константа $A > 0$ такая, что $\|\hat{u}(a_k, V_k)\|_\infty < A$.

Выделим из последовательности $\{\hat{u}(a_k, V_k)\}_k$ сходящуюся подпоследовательность. Пусть $\lim\limits_{j\to\infty} \hat{u}(a_{k_j}, V_{k_j}) = \tilde{u}$. Последовательность $\{V_{k_j}\}_j$ ограничена, поэтому выделим из нее сходящуюся подпоследовательность, которую снова обозначим $\{V_{k_j}\}_j$. Тогда имеем $\lim\limits_{j\to\infty} V_{k_j}\hat{u}(a_{k_j}, V_{k_j}) = V\tilde{u}$. Заметим, что для любого $u \in \mathbb{C}^r$ выполнено $\|a_{k_j} - V_{k_j}\hat{u}(a_{k_j}, V_{k_j})\|_\infty \leqslant \|a_{k_j} - V_{k_j}u\|_\infty$. Переходя к пределу, имеем $\|a - V\tilde{u}\|_\infty \leqslant \|a - Vu\|_\infty$ для любого $u \in \mathbb{C}^r$. Но в силу единственности решения (см. Теорему 1.5), $\tilde{u} = \hat{u}(a, V)$. Следовательно,

$$\lim_{j\to\infty} V_{k_j}\hat{u}(a_{k_j}, V_{k_j}) = V\hat{u}(a, V).$$

Докажем, что вся последовательность $V_k\hat{u}(a_k, V_k)$ сходится к $V\hat{u}(a, V)$. Пусть это не так. Тогда существует $\delta_2 > 0$ и подпоследовательность $\{V_{k_j}\hat{u}(a_{k_j}, V_{k_j})\}_j$ такая, что

$$\|V_{k_j}\hat{u}(a_{k_j}, V_{k_j}) - V\hat{u}(a, V)\|_\infty \geqslant \delta_2.$$

Выделим аналогично описанному выше из $\{\hat{u}(a_{k_j}, V_{k_j})\}_j$ и $\{V_{k_j}\}_j$ сходящиеся подпоследовательности (для которых сохраним обозначения) и получим снова

$$\lim_{j\to\infty} V_{k_j}\hat{u}(a_{k_j}, V_{k_j}) = V\hat{u}(a, V).$$

Таким образом, приходим к противоречию и получаем

$$\lim_{k\to\infty} V_k\hat{u}(a_k, V_k) = V\hat{u}(a, V).$$

Тогда

$$\|V_k\hat{u}(a_k,V_k)-V\hat{u}(a,V)\|_\infty = \\ \|V_k\hat{u}(a_k,V_k)-V\hat{u}(a_k,V_k)+V\hat{u}(a_k,V_k)-V\hat{u}(a,V)\|_\infty \geqslant \\ \Big|\|(V_k-V)\hat{u}(a_k,V_k)\|_\infty - \|V(\hat{u}(a_k,V_k)-\hat{u}(a,V))\|_\infty\Big|.$$

Заметим, что

$$\lim_{k\to\infty}\|V_k\hat{u}(a_k,V_k)-V\hat{u}(a,V)\|_\infty = 0$$

и

$$\lim_{k\to\infty}\|(V_k-V)\hat{u}(a_k,V_k)\|_\infty = 0,$$

поскольку $\{V_k\}_k$ стремится к V, а последовательность $\{\hat{u}(a_k,V_k)\}_k$ ограничена. Но тогда

$$\lim_{k\to\infty}\|V(\hat{u}(a_k,V_k)-\hat{u}(a,V))\|_\infty = 0.$$

Пусть J — произвольное множество номеров строк матрицы V. Обозначим

$$g_k = \hat{u}(a_k,V_k)-\hat{u}(a,V),$$

$$\tau_k = V\left(\hat{u}(a_k,V_k)-\hat{u}(a,V)\right).$$

Тогда $g_k(J) = V(J)^{-1}\tau_k(J)$. Поскольку существует лишь конечное число способов выбрать r строк матрицы V, мы имеем, что $\|g_k\|_\infty \leqslant C\|\tau_k\|_\infty$. Выше было показано, что $\tau_k \to 0$, откуда

$$\lim_{k\to\infty}\hat{u}(a_k,V_k) = \hat{u}(a,V).$$

□

Заметим, что все результаты данного раздела верны как в вещественном, так и комплексном случае (доказательства в вещественном случае полностью совпадают с доказательствами в комплексном).

1.4 Характеристические множества

В теории равномерного приближения непрерывных функций ключевую роль играет понятие характеристического множества. Можно показать, что при

равномерном приближении непрерывной функции на отрезке полиномами существует набор из $d+2$ точек, где d — степень полинома, таких, что если полином оптимально приближает функцию в этих точках, то он оптимально приближает функцию на всем отрезке (см. например [26]). Аналогичную теорию можно построить в случае задачи о наилучшем равномерном приближении векторов.

Обозначим $J = (1, 2, \ldots, n)$ и пусть $J' \subset J$. Тогда обозначим

$$\mu(J') = \min_{u \in \mathbb{C}^r} \|a(J') - V(J')u\|_\infty .$$

При этом $\mu = \mu(J)$.

Определение 1.8. *Множество J' называется характеристическим множеством, если $\mu(J) = \mu(J')$ и для любого подмножества $J'' \subsetneq J'$ выполнено $\mu(J'') < \mu(J)$.*

Далее будет показано, что если столбцы матрицы $V \in \mathbb{C}^{n\times r}$ линейно независимы, то существует по крайней мере одно характеристическое множество, содержащее не более $2r+1$ точек в комплексном случае и не более $r+1$ точек в вещественном.

Обозначим

$$\mu_k = \max_{J_k \subset J, |J|=k} \mu(J_k) = \max_{J_k \subset J, |J|=k} \min_{u \in \mathbb{C}^r} \|a(J_k) - V(J_k)u\|_\infty .$$

Легко доказать следующее

Утверждение 1.9. $\mu_k \leqslant \mu_{k+1} \leqslant \mu, \quad k = 1, \ldots, n-1.$

Для дальнейшего анализа понадобится следующая классическая

Теорема 1.10 (Хелли)**.** *Если конечное множество S выпуклых множеств в $\mathbb{R}^r$ содержит не менее $r+1$ множеств (среди которых могут быть одинаковые), пересечение любых $r+1$ множеств из S не пусто и пересечение некоторого числа множеств из S ограничено, то пересечение всех множеств из S не пусто.*

Основываясь на этой теореме докажем следующий результат

Теорема 1.11. *Пусть существует такое $\lambda_0 > \mu_{2r+1}$ ($\lambda_0 > \mu_{r+1}$ в вещественном случае), что для любого j и любого λ такого, что $\mu_{2r+1} < \lambda < \lambda_0$ ($\mu_{r+1} < \lambda < \lambda_0$ в вещественном случае), множество $K(j, \lambda)$ замкнуто и выпукло, и пересечение некоторого конечного числа множеств $K(j, \lambda)$ ограничено. Тогда $\mu_{2r+1} = \mu$ ($\mu_{r+1} = \mu$ в вещественном случае).*

Доказательство. Пусть $k = r + 1$ в вещественном случае и $k = 2r + 1$ в комплексном. При любом $\lambda > \mu_k$ множество $K(J_k, \lambda)$ не пусто по определению μ_k для любого $J_k \subset J$ такого, что $|J_k| = k$. Кроме того, так как $K(j, \lambda)$ по условию выпукло и замкнуто, то

$$K(J_k, \lambda) = \bigcap_{j \in J_k} K(j, \lambda)$$

не пусто, замкнуто и выпукло для любого $J_k \subset J$ такого, что $|J_k| = k$.

Убедимся, что выполнены все условия Теоремы 1.10. В качестве совокупности множеств S возьмем множества $K(j, \lambda)$, $j \in J$. По условию теоремы они замкнуты и выпуклы, и пересечение некоторого конечного числа этих множеств ограничено. В вещественном случае то, что пересечение любых $r + 1$ множеств не пусто эквивалентно тому, что $K(J_{r+1}, \lambda)$ не пусто, что было показано выше. В комплексном случае нам нужно работать с пространством $\mathbb{C}^r$, которое мы отождествим с $\mathbb{R}^{2r}$, поэтому нам требуется, чтобы пересечение любых $2r+1$ множеств было не пусто, что также было показано выше. Итак, все условия Теоремы 1.10 выполнены и мы имеем, что

$$K(J, \lambda) = \bigcap_{j \in J} K(j, \lambda)$$

не пусто, замкнуто, выпукло и ограничено.

Пусть последовательность $\{\lambda_t\}_t$ убывающая, $\mu_k < \lambda_t < \lambda_0$ и $\lambda_t \to \mu_k$. При этом $K(J, \lambda_{t+1}) \subset K(J, \lambda_t)$, следовательно пересечение

$$K = \bigcap_{t=1}^{\infty} K(J, \lambda_t)$$

не пусто, замкнуто и ограничено. Пусть $\hat{u} \in K$. Это значит, что $\hat{u} \in K(J, \lambda_t)$, то есть

$$\|a - V\hat{u}\|_\infty \leqslant \lambda_t, \quad t = 1, 2, \ldots.$$

В пределе при $t \to \infty$ получаем, что

$$\mu \leqslant \|a - V\hat{u}\|_\infty \leqslant \mu_k.$$

Но, как было отмечено выше, $\mu_k \leqslant \mu$. □

Доказанная теорема позволяет сформулировать следующий результат.

Теорема 1.12. *Пусть матрица $V \in \mathbb{C}^{n\times r}$, где $n > 2r$ ($n > r$ в вещественном случае) и вектор $a \in \mathbb{C}^n$ не принадлежит образу матрицы V. Тогда существует по крайней мере одно характеристическое множество, состоящее не более чем из $2r+1$ элементов ($r+1$ элементов в вещественном случае). Кроме того, если матрица V является чебышевской, то любое характеристическое множество состоит не менее чем из $r+1$ точек.*

Доказательство. Результат для произвольной системы сразу следует из Леммы 1.1 и Теоремы 1.11. В случае чебышевской матрицы на любом множестве из r и менее точек можно решить систему и точно приблизить вектор в этих точках, а поскольку множество является характеристическим, то это противоречит условию, что a не принадлежит образу V. □

Полезным для дальнейшего анализа является следующий результат.

Теорема 1.13. *Пусть $V \in \mathbb{C}^{n\times r}$ и $a \in \mathbb{C}^n$. Пусть $\hat{u} \in \mathbb{C}^r$ является решением задачи*

$$\|a - Vu\|_\infty \to \min_{u\in\mathbb{C}^r},$$

причем $J = (j_1, j_2, \ldots, j_k)$ является характеристическим множеством. Тогда

$$|\hat{u}^T v^{j_1} - a_{j_1}| = |\hat{u}^T v^{j_2} - a_{j_2}| = \cdots = |\hat{u}^T v^{j_k} - a_{j_k}| = \|a - V\hat{u}\|_\infty.$$

Доказательство. Согласно определению характеристического множества,

$$\max_{t\in\{1,\ldots,k\}} |\hat{u}^T v^{j_t} - a_{j_t}| = \|a - V\hat{u}\|_\infty.$$

Пусть существует номер p такой, что $|\hat{u}^T v^{j_p} - a_{j_p}| < \|a - V\hat{u}\|_\infty$. Обозначим $J' = J \setminus \{j_p\}$. Обозначим через $\hat{u}'$ решение задачи

$$\|a(J') - V(J')u\|_\infty \to \min_{u\in\mathbb{C}^r}.$$

Поскольку множество J характеристическое, согласно определению, $\|a(J') - V(J')\hat{u}'\|_\infty < \|a - V\hat{u}\|_\infty$.

Обозначим $\mu = \|a - V\hat{u}\|_\infty$, $\eta = |\hat{u}^T v^{j_p} - a_{j_p}|$, $\mu' = \|a(J') - V(J')\hat{u}'\|_\infty$, $\gamma = |(\hat{u}')^T v^{j_p} - a_{j_p}|$. Как было показано выше, $\eta < \mu$, $\mu' < \mu$. Кроме того, $\gamma \geqslant \mu$, поскольку в противном случае $\hat{u}$ не является решением на характеристическом множестве J. Заметим, что при данных условиях величина $\dfrac{\mu-\eta}{\gamma-\eta} > 0$. Возьмем $0 < \delta < \dfrac{\mu-\eta}{\gamma-\eta}$. Тогда

$$(\gamma - \eta)\delta < \mu - \eta.$$

$$(1-\delta)\eta + \delta\gamma < \mu. \tag{1.3}$$

Кроме того,

$$(1-\delta)\mu + \delta\mu' < \mu \tag{1.4}$$

при любом $\delta \in (0,1)$. Возьмем $\tilde{u} = (1-\delta)\hat{u} + \delta\hat{u}'$. Тогда

$$\tilde{w} = a(J) - V(J)\tilde{u} = (1-\delta)(a(J) - V(J)\hat{u}) + \delta(a(J) - V(J)\hat{u}').$$

Пусть $k \in J'$. Тогда $|(a_k - \hat{u}^T v^k)| \leqslant \mu$, $|a_k - (\hat{u}')^T v^k| \leqslant \mu'$, откуда согласно (1.4) имеем $|\tilde{w}_k| < \mu$. Пусть $k = j_p$. Тогда согласно (1.3), $|\tilde{w}_k| \leqslant (1-\delta)\eta + \delta\gamma < \mu$. Следовательно, $\|\tilde{w}\|_\infty < \mu$ и $\hat{u}$ не является решением на множестве J, и J не является характеристическом множеством. Получаем противоречие. □

1.5 Критерии оптимальности

В теории чебышевских приближений непрерывных функций широко известны критерии Колмогорова и Ремеза для оптимальности приближения (см. например [26]). Приведем их аналоги для векторной задачи о наилучшем равномерном приближении.

Теорема 1.14. *Пусть матрица $V \in \mathbb{C}^{n\times r}$ и вектор $a \in \mathbb{C}^n$. Пусть $\hat{u} \in \mathbb{C}^r$ и пусть*

$$J = \{j \in \{1, 2, \ldots, n\} : |w_j| = \|w\|_\infty\},$$

где $w = a - V\hat{u}$. Тогда для того чтобы вектор $\hat{u}$ был решением задачи

$$\|a - Vu\|_\infty \to \min_{u\in\mathbb{C}^r}$$

необходимо и достаточно, чтобы на множестве J выполнялось

$$\min_{j\in J} \operatorname{Re}\left((Vu)_j \overline{w_j}\right) \leqslant 0$$

для любого $u \in \mathbb{C}^r$.

Доказательство. Докажем необходимость. Пусть $\hat{u}$ является решением задачи о наилучшем равномерном приближении. От противного, пусть

$$\min_{j\in J} \operatorname{Re}\left((Vu)_j \overline{w_j}\right) > c > 0$$

для некоторого вектора $u \in \mathbb{C}^r$. Обозначим

$$G = \max_{j \in J} |w_j|, \quad G' = \max_{j \notin J} |w_j|,$$

$$h = G - G' > 0, \quad M = \max_j |(Vu)_j|, \quad \lambda = \max\left\{\frac{c}{M^2}, \frac{h}{2M}\right\} > 0.$$

Покажем, что в этом случае вектор $\hat{u} + \lambda u$ дает лучшее решение, чем $\hat{u}$.

1. Пусть $j \in J$. Тогда

$$|a_j - (\hat{u} + \lambda u)^T v^j|^2 = (a_j - \hat{u}^T v^j - \lambda u^T v^j) \cdot (\overline{(a_j - \hat{u}^T v^j)} - \lambda \overline{(u^T v^j)}) =$$
$$|a_j - \hat{u}^T v^j|^2 + \lambda^2 |u^T v^j|^2 - 2\lambda \operatorname{Re}\left((Vu)_j \overline{(a_j - (V\hat{u})_j)}\right) \leqslant$$
$$G^2 + \lambda^2 M^2 - 2\lambda \operatorname{Re}\left((Vu)_j \overline{(a_j - (V\hat{u})_j)}\right) <$$
$$G^2 + \lambda^2 M^2 - 2\lambda c \leqslant G^2 + \lambda \frac{c}{M^2} M^2 - 2\lambda c = G^2 - \lambda c < G^2.$$

2. Пусть $j \notin J$. Тогда

$$|a_j - (\hat{u} + \lambda u)^T v^j| \leqslant |a_j - \hat{u}^T v^j| + \lambda |u^T v^j| \leqslant$$
$$G' + \lambda M \leqslant G - h + \frac{h}{2M} M = G - h/2 < G.$$

Отсюда следует, что вектор $\hat{u} + \lambda u$ дает лучшее решение, чем $\hat{u}$. Получили противоречие.

Докажем достаточность. Пусть выполнено условие теоремы с вектором $\hat{u} \in \mathbb{C}^r$ и пусть $u \in \mathbb{C}^r$ произвольный. Выберем индекс j_0, для которого выполняется неравенство

$$\operatorname{Re}\left((V(u - \hat{u}))_{j_0} \overline{(a_{j_0} - (V\hat{u})_{j_0})}\right) \leqslant 0.$$

Тогда

$$|a_{j_0} - (Vu)_{j_0}|^2 = |a_{j_0} - (V\hat{u})_{j_0} - ((Vu)_{j_0} - (V\hat{u})_{j_0})|^2 =$$
$$|a_{j_0} - (V\hat{u})_{j_0}|^2 + |(Vu)_{j_0} - (V\hat{u})_{j_0}|^2 -$$
$$- 2 \operatorname{Re}\left((V(u - \hat{u}))_{j_0} \overline{(a_{j_0} - (V\hat{u})_{j_0})}\right) \geqslant |a_{j_0} - (V\hat{u})_{j_0}|^2.$$

Отсюда видно, что для любого вектора $u \in \mathbb{C}^r$, вектор $\hat{u}$ дает приближение не хуже, то есть является оптимальным. □

Приведем другой, в некоторых ситуациях более удобный, критерий оптимальности. В некотором смысле он является переформулировкой Теоремы 1.14 с использованием следующей леммы.

Лемма 1.15. *Пусть $U \in \mathbb{C}^{m\times n}$ — некоторая матрица. Для того, чтобы существовал ненулевой вектор $\delta \in \mathbb{R}^m$ с неотрицательными компонентами такой, что $U^T\delta = 0$, необходимо и достаточно, чтобы для любого вектора $c \in \mathbb{C}^n$ неравенства*

$$\operatorname{Re}\sum_{j=1}^{n} c_j u_{ij} > 0, \quad i = 1,\dots,m$$

не выполнялись одновременно.

Доказательство. Докажем необходимость. Пусть $\delta \in \mathbb{R}^m$ является ненулевым, имеет неотрицательные компоненты и $U^T\delta = 0$. Тогда

$$\sum_{i=1}^{m} \delta_i \operatorname{Re}\sum_{j=1}^{n} c_j u_{ij} = \operatorname{Re}\sum_{j=1}^{n} c_j \sum_{i=1}^{m} \delta_i u_{ij} = 0,$$

откуда следует, что для любой системы c_j условия

$$\operatorname{Re}\sum_{j=1}^{n} c_j u_{ij} > 0, \quad i = 1,\dots,m$$

не могут быть выполнены одновременно.

Докажем достаточность. Введем величину

$$v = v(\delta) = \sum_{j=1}^{n} \left|\sum_{i=1}^{m} \delta_i u_{ij}\right|^2, \quad \delta_i \geqslant 0, \quad \sum_{i=1}^{m} \delta_i = 1.$$

Так как функция v непрерывна на компактном множестве, то она принимает минимальное значение v_0 при $\delta = \delta^{(0)}$. Легко видеть, что условие леммы эквивалентно тому, что если неравенства $\operatorname{Re}\sum\limits_{j=1}^{n} c_j u_{ij} > 0$ не могут быть выполнены одновременно, то $v_0 = 0$. Докажем это от противного. Пусть неравенства не могут быть выполнены одновременно ни при каком выборе c, но $v_0 > 0$. Возьмем

$$c_j = \sum_{i=1}^{m} \delta_i^{(0)} \overline{u}_{ij}$$

и пусть при таком выборе c не ограничивая общности не выполнено неравенство при $i = m$:

$$\operatorname{Re}\sum_{j=1}^{n} \left(\sum_{i=1}^{m} \delta_i^{(0)} \overline{u}_{ij}\right) u_{mj} \leqslant 0.$$

Обозначим

$$v_* = \sum_{j=1}^{m} |u_{mj}|^2, \quad \lambda = \frac{v_*}{v_* + v_0} < 1.$$

Выберем δ следующим образом:

$$\delta_i = \begin{cases} \lambda\delta_i^{(0)}, & i = 1,2,\ldots,m-1 \\ (1-\lambda) + \lambda\delta_m^{(0)}, & i = m \end{cases}$$

и покажем, что $v(\delta) < v_0$. Действительно,

$$v = \sum_{j=1}^{n} \left| \sum_{i=1}^{m} \delta_i u_{ij} \right|^2 = \sum_{j=1}^{n} \left| (1-\lambda)u_{mj} + \lambda \sum_{i=1}^{m} \delta_i^{(0)} u_{ij} \right|^2 =$$

$$(1-\lambda)^2 \sum_{j}^{n} |u_{mj}|^2 + \lambda^2 \sum_{j=1}^{n} \left| \sum_{i=1}^{m} \delta_i^{(0)} u_{ij} \right|^2 + 2\lambda(1-\lambda)\,\mathrm{Re} \left(\sum_{j=1}^{n} \sum_{i=1}^{m} \delta_i^{(0)} \overline{u}_{ij} u_{mj} \right).$$

Как было отмечено выше,

$$\mathrm{Re} \left(\sum_{j=1}^{n} \sum_{i=1}^{m} \delta_i^{(0)} \overline{u}_{ij} u_{mj} \right) \leqslant 0,$$

но $\lambda \geqslant 0$, $1 - \lambda > 0$, откуда

$$v \leqslant (1-\lambda)^2 v_* + \lambda^2 v_0 = \frac{v_0^2}{(v_* + v_0)^2} v_* + \frac{v_*^2}{(v_* + v_0)^2} v_0 =$$

$$v_0 v_* \frac{v_* + v_0}{(v_* + v_0)^2} = \frac{v_*}{v_* + v_0} v_0 = \lambda v_0 < v_0.$$

Пришли к противоречию с оптимальностью v_0, следовательно $v_0 = 0$. □

Замечание. *Если все $U \in \mathbb{R}^{m\times n}$, то c достаточно выбирать вещественными.*

Используя доказанную лемму и Теорему 1.14, докажем другой критерий оптимальности равномерного приближения.

Теорема 1.16. *Пусть $V \in \mathbb{C}^{n\times r}$ и $a \in \mathbb{C}^n$. Пусть $\hat{u} \in \mathbb{C}^r$ и*

$$J = \{j \in \{1,2,\ldots,n\} : |w_j| = \|w\|_\infty\},$$

где $w = a - V\hat{u}$. Тогда $\hat{u}$ является решением задачи наилучшего равномерного приближения

$$\|a - Vu\|_\infty \to \min_{u\in\mathbb{C}^r}$$

тогда и только тогда, когда существует ненулевой вектор $\delta \in \mathbb{R}^{|J|}$ *с неотрицательными компонентами такой, что*

$$V(J)^T(\operatorname{diag}\overline{w(J)})\delta = 0.$$

Доказательство. Докажем достаточность. Пусть выполнено условие

$$V(J)^T(\operatorname{diag}\overline{w(J)})\delta = 0.$$

Обозначим $u_{kj} = \overline{w_k}v_{kj}$. Тогда согласно Лемме 1.15 условия

$$\operatorname{Re}\sum_{j\in J} c_j\overline{w_k}v_{kj} > 0$$

не выполнены одновременно для любого выбора $c \in \mathbb{C}^{|J|}$. Принимая во внимание, что $\sum\limits_{j\in J} c_j v_{kj}$ задает произвольный вектор вида Vc на строках из J, получаем, что выполнено условие Теоремы 1.14 и $\hat{u}$ является решением.

Докажем необходимость. Если $\hat{u}$ является решением задачи о наилучшем равномерном приближении, то по Теореме 1.14

$$\min_{k\in J}\operatorname{Re}\sum_{j\in J} c_j v_{kj}\overline{w_k} \leqslant 0,$$

следовательно условия

$$\operatorname{Re}\sum_{j\in J} c_j v_{kj}\overline{w_k} > 0$$

не выполнены одновременно и тогда по Лемме 1.15 существует ненулевой вектор $\delta \in \mathbb{R}^{|J|}$ с неотрицательными компонентами, удовлетворяющий условиям теоремы. □

Замечание. *Условие теоремы эквивалентно* $V(J)^T(\operatorname{diag}\operatorname{sign}\overline{w(J)})\delta = 0$.

Следствие 1.17. *Пусть* $V \in \mathbb{C}^{n\times r}$ *и* $a \in \mathbb{C}^n$. *Пусть* $J \subset \{1,\dots,n\}$ *и пусть* $\tilde{u} \in \mathbb{C}^r$ *является решением задачи*

$$\|a(J) - V(J)u\|_\infty \to \min_{u\in\mathbb{C}^r}.$$

Пусть также $\|a(J) - V(J)\tilde{u}\|_\infty = \|a - V\tilde{u}\|_\infty$. *Тогда* $\tilde{u}$ *является решением задачи*

$$\|a - Vu\|_\infty \to \min_{u\in\mathbb{C}^r}.$$

Доказательство. Поскольку $\tilde{u}$ является решением для подзадачи на множестве J, то выполнено условие Теоремы 1.16 на некотором множестве $\tilde{J} \subset J$ для некоторого ненулевого вектора $\tilde{\delta}$ с неотрицательными компонентами. Обозначим $\hat{J} = \{j \in \{1, 2, \ldots, n\} : |w_j| = \|w\|_\infty\}$, где $w = a - V\tilde{u}$. В силу условия $\|a(J) - V(J)\tilde{u}\|_\infty = \|a - V\tilde{u}\|_\infty$, $\tilde{J} \subset \hat{J}$. Дополняя $\tilde{\delta}$ нулями для всех позиций из $\hat{J} \setminus \tilde{J}$, получаем, что условие Теоремы 1.16 выполнено для всей задачи. □

1.6 О задаче поиска равноудаленных точек

Из результатов Раздела 1.4 следует, что задача о построении наилучшего равномерного приближения может быть сведена к набору аналогичных задач с матрицей размера $(r+1) \times r$. Приведем результаты о решении соответствующей задачи. Теорема 1.19 и Теорема 1.20 впервые были доказаны в [28] и приведены здесь с доказательствами для полноты изложения.

Введем следующее

Определение 1.18. *Пусть $V \in \mathbb{C}^{(r+1)\times r}$ и $a \in \mathbb{C}^{r+1}$. Пусть система $Vu = a$ не совместна. Будем называть точку u равноудаленной точкой системы, если*

$$\rho(u) = |u^T v^1 - a_1| = |u^T v^2 - a_2| = \cdots = |u^T v^{r+1} - a_{r+1}|.$$

Будем называть точку u наилучшей равноудаленной точкой системы, если она является равноудаленной и величина $\rho(u)$ минимальна.

Заметим, что из Утверждения 1.6 следует, что решение задачи (1.1) при $n = r + 1$ является наилучшей равноудаленной точкой данной системы.

Пусть $V \in \mathbb{C}^{(r+1)\times r}$ и $a \in \mathbb{C}^{r+1}$. Обозначим через $\hat{u}^{(j)} \in \mathbb{C}^r$ решение системы $V^{\setminus j}\hat{u}^{(j)} = a_{\setminus j}$. Верна следующая теорема о множестве равноудаленных точек системы.

Теорема 1.19. *Пусть задана несовместная система уравнений $Vu = a$, где $V \in \mathbb{C}^{(r+1)\times r}$ и $a \in \mathbb{C}^{r+1}$. Тогда*

1. При каждом $j = 1, \ldots, r+1$ имеет место равенство

$$(\hat{u}^{(j)})^T v^j - a_j = \frac{(-1)^{j+1}}{D_j(V)} \sum_{\nu=1}^{r+1} (-1)^\nu a_\nu D_\nu(V).$$

2. *Для любых действительных k_j, $j = 1, \ldots, r+1$ таких, что*

$$\sum_{j=1}^{r+1} |D_j(V)| e^{ik_j} \neq 0,$$

точка u, определяемая по формуле

$$u = \rho \sum_{j=1}^{r+1} \frac{\hat{u}^{(j)} e^{ik_j}}{|(\hat{u}^{(j)})^T v^j - a_j|} = \rho \frac{\sum\limits_{j=1}^{r+1} |D_j(V)| \hat{u}^{(j)} e^{ik_j}}{\left| \sum\limits_{j=1}^{r+1} (-1)^j D_j(V) a_j \right|},$$

где

$$\rho = \left(\sum_{j=1}^{r+1} \frac{e^{ik_j}}{|(\hat{u}^{(j)})^T v^j - a_j|} \right)^{-1} = \frac{\left| \sum\limits_{j=1}^{r+1} (-1)^j D_j(V) a_j \right|}{\sum\limits_{j=1}^{r+1} |D_j(V)| e^{ik_j}}$$

является равноудаленной точкой системы $Vu = a$, при этом $|\rho| = \|Vu - a\|_\infty$.

3. *Всякая равноудаленная точка u системы $Vu = a$ может быть при некоторых действительных k_j представлена по формуле выше. При этом k_j могут, в частности, быть выражены по формуле*

$$k_j = \arg(u^T v^j - a_j) - \arg((\hat{u}^{(j)})^T v^j - a_j).$$

Доказательство. Обозначим через $A^j_{\nu k}$ алгебраическое дополнение в матрице $V^{\setminus j}$ к элементу $v_{\nu k}$, а через $M^j_{\nu k}$ соответствующий минор. Представляя решение по формулам Крамера и раскладывая по столбцу, получаем

$$\hat{u}^{(j)}_k = \frac{\sum\limits_{\nu=1, \nu \neq j}^{r+1} a_\nu A^j_{\nu k}}{D_j(V)} = \frac{\sum\limits_{\nu=1, \nu \neq j}^{r+1} a_\nu (-1)^{\nu+k} \operatorname{sign}(j - \nu) M^j_{\nu k}}{D_j(V)}.$$

Заметим, что $M^j_{\nu k} = M^\nu_{jk}$, поскольку оба получаются удалением из матрицы V j–ой и ν–ой строк и k–го столбца.

$$A^j_{\nu k} = (-1)^{\nu+k} \operatorname{sign}(j - \nu) M^j_{\nu k} = (-1)^{\nu+k} (-\operatorname{sign}(\nu - j)) M^\nu_{jk} = (-1)^{\nu+k} (-1)(-1)^{j+k} A^\nu_{jk} = -(-1)^{\nu+j} A^\nu_{jk}.$$

Тогда

$$(\hat{u}^{(j)})^T v^j - a_j = \sum_{k=1}^{r} v_{jk}\hat{u}_k^{(j)} - a_j = \sum_{k=1}^{r} v_{jk}\frac{\sum\limits_{\nu=1,\nu\neq j}^{r+1} a_\nu A_{\nu k}^j}{D_j(V)} - a_j =$$

$$\frac{1}{D_j(V)}\sum_{k=1}^{r} v_{jk}\sum_{\nu=1,\nu\neq j}^{r+1} a_\nu(-1)^{\nu+j+1}A_{jk}^\nu - a_j = \frac{(-1)^{j+1}}{D_j(V)}\sum_{\nu=1,\nu\neq j}^{r+1}(-1)^\nu a_\nu\sum_{k=1}^{r} v_{jk}A_{jk}^\nu - a_j.$$

Заметим, что

$$\sum_{k=1}^{r} v_{jk}A_{jk}^\nu = D_\nu(V),$$

откуда

$$(\hat{u}^{(j)})^T v^j - a_j = \frac{(-1)^{j+1}}{D_j(V)}\sum_{\nu=1,\nu\neq j}^{r+1}(-1)^\nu a_\nu D_\nu(V) - a_j = \frac{(-1)^{j+1}}{D_j(V)}\sum_{\nu=1}^{r+1}(-1)^\nu a_\nu D_\nu(V),$$

тем самым первое утверждение теоремы доказано.

Далее докажем третью часть теоремы. Пусть $u \in \mathbb{C}^r$ — произвольная равноудаленная точка системы $Vu = a$. Докажем, что u может быть представлен по формулам из формулировки теоремы. Пусть

$$k_j = \arg(u^T v^j - a_j) - \arg((\hat{u}^{(j)})^T v^j - a_j).$$

Так как u равноудаленная,

$$u^T v^j - a_j = \rho b_j, \quad \rho > 0, b_j = e^{i\arg(u^T v^j - a_j)}.$$

Тогда

$$b_j = e^{i(\arg(u^T v^j - a_j) - \arg((\hat{u}^{(j)})^T v^j - a_j))}e^{i\arg((\hat{u}^{(j)})^T v^j - a_j)} = \frac{(\hat{u}^{(j)})^T v^j - a_j}{|(\hat{u}^{(j)})^T v^j - a_j|}e^{ik_j}. \qquad (1.5)$$

Заметим, что система

$$u^T v^j - a_j = \rho b_j, \quad j = 1, \ldots, r+1$$

совместна, поэтому

$$\det\begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1r} & a_1 + \rho b_1 \\ v_{21} & v_{22} & \ldots & v_{2r} & a_2 + \rho b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{r+1,1} & v_{r+1,2} & \ldots & v_{r+1,r} & a_{r+1} + \rho b_{r+1} \end{bmatrix} = 0.$$

Разложим этот определитель по последнему столбцу.

$$\sum_{j=1}^{r+1}(-1)^{r+1+j}(a_j+\rho b_j)D_j(V)=0,$$

откуда

$$\sum_{j=1}^{r+1}(-1)^{j}(a_j+\rho b_j)D_j(V)=0.$$

$$\rho\sum_{j=1}^{r+1}(-1)^{j}b_jD_j(V)=-\sum_{j=1}^{r+1}(-1)^{j}a_jD_j(V).$$

$$\rho=-\frac{\sum\limits_{j=1}^{r+1}(-1)^{j}a_jD_j(V)}{\sum\limits_{j=1}^{r+1}(-1)^{j}b_jD_j(V)}. \tag{1.6}$$

По доказанной первой части теоремы

$$(\hat{u}^{(j)})^Tv^j-a_j=\frac{(-1)^{j+1}}{D_j(V)}\sum_{\nu=1}^{r+1}(-1)^{\nu}a_\nu D_\nu(V),$$

откуда и из (1.5) имеем

$$b_jD_j(V)=D_j(V)\left(\frac{(-1)^{j+1}}{D_j(V)}\sum_{\nu=1}^{r+1}(-1)^{\nu}w_\nu D_\nu(V)\right)\frac{e^{ik_j}}{|(\hat{u}^{(j)})^Tv^j-a_j|}=$$

$$(-1)^{j+1}\sum_{\nu=1}^{r+1}(-1)^{\nu}a_\nu D_\nu(V)\cdot\frac{e^{ik_j}}{|(\hat{u}^{(j)})^Tv^j-a_j|}.$$

Тогда, подставляя в (1.6), получаем

$$\rho=-\frac{\sum\limits_{j=1}^{r+1}(-1)^{j}a_jD_j(V)}{\sum\limits_{j=1}^{r+1}(-1)^{j}(-1)^{j+1}\left(\sum\limits_{\nu=1}^{r+1}(-1)^{\nu}a_\nu D_\nu(V)\right)\cdot\frac{e^{ik_j}}{|(\hat{u}^{(j)})^Tv^j-a_j|}}=$$

$$\frac{1}{\sum\limits_{j=1}^{r+1}\frac{e^{ik_j}}{|(\hat{u}^{(j)})^Tv^j-a_j|}}.$$

Обозначим

$$\tilde{u}=\rho\sum_{j=1}^{r+1}\frac{\hat{u}^{(j)}e^{ik_j}}{|(\hat{u}^{(j)})^Tv^j-a_j|} \tag{1.7}$$

и докажем, что $u = \tilde{u}$. Поскольку вся система не совместна, то $(\hat{u}^{(j)})^T v^j - a_j \neq 0$ и $\tilde{u}$ определено корректно. Из (1.7) имеем

$$\tilde{u}^T v^j - a_j = \rho \sum_{\nu=1}^{r+1} \frac{(\hat{u}^{(\nu)})^T v^j e^{ik_\nu}}{|(\hat{u}^{(\nu)})^T v^\nu - a_\nu|} - a_\nu \rho \frac{1}{\rho}.$$

Из последней формулы для ρ получаем

$$\frac{1}{\rho} = \sum_{j=1}^{r+1} \frac{e^{ik_j}}{|(\hat{u}^{(j)})^T v^j - a_j|}.$$

Отсюда имеем, что

$$\tilde{u}^T v^j - a_j = \rho \sum_{\nu=1}^{r+1} \frac{((\hat{u}^{(\nu)})^T v^j - a_j) e^{ik_\nu}}{|(\hat{u}^{(\nu)})^T v^\nu - a_\nu|}.$$

Заметим, что по определению решения $\hat{u}^{(\nu)}$ только одно из слагаемых не равно 0, поэтому

$$\tilde{u}^T v^j - a_j = \rho \frac{((\hat{u}^{(j)})^T v^j - a_j) e^{ik_j}}{|(\hat{u}^{(j)})^T v^j - a_j|} = \rho b_j.$$

Тогда получаем, что

$$\tilde{u}^T v^j - a_j = \rho b_j,$$
$$u^T v^j - a_j = \rho b_j,$$

откуда

$$(u - \tilde{u})^T v^j = 0, \quad j = 1, \ldots, r+1.$$

В силу невырожденности определителей системы получаем, что $u = \tilde{u}$, откуда третья часть теоремы доказана.

Вторая часть утверждения теоремы получается аналогично, а именно, если u задается согласно формулам из формулировки, то

$$u^T v^j - a_j = \rho b_j.$$

Остается применить к этим формулам результат первой части теоремы. □

Теорема 1.20. *Пусть задана несовместная система уравнений $Vu = a$, где $V \in \mathbb{C}^{(r+1)\times r}$ и $a \in \mathbb{C}^{r+1}$. Тогда наилучшая равноудаленная точка системы $\hat{u}$ определяется по формуле*

$$\hat{u} = \hat{\rho} \sum_{j=1}^{r+1} \frac{\hat{u}^{(j)}}{|(\hat{u}^{(j)})^T v^j - a_j|} = \frac{\sum\limits_{j=1}^{r+1} |D_j(V)| \hat{u}^{(j)}}{\sum\limits_{j=1}^{r+1} |D_j(V)|},$$

где

$$\hat{\rho} = \left(\sum_{j=1}^{r+1} \frac{1}{|(\hat{u}^{(j)})^T v^j - a_j|} \right)^{-1} = \frac{\left| \sum\limits_{j=1}^{r+1} (-1)^j D_j(V) a_j \right|}{\sum\limits_{j=1}^{r+1} |D_j(V)|}$$

Доказательство. Достаточно заметить, что величина

$$\rho = \left(\sum_{j=1}^{r+1} \frac{e^{ik_j}}{|(\hat{u}^{(j)})^T v^j - a_j|} \right)^{-1}$$

принимает наименьшее значение, когда все слагаемые сонаправлены, то есть $e^{ik_1} = e^{ik_2} = \cdots = e^{ik_{r+1}}$. □

Приведенные результаты удобны для доказательств, однако если явно вычислять решения по доказанным формулам, то решение задачи о наилучшем равномерном приближении потребует $O(r^4)$ операций. В Разделе 1.10 будет построен алгоритм, которых находит наилучшую равноудаленную точку системы за $O(r^2)$ операций, если известно QR разложение матрицы V.

1.7 Комбинаторная формула решения в вещественном случае

Рассмотрим задачу

$$\|a - Vu\|_\infty \to \min_{u \in \mathbb{R}^r},$$

где $V \in \mathbb{R}^{n \times r}$ и $a \in \mathbb{R}^n$, причем V является чебышевской. Пусть $\hat{V} \in \mathbb{R}^{(r+1) \times r}$ и $\hat{a} \in \mathbb{R}^{r+1}$. Тогда через $\begin{bmatrix} \hat{V} & \hat{a} \end{bmatrix} \in \mathbb{R}^{(r+1) \times (r+1)}$ обозначим матрицу, первые r столбцов которой являются столбцами матрицы $\hat{V}$, а последний столбец совпадает с $\hat{a}$.

Теорема 1.21. *Пусть $V \in \mathbb{R}^{n \times r}$ является чебышевской и $a \in \mathbb{R}^n$. Тогда*

$$\inf_{u \in \mathbb{R}^r} \|a - Vu\|_\infty = \max_{J' \subset \{1,2,\ldots,n\}, |J'| = r+1} \frac{\left| \det \begin{bmatrix} V(J') & a(J') \end{bmatrix} \right|}{\sum\limits_{j \in J'} |\det (V(J' \setminus \{j\}))|},$$

причем максимум достигается на характеристическом множестве задачи.

Доказательство. Обозначим $\mu = \inf\limits_{u\in\mathbb{R}^r} \|a - Vu\|_\infty$. По Теореме 1.11 имеем, что $\mu = \mu_{r+1}$, а согласно определению

$$\mu_{r+1} = \max_{J'\subset\{1,2,\ldots,n\},|J'|=r+1} \mu(J').$$

Остается применить Теорему 1.20 для получения явного вида $\mu(J')$. □

1.8 Теорема об альтернансе

Одним из фундаментальных результатов в теории равномерных приближений непрерывных функций является теорема Чебышева об альтернансе (см., например, [26]).

Теорема 1.22. *Пусть f является непрерывной вещественной функцией, определенной на отрезке $[a, b]$. Среди всех полиномов степени не выше d, полином g доставляет минимальное значение функционала $\|f - g\|_C$ тогда и только тогда, когда существует $d + 2$ точки $a \leqslant x_0 < x_1 < \cdots < x_{d+1} \leqslant b$ такие, что*

$$f(x_j) - g(x_j) = \sigma(-1)^j \|f - g\|_C,$$

где $\sigma \in \{-1, 1\}$.

Рассмотрим задачу наилучшего равномерного приближения

$$\|Vu - a\|_\infty \to \min_{u\in\mathbb{R}^r}, \tag{1.8}$$

где $V \in \mathbb{R}^{n\times r}$ и $a \in \mathbb{R}^n$. В этом случае можно доказать результат, аналогичный приведенному в Теореме 1.22. Для начала докажем следующую лемму.

Лемма 1.23. *Рассмотрим задачу (1.8) с чебышевской матрицей $V \in \mathbb{R}^{(r+1)\times r}$ и вектором $a \in \mathbb{R}^{r+1}$, который не принадлежит образу матрицы V. Пусть $\hat{u} \in \mathbb{R}^r$. Обозначим невязку через $w = a - V\hat{u}$. Тогда $\hat{u}$ является решением задачи (1.8) тогда и только тогда, когда*

$$|w_1| = |w_2| = \cdots = |w_{r+1}| = \|w\|_\infty$$

и знаки в последовательности

$$w_1 D_1(V), w_2 D_2(V), \ldots, w_{r+1} D_{r+1}(V)$$

чередуются.

Доказательство. Пусть $\hat{u} \in \mathbb{R}^r$ является решением (1.8). По Теореме 1.20 мы имеем

$$\hat{u} = \sum_{j=1}^{r+1} \frac{|D_j(V)|}{\sum\limits_{k=1}^{r+1} |D_k(V)|} \hat{u}^{(j)},$$

где $\hat{u}^{(j)}$ является решением задачи $V^{\setminus j} u = a_{\setminus j}$. Более того, по Теореме 1.19 мы имеем

$$(\hat{u}^{(j)})^T v^j - a_j = \frac{(-1)^{j+1}}{D_j(V)} x, \quad \text{где} \quad x = \sum_{k=1}^{r+1} (-1)^k a_k D_k(V). \tag{1.9}$$

По определению $\hat{u}^{(j)}$ мы имеем

$$V\hat{u}^{(j)} = \begin{bmatrix} a_1 & a_2 & \dots & a_{j-1} & \tilde{a}_j & a_{j+1} & \dots & a_{n+1} \end{bmatrix}^T,$$

где $\tilde{a}_j = (\hat{u}^{(j)})^T v^j$. Обозначим $D = \operatorname{diag}(D_1(V), \dots, D_{r+1}(V))$ и $\hat{D} = \operatorname{diag}(|D_1(V)|, \dots, |D_{r+1}(V)|)$. Тогда

$$V\hat{u} = \sum_{j=1}^{r+1} \frac{|D_j(V)|}{\sum\limits_{k=1}^{r+1} |D_k(V)|} (a + (\tilde{a}_j - a_j) e_j) = a + \frac{\hat{D}(\tilde{a} - a)}{\sum\limits_{k=1}^{r+1} |D_k(V)|}.$$

Тогда из (1.9) мы имеем

$$w = a - V\hat{u} = \frac{\hat{D}(\tilde{a} - a)}{\sum\limits_{k=1}^{r+1} |D_k(V)|} = \frac{x}{\sum\limits_{k=1}^{r+1} |D_k(V)|} \hat{D} S D^{-1} e,$$

где $S = \operatorname{diag}((-1)^1, (-1)^2, \dots, (-1)^{r+1})$ и $e = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$, откуда

$$w_j D_j = \frac{x}{\sum\limits_{k=1}^{r+1} |D_k(V)|} (-1)^j |D_j|$$

и знаки в последовательности

$$w_1 D_1, w_2 D_2, \dots, w_{r+1} D_{r+1}$$

чередуются. Более того, из Утверждения 1.6, модули всех элементов вектора w равны.

Докажем обратное. Пусть $\hat{u} \in \mathbb{R}^r$ является таким вектором, что для $w = a - V\hat{u}$ мы имеем

$$|w_1| = |w_2| = \dots = |w_{r+1}| = \|w\|_\infty$$

и знаки в последовательности

$$w_1 D_1(V), w_2 D_2(V), \ldots, w_{n+1} D_{r+1}(V)$$

чередуются. Покажем, что $\hat{u}$ является решением (1.8). Из предположения на w мы имеем, что

$$w_j = c(-1)^j \operatorname{sign} D_j(V). \tag{1.10}$$

По Теореме 1.16 $\hat{u}$ является решением (1.8) тогда и только тогда, когда существует ненулевой вектор $\delta \in \mathbb{R}^{r+1}$ с неотрицательными компонентами такой, что

$$V^T \operatorname{diag}(\operatorname{sign} w)\delta = 0.$$

Подставляя (1.10), получаем

$$V^T S D \hat{D}^{-1} \delta = 0.$$

Рассмотрим матрицу $\tilde{V}^{(k)} \in \mathbb{R}^{(r+1)\times(r+1)}$, которая соответствует матрице V с удвоенным k–ым столбцом:

$$\tilde{V}^{(k)} = \begin{bmatrix} v_1 & v_2 & \ldots & v_{k-1} & v_k & v_k & v_{k+1} & \ldots & v_{r+1} \end{bmatrix}.$$

Очевидно, что $\tilde{V}_k$ является вырожденной. Применим разложение Лапласа к k–му столбцу $\tilde{V}_k$. Тогда

$$\sum_{j=1}^{r+1} (-1)^{k+j} D_j(V) v_{jk} = 0. \tag{1.11}$$

Выбирая $\delta_j = |D_j|$, получаем, что

$$V^T S D \hat{D}^{-1} \delta = V^T S D e,$$

что равно нулю по (1.11). □

Наконец, мы готовы к тому, чтобы доказать теорему об альтернансе для задачи (1.8).

Теорема 1.24. *Рассмотрим задачу (1.8) с чебышевской матрицей $V \in \mathbb{R}^{n\times r}$ и вектором $a \in \mathbb{R}^n$, который не принадлежит образу матрицы V. Пусть $\hat{u} \in \mathbb{R}^r$. Обозначим невязку через $w = a - V\hat{u}$. Тогда $\hat{u}$ является решением (1.8) тогда и только тогда, когда существует множество целых чисел $1 \leqslant j_1 < j_2 < \cdots < j_{r+1} \leqslant n$ такое, что*

$$|w_{j_1}| = |w_{j_2}| = \cdots = |w_{j_{r+1}}| = \|w\|_\infty$$

и знаки в последовательности

$$w_{j_1}\Delta_1, w_{j_2}\Delta_2, \ldots, w_{j_{r+1}}\Delta_{r+1}$$

чередуются, где $\Delta_k = \det V((j_1, \ldots, j_{k-1}, j_{k+1}, \ldots, j_{r+1}))$.

Доказательство. Пусть $\hat{u}$ является решением (1.8). Тогда по Теореме 1.12 существует характеристическое множество, состоящее из $r+1$ элементов $j_1 < j_2 < \cdots < j_{r+1}$. Обозначим $J' = (j_1, j_2, \ldots, j_{r+1})$. По определению характеристического множества $\hat{u}$ является решением задачи

$$\|a(J') - V(J')u\|_\infty \to \min_{u \in \mathbb{C}^r},$$

следовательно выполнены условия Леммы 1.23 для $\hat{u}$, откуда следует утверждение теоремы.

Пусть $J' = (j_1, j_2, \ldots, j_{r+1})$ удовлетворяет условиям теоремы. Тогда по Лемме 1.23 вектор $\hat{u}$ является решением задачи

$$\|a(J') - V(J')u\|_\infty \to \min_{u \in \mathbb{R}^r},$$

откуда по Теореме 1.16 существует ненулевой вектор $\delta \in \mathbb{R}^{r+1}$ с неотрицательными компонентами такой, что

$$V(J')^T \operatorname{diag}(\operatorname{sign} w(J'))\delta = 0.$$

Пусть

$$\hat{J} = \{j \in \{1, 2, \ldots, n\} : |w_j| = \|w\|_\infty\}.$$

Очевидно, $J' \subset \hat{J}$. Обозначим через $E_{(k)}$ и $J'_{(k)}$ k–ый наименьший элемент множеств $\hat{J}$ и J' соответственно. Обозначим через $\hat{\delta} \in \mathbb{R}^{|\hat{J}|}$ вектор такой, что

$$\hat{\delta}_j = \begin{cases} \delta_k, & \hat{J}_{(j)} \in J' \text{ и } J'_{(k)} = \hat{J}_{(j)} \\ 0, & \text{иначе.} \end{cases}$$

Тогда

$$V(\hat{J})^T \operatorname{diag}(\operatorname{sign} w(\hat{J}))\hat{\delta} = V(J')^T \operatorname{diag}(\operatorname{sign} w(J'))\delta = 0,$$

откуда по Теореме 1.16 $\hat{u}$ является решением задачи (1.8). □

В дальнейшем нас будет особо интересовать случай $r = 1$, поэтому сформулируем соответствующий результат отдельно.

Следствие 1.25. *Пусть* $a,v \in \mathbb{R}^n$ *и все компоненты вектора* v *не равны нулю. Тогда существует единственное* $t \in \mathbb{R}$ *такое, что*

$$\|a - tv\|_\infty = \inf_{u\in\mathbb{R}} \|a - uv\|.$$

Более того, $u = t$ *тогда и только тогда, когда существует пара различных индексов* $i,j \in \{1,\dots,n\}$ *таких, что*

$$\begin{aligned} |a_i - uv_i| = |a_j - uv_j| &= \|a - uv\|_\infty, \\ \operatorname{sign}(v_i(a_i - uv_i)) &= -\operatorname{sign}(v_j(a_j - uv_j)). \end{aligned}$$

Рассмотрим систему точек $x_0,\dots,x_{d+1}$ на отрезке $[a,b] \subset \mathbb{R}$, а именно,

$$a \leqslant x_0 < x_1 < \cdots < x_{d+1} \leqslant b$$

и матрицу Вандермонда, построенную на этих точках $W(x_0,\dots,x_{d+1}) \in \mathbb{R}^{(d+2)\times(d+1)}$. Известно, что

$$\det W(y_1,\dots,y_k) = \prod_{i<j}(y_i - y_j)$$

для квадратной матрицы Вандермонда, откуда определители для всех подматриц $W(x_0,\dots,x_{d+1})^{\setminus j}$ имеют один и тот же знак, поэтому Теорема 1.24 упрощается до Теоремы 1.22.

В дальнейшем нам понадобится утверждение о том, что если система обладает альтернансом, то в ней всегда можно заменить одно из уравнений таким образом, что чередование знаков сохранится.

Лемма 1.26. *Пусть матрица* $\hat{V} \in \mathbb{R}^{(r+1)\times r}$ *и* $h \in \mathbb{R}^r$ *таковы, что* $\begin{bmatrix}\hat{V}\\ h^T\end{bmatrix} \in \mathbb{R}^{(r+2)\times r}$ *является чебышевской. Пусть вектор* $\hat{w} \in \mathbb{R}^{r+1}$ *имеет ненулевые компоненты и* $\xi \in \mathbb{R}$, $\xi \neq 0$. *Пусть знаки в последовательности*

$$\hat{w}_1 D_1(\hat{V}), \hat{w}_2 D_2(\hat{V}), \dots, \hat{w}_{r+1} D_{r+1}(\hat{V})$$

чередуются. Тогда существует такое $k \in \{1,\dots,r+1\}$, *что если матрица* $\tilde{V}$ *получена из* $\hat{V}$ *заменой* k*–ой строки на вектор* h *и вектор* $\tilde{w}$ *получен из* $\hat{w}$ *заменой* k*–го элемента на* ξ, *то знаки в последовательности*

$$\tilde{w}_1 D_1(\tilde{V}), \tilde{w}_2 D_2(\tilde{V}), \dots, \tilde{w}_{r+1} D_{r+1}(\tilde{V})$$

чередуются.

Для доказательства леммы нам понадобится ряд технических утверждений. Введем необходимые обозначения. Пусть $V \in \mathbb{R}^{(r+2)\times r}$ является чебышевской матрицей и $w \in \mathbb{R}^{r+2}$. Обозначим

$$D_j = \det V((1,\dots,j-1,j+1,\dots,r+1)),$$

$$\hat{D}_j^k = \begin{cases} \det V(\{1,2,\dots,k-1,r+2,k+1,\dots,j-1,j+1,\dots,r+1\}), & k<j, \\ \det V(\{1,2,\dots,j-1,j+1,\dots,k-1,r+2,k+1,\dots,r+1\}), & k>j, \\ D_j, & k=j. \end{cases}$$

Легко видеть, что

$$\hat{D}_i^j = (-1)^{i-j+1}\hat{D}_j^i, \quad i \neq j.$$

Обозначим через $\hat{w}^k \in \mathbb{R}^{r+1}$ вектор такой, что

$$\hat{w}_j^k = \begin{cases} w_j, & j \neq k \\ w_{r+2}, & j = k. \end{cases}$$

Лемма 1.27. *Для любых попарно различных i, j и k условия* $\operatorname{sign}(\hat{D}_i^k D_i \hat{D}_j^k D_j) = -1$ *и* $\operatorname{sign}(\hat{D}_k^i D_k \hat{D}_j^i D_j) = -1$ *не могут быть выполнены одновременно.*

Доказательство. Предположим, что $\operatorname{sign}(\hat{D}_i^k D_i \hat{D}_j^k D_j) = -1$ и $\operatorname{sign}(\hat{D}_k^i D_k \hat{D}_j^i D_j) = -1$. Поскольку $\hat{D}_k^i = (-1)^{k-i+1}\hat{D}_i^k$, мы имеем

$$\operatorname{sign}(\hat{D}_i^k D_i \hat{D}_j^k D_j) = -1$$
$$\operatorname{sign}(\hat{D}_i^k D_k \hat{D}_j^i D_j) = (-1)^{k-i}.$$

Пусть $\operatorname{sign}(D_j \hat{D}_i^k) = \delta$, тогда

$$\operatorname{sign}(D_i \hat{D}_j^k) = -\delta,$$
$$\operatorname{sign}(D_k \hat{D}_j^i) = (-1)^{k-i}\delta,$$
$$\operatorname{sign}(D_j \hat{D}_i^k) = \delta.$$

Пусть $\mathcal{L}$ является линейной оболочкой $v^1,\dots,v^{r+1}$, среди которых не содержатся v^i, v^j и v^k. Поскольку матрица V является чебышевской, вектор v^k может быть единственным образом представлен как

$$v^k = \alpha v^i + \gamma v^j + z_1, \tag{1.12}$$

где $z_1 \in \mathcal{L}$. Представим также

$$v^i = \beta v^k + \tau v^{r+2} + z_2, \tag{1.13}$$

где $z_2 \in \mathcal{L}$.

Из (1.12) и свойств определителя получаем, что $D_i = \alpha(-1)^{k-i+1} D_k$. Аналогично из (1.13) и свойств определителя следует, что $\hat{D}^k_j = -\beta \hat{D}^i_j$. Следовательно

$$D_i \hat{D}^k_j = \alpha(-1)^{k-i+1} D_k (-1)\beta \hat{D}^i_j = \alpha\beta(-1)^{k-i} D_k \hat{D}^i_j.$$

Возьмем знаки с обоих сторон и получим $-\delta = \operatorname{sign}(\alpha\beta)(-1)^{k-i}(-1)^{k-i}\delta$, откуда $\operatorname{sign}(\alpha\beta) = -1$.

Заметим, что коэффициенты в (1.12) могут быть вычислены по формулам Крамера, и, благодаря тому, что V является чебышевской, получаем $\alpha, \gamma \neq 0$. Тогда вектор v^j может быть выражен из (1.12) как $v^j = \dfrac{1}{\gamma} v^k - \dfrac{\alpha}{\gamma} v^i - \dfrac{1}{\gamma} z_1$. Тогда

$$D_i = -\frac{\alpha}{\gamma}(-1)^{j-i+1} D_j = \frac{\alpha}{\gamma}(-1)^{j-i} D_j. \tag{1.14}$$

Из (1.12) и (1.13) получаем $v^i = \beta(\alpha v^i + \gamma v^j + z_1) + \tau v^{r+2} + z_2$, следовательно

$$(1 - \alpha\beta) v^i = \beta\gamma v^j + \beta z_1 + \tau v^{r+2} + z_2. \tag{1.15}$$

Заметим, что $z = \beta z_1 + z_2 \in \mathcal{L}$. Тогда векторы v^j, v^{r+2} и z линейно независимы и, поскольку $\tau \neq 0$, в правой части (1.15) находится нетривиальная линейная комбинация линейно независимых векторов, которая является ненулевой. Таким образом, $1 - \alpha\beta \neq 0$. Тогда

$$v^i = \frac{\beta\gamma}{1 - \alpha\beta} v^j + \frac{\beta z_1 + \tau v^{r+2} + z_2}{1 - \alpha\beta}.$$

и мы получаем, что

$$\hat{D}^k_j = \frac{\beta\gamma}{1 - \alpha\beta}(-1)^{j-i+1} \hat{D}^k_i, \tag{1.16}$$

поэтому мы имеем из (1.14) и (1.16), что

$$D_i \hat{D}^k_j = \frac{\alpha}{\gamma}(-1)^{j-i} D_j \frac{\beta\gamma}{1 - \alpha\beta}(-1)^{j-i+1} \hat{D}^k_i = \frac{\alpha\beta}{1 - \alpha\beta}(-1) D_j \hat{D}^k_i.$$

Вычисляя с обоих сторон знаки, получаем

$$-\delta = \frac{\operatorname{sign}(\alpha\beta)}{\operatorname{sign}(1 - \alpha\beta)}(-1)\delta,$$

следовательно $\operatorname{sign}(\alpha\beta) = \operatorname{sign}(1 - \alpha\beta) = -1$. Таким образом, приходим к противоречию. □

Наша цель состоит в том, чтобы показать, что существует такое k, что знаки в последовательности

$$\hat{w}_1^k \hat{D}_1^k, \hat{w}_2^k \hat{D}_2^k, \ldots, \hat{w}_{r+1}^k \hat{D}_{r+1}^k$$

чередуются. Это эквивалентно тому свойству, что $\operatorname{sign} \hat{w}_i^k \hat{D}_i^k \hat{w}_j^k \hat{D}_j^k = (-1)^{i-j}$ для любых i и j. Следующая лемма показывает, что если это свойство не выполнено для некоторого k и пары позиций (i, j), то оно выполнено для i или j (или даже для обоих) и другой пары позиций.

Лемма 1.28. *Пусть знаки в последовательности*

$$w_1 D_1, w_2 D_2, \ldots, w_{r+1} D_{r+1}$$

чередуются. Пусть

$$\operatorname{sign}(\hat{w}_i^k \hat{D}_i^k \hat{w}_j^k \hat{D}_j^k) = (-1)^{i-j+1}.$$

Тогда

- *если* $i = k$, *то* $\operatorname{sign}(\hat{w}_i^j \hat{D}_i^j \hat{w}_j^j \hat{D}_j^j) = (-1)^{i-j}$;
- *если* $j = k$, *то* $\operatorname{sign}(\hat{w}_j^i \hat{D}_j^i \hat{w}_i^i \hat{D}_i^i) = (-1)^{i-j}$;
- *если* $i \neq k$ *и* $j \neq k$, *то* $\operatorname{sign}(\hat{w}_k^i \hat{D}_k^i \hat{w}_j^i \hat{D}_j^i) = (-1)^{k-j}$ *и* $\operatorname{sign}(\hat{w}_k^j \hat{D}_k^j \hat{w}_i^j \hat{D}_i^j) = (-1)^{k-i}$.

Доказательство. Пусть $i = j$. Тогда условие леммы не может быть выполнено, поскольку

$$\operatorname{sign}(\hat{w}_i^k \hat{D}_i^k \hat{w}_i^k \hat{D}_i^k) = 1 \neq (-1)^{i-j+1} = -1.$$

Пусть $i = k \neq j$. Тогда условие леммы может быть записано как $\operatorname{sign}(\hat{w}_k^k \hat{D}_k^k \hat{w}_j^k \hat{D}_j^k) = (-1)^{k-j+1}$, что по определению $\hat{D}_k^k$ и $\hat{w}^k$ эквивалентно

$$\operatorname{sign}(w_{r+2} D_k w_j \hat{D}_j^k) = (-1)^{k-j+1}. \tag{1.17}$$

Поскольку знаки в последовательности $w_1 D_1, w_2 D_2, \ldots, w_{r+1} D_{r+1}$ чередуются,

$$\operatorname{sign}(w_k D_k w_j D_j) = (-1)^{k-j}. \tag{1.18}$$

Перемножим (1.17) и (1.18). Тогда

$$\operatorname{sign}(w_{r+2} w_k D_j \hat{D}_j^k) = \operatorname{sign}(w_k D_k w_j D_j) \operatorname{sign}(w_{r+2} D_k w_j \hat{D}_j^k) = (-1)^{k-j}(-1)^{k-j+1} = -1.$$

Поскольку $\hat{D}_k^j = (-1)^{k-j+1}\hat{D}_j^k$, мы имеем, что

$$\operatorname{sign}(w_{r+2}D_j w_k \hat{D}_k^j) = (-1)(-1)^{k-j+1} = (-1)^{k-j}.$$

Остается заметить, что $\hat{w}_j^j = w_{r+2}$, $\hat{w}_k^j = w_k$ и $\hat{D}_j^j = D_j$, откуда $\operatorname{sign}(\hat{w}_j^j \hat{D}_j^j \hat{w}_k^j \hat{D}_k^j) = (-1)^{k-j}$. Таким образом, первая часть леммы доказана. Вторая часть может быть получена из первой сменой обозначений.

Докажем третью часть. Пусть i, j и k попарно различны. Тогда условие леммы может быть записано как

$$\operatorname{sign}(w_i \hat{D}_i^k w_j \hat{D}_j^k) = (-1)^{i-j+1}. \tag{1.19}$$

Докажем, что в этом случае $\operatorname{sign}(w_k \hat{D}_k^i w_j \hat{D}_j^i) = (-1)^{k-j}$. От противного, пусть

$$\operatorname{sign}(w_k \hat{D}_k^i w_j \hat{D}_j^i) = (-1)^{k-j+1}. \tag{1.20}$$

Умножая (1.19) и условие альтернанса $\operatorname{sign}(w_i D_i w_j D_j) = (-1)^{i-j}$, получаем, что

$$\operatorname{sign}(\hat{D}_i^k D_i \hat{D}_j^k D_j) = \operatorname{sign}(w_i \hat{D}_i^k w_j \hat{D}_j^k)\operatorname{sign}(w_i D_i w_j D_j) = (-1)^{i-j+1}(-1)^{i-j} = -1. \tag{1.21}$$

Аналогично, умножая (1.20) и $\operatorname{sign}(w_k D_k w_j D_j) = (-1)^{k-j}$ получаем, что

$$\operatorname{sign}(\hat{D}_k^i D_k \hat{D}_j^i D_j) = \operatorname{sign}(w_k \hat{D}_k^i w_j \hat{D}_j^i)\operatorname{sign}(w_k D_k w_j D_j) = (-1)^{k-j+1}(-1)^{k-j} = -1. \tag{1.22}$$

Остается заметить, что (1.21) и (1.22) противоречит Лемме 1.27, откуда

$$\operatorname{sign}(w_k \hat{D}_k^i w_j \hat{D}_j^i) = (-1)^{k-j}. \tag{1.23}$$

С точностью до обозначений (1.23) соответствует первой части третьего утверждения леммы. Вторая часть следует из первой, поскольку i и j симметричны в условии и утверждении леммы. □

Следующая лемма показывает, что если условие Леммы 1.28 выполнено, то существует такое k, что знаки чередуются.

Лемма 1.29. *Пусть матрица $S \in \mathbb{R}^{n\times n}$ такова, что $s_{ki} \in \{-1, 1\}$ и удовлетворяет следующему свойству: если для тройки (i, j, k), где $1 \leqslant i, j, k \leqslant n$, выполнено $s_{ki}s_{kj} = (-1)^{i-j+1}$, то*

- *если $i = k$, то $s_{jk}s_{jj} = (-1)^{k-j}$;*
- *если $j = k$, то $s_{ik}s_{ii} = (-1)^{k-i}$;*

– *если* i, j *и* k *попарно различны, то* $s_{ik}s_{ij} = (-1)^{k-j}$ *и* $s_{jk}s_{ji} = (-1)^{k-i}$.

Тогда матрица S *имеет строку с чередующимися знаками.*

Доказательство. Докажем лемму по индукции. Пусть $n = 2$. Если $s_{11}s_{12} = -1$, то знаки чередуются в первой строке. В противном случае $s_{11}s_{12} = 1$ и условие леммы выполнено для $i = 1$, $j = 2$, $k = 1$. Тогда $s_{21}s_{22} = -1$ и знаки чередуются во второй строке.

Предположим, что утверждение выполнено при $n - 1$ и докажем его для матрицы размера n. Если условие леммы выполнено для матрицы S, оно также выполнено для ее ведущей подматрицы $\hat{S}$. По предположению индукции в $\hat{S}$ есть строка с номером t, в которой чередуются знаки элементов. Если $s_{t,n-1}s_{t,n} = -1$, то знаки чередуются в строке с номером t в матрице S. Пусть $s_{t,n-1}s_{t,n} = 1$. Тогда, поскольку знаки в строке t чередуются, мы имеем $s_{t,k}s_{t,n} = (-1)^{k-n+1}$ при $k = 1, \ldots, n$. Тогда из условия леммы при $k = t$ имеем $s_{n,t}s_{n,n} = (-1)^{t-n}$, а при $k \neq t$ получаем $s_{n,t}s_{n,k} = (-1)^{t-k}$, таким образом,

$$s_{n,t}s_{n,k} = (-1)^{t-k}, \quad k = 1, \ldots, n,$$

что означает, что в матрице S знаки чередуются в последней строке. □

Доказательство Леммы 1.26. Пусть $\hat{V} \in \mathbb{R}^{(r+1)\times r}$ и $h \in \mathbb{R}^r$ таковы, что $V = \begin{bmatrix} \hat{V} \\ h \end{bmatrix}$ является чебышевской. Пусть также $\hat{w} \in \mathbb{R}^{r+1}$ имеет ненулевые компоненты и $\xi \in \mathbb{R}$, $\xi \neq 0$. Обозначим $w = \begin{bmatrix} \hat{w}^T & \xi \end{bmatrix}^T$. Определим матрицу $S \in \mathbb{R}^{(r+1)\times(r+1)}$ такую, что $s_{ki} = \operatorname{sign}(\hat{w}_i^k \hat{D}_i^k)$. Легко видеть, что выполнены условия Леммы 1.28, откуда следует, что выполнены требования Леммы 1.29, что в свою очередь приводит к утверждению Леммы 1.26. □

1.9 Обобщенный алгоритм Ремеза

Теорема 1.21 позволяет находить решение задачи (1.1) за конечное число операций. Однако сложность вычислений по формулам из Теоремы 1.21 растет экспоненциально с размером задачи. В данном разделе предлагается алгоритм, который способен решать задачу быстрее. Идея алгоритма вдохновлена классическим алгоритмом Ремеза (см. например [26]) для построения наилучшего

равномерного приближения непрерывной функции на отрезке, однако не является его копией, поскольку свойства векторной и функциональной задач отличаются (см. например Теорему 1.24). Предложенный алгоритм был назван обобщенным алгоритмом Ремеза. Он пытается построить вектор, обеспечивающий свойства альтернанса, описанные в Теореме 1.24.

Приведем алгоритм решения задачи о наилучшем равномерном приближении в вещественном случае. Пусть $V \in \mathbb{R}^{n\times r}$ является чебышевской и $a \in \mathbb{R}^n$.

1. Выберем произвольное множество из $r+1$ различных индексов строк матрицы V. Обозначим это множество через J_1 и положим $t = 1$.
2. Найдем решение задачи наилучшего равномерного приближения для матрицы $V(J_t)$ и вектора $a(J_t)$. Обозначим решение через u_t.
3. Вычислим невязку $w_t = a - Vu_t$ и найдем позицию $\hat{j}_t$ максимального по модулю значения в векторе w_t. Если $|(w_t)_{\hat{j}_t}| = \|w_t(J_t)\|_\infty$, то u_t является решением задачи по Теореме 1.16.
4. Если $|(w_t)_{\hat{j}_t}| > \|w_t(J_t)\|_\infty$, то попробуем заменить каждый из элементов J_t на $\hat{j}_t$. Пусть $J_t = (j_1^t, j_2^t, \ldots, j_{r+1}^t)$ и обозначим $J_t^k = (j_1^t, \ldots, j_{k-1}^t, \hat{j}_t, j_{k+1}^t, \ldots, j_{r+1}^t)$. Пусть u_t^k является решением задачи наилучшего равномерного приближения для матрицы $V(J_t^k)$ и вектора $a(J_t^k)$ и обозначим $w_t^k = a(J_t^k) - V(J_t^k)u_t^k$. Можно показать (см. Теорему 1.30), что существует номер $\hat{k}$ такой, что $\|w_t^{\hat{k}}\|_\infty > \|w_t(J_t)\|_\infty$.
5. $J_{t+1} = J_t^{\hat{k}}$, $t = t+1$ и перейдем к шагу 2.

Докажем результат о сходимости обобщенного алгоритма Ремеза.

Теорема 1.30. *Пусть матрица $V \in \mathbb{R}^{n\times r}$ является чебышевской и вектор $a \in \mathbb{R}^n$. Обозначим минимальное по модулю значение $r \times r$ определителя матрицы V через $m > 0$, а максимальное по модулю значение $r \times r$ определителя матрицы V через $M > 0$. Пусть u_t является решением на t–ой итерации обобщенного алгоритма Ремеза и $w_t = a - Vu_t$. Пусть также J_t обозначает текущее множество индексов на t–ой итерации. Обозначим также*

$$\overline{E}_t = \|w_t\|_\infty, \quad \underline{E}_t = \|w_t(J_t)\|_\infty,$$

$$\mu = \inf_{u\in\mathbb{C}^r} \|a - Vu\|_\infty.$$

Тогда обобщенный алгоритм Ремеза находит решение задачи наилучшего равномерного приближения за конечное число операций, причем

$$0 \leqslant \overline{E}_t - \mu \leqslant \frac{(r+1)M}{m}\left(1 - \frac{m}{(r+1)M}\right)^{t-1}(\mu - \underline{E}_1).$$

Доказательство. По Теореме 1.24, знаки в последовательности

$$(w_t(J_t))_1 D_1(V(J_t)), (w_t(J_t))_2 D_2(V(J_t)), \ldots, (w_t(J_t))_{r+1} D_{r+1}(V(J_t))$$

чередуются. Если u_t не является решением, то найдем позицию $\hat{j}_t$ максимального по модулю элемента в векторе w_t и заменим один из элементов в множестве J_t на $\hat{j}_t$. По Лемме 1.26, найдется множество индексов $\hat{J}_{t+1}$, полученное из J_t заменой одного элемента на $\hat{j}_t$, такое, что знаки в последовательности

$$(w_t(\hat{J}_{t+1}))_1 D_1(V(\hat{J}_{t+1})), (w_t(\hat{J}_{t+1}))_2 D_2(V(\hat{J}_{t+1})), \ldots, (w_t(\hat{J}_{t+1}))_{r+1} D_{r+1}(V(\hat{J}_{t+1})) \tag{1.24}$$

чередуются. Поскольку замена в множестве J_t выполняется таким образом, чтобы величина $\underline{E}_{t+1}$ была максимальна, из Теоремы 1.20 имеем

$$\underline{E}_{t+1} \geqslant \frac{\left|\sum\limits_{j=1}^{r+1} (-1)^j D_j(V(\hat{J}_{t+1}))(a(\hat{J}_{t+1}))_j\right|}{\sum\limits_{j=1}^{r+1} |D_j(V(\hat{J}_{t+1}))|}.$$

Нормы невязки являются одинаковыми для решений всех задач вида

$$\|V(\hat{J}_{t+1})u - (a(\hat{J}_{t+1}) + V(\hat{J}_{t+1})p)\|_\infty \to \min_{u \in \mathbb{R}^r}$$

для любого $p \in \mathbb{R}^r$, в частности, для правой части $w_t(\hat{J}_{t+1}) = a(\hat{J}_{t+1}) - V(\hat{J}_{t+1})u_t$. Поэтому

$$\underline{E}_{t+1} \geqslant \frac{\left|\sum\limits_{j=1}^{r+1} (-1)^j D_j(V(\hat{J}_{t+1}))(w_t(\hat{J}_{t+1}))_j\right|}{\sum\limits_{j=1}^{r+1} |D_j(V(\hat{J}_{t+1}))|} = \sum_{j=1}^{r+1} \frac{|D_j(V(\hat{J}_{t+1}))|}{\sum\limits_{k=1}^{r+1} |D_k(V(\hat{J}_{t+1}))|} |(w_t(\hat{J}_{t+1}))_j|,$$

где последнее равенство выполнено в силу чередования знаков в (1.24).

Поскольку индекс $\hat{j}_t$ соответствует элементу с максимальным по модулю значением, $\overline{E}_t = \|w_t(\hat{J}_{t+1})\|_\infty$. Если u_t не является решением, $\overline{E}_t > \mu \geqslant \underline{E}_t$. Тогда

$$\underline{E}_{t+1} - \underline{E}_t \geqslant \sum_{j=1}^{r+1} \frac{|D_j(V(\hat{J}_{t+1}))|}{\sum\limits_{k=1}^{r+1} |D_k(V(\hat{J}_{t+1}))|} (|(w_t(\hat{J}_{t+1}))_j| - \underline{E}_t).$$

Заметим, что по построению $\hat{J}_{t+1}$ и из Теоремы 1.24, значения $|(w_t(\hat{J}_{t+1}))_j| - \underline{E}_t$ равны нулю для всех j за исключением одного, для которого $|w_t(\hat{J}_{t+1}))_j| = \overline{E}_t$. Следовательно

$$\underline{E}_{t+1} - \underline{E}_t \geqslant \frac{\min\limits_j |D_j(V(\hat{J}_{t+1}))|}{\sum\limits_{k=1}^{r+1} |D_k(V(\hat{J}_{t+1}))|}(\overline{E}_t - \underline{E}_t) \geqslant \frac{m}{(r+1)M}(\overline{E}_t - \underline{E}_t) > 0. \qquad (1.25)$$

Тогда

$$\mu - \underline{E}_t - (\mu - \underline{E}_{t+1}) = \underline{E}_{t+1} - \underline{E}_t \geqslant \frac{m}{(r+1)M}(\overline{E}_t - \underline{E}_t) > \frac{m}{(r+1)M}(\mu - \underline{E}_t).$$

$$\mu - \underline{E}_{t+1} < \mu - \underline{E}_t - \frac{m}{(r+1)M}(\mu - \underline{E}_t) = \left(1 - \frac{m}{(r+1)M}\right)(\mu - \underline{E}_t),$$

откуда

$$\mu - \underline{E}_{t+1} < \left(1 - \frac{m}{(r+1)M}\right)^t (\mu - \underline{E}_1).$$

Тогда из (1.25) получаем, что

$$\overline{E}_t - \underline{E}_t \leqslant \frac{(r+1)M}{m}(\underline{E}_{t+1} - \underline{E}_t) = \frac{(r+1)M}{m}(\mu - \underline{E}_t - (\mu - \underline{E}_{t+1})) \leqslant \frac{(r+1)M}{m}\left(1 - \frac{m}{(r+1)M}\right)^{t-1}(\mu - \underline{E}_1).$$

□

Заметим, что из доказанной теоремы следует, что скорость сходимости геометрическая, однако полученные оценки не позволяют оценить число операций, требуемых для построения точного решения. Сложность одной итерации обобщенного метода Ремеза равна $O(nr + r^4)$. Число итераций в настоящее время оценить не удается, однако в Разделе 2.8 приведено численное исследование этого вопроса.

1.10 Ускоренный алгоритм решения

В данном разделе приводится ускоренный алгоритм для решения задачи наилучшего равномерного приближения

$$\|Vu - a\|_\infty \to \min_{u \in \mathbb{R}^r}. \qquad (1.26)$$

для чебышевской матрицы $V \in \mathbb{R}^{n\times r}$ и вектора $a \in \mathbb{R}^n$. Как было отмечено выше, сложность обобщенного алгоритма Ремеза составляет $O(I(nr + r^4))$, где I — число итераций метода. В данном разделе предлагается алгоритм со сложностью $O(r^3 + Inr)$, который эквивалентен методу, описанному в Разделе 1.9, в точной арифметике. Идея алгоритма состоит в том, чтобы поддерживать QR разложение для текущей подматрицы размера $(r + 1) \times r$. Эта идея основана на аналогичном подходе для метода максимального объема [29].

1.10.1 О решении задачи размера $(r + 1) \times r$

Если известно QR разложение матрицы, то решение задачи наименьших квадратов с этой матрицей может быть найдено за $O(r^2)$ операций, где r — размер матрицы. Аналогично решение задачи наилучшего равномерного приближения может быть найдено за $O(r^2)$ операций для матрицы размера $(r + 1) \times r$, если известно ее QR разложение. Пусть матрица $\hat{V} \in \mathbb{R}^{(r+1)\times r}$ является чебышевской и $\hat{a} \in \mathbb{R}^{r+1}$ является правой частью задачи о наилучшем равномерном приближении. Пусть $\hat{V} = \hat{Q}\hat{R}$, где $\hat{Q} \in \mathbb{R}^{(r+1)\times r}$ имеет ортонормированные столбцы и $\hat{R}$ является невырожденной верхней треугольной матрицей. Обозначим также через $\hat{q}'$ вектор такой, что $\begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix} \in \mathbb{R}^{(r+1)\times(r+1)}$ ортогональна. Докажем несколько вспомогательных утверждений.

Лемма 1.31. *Пусть $Q \in \mathbb{R}^{(r+1)\times(r+1)}$ является ортогональной матрицей вида $Q = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}$, где $\hat{Q} \in \mathbb{R}^{(r+1)\times r}$ и $\hat{q}' \in \mathbb{R}^{r+1}$. Тогда*

$$\hat{q}' = (-1)^{r+1} \operatorname{sign}\det Q \begin{bmatrix}(-1)D_1(\hat{Q}) & (-1)^2 D_2(\hat{Q}) & \dots & (-1)^{r+1} D_{r+1}(\hat{Q})\end{bmatrix}^T.$$

Доказательство. Рассмотрим систему $Qy^{(k)} = e_k$, где e_k является стандартным базисным вектором. Тогда $y_n^{(k)} = (Q^T e_k)_{r+1} = q'_k$. Однако по формулам Крамера

$$y_{r+1}^{(k)} = \frac{\det Q^{(k)}}{\det Q},$$

где $Q^{(k)}$ является матрицей, полученной из Q заменой последнего столбца на e_k. Ясно, что $\det Q^{(k)} = (-1)^{r+1+k} D_j(\hat{Q})$. □

Следующая лемма дает удобный способ вычисления ошибки наилучшего равномерного приближения, который нам понадобится для построения ускоренного алгоритма.

Лемма 1.32. *Пусть $\hat{V} \in \mathbb{R}^{(r+1)\times r}$ является чебышевской и $\hat{a} \in \mathbb{R}^n$. Пусть $q \in \mathbb{R}^{r+1}$ является ненулевым вектором таким, что $V^T q = 0$. Тогда*

$$\min_{u\in\mathbb{R}^r} \left\|\hat{a} - \hat{V}u\right\|_\infty = \frac{|q^T\hat{a}|}{\|q\|_1}.$$

Доказательство. Следует из Теоремы 1.20 и Леммы 1.31. □

Для решения задачи (1.26) может быть использована Теорема 1.16, которая дает критерий того, что вектор является решением задачи. Из Утверждения 1.6 и Теоремы 1.16 вектор $\hat{u} \in \mathbb{R}^r$ является решением задачи наилучшего равномерного приближения тогда и только тогда, когда существует ненулевой вектор $\delta \in \mathbb{R}^{r+1}$ с неотрицательными компонентами такой, что

$$\hat{V}^T \operatorname{diag}(\operatorname{sign}\hat{w})\delta = 0, \tag{1.27}$$

где $\hat{w} = \hat{a} - \hat{V}\hat{u}$. Поскольку $\hat{V}$ является чебышевской, размерность $\ker\hat{V}^T$ равна 1 и базисным вектором пространства $\ker\hat{V}^T$ является $\hat{q}'$ (заметим, что $\hat{q}'$ имеет ненулевые компоненты по Лемме 1.31). Тогда уравнение (1.27) может быть выполнено только если $\delta_k = |\hat{q}'_k|$ и $\operatorname{sign}\hat{w}_k = c\operatorname{sign}\hat{q}'_k$ при $k = 1,2,\ldots,r+1$, где $c = \pm 1$ и не зависит от k. Если $\hat{u}$ является решением, то из Утверждения 1.6 мы имеем $|\hat{w}_k| = \|\hat{w}\|_\infty$ при $k = 1,2,\ldots,r+1$. Следовательно $\hat{w}_j = \hat{c}\operatorname{sign}\hat{q}'_j$, где $\hat{w}$ является невязкой для оптимального решения. Поскольку $\hat{w}$ является невязкой, уравнение $\hat{V}\hat{u} = \hat{a} - \hat{w}$ должно иметь решение.

$$\hat{a} - \hat{w} = \hat{Q}\hat{R}\hat{u} = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}\begin{bmatrix}\hat{R}\\ 0\end{bmatrix}\hat{u}. \tag{1.28}$$

Тогда для того чтобы (1.28) имело решение, необходимо, чтобы $(\hat{a}-\hat{w})^T\hat{q}' = 0$, следовательно $\hat{c}(\operatorname{sign}\hat{q}')^T\hat{q}' = \hat{a}^T\hat{q}'$ и $\hat{c} = \dfrac{\hat{a}^T\hat{q}'}{\|\hat{q}'\|_1}$. Таким образом может быть вычислена невязка для оптимального решения. Как только невязка вычислена, решение может быть найдено как $\hat{u} = \hat{R}^{-1}\hat{Q}^T(\hat{a}-\hat{w})$.

Итоговая процедура представлена в Алгоритме 1. Ясно, что сложность описанного метода составляет $O(r^2)$ операций.

Входные данные: Матрица с ортонормированными столбцами $\hat{Q} \in \mathbb{R}^{(r+1)\times r}$, дополнение до ортогональной матрицы $\hat{q}' \in \mathbb{R}^{r+1}$, верхняя треугольная матрица $\hat{R} \in \mathbb{R}^{r\times r}$, вектор правой части $\hat{a} \in \mathbb{R}^{r+1}$.

Результат: $\hat{u} \in \mathbb{R}^r$ — решение задачи $\|\hat{Q}\hat{R}u - a\|_\infty \to \min\limits_{u\in\mathbb{R}^r}$.

$\hat{c} = \hat{a}^T\hat{q}'/\|\hat{q}'\|_1$;

$\hat{w} = \hat{c}\operatorname{sign}\hat{q}'$;

$\hat{u} = \texttt{solve\_triangular}(\hat{R}, \hat{Q}^T(\hat{a} - \hat{w}))$;

Алгоритм 1: Ускоренный алгоритм для решения задачи размера $(r+1)\times r$.

1.10.2 Обновление множества индексов

Пусть $V \in \mathbb{R}^{n\times r}$ является чебышевской матрицей и $a \in \mathbb{R}^n$. Пусть $\hat{J}$ является множеством номеров строк матрицы V и $|\hat{J}| = r+1$. Обозначим $\hat{V} = V(\hat{J})$ и $\hat{a} = a(\hat{J})$. При помощи Алгоритма 1 может быть найдено решение задачи наилучшего равномерного приближения с матрицей $\hat{V}$ и правой частью $\hat{a}$. Обозначим решение через $\hat{u}$ и $w = a - V\hat{u}$. Из Теоремы 1.16 следует, что если $\|w(\hat{J})\|_\infty = \|w\|_\infty$, то $\hat{u}$ является решением задачи наилучшего равномерного приближения с матрицей V и правой частью a. В противном случае обозначим через $\tilde{\tilde{j}}$ позицию максимального по модулю элемента в w. В этом случае (см. Теорему 1.30) существует множество $\tilde{J}$, полученное из $\hat{J}$ заменой одного элемента на $\tilde{\tilde{j}}$ такое, что

$$\min_{u\in\mathbb{R}^r}\|V(\hat{J})u - a(\hat{J})\|_\infty < \min_{u\in\mathbb{R}^r}\|V(\tilde{J})u - a(\tilde{J})\|_\infty.$$

Поскольку величина ошибки не может возрастать бесконечно, за конечное число шагов итерационная процедура сходится к решению задачи наилучшего равномерного приближения.

В данном разделе предлагается эффективная процедура обновления множества индексов $\hat{J}$. Пусть матрица $\tilde{V} \in \mathbb{R}^{(r+1)\times r}$ получена из матрицы $\hat{V}$ заменой k–ой строки на вектор h, а вектор $\tilde{a} \in \mathbb{R}^{r+1}$ получен из вектора $\hat{a}$ заменой k–го элемента на $\xi \in \mathbb{R}$. Тогда $\tilde{V} = \hat{V} + e_k(h - \hat{V}^T e_k)^T$. Построим ненулевой вектор $\tilde{q}' \in \mathbb{R}^{r+1}$ такой, что $\tilde{V}^T\tilde{q}' = 0$. Пусть $\tilde{V} = \tilde{Q}\tilde{R}$, где $\tilde{Q} \in \mathbb{R}^{(r+1)\times r}$ имеет ортонормированные столбцы и $\tilde{R}$ является невырожденной верхней треугольной

матрицей. Тогда

$$\hat{Q}\hat{R} + e_k(h - \hat{V}^T e_k)^T = \tilde{Q}\tilde{R}.$$

Домножим последнее уравнение на $\begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}^T$ слева и на $\hat{R}^{-1}$ справа.

$$\begin{bmatrix} I \\ 0 \end{bmatrix} + \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}^T e_k(h - \hat{V}^T e_k)^T \hat{R}^{-1} = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}^T \tilde{Q}\tilde{R}\hat{R}^{-1}.$$

Обозначим $z = \hat{R}^{-T}(h - \hat{V}^T e_k) = \hat{R}^{-T}h - \hat{q}^k$, где через $\hat{q}^k \in \mathbb{R}^r$ обозначена k–ая строка матрицы $\hat{Q}$. Тогда

$$\begin{bmatrix} I \\ 0 \end{bmatrix} + \begin{bmatrix} \hat{q}^k \\ \hat{q}'_k \end{bmatrix} z^T = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}^T \tilde{Q}\tilde{R}\hat{R}^{-1}. \tag{1.29}$$

Заметим, что $\tilde{V}^T\tilde{q}' = 0$ эквивалентно

$$(\tilde{q}')^T \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix} \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}^T \tilde{Q}\tilde{R}\hat{R}^{-1} = 0.$$

Подставляя (1.29) в последнее уравнение, получаем

$$(\tilde{q}')^T \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix} \left(\begin{bmatrix} I \\ 0 \end{bmatrix} + \begin{bmatrix} \hat{q}^k \\ \hat{q}'_k \end{bmatrix} z^T \right) = 0. \tag{1.30}$$

Обозначим $x = \hat{Q}^T\tilde{q}'$ и $\alpha = (\hat{q}')^T\tilde{q}'$. Тогда

$$\begin{bmatrix} x^T & \alpha \end{bmatrix} \left(\begin{bmatrix} I \\ 0 \end{bmatrix} + \begin{bmatrix} \hat{q}^k \\ \hat{q}'_k \end{bmatrix} z^T \right) = 0, \tag{1.31}$$

что эквивалентно

$$x + (x^T\hat{q}^k + \alpha\hat{q}'_k)z = 0,$$

откуда $x = cz$, где $c \in \mathbb{R}$. Тогда $cz + (cz^T\hat{q}^k + \alpha\hat{q}'_k)z = 0$ и если $z \neq 0$, то

$$c + cz^T\hat{q}^k + \alpha\hat{q}'_k = 0.$$

Следовательно $\alpha\hat{q}'_k = -c(1 + z^T\hat{q}^k)$. Тогда

$$\begin{bmatrix} \hat{q}'_k x \\ \hat{q}'_k \alpha \end{bmatrix} = \begin{bmatrix} \hat{q}'_k cz \\ -c(1 + z^T\hat{q}^k) \end{bmatrix} = c \begin{bmatrix} -\hat{q}'_k z \\ 1 + z^T\hat{q}^k \end{bmatrix}.$$

Когда мы ищем $\tilde{q}' \neq 0$ такой, что $\tilde{V}^T\tilde{q}' = 0$, нам не важна норма вектора $\tilde{q}'$, поэтому мы можем взять

$$\tilde{q}' = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}\begin{bmatrix}-\hat{q}'_k z \\ 1 + z^T\hat{q}^k\end{bmatrix}. \tag{1.32}$$

Заметим, что подстановка (1.32) в (1.30) дает тождественный ноль, поэтому (1.32) верно для любых значений z. Как только значение $\tilde{q}'$ найдено, по Лемме 1.32 мы можем вычислить

$$\min_{u\in\mathbb{R}^r}\|V(\tilde{J})u - a(\tilde{J})\|_\infty = \frac{|\tilde{a}^T\tilde{q}'|}{\|\tilde{q}'\|_1}.$$

Явное вычисление (1.32) требует $O(r^2)$ операций, однако сложность может быть уменьшена с помощью предподсчета. Обозначим $g = \hat{R}^{-T}h$ и $y = \hat{Q}g$. Оба вектора могут быть предподсчитаны за $O(r^2)$ операций и не зависят от номера строки k. Тогда $z = g - \hat{q}^k$ и

$$\tilde{q}' = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}\begin{bmatrix}-\hat{q}'_k(g - \hat{q}^k) \\ 1 + (g - \hat{q}^k)^T\hat{q}^k\end{bmatrix} = -\hat{q}'_k y + \hat{q}'_k\hat{Q}\hat{q}^k + (1 + (g - \hat{q}^k)^T\hat{q}^k)\hat{q}'.$$

Заметим, что $\hat{Q}\hat{q}^k = e_k - \hat{q}'_k\hat{q}'$, откуда

$$\tilde{q}' = -\hat{q}'_k y + \hat{q}'_k(e_k - \hat{q}'_k\hat{q}') + \hat{q}' + (g^T\hat{q}^k)\hat{q}' - ((\hat{q}^k)^T\hat{q}^k)\hat{q}'.$$

Поскольку $(\hat{q}^k)^T\hat{q}^k = 1 - (\hat{q}'_k)^2$ и $g^T\hat{q}^k = y_k$, получаем

$$\tilde{q}' = \hat{q}'_k(e_k - y) + y_k\hat{q}'.$$

Таким образом, вектор $\tilde{q}'$ может быть вычислен за $O(r)$ операций, если известен вектор y. Тогда мы можем вычислить норму невязки при замене k–го элемента множества $\hat{J}$ за $O(r)$ операций, а следовательно выбрать замену, которая максимизирует $\min\limits_{u\in\mathbb{R}^r}\|V(\tilde{J})u - a(\tilde{J})\|_\infty$ за $O(r^2)$ операций. Итоговая процедура нахождения оптимальной замены представлена в Алгоритме 2.

После того как элемент, который требуется заменить, найден, требуется обновить текущее множество индексов и QR разложение для соответствующей подматрицы. Пусть $\hat{V} = \hat{Q}\hat{R}$ и матрица $\begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}$ ортогональная. Пусть матрица $\tilde{V}$ получена из матрицы $\hat{V}$ заменой $\hat{k}$–ой строки на $h \in \mathbb{R}^r$. Требуется построить представление $\tilde{V} = \tilde{Q}\tilde{R}$ и вектор $\tilde{q}'$ такие, что $\begin{bmatrix}\tilde{Q} & \tilde{q}'\end{bmatrix}$ ортогональная и $\tilde{R}$ является

Входные данные: Матрица с ортонормированными столбцами $\hat{Q} \in \mathbb{R}^{(r+1)\times r}$, дополнение до ортогональной матрицы $\hat{q}' \in \mathbb{R}^{r+1}$, верхняя треугольная матрица $\hat{R} \in \mathbb{R}^{r\times r}$ такая, что $\hat{V} = \hat{Q}\hat{R}$. Вектор правой части $\hat{a} \in \mathbb{R}^{r+1}$, новая строка $h \in \mathbb{R}^r$, новый элемент в векторе правой части $\xi \in \mathbb{R}$.

Результат: $\hat{k}$ — номер строки в матрице $\hat{V}$, который следует заменить на h и $\hat{\mu}$ — новое значение ошибки.

$y = \hat{Q} \cdot \texttt{solve\_triangular}(\hat{R}^T, h)$;
$\hat{\mu} = -1$, $\hat{k} = -1$;
for $k = 1,2,\ldots,r+1$ **do**
  $\tilde{a} = \hat{a}$;
  $\tilde{a}_k = \xi$;
  $\tilde{q}' = \hat{q}'_k(e_k - y) + y_k\hat{q}'$;
  $\mu = \dfrac{|(\tilde{q}')^T\tilde{a}|}{\|\tilde{q}'\|_1}$;
  if $\mu > \hat{\mu}$ **then**
    $\hat{\mu} = \mu$;
    $\hat{k} = k$;
  end
end

Алгоритм 2: Поиск оптимальной замены строки в матрице размера $(r+1)\times r$.

невырожденной верхней треугольной матрицей. Заметим, что $\tilde{V} = \hat{V} + e_{\hat{k}}(h - \hat{v}^{\hat{k}})^T$ и $\hat{V} = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}\begin{bmatrix}\hat{R}\\ 0\end{bmatrix}$. Таким образом, задача сводится к классической задаче одноранговой обновления QR разложения, что может быть сделано за $O(r^2)$ операций (см. [30, Раздел 6.5.1]).

1.10.3 Итоговый алгоритм

В этом разделе мы приводим описание ускоренного алгоритма для решения задачи наилучшего равномерного приближения с чебышевской матрицей

$V \in \mathbb{R}^{n\times r}$ и вектором $a \in \mathbb{R}^n$ за $O(r^3 + Inr)$ операций. Выберем произвольное множество $\hat{J}_1$ из $r+1$ различных индексов строк матрицы V. Пусть $\hat{V} = V(\hat{J}_1)$ и $\hat{a} = a(\hat{J}_1)$. Построим полное QR разложение матрицы $\hat{V} = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}\begin{bmatrix}\hat{R}\\ 0\end{bmatrix}$. Это может быть сделано за $O(r^3)$ операций.

Найдем решение $\hat{u} \in \mathbb{R}^r$ задачи $\|V(\hat{J}_1)u - a(\hat{J}_1)\|_\infty \to \min\limits_{u\in\mathbb{R}^r}$ и вычислим невязку $w = a - V\hat{u}$. Этот шаг требует $O(nr)$ операций. Если $\|w(\hat{J}_1)\|_\infty = \|w\|_\infty$, то $\hat{u}$ является решением задачи наилучшего равномерного приближения с матрицей V и правой частью a по Теореме 1.16. В противном случае мы имеем $\|w(\hat{J}_1)\|_\infty < \|w\|_\infty$. Найдем позицию $\hat{j}$ максимального по модулю элемента в векторе w. Заметим, что $\hat{j} \notin \hat{J}_1$. Используя Алгоритм 2, найдем номер $\hat{k}$ элемента в множестве $\hat{J}_1$, который нужно заменить на $\hat{j}$ за $O(r^2)$ операций. Обозначим через $\hat{J}_2$ множество, полученное из $\hat{J}_1$ заменой $\hat{k}$–го элемента на $\hat{j}$. В этом случае

$$\min_{u\in\mathbb{R}^r}\|V(\hat{J}_1)u - a(\hat{J}_1)\|_\infty < \min_{u\in\mathbb{R}^r}\|V(\hat{J}_2)u - a(\hat{J}_2)\|_\infty.$$

Остается построить QR разложение для матрицы $V(\hat{J}_2)$, которое может быть получено из QR разложения матрицы $V(\hat{J}_1)$ при помощи обновления ранга 1, как это описано в Разделе 1.10.2. Тогда повторим описанную процедуру для множества $\hat{J}_2$ и получим $\hat{J}_3, \hat{J}_4$, и т.д. Заметим, что поскольку $\hat{\mu}_t = \min\limits_{u\in\mathbb{R}^r}\|V(\hat{J}_t)u - a(\hat{J}_t)\|_\infty$ не может возрастать бесконечно, за конечное число операций итерационная процедура сойдется к решению задачи наилучшего равномерного приближения. Итоговая процедура представлена в Алгоритме 3. Легко понять, что сложность построенного метода равна $O(r^3 + Inr)$, где I — число итераций алгоритма.

1.11 Замечания о комплексном случае

В комплексном случае было доказано (Теорема 1.12), что размер характеристического множество лежит в интервале от $r+1$ до $2r+1$. В данном разделе исследуется вопрос все ли значения от $r+1$ до $2r+1$ могут встречаться или эта оценка является проблемой доказательства, а также обсуждается вопрос построения решения в комплексном случае.

Входные данные: Чебышевская матрица $V \in \mathbb{R}^{n \times r}$, правая часть $a \in \mathbb{R}^n$, начальное множество $\hat{J}$.

Результат: Решение $\hat{u} \in \mathbb{R}^r$ задачи наилучшего равномерного приближения $\|Vu - a\|_\infty \to \min\limits_{u \in \mathbb{R}^r}$, характеристическое множество задачи $\hat{J}$.

$\hat{V} = V(\hat{J})$, $\hat{a} = a(\hat{J})$, $t = 1$;

$\hat{Q}, \hat{q}', \hat{R} \leftarrow$ `qr_decomposition`$(\hat{V})$; `//` $\hat{V} = \begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix} \begin{bmatrix}\hat{R} \\ 0\end{bmatrix}$, $\begin{bmatrix}\hat{Q} & \hat{q}'\end{bmatrix}$ `ортогональная`

while *true* **do**

  $\hat{u} \leftarrow$ `uniform_approximation`$(\hat{Q}, \hat{q}', \hat{R}, \hat{a})$; `// Алгоритм 1`

  $w = a - V\hat{u}$;

  if $\|w(\hat{J})\|_\infty = \|w\|_\infty$ **then**

    break

  end

  $\hat{j} \leftarrow \arg\max\limits_{j \in \{1,\ldots,n\}} |w_j|$;

  $\hat{k} \leftarrow$ `best_replacement`$(\hat{Q}, \hat{q}', \hat{R}, v^{\hat{j}}, a_{\hat{j}})$; `// Алгоритм 2`

  $\hat{a}_{\hat{k}} = a_{\hat{j}}$;

  Заменить $\hat{k}$–ую строку матрицы $\hat{V}$ на $v^{\hat{j}}$;

  Обновить факторы QR разложения $\hat{Q}$, $\hat{q}'$ и $\hat{R}$ для матрицы $\hat{V}$;

  `// Обновление ранга 1`

  Заменить $\hat{k}$–ый элемент множества $\hat{J}$ на $\hat{j}$;

end

Алгоритм 3: Ускоренный алгоритм решения задачи наилучшего равномерного приближения.

Утверждение 1.33. *Для любого $r \geqslant 1$ и любого n существуют чебышевская матрица $V \in \mathbb{C}^{n\times r}$ и вектор $a \in \mathbb{C}^n$ такие, что размер характеристического множества равен $r+1$ (при условии $n \geqslant r+1$) и такие, что размер характеристического множества равен $2r+1$ (при условии $n \geqslant 2r+1$).*

Доказательство. Покажем, что существует система, для которой размер характеристического множества равен $r+1$. Рассмотрим произвольную чебышевскую матрицу $\hat{V} \in \mathbb{C}^{(r+1)\times r}$ и вектор $\hat{a} \in \mathbb{C}^{r+1}$, не лежащий в образе системы. Пусть $\hat{u}$ является решением задачи

$$\|\hat{a} - \hat{V}u\|_\infty \to \min_{u\in\mathbb{C}^r}.$$

и $\mu = \|\hat{a} - \hat{V}\hat{u}\|_\infty > 0$. Дополним матрицу $\hat{V}$ произвольным образом до чебышевской матрицы размера $n \times r$, а вектор $\hat{a}$ продолжим до размера n произвольным образом.

$$\begin{bmatrix}\hat{V}\\ \tilde{V}\end{bmatrix} \in \mathbb{C}^{n\times r}, \quad \begin{bmatrix}\hat{a}\\ \tilde{a}\end{bmatrix} \in \mathbb{C}^n.$$

Возьмем $\varepsilon > 0$ такой, что $\varepsilon\|\tilde{V}\hat{u} - \tilde{a}\|_\infty < \mu$. Тогда возьмем

$$V = \begin{bmatrix}\hat{V}\\ \varepsilon\tilde{V}\end{bmatrix} \in \mathbb{C}^{n\times r}, \quad a = \begin{bmatrix}\hat{a}\\ \varepsilon\tilde{a}\end{bmatrix} \in \mathbb{C}^n.$$

С одной стороны ясно, что $\|Vu - a\|_\infty \geqslant \mu$ для любого $u \in \mathbb{C}^r$, поскольку эта оценка верна уже для подсистемы. С другой стороны, $\|V\hat{u} - a\|_\infty = \mu$, откуда $\mu = \inf\limits_{u\in\mathbb{C}^r} \|a - Vu\|_\infty$, а следовательно по определению множество $\{1, 2, \ldots, r+1\}$ является характеристическим (оно не может быть уменьшено, потому что размер характеристического множество не меньше $r+1$ для любой несовместной системы).

Покажем, что существует система, для которой размер характеристического множества равен $2r+1$. Мы покажем, что существует система с матрицей размера $(2r+1) \times r$, для которой размер характеристического множества равен $2r+1$. Дополняя ее до размера n по описанной выше схеме, можно получить результат для произвольного $n \geqslant 2r+1$.

Возьмем $a = \begin{bmatrix}1 & 1 & \ldots & 1\end{bmatrix}^T \in \mathbb{C}^{2r+1}$. Рассмотрим две матрицы $V^{(re)} \in \mathbb{R}^{(2r+1)\times r}$, $V^{(im)} \in \mathbb{R}^{(2r+1)\times r}$. Пусть эти матрицы таковы, что все миноры размера $r \times r$ матрицы $V = V^{(re)} + iV^{(im)}$ не равны 0 и все миноры размера $2r \times 2r$

матрицы

$$W = \begin{bmatrix} (V^{(re)})^T \\ (V^{(im)})^T \end{bmatrix} \in \mathbb{R}^{2r\times(2r+1)}$$

не равны 0. Покажем, что такие матрицы существуют. Действительно, рассмотрим произвольный набор из r строк матрицы V. Множество матриц $V^{(re)}$ и $V^{(im)}$, для которых минор на выбранных строках равен нулю, имеет меру нуль. Поскольку в матрице V конечное число миноров и объедение множеств меры нуль имеет меру нуль, имеем, что множество матриц $V^{(re)}$ и $V^{(im)}$, для которых какой-либо минор размера $r \times r$ в V равен нулю, имеет меру нуль. Повторяя проведенные рассуждения с матрицей W и объединяя с полученным ранее множеством меры нуль получаем, что множество матриц $V^{(re)}$ и $V^{(im)}$, для которых какой-либо из требуемых миноров в матрице V или W равен нулю, имеет меру нуль. Следовательно, матрицы $V^{(re)}$ и $V^{(im)}$ с искомыми требованиями существуют (причем их достаточно много).

Поскольку ранг матрицы W равен $2r$, размерность ядра этой матрицы равна 1. Возьмем произвольный нетривиальный вектор из ядра этой матрицы и обозначим его $d \in \mathbb{R}^{2r+1}$. Умножим строки матриц $V^{(re)}$ и $V^{(im)}$ на -1 для тех номеров i для которых $d_i < 0$. Обозначим полученные матрицы $\hat{V}^{(re)}$ и $\hat{V}^{(im)}$. Тогда снова обозначим $\hat{V} = \hat{V}^{(re)} + i\hat{V}^{(im)}$ и

$$\hat{W} = \begin{bmatrix} (\hat{V}^{(re)})^T \\ (\hat{V}^{(im)})^T \end{bmatrix} \in \mathbb{R}^{2r\times(2r+1)}.$$

Заметим, что поскольку умножение строк и столбцов на -1 не влияет на вырожденность, все миноры размера $r \times r$ матрицы $\hat{V}$ и все миноры размера $2r \times 2r$ матрицы $\hat{W}$ не равны 0. Кроме того, матрица $\hat{W}$ имеет в ядре нетривиальный вектор $\hat{d}$ с неотрицательными компонентами. Возьмем $\hat{u} = 0$ и запишем критерий оптимальности из Теоремы 1.16. Заметим, что для вектора решения $\hat{u} = 0$ и вектора правой части $a = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$ множество

$$J = \{j \in \{1, 2, \dots, 2r+1\} : |a_j - (\hat{V}\hat{u})_j| = \|a - \hat{V}\hat{u}\|_\infty\} = \{1, 2, \dots, 2r+1\}.$$

Тогда критерий записывается как

$$\sum_{k=1}^{2r+1} \delta_k \overline{(a_k - \hat{u}^T\hat{v}^k)}\hat{v}_{kj} = \sum_{k=1}^{2r+1} \delta_k \hat{v}_{kj} = 0,$$

что эквивалентно $\hat{V}^T\delta = 0$, где $\delta_k \geqslant 0$ для любого k, причем $\delta \neq 0$. Поскольку δ является вещественным вектором,

$$0 = \hat{V}^T\delta = (\hat{V}^{(re)})^T\delta + i(\hat{V}^{(im)})^T\delta,$$

что в свою очередь эквивалентно $\hat{W}d = 0$. Выбирая $\delta = \hat{d}$ получаем, что критерий выполнен и $\hat{u} = 0$ является решением задачи. Однако если мы рассмотрим произвольное собственное подмножество строк матрицы $\hat{V}$, то для того чтобы был выполнен критерий для вектора $\hat{u} = 0$ необходимо и достаточно, чтобы для некоторого собственного подмножества столбцов матрицы $\hat{W}$ существовал неотрицательный нетривиальный вектор из ядра. Однако это невозможно, так как все миноры размера $2r \times 2r$ не равны 0 и ядро любого собственного подмножества столбцов $\hat{W}$ тривиально. Отсюда следует, что $\hat{u} = 0$ является решением для системы с матрицей $\hat{V}$ и правой частью a и не является решением ни для какой собственной подсистемы. Следовательно, размер характеристического множества равен $2r + 1$.

□

Заметим, что из приведенных доказательств следует, что и систем для которых размер характеристического множества равен $r + 1$, и систем для которых размер характеристического множества равен $2r + 1$, достаточно много.

Рассмотрим подробнее задачу при $r = 1$. Тогда если матрица является чебышевской, то могут существовать характеристические множества размера 2 или 3. Пусть $V = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}^T \in \mathbb{C}^{3\times 1}$ является чебышевской и $a = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}^T \in \mathbb{C}^3$. Пусть $\hat{u} \in \mathbb{C}$ является решением. Обозначим невязку $\hat{w} = V\hat{u} - a$. Если размер характеристического множества равен 2, то $\hat{u}$ должно быть оптимально на множествах индексов $\{1, 2\}$, $\{1, 3\}$ или $\{2, 3\}$. В этом случае решение задается формулами из Теоремы 1.20. В случае, если размер характеристического множества равен 3, согласно Теореме 1.13, должно быть выполнено условие $|\hat{w}_1| = |\hat{w}_2| = |\hat{w}_3|$. Разберем подробнее этот случай. Множество решений, при котором $|\hat{w}_1| = |\hat{w}_2|$, согласно Теореме 1.19, описывается следующим образом:

$$u = \frac{|D_1|\hat{u}_1 + |D_2|\hat{u}_2 e^{ik}}{|v_2| + |v_1|e^{ik}}, \tag{1.33}$$

где k — произвольное вещественное число такое, что $|v_2| + |v_1|e^{ik} \neq 0$,

$$D_1 = v_2, \quad D_2 = v_1,$$

$$\hat{u}_1 = a_2/v_2, \quad \hat{u}_2 = a_1/v_1.$$

Тогда

$$u = \frac{a_2 e^{-i\arg v_2} + a_1 e^{-i\arg v_1} e^{ik}}{|v_2| + |v_1| e^{ik}}.$$

При этом

$$|v_1 u - a_1| = |v_2 u - a_2| = \frac{|-D_1 a_1 + D_2 a_2|}{||D_1| + |D_2| e^{ik}|} = \frac{|a_2 v_1 - a_1 v_2|}{||v_2| + |v_1| e^{ik}|}.$$

$$|v_3 u - a_3| = \left| v_3 \frac{a_2 e^{-i\arg v_2} + a_1 e^{-i\arg v_1} e^{ik}}{|v_2| + |v_1| e^{ik}} - a_3 \right| = \left| \frac{v_3 a_2 e^{-i\arg v_2} + v_3 a_1 e^{-i\arg v_1} e^{ik} - v_2 a_3 e^{-i\arg v_2} - v_1 a_3 e^{-i\arg v_1} e^{ik}}{|v_2| + |v_1| e^{ik}} \right| = \left| \frac{e^{-i\arg v_2}(v_3 a_2 - v_2 a_3) + e^{-i\arg v_1}(v_3 a_1 - v_1 a_3) e^{ik}}{|v_2| + |v_1| e^{ik}} \right|.$$

Обозначим

$$A = a_2 v_1 - a_1 v_2, \quad B = (a_2 v_3 - a_3 v_2) e^{-i\arg v_2}, \quad C = (a_1 v_3 - a_3 v_1) e^{-i\arg v_1}.$$

Тогда для того, чтобы $|\hat{w}_1| = |\hat{w}_2| = |\hat{w}_3|$ необходимо и достаточно, чтобы было выполнено условие

$$|A| = |B + C e^{ik}|. \tag{1.34}$$

Найдем решения этого уравнения.

$$|B + C e^{ik}|^2 = \left| |B| e^{i\arg B} + |C| e^{i(\arg C + k)} \right|^2 = \left| |B| e^{i\frac{\arg B - \arg C - k}{2}} + |C| e^{-i\frac{\arg B - \arg C - k}{2}} \right|^2,$$

где при последнем переходе мы домножили выражение под модулем на $e^{i\frac{\arg B + \arg C + k}{2}}$. Заменим $\alpha = \dfrac{\arg B - \arg C - k}{2}$. Тогда имеем

$$|B + C e^{ik}|^2 = \left| |B| e^{i\alpha} + |C| e^{-i\alpha} \right|^2 = \left(|B| e^{i\alpha} + |C| e^{-i\alpha} \right) \left(|B| e^{-i\alpha} + |C| e^{i\alpha} \right) = |B|^2 + |C|^2 + |B||C|(e^{i2\alpha} + e^{-i2\alpha}) = |B|^2 + |C|^2 + 2\cos(2\alpha)|B||C| = |A|^2.$$

$$\cos(2\alpha) = \frac{|A|^2 - |B|^2 - |C|^2}{2|B||C|}.$$

Тогда получаем, что если $\left| \dfrac{|A|^2 - |B|^2 - |C|^2}{2|B||C|} \right| > 1$, то уравнение (1.34) не имеет решений. В противном случае

$$2\alpha = \pm \arccos\left(\frac{|A|^2 - |B|^2 - |C|^2}{2|B||C|} \right) + 2\pi t = \arg B - \arg C - k,$$

откуда

$$k = \arg B - \arg C \pm \arccos\left(\frac{|A|^2 - |B|^2 - |C|^2}{2|B||C|}\right) + 2\pi t, \quad t \in \mathbb{Z}.$$

Поскольку в выражении для решения (1.33) величина k стоит только в экспоненте, достаточно рассматривать только 2 значения k. В результате выбирая наилучшее решение из оптимальных решений для множеств $\{1,2\}$, $\{1,3\}$ и $\{2,3\}$ и двух решений когда $|\hat{w}_1| = |\hat{w}_2| = |\hat{w}_3|$, можно найти оптимальное решение при $n = 3$, $r = 1$.

Проиллюстрируем задачу для $r = 1$ геометрически. Если μ является решением, то существует $u \in \mathbb{C}$, являющееся решением системы неравенств

$$\begin{cases} |v_1 u - a_1| \leqslant \mu, \\ |v_2 u - a_2| \leqslant \mu, \\ |v_3 u - a_3| \leqslant \mu, \end{cases}$$

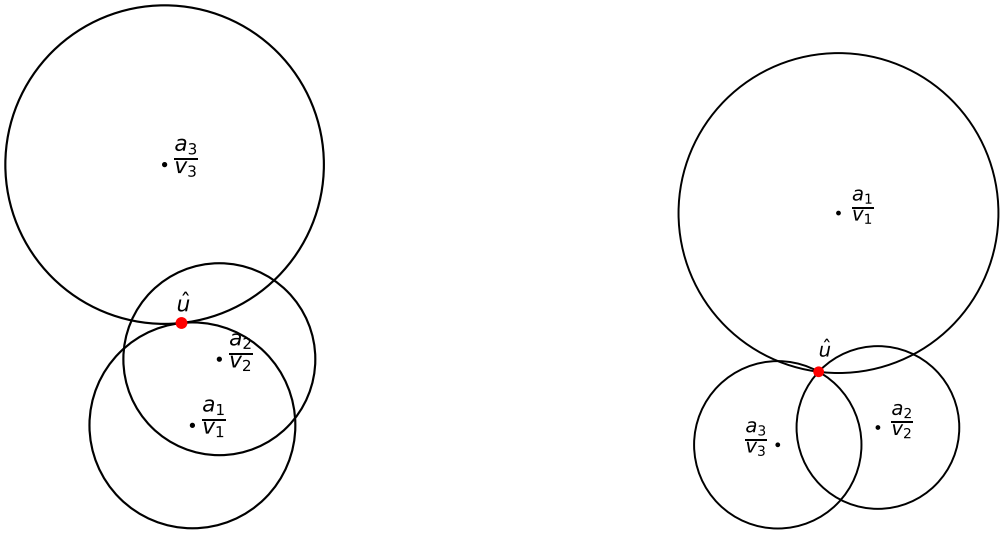
что эквивалентно

$$\begin{cases} |u - \dfrac{a_1}{v_1}| \leqslant \dfrac{\mu}{|v_1|}, \\ |u - \dfrac{a_1}{v_1}| \leqslant \dfrac{\mu}{|v_2|}, \\ |u - \dfrac{a_1}{v_1}| \leqslant \dfrac{\mu}{|v_3|}. \end{cases}$$

Геометрически это соответствует трем кругам с центрами $\dfrac{a_1}{v_1}$, $\dfrac{a_2}{v_2}$ и $\dfrac{a_3}{v_3}$ и радиусами $\dfrac{\mu}{|v_1|}$, $\dfrac{\mu}{|v_2|}$ и $\dfrac{\mu}{|v_3|}$ соответственно. Требуется найти минимальное μ, при котором все 3 круга имеют общую точку (Рисунок 1.1). Ясно, что любые два круга должны пересекаться. Тогда рассмотрим величину μ при которой какие-то два круга касаются. В этом случае возможны 2 варианта:

1. точка касания лежит внутри третьего круга (см. Рисунок 1.1(а)). Это соответствует тому, что если задача оптимально решена для двух неравенств, то третье выполнено автоматически, что соответствует ситуации, когда характеристическое множество состоит из 2 точек;
2. если ни для каких двух кругов не выполнено, что точка их касания лежит в третьем круге, то это ситуация, когда характеристическое множество состоит из 3 точек (см. Рисунок 1.1(б));

Стоит отметить, что обобщенный алгоритм Ремеза, предложенный в Разделе 1.9 работает и в комплексном случае. В этом случае достаточно искать



а) б)

Рисунок 1.1 — Три круга с центрами $\frac{a_1}{v_1}$, $\frac{a_2}{v_2}$ и $\frac{a_3}{v_3}$ и радиусами $\frac{\mu}{|v_1|}$, $\frac{\mu}{|v_2|}$ и $\frac{\mu}{|v_3|}$ соответственно. Красной точкой $\hat{u}$ обозначена общая точка для всех трех кругов. На Рисунке а) точка касания двух кругов лежит в третьем круге, на Рисунке б) точка касания никаких двух кругов не лежит в третьем круге.

множество индексов J размера $2r+1$, максимизирующее величину $\mu(J)$. Для полноты изложения приведем алгоритим в комплексном случае (см. Алгоритм 4).

Теорема 1.34. *Пусть матрица $V \in \mathbb{C}^{n\times r}$ является чебышевской и $a \in \mathbb{C}^n$. Тогда обобщенный алгоритм Ремеза (Алгоритм 4) находит решение задачи о наилучшем равномерном приближении за конечное число операций.*

Доказательство. Пусть $j_t \notin J_t$. Тогда $|(w_t)_{j_t}| > \|w_t(J_t)\|_\infty$. Следовательно J_t не содержит характеристическое множество для задачи

$$\|a(\hat{J}) - V(\hat{J})u\|_\infty \to \min_{u\in\mathbb{C}^r}, \tag{1.35}$$

где $\hat{J} = J_t \cup \{j_t\}$. Тогда характеристическое множество для (1.35) содержит элемент j_t. При этом размер характеристического множества для задачи (1.35) содержит не более $2r+1$ элементов. Следовательно, существует такое i, что характеристическое множество для (1.35) содержится в $J' = J_t \setminus \{i\} \cup \{j_t\}$. А тогда $\mu(J') = \mu(\hat{J}) > \mu(J_t)$. Следовательно, $\mu(J_1) < \mu(J_2) < \dots$. При этом если множество J_t не содержит характеристическое множество для всей задачи, то всегда можно заменить один элемент так, чтобы для нового множества J_{t+1} было выполнено $\mu(J_t) < \mu(J_{t+1})$. Остается применить Теорему 1.11. □

Входные данные: Чебышевская матрица $V \in \mathbb{C}^{n\times r}$, вектор $a \in \mathbb{C}^n$.
Результат: $\hat{u} \in \mathbb{C}^r$ — решение задачи наилучшего равномерного приближения.

J_1 — произвольное подмножество $\{1, \ldots, n\}$, состоящее из $\min\{2r+1, n\}$ элементов;
$t = 1$;
repeat
  $u_t = \arg\min\limits_{u\in\mathbb{R}^r} \|a(J_t) - V(J_t)u\|_\infty$;
  $w_t = Vu_t - a$;
  $j_t = \arg\max\limits_{j=1,\ldots,n} |(w_t)_j|$;
  if $j_t \in J_t$ **then**
    return u_t;
  else
    $\hat{l} = \|a(J_t) - V(J_t)u_t\|_\infty$;
    for $i \in J_t$ **do**
      $J' = J_t \setminus \{i\} \cup \{j_t\}$;
      $\tilde{u} = \arg\min\limits_{u\in\mathbb{R}^r} \|a(J') - V(J')u\|_\infty$;
      $\tilde{w} = a(J') - V(J')\tilde{u}$;
      if $\|\tilde{w}\|_\infty > \hat{l}$ **then**
        $\hat{l} = \|\tilde{w}\|_\infty$;
        $J_{t+1} = J'$;
      end
    end
  end
  $t = t + 1$;
until *решение не найдено*;

Алгоритм 4: Обобщенный алгоритм Ремеза для комплексной задачи.

Заметим, что Алгоритм 4 требует для своей работы метода решения задач с матрицами размера $(2r + 1) \times r$. К сожалению, в настоящий момент вопрос о построении решения для таких задач остается открытым при $r > 1$ (случай $r = 1$ разобран выше).

Глава 2. Построение малоранговых приближений матриц в чебышевской норме

2.1 Постановка задачи

Матрица обладает хорошим малоранговым приближением в унитарно-инвариантных нормах, если ее сингулярные числа убывают быстро. В противном случае в таких нормах не существует разумных приближений. Ситуация меняется, если рассматривать приближения в чебышевской норме. В работе [5] было доказано, что для любой матрицы $X \in \mathbb{R}^{m\times n}$, где $m \geqslant n$, и любого $\varepsilon > 0$ существует матрица $Y \in \mathbb{R}^{m\times n}$ ранга r, где

$$r \geqslant \lceil 72 \log(2n+1)/\varepsilon^2 \rceil, \tag{2.1}$$

такая, что $\|X - Y\|_C \leqslant \varepsilon \|X\|_2$. Это означает, что для любого фиксированного $\varepsilon > 0$ и последовательности матриц $\{X_n\}_n$ с ограниченной спектральной нормой и возрастающими размерами n, существует последовательность матриц $\{Y_n\}_n$ ранга $O(\log n)$ такая, что $\|X_n - Y_n\|_C \leqslant \varepsilon$. Заметим, что по крайней мере константы в (2.1) завышены (см. также [31] для более тонких оценок с использованием понятия μ–когерентности). Например, в Разделе 2.8 будет показано, что единичная матрицы размера $16{,}384$ может быть приближена с точностью 0.1 матрицей ранга 333.

Сформулируем интересующую нас задачу формально. Пусть матрица $A \in \mathbb{C}^{m\times n}$ и ранг $r > 0$. Рассмотрим задачу о построении наилучшего приближения ранга r в чебышевской норме, а именно, требуется найти такие матрицы $\hat{U} \in \mathbb{C}^{m\times r}$ и $\hat{V} \in \mathbb{C}^{n\times r}$, что

$$\|A - \hat{U}\hat{V}^T\|_C = \inf_{U\in\mathbb{C}^{m\times r}, V\in\mathbb{C}^{n\times r}} \|A - UV^T\|_C. \tag{2.2}$$

Задача построения малоранговых приближений в чебышевской норме является сложной. В [21] было показано. что даже для ранга $r = 1$ задача проверки существуют ли для матрицы $A \in \mathbb{R}^{m\times n}$ и числа $\varepsilon > 0$ такие векторы $u \in \mathbb{R}^m$ и $v \in \mathbb{R}^n$, что $\|A - uv^T\|_C < \varepsilon$, является NP-полной.

Первым методом для решения задачи (2.2) был *метод переменных направлений*, предложенный в [20] для построения приближений ранга 1. Пусть $A \in$

$\mathbb{R}^{m\times n}$ — матрица, которую требуется приблизить и $v^{(0)} \in \mathbb{R}^n$. Метод переменных направлений решает попеременно задачи

$$u^{(t)} \leftarrow \arg\min_{u\in\mathbb{R}^m} \|A - uv^{(t)}\|_C, \quad v^{(t+1)} \leftarrow \arg\min_{v\in\mathbb{R}^n} \|A - u^{(t)}v\|_C$$

при $t = 0, 1, 2, \ldots$. Также в [20] вводится понятие *двумерного альтернанса* и доказывается, что все предельные точки последовательности $\{(u^{(t)}, v^{(t)})\}_t$ обладают этим альтернансом. Кроме того, доказывается, что если матрица $A \in \mathbb{R}^{m\times n}$ и векторы $\hat{u} \in \mathbb{R}^m$ и $\hat{v} \in \mathbb{R}^n$ обладают двумерным альтернансом, то для любого $u \in \mathbb{R}^m$ такого, что $\operatorname{sign} u = \operatorname{sign} \hat{u}$ выполнено $\|A - \hat{u}\hat{v}^T\|_C \leqslant \|A - uv^T\|_C$ для любого $v \in \mathbb{R}^n$. Аналогичное свойство также выполнено для $v \in \mathbb{R}^n$ таких, что $\operatorname{sign} v = \operatorname{sign} \hat{v}$.

В настоящей работе метод переменных направлений обобщается на случай приближений произвольного ранга, анализируются его свойства, вводится понятие *двумерного альтернанса ранга* r, доказывается, что наличие такого альтернанса является необходимым условием решения задачи (2.2), а также что предельные точки метода переменных направлений удовлетворяют этому свойству. Кроме того, проводится детальный анализ метода переменных направлений для построения приближений ранга 1 и предлагается метод, позволяющий находить оптимальные аппроксимации.

2.2 Основные свойства

Нетрудно показать, что решение задачи (2.2) существует.

Утверждение 2.1. *Решение задачи*

$$\|A - UV^T\|_C \to \min_{U\in\mathbb{C}^{m\times r}, V\in\mathbb{C}^{m\times r}}$$

существует для любой матрицы $A \in \mathbb{C}^{m\times n}$ *и ранга* r.

Доказательство. Пусть $B \in \mathbb{C}^{m\times n}$ и $\|B\|_C \geqslant 3\|A\|_C$. Тогда

$$\|A - B\|_C \geqslant \|B\|_C - \|A\|_C \geqslant 2\|A\|_C.$$

Следовательно матрица B с таким свойством дает ошибку хуже, чем нулевое приближение. Поэтому достаточно искать решение среди матриц B таких, что

$\|B\|_C \leqslant 3\|A\|_C$. Это множество замкнуто и ограничено, а множество матриц ранга не выше r замкнуто, поэтому достаточно искать решение на замкнутом ограниченном множестве. Поскольку функционал $\|A - B\|_C$ непрерывен по B, из теоремы Вейерштрасса имеем, что существует матрица $\hat{B}$ ранга не выше r такая, что

$$\|A - \hat{B}\|_C = \inf_{B \in \mathbb{C}^{m\times n}, \operatorname{rank} B \leqslant r} \|A - B\|_C = \inf_{U \in \mathbb{C}^{m\times r}, V \in \mathbb{C}^{m\times r}} \|A - UV^T\|_C.$$

Остается заметить, что если $\operatorname{rank} \hat{B} \leqslant r$, то $\hat{B} = \hat{U}\hat{V}^T$, $\hat{U} \in \mathbb{C}^{m\times r}$, $\hat{V} \in \mathbb{C}^{n\times r}$. $\square$

Однако стоит отметить, что решение задачи (2.2) может быть не единственным. Во-первых, если $\hat{U}$ и $\hat{V}$ — решение задачи, то $\hat{U}R$ и $\hat{V}R^{-T}$ также являются решением задачи для любой невырожденной матрицы R. Однако даже произведение $\hat{U}\hat{V}^T$ может быть не единственным. Так, например, рассмотрим задачу приближения единичной матрицы 2×2 рангом 1. Эта задача имеет как минимум два решения

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \approx \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

и

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \approx \begin{bmatrix} 1/2 \\ -1/2 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}.$$

Кроме этого, даже для вещественных матриц решение может различаться в зависимости от того ищем мы решение над вещественным или комплексным полем. Простым примером является единичная матрица 4×4, которую требуется приблизить рангом 2. В вещественном случае решением является матрица

$$\begin{bmatrix} 2-\sqrt{2} & 0 & 1-\sqrt{2} & 1-\sqrt{2} \\ 0 & 2-\sqrt{2} & 1-\sqrt{2} & \sqrt{2}-1 \\ 1-\sqrt{2} & 1-\sqrt{2} & 2-\sqrt{2} & 0 \\ 1-\sqrt{2} & \sqrt{2}-1 & 0 & 2-\sqrt{2} \end{bmatrix},$$

которая дает точность $\sqrt{2} - 1 \approx 0.414213562373095...$ [32] (стоит отметить, что существует по крайней мере 20 различных решений с такой точностью). При приближении над комплексным полем решением является матрица

$$\frac{1}{1+\sqrt{3}} \begin{bmatrix} \sqrt{3} & i & i & i \\ -i & \sqrt{3} & i & -i \\ -i & -i & \sqrt{3} & i \\ -i & i & -i & \sqrt{3} \end{bmatrix},$$

которая дает точность $\frac{1}{1+\sqrt{3}} \approx 0.36602540378444...$ [32]. Вся оставшаяся часть главы посвящена задаче вещественного приближения, а именно, требуется найти матрицы $\hat{U} \in \mathbb{R}^{m\times r}$ и $\hat{V} \in \mathbb{R}^{n\times r}$, являющиеся решением задачи

$$\|A - UV^T\|_C \to \min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}}, \tag{2.3}$$

где $A \in \mathbb{R}^{m\times n}$.

2.3 Метод переменных направлений

В данном разделе приводится описание метода переменных направлений для решения задачи построения малоранговых приближений матриц в чебышевской норме. Пусть $A \in \mathbb{R}^{m\times n}$. Здесь и далее будем предполагать, что размеры m и n строго больше 1. Также будем предполагать, что $\operatorname{rank} A > r$. Задача (2.3) трудна для явного решения [21], поэтому предположим, что одна из матриц (U или V) известна. Рассмотрим задачу

$$\|A - UV^T\|_C \to \min_{U\in\mathbb{R}^{m\times r}}, \tag{2.4}$$

которая может быть разбита на множество независимых задач вида

$$\|a - Vu\|_\infty \to \min_{u\in\mathbb{R}^r},$$

где $V \in \mathbb{R}^{n\times r}$ и $a \in \mathbb{R}^n$. Пусть $V \in \mathbb{R}^{n\times r}$ является чебышевской матрицей. Тогда существует единственное отображение (см. Теорему 1.5) $\varphi : \mathbb{R}^{m\times n}\times\mathbb{R}^{n\times r} \to \mathbb{R}^{m\times r}$ такое, что

$$\varphi(A, V)^i = \arg\min_{x\in\mathbb{R}^r} \|a^i - Vx\|_\infty,$$

где верхним индексом обозначена строка матрицы. Заметим, что $\varphi(A, V)$ является решением задачи (2.4) (однако решение задачи (2.4) может быть не единственно). Аналогично определим отображение $\psi : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times r} \to \mathbb{R}^{n\times r}$ такое, что

$$\psi(A, U)^j = \arg\min_{x\in\mathbb{R}^r} \|a_j - Ux\|_\infty.$$

$\psi(A, U)$ является решением задачи

$$\|A - UV^T\|_C \to \min_{V\in\mathbb{R}^{n\times r}}.$$

Определение 2.2. *Пусть $A \in \mathbb{R}^{m\times n}$. Будем говорить, что пара последовательностей чебышевских матриц $\{U^{(t)} \in \mathbb{R}^{m\times r}\}_{t\in\mathbb{N}}$ и $\{V^{(t)} \in \mathbb{R}^{n\times r}\}_{t\in\mathbb{N}}$ получена методом переменных направлений для матрицы A и начальной точки $V^{(0)}$, где $V^{(0)} \in \mathbb{R}^{n\times r}$ является чебышевской матрицей, если*

$$\begin{cases} U^{(t)} = \varphi(A, V^{(t-1)}), \\ V^{(t)} = \psi(A, U^{(t)}) \end{cases}$$

при всех $t \in \mathbb{N}$.

Заметим, что если матрица V является чебышевской, то $\varphi(A, V)$ не обязательно будет чебышевской. В дальнейшем мы покажем, что в случае построения приближений ранга 1 для почти всех матриц A, если V является чебышевской, то $\varphi(A, V)$ также является чебышевской. Однако это не верно при построении приближений произвольного ранга. Более того, мы предполагаем, что при $r \geqslant 2$ для почти всех матриц $A \in \mathbb{R}^{m\times n}$ существует чебышевская матрица $V \in \mathbb{R}^{n\times r}$ такая, что $\varphi(A, V)$ не является чебышевской. К счастью, наши численные эксперименты показывают, что такие ситуации редки. Тем не менее, при применении Определения 2.2 в теоретических выкладках, нам требуется явно предполагать, что построенные матрицы являются чебышевскими.

Сформулируем базовые свойства метода переменных направлений.

Лемма 2.3. *Пусть $A \in \mathbb{R}^{m\times n}$ и матрица $V^{(0)} \in \mathbb{R}^{n\times r}$ является чебышевской. Пусть пара последовательностей $\{U^{(t)} \in \mathbb{R}^{m\times r}\}_{t\in\mathbb{N}}$ и $\{V^{(t)} \in \mathbb{R}^{n\times r}\}_{t\in\mathbb{N}}$ порождена методом переменных направлений для матрицы A и начальной точки $V^{(0)}$. Тогда выполнены следующие утверждения.*

(i) Выполнены неравенства

$$\|A - U^{(t)}(V^{(t-1)})^T\|_C \geqslant \|A - U^{(t)}(V^{(t)})^T\|_C \geqslant \|A - U^{(t+1)}(V^{(t)})^T\|_C$$

при всех $t \in \mathbb{N}$.

(ii) Если пара последовательностей $\{\tilde{U}^{(t)}\}_{t\in\mathbb{N}}$ и $\{\tilde{V}^{(t)}\}_{t\in\mathbb{N}}$ получена методом переменных направлений для матрицы A и начальной точки $\alpha V^{(0)}$, где $\alpha \neq 0$, то $\tilde{U}^{(t)} = 1/\alpha\, U^{(t)}$ и $\tilde{V}^{(t)} = \alpha V^{(t)}$.

Доказательство. По построению φ,

$$\inf_{U\in\mathbb{R}^{m\times r}} \|A - U(V^{(t)})^T\|_C = \|A - \varphi(A,V^{(t)})(V^{(t)})^T\|_C,$$

откуда получаем

$$\|A - U^{(t)}(V^{(t)})^T\|_C \geqslant \|A - U^{(t+1)}(V^{(t)})^T\|_C$$

поскольку $U^{(t+1)} = \varphi(T,V^{(t)})$. Второе неравенство в (i) доказывается аналогично.

Утверждение (ii) следует из единственности решения задачи наилучшего равномерного приближения (см. Теорему 1.5). □

Пусть матрица $V \in \mathbb{R}^{n\times r}$ является чебышевской и пара последовательностей $\{U^{(t)}\}_{t\in\mathbb{N}}$ и $\{V^{(t)}\}_{t\in\mathbb{N}}$ получена методом переменных направлений для матрицы A и начальной точки $V^{(0)} = V$. По Лемме 2.3 (i) последовательность $\|A - U^{(t)}(V^{(t)})^T\|_C$ не возрастает и, поскольку состоит только из неотрицательных чисел, сходится. Обозначим предел этой последовательности через $E(A,V)$. Следующая лемма содержит элементарные свойства этой функции.

Лемма 2.4. *Пусть $A \in \mathbb{R}^{m\times n}$ и матрица $V \in \mathbb{R}^{n\times r}$ является чебышевской. Пусть также метод переменных направлений для матрицы A и начальной точки $V^{(0)} = V$ является корректным (то есть порожденные матрицы являются чебышевскими). Тогда выполнены следующие утверждения.*

(i) $E(A,V) \geqslant 0$ *и* $E(A,V) = E(A,\alpha V)$ *при* $\alpha \neq 0$.

(ii) $E(A,V) = E(A,\tilde{V})$, *где* $\tilde{V} = \psi(A, \varphi(A,V))$.

(iii) Функция $E(A,V)$ является полунепрерывной сверху по V.

Доказательство. Утверждения (i) и (ii) сразу следуют из определения $E(T,V)$ и Леммы 2.3. Полунепрерывность сверху имеет место поскольку $E(A,V)$ является пределом убывающей последовательности непрерывных функций (φ и ψ непрерывны по Теореме 1.7). □

Итоговая процедура метода переменных направлений представлена в Алгоритме 5. Заметим, что в алгоритме также используются перенормировки на каждой итерации, поскольку по Лемме 2.4 они не влияют на итоговый результат, но улучшают численную устойчивость.

Вопрос о сходимости последовательностей $\{U^{(t)}\}_{t\in\mathbb{N}}$ и $\{V^{(t)}\}_{t\in\mathbb{N}}$ остается открытым, поэтому все дальнейшие результаты будут сформулированы для предельных точек последовательностей. Во всех проведенных численных экспериментах последовательности, порождаемые методом переменных направлений, сходятся, однако нам не удается доказать или опровергнуть наличие сходимости теоретически. Стоит отметить, что аналогичная ситуация известна для популярного метода переменных наименьших квадратов [33].

Входные данные: Матрица $A \in \mathbb{R}^{m\times n}$, ранг $r \geqslant 1$, начальная матрица $V^{(0)} \in \mathbb{R}^{n\times r}$.

Результат: Факторы аппроксимации ранга r: $\hat{U} \in \mathbb{R}^{m\times r}$ и $\hat{V} \in \mathbb{R}^{n\times r}$.

$t = 1$;

repeat

$U^{(t)} = \varphi(A, V^{(t-1)})$;

$V^{(t)} = \psi(A, U^{(t)})$;

$C = \|U^{(t)}\|_C \|V^{(t)}\|_C$;

$U^{(t)} = U^{(t)}/\|U^{(t)}\|_C \cdot C^{1/2}$;

$V^{(t)} = V^{(t)}/\|V^{(t)}\|_C \cdot C^{1/2}$;

$t = t + 1$

until *сходимость*;

$\hat{U} = U^{(t-1)}, \quad \hat{V} = V^{(t-1)}$;

Алгоритм 5: Метод переменных направлений.

2.4 Теорема об альтернансе

В данном разделе вводится понятие двумерного альтернанса ранга r и доказывается, что оптимальное решение задачи (2.3) обладает введенной структурой. Кроме того, все предельные точки метода переменных направлений также обладают двумерным альтернансом ранга r. Введем необходимые обозначения. Пусть $A \in \mathbb{R}^{m\times n}$ и матрицы $U \in \mathbb{R}^{m\times r}$ и $V \in \mathbb{R}^{n\times r}$ являются чебышевскими. Обозначим $G = A - UV^T$. Обозначим также

$$S(A, U, V) = \{(i, j) : |g_{ij}| = \|G\|_C\},$$

$$\mathcal{I}(A, U, V) = \{i : \exists j \text{ такие, что } (i, j) \in S(A, U, V)\},$$

$$\mathcal{J}(A, U, V) = \{j : \exists i \text{ такие, что } (i, j) \in S(A, U, V)\}.$$

Определение 2.5. *Пусть $A \in \mathbb{R}^{m\times n}$ и матрицы $U \in \mathbb{R}^{m\times r}$ и $V \in \mathbb{R}^{n\times r}$ являются чебышевскими. Будем говорить, что тройка (A, U, V) обладает двумерным альтернансом ранга r, если существует непустое множество $\mathcal{A} \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$ такое, что $\mathcal{A} \subset S(T, U, V)$ и если $(i, j) \in \mathcal{A}$, то существует множество I, состоящее из $r + 1$ различных целых чисел*

$$1 \leqslant i_1 < i_2 < \cdots < i_{r+1} \leqslant m$$

таких, что $i \in I$, а также множество J, состоящее из $r+1$ различных целых чисел

$$1 \leqslant j_1 < j_2 < \cdots < j_{r+1} \leqslant n$$

таких, что $j \in J$, причем выполнены следующие свойства.

1. $(i, j_1), (i, j_2), \ldots, (i, j_{r+1}), (i_1, j), (i_2, j), \ldots, (i_{r+1}, j) \in \mathcal{A}$.
2. *Знаки в последовательности*

$$g_{ij_1} D_1(\mathcal{V}), g_{ij_2} D_2(\mathcal{V}), \ldots, g_{ij_{r+1}} D_{r+1}(\mathcal{V})$$

и знаки в последовательности

$$g_{i_1 j} D_1(\mathcal{U}), g_{i_2 j} D_2(\mathcal{U}), \ldots, g_{i_{r+1} j} D_{r+1}(\mathcal{U})$$

чередуются, где $\mathcal{U} = U(I)$ и $\mathcal{V} = V(J)$.

Основные результаты данного раздела следующие:

1. если $\hat{U} \in \mathbb{R}^{m \times r}$ и $\hat{V} \in \mathbb{R}^{n \times r}$ таковы, что $\hat{U}\hat{V}^T$ является наилучшим приближением ранга r для матрицы A в чебышевской норме, то $(A, \hat{U}, \hat{V})$ обладает двумерным альтернансом ранга r;
2. предельные точки метода переменных направлений обладают двумерным альтернансом ранга r.

Приведем строгие формулировки этих утверждений.

Теорема 2.6. *Пусть $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank} A > r$. Пусть $\hat{U} \in \mathbb{R}^{m \times r}$ и $\hat{V} \in \mathbb{R}^{n \times r}$ являются решением задачи*

$$\|A - UV^T\|_C \to \min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}}.$$

Пусть матрица $\hat{V}$ является чебышевской и метод переменных направлений для матрицы A и начальной точки $V^{(0)} = \hat{V}$ является корректным. Тогда тройка $(A, \hat{U}, \hat{V})$ обладает двумерным альтернансом ранга r.

Теорема 2.7. *Пусть $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank} A > r$ и матрица $V \in \mathbb{R}^{n \times r}$ является чебышевской. Пусть метод переменных направлений для матрицы A и начальной точки $V^{(0)} = V$ является корректным и последовательности $\{U^{(t)}\}_{t \in \mathbb{N}}$ и $\{V^{(t)}\}_{t \in \mathbb{N}}$ получены методом переменных направлений. Пусть Ξ является чебышевской матрицей и предельной точкой последовательности Ξ_t, где $\Xi_t = V^{(t)} / \|V^{(t)}\|_C$ являются чебышевскими. Тогда $(A, \varphi(A, \Xi), \Xi)$ обладает двумерным альтернансом ранга r.*

Для доказательства этим теорем нам понадобится несколько подготовительных лемм.

Лемма 2.8. *Пусть $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank} A > r$ и матрицы $V \in \mathbb{R}^{n\times r}$ и $U = \varphi(A, V)$ являются чебышевскими. Тогда для любого $i \in \mathcal{I}(A, U, V)$ существует множество $J = (j_1, \dots, j_{r+1})$, где*

$$1 \leqslant j_1 < j_2 < \cdots < j_{r+1} \leqslant n,$$

такое, что $(i, j_1), \dots, (i, j_{r+1}) \in S(A, U, V)$ и знаки в последовательности

$$g_{ij_1} D_1(\mathcal{V}), g_{ij_2} D_2(\mathcal{V}), \dots, g_{ij_{r+1}} D_{r+1}(\mathcal{V})$$

чередуются, где $\mathcal{V} = V(J)$.

Доказательство. Пусть $i \in \mathcal{I}(A, U, V)$. По определению φ, i–ая строка матрицы U является решением задачи

$$\|a^i - Vx\|_\infty \to \min_{x\in\mathbb{R}^r},$$

откуда по Теореме 1.24 существует множество из $r + 1$ целых чисел

$$1 \leqslant j_1 < j_2 < \cdots < j_{r+1} \leqslant n$$

таких, что знаки в последовательности

$$(a_{ij_1} - (v^{j_1})^T u^i) D_1(\mathcal{V}), (a_{ij_2} - (v^{j_2})^T u^i) D_2(\mathcal{V}), \dots, (a_{ij_{r+1}} - (v^{j_{r+1}})^T u^i) D_1(\mathcal{V})$$

чередуются и

$$|a_{ij_1} - (v^{j_1})^T u^i| = \cdots = |a_{ij_{r+1}} - (v^{j_{r+1}})^T u^i| = \|a^i - Vu^i\|_\infty = \|G\|_C,$$

где последнее равенство выполнено поскольку $i \in \mathcal{I}(A, U, V)$. □

Лемма 2.9. *Пусть $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank} A > r$ и матрицы $V \in \mathbb{R}^{n\times r}$, $U = \varphi(A, V)$ и $\tilde{V} = \psi(A, U)$ являются чебышевскими. Пусть также $\|A - UV^T\|_C = \|A - U\tilde{V}^T\|_C$. Тогда $S(A, U, \tilde{V}) \subset S(A, U, V)$. Более того, $j \in \mathcal{J}(A, U, \tilde{V})$ тогда и только тогда, когда $j \in \mathcal{J}(A, U, V)$ и $v^j = \tilde{v}^j$, что имеет место тогда и только тогда, когда существует множество $I = (i_1, \dots, i_{r+1})$ такое, что*

$$1 \leqslant i_1 < i_2 < \cdots < i_{r+1} \leqslant m,$$

$(i_1, j), \dots, (i_{r+1}, j) \in S(A, U, V)$ и знаки в последовательности

$$g_{i_1 j} D_1(\mathcal{U}), g_{i_2 j} D_2(\mathcal{U}), \dots, g_{i_{r+1} j} D_{r+1}(\mathcal{U})$$

чередуются, где $\mathcal{U} = U(I)$.

Доказательство. Рассмотрим произвольный индекс j. Поскольку $\tilde{V} = \psi(A, U)$,

$$\|a_j - Uv^j\|_\infty \geqslant \|a_j - U\tilde{v}^j\|_\infty.$$

Пусть $j \notin \mathcal{J}(A, U, V)$. Тогда $\|a_j - Uv^j\|_\infty < \|A - UV^T\|_C$, откуда $\|a_j - Uv^j\|_\infty < \|A - U\tilde{V}^T\|_C$, следовательно $j \notin \mathcal{J}(A, U, \tilde{V})$.

Пусть теперь $\tilde{v}^j \neq v^j$. Благодаря единственности решения задачи наилучшего равномерного приближения (см. Теорему 1.5),

$$\|a_j - U\tilde{v}^j\|_\infty < \|a_j - Uv^j\|_\infty \leqslant \|A - UV^T\|_C = \|A - U\tilde{V}^T\|_C,$$

откуда следует, что $j \notin \mathcal{J}(A, U, \tilde{V})$. Таким образом мы показали, что если $j \in \mathcal{J}(A, U, \tilde{V})$, то $j \in (A, U, V)$ и $\tilde{v}^j = v^j$.

Докажем обратное, пусть $j \in \mathcal{J}(A, U, V)$ и $\tilde{v}^j = v^j$. Из последнего условия получаем

$$\|a_j - Uv^j\|_\infty = \|a - U\tilde{v}^j\|_\infty,$$

а из $j \in \mathcal{J}(A, U, V)$ имеем

$$\|a_j - Uv^j\|_\infty = \|A - UV^T\|_C = \|A - U\tilde{V}^T\|_C.$$

Следовательно $j \in \mathcal{J}(A, U, \tilde{V})$.

Поскольку $\tilde{v}^j$ является решением задачи

$$\|a_j - Ux\|_\infty \to \min_{x \in \mathbb{R}^r},$$

по Теореме 1.24 существует множество $I = (i_1, i_2, \ldots, i_{r+1})$, где

$$1 \leqslant i_1 < i_2 < \cdots < i_{r+1} \leqslant m,$$

такое, что

$$|a_{i_1 j} - (\tilde{v}^j)^T u^{i_1}| = |a_{i_1 j} - (\tilde{v}^j)^T u^{i_1}| = \|a_j - U\tilde{v}^j\|_\infty,$$

и знаки в последовательности

$$(a_{i_1 j} - (\tilde{v}^j)^T u^{i_1}) D_1(\mathcal{U}), (a_{i_2 j} - (\tilde{v}^j)^T u^{i_2}) D_2(\mathcal{U}), \ldots, (a_{i_{r+1} j} - (\tilde{v}^j)^T u^{i_{r+1}}) D_{r+1}(\mathcal{U})$$

чередуются, где $\mathcal{U} = U(I)$. Поскольку $j \in \mathcal{J}(A, U, \tilde{V})$,

$$\|a_j - U\tilde{v}^j\|_\infty = \|A - U\tilde{V}^T\|_C = \|A - UV^T\|_C,$$

следовательно $(i_1, j), (i_2, j), \ldots, (i_{r+1}, j) \in S(A, U, \tilde{V})$, но поскольку $v^j = \tilde{v}^j$, $(i_1, j), (i_2, j), \ldots, (i_{r+1}, j) \in S(A, U, V)$ и знаки в последовательности

$$(a_{i_1 j} - (v^j)^T u^{i_1}) D_1(\mathcal{U}), (a_{i_2 j} - (v^j)^T u^{i_2}) D_2(\mathcal{U}), \ldots, (a_{i_{r+1} j} - (v^j)^T u^{i_{r+1}}) D_{r+1}(\mathcal{U})$$

чередуются.

Обратно, пусть существует множество $I = (i_1, i_2, \ldots, i_{r+1})$, где

$$1 \leqslant i_1 < i_2 < \cdots < i_{r+1} \leqslant m,$$

такое, что $(i_1, j), \ldots, (i_{r+1}, j) \in S(A, U, V)$ и знаки в последовательности

$$g_{i_1 j} D_1(\mathcal{U}), g_{i_2 j} D_2(\mathcal{U}), \ldots, g_{i_{r+1} j} D_{r+1}(\mathcal{U})$$

чередуются, где $\mathcal{U} = U(I)$. Тогда по Теореме 1.24 v^j является решением задачи

$$\|a_j - Ux\|_\infty \to \min_{x \in \mathbb{R}^r},$$

но поскольку по Теореме 1.5 решение единственно, получаем $v^j = \tilde{v}^j$ и, очевидно, $j \in \mathcal{J}(A, U, V)$.

Докажем, наконец, что $S(A, U, \tilde{V}) \subset S(A, U, V)$. Пусть $(i, j) \in S(A, U, \tilde{V})$. Тогда $j \in \mathcal{J}(A, U, \tilde{V})$, откуда $\tilde{v}^j = v^j$. Следовательно

$$a_{ij} - (v^j)^T u^i = a_{ij} - (\tilde{v}^j)^T u^i.$$

Однако

$$|a_{ij} - (\tilde{v}^j)^T u^i| = \|A - U\tilde{V}^T\|_C = \|A - UV^T\|_C,$$

откуда следует, что $(i, j) \in S(A, U, V)$. □

Лемма 2.10. *Пусть $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank} A > r$ и матрицы $V \in \mathbb{R}^{n \times r}$, $U = \varphi(A, V)$ и $\tilde{V} = \psi(A, U)$ являются чебышевскими. Пусть также*

$$\|A - UV^T\|_C = \|A - U\tilde{V}^T\|_C$$

и

$$S(A, U, V) = S(A, U, \tilde{V}).$$

Тогда (A, U, V) обладает двумерным альтернансом ранга r.

Доказательство. Покажем, что (A, U, V) обладает двумерным альтернансом ранга r с множеством пар индексов $\mathcal{A} = S(A, U, V)$. Пусть $(i, j) \in \mathcal{A}$, тогда $i \in \mathcal{I}(A, U, V)$ и по Лемме 2.8 существует множество $J = (j_1, \ldots, j_{r+1})$, где

$$1 \leqslant j_1 < j_2 < \cdots < j_{r+1} \leqslant n$$

такое, что $(i, j_1), \ldots, (i, j_{r+1}) \in S(A, U, V)$ и знаки в последовательности

$$g_{ij_1} D_1(\mathcal{V}), g_{ij_2} D_2(\mathcal{V}), \ldots, g_{ij_{r+1}} D_{r+1}(\mathcal{V})$$

чередуются, где $\mathcal{V} = V(J)$. Может случиться, что $j \notin J$. Но по Лемме 1.26 существует множество J', полученное из J заменой одного из элементов на j, такое, что знаки в последовательности

$$g_{ij'_1} D_1(\mathcal{V}'), g_{ij'_2} D_2(\mathcal{V}'), \ldots, g_{ij'_{r+1}} D_{r+1}(\mathcal{V}') \tag{2.5}$$

чередуются, где $J' = (j'_1, \ldots, j'_{r+1})$ и $\mathcal{V}' = V(J')$. Множество J' может не быть упорядоченным, но заметим, что перестановка двух соседних элементов в J' сохраняет альтернанс в (2.5). Таким образом, существует множество $J' = (j'_1, \ldots, j'_{r+1})$, где

$$1 \leqslant j'_1 < j'_2 < \cdots < j'_{r+1} \leqslant n,$$

такое, что $j \in J'$, $(i, j'_1), \ldots, (i, j'_{r+1}) \in S(A, U, V)$ и знаки в последовательности (2.5) чередуются.

Пусть $(i, j) \in S(A, U, \tilde{V}) = S(A, U, V) = \mathcal{A}$. Тогда $j \in \mathcal{J}(A, U, \tilde{V})$, откуда и из Леммы 2.9 существует множество $I = (i_1, \ldots, i_{r+1})$ такое, что

$$1 \leqslant i_1 < i_2 < \cdots < i_{r+1} \leqslant m,$$

$(i_1, j), \ldots, (i_{r+1}, j) \in S(A, U, V)$ и знаки в последовательности

$$g_{i_1 j} D_1(\mathcal{U}), g_{i_2 j} D_2(\mathcal{U}), \ldots, g_{i_{r+1} j} D_{r+1}(\mathcal{U})$$

чередуются, где $\mathcal{U} = U(I)$. Если $i \notin I$, мы можем повторить рассуждения выше и получить, что существует множество $I' = (i'_1, \ldots, i'_{r+1})$, где

$$1 \leqslant i'_1 < i'_2 < \cdots < i'_{r+1} \leqslant m,$$

такое, что $i \in I'$, $(i'_1, j), \ldots, (i'_{r+1}, j) \in S(A, U, V)$ и

$$g_{i'_1 j} D_1(\mathcal{U}'), g_{i'_2 j} D_2(\mathcal{U}'), \ldots, g_{i'_{r+1} j} D_{r+1}(\mathcal{U}')$$

чередуются, где $\mathcal{U}' = U(I')$. Таким образом, все свойства из определения двумерного альтернанса ранга r выполнены. □

Лемма 2.11. *Пусть* $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank} A > r$ *и матрицы* $V \in \mathbb{R}^{n\times r}$, $U = \varphi(A,V)$, $\tilde V = \psi(A,U)$ *и* $\tilde U = \varphi(A,\tilde V)$ *являются чебышевскими. Пусть также* $\|A - \tilde U\tilde V^T\|_C = \|A - UV^T\|_C$ *и* $(A,\tilde U,\tilde V)$ *обладает двумерным альтернансом ранга* r. *Тогда* (A,U,V) *также обладает двумерным альтернансом ранга* r.

Доказательство. По Лемме 2.3 (i)

$$\|A - UV^T\|_C \geqslant \|A - U\tilde V^T\|_C \geqslant \|A - \tilde U\tilde V^T\|_C,$$

однако первый и последний члены равны, поэтому

$$\|A - UV^T\|_C = \|A - U\tilde V^T\|_C = \|A - \tilde U\tilde V^T\|_C.$$

Применяя Лемму 2.9, получаем $S(A,U,\tilde V) \subset S(T,U,V)$. Поскольку $\|A^T - \tilde V U^T\|_C = \|A^T - \tilde V\tilde U^T\|_C$, мы имеем $S(A^T,\tilde V,\tilde U) \subset S(A^T,\tilde V,U)$, что эквивалентно $S(A,\tilde U,\tilde V) \subset S(A,U,\tilde V)$, следовательно $S(A,\tilde U,\tilde V) \subset S(A,U,V)$.

Пусть (i,j) принадлежит двумерному альтернансу ранга r для $(A,\tilde U,\tilde V)$ и $I = (i_1,\dots,i_{r+1})$ и $J = (j_1,\dots,j_{r+1})$ являются множествами индексов, соответствующими (i,j) из определения альтернанса. Из Леммы 2.9 получаем, что $u^i = \tilde u^i$ для любого $i \in \mathcal{I}(A,\tilde U,\tilde V)$, следовательно $U(I) = \tilde U(I)$. Аналогично получаем, что $V(J) = \tilde V(J)$. Таким образом,

$$(a_{i_k j} - (u^{i_k})^T v^j) D_k(U(I)) = (a_{i_k j} - (\tilde u^{i_k})^T v^j) D_k(\tilde U(I)), \quad k = 1,2,\dots,r+1$$

поскольку $j \in J$. Также получаем, что

$$(a_{i j_k} - (u^i)^T v^{j_k}) D_k(V(J)) = (a_{i j_k} - (u^i)^T \tilde v^{j_k}) D_k(\tilde V(J)), \quad k = 1,2,\dots,r+1$$

поскольку $i \in I$. Таким образом, двумерный альтернанс ранга r для $(A,\tilde U,\tilde V)$ является двумерным альтернансом ранга r для (A,U,V). □

Лемма 2.12. *Пусть* $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank} A > r$ *и матрица* $V \in \mathbb{R}^{n\times r}$ *является чебышевской. Пусть последовательности* $\{U^{(t)}\}_{t\in\mathbb{N}}$ *и* $\{V^{(t)}\}_{t\in\mathbb{N}}$ *являются чебышевскими и построены методом переменных направлений для матрицы* A *и начальной точки* $V^{(0)} = V$. *Пусть также чебышевская матрица* $\Xi \in \mathbb{R}^{n\times r}$ *является предельной точкой последовательности* $\Xi^{(t)} = V^{(t)}/\|V^{(t)}\|_C$. *Тогда*

$$E(A,\Xi) = \|A - \varphi(A,\Xi)\Xi^T\|_C = E(A,V).$$

Доказательство. Из Леммы 2.4 (ii) следует, что $E(T,\Xi^{(t)}) = E(T,V)$ для любых $t \in \mathbb{N}$. Из полунепрерывности сверху для E (см. Лемму 2.4 (iii)) получаем, что $E(A,\Xi) \geqslant E(A,V)$. Более того, $\|A - \varphi(A,\Xi)\Xi^T\|_C \geqslant E(T,\Xi)$.

$$\|A - \varphi(A,\Xi)\Xi^T\|_C = \lim_{t\to\infty} \|A - \varphi(A,\Xi^{(l_t)})(\Xi^{(l_t)})^T\|_C = \lim_{t\to\infty} \|A - \varphi(A,V^{(l_t)})(V^{(l_t)})^T\|_C = E(A,V).$$

Таким образом,

$$E(A,V) = \|A - \varphi(A,\Xi)\Xi^T\|_C \geqslant E(A,\Xi) \geqslant E(A,V)$$

и лемма доказана. □

Лемма 2.13. *Пусть $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank} A > r$ и матрица $V \in \mathbb{R}^{n\times r}$ является чебышевской. Пусть метод переменных направлений для матрицы A и начальной точки V является корректным. Тогда если $\|A - \varphi(A,V)V^T\|_C = E(A,V)$, то $(A,\varphi(A,V),V)$ обладает двумерным альтернансом ранга r.*

Доказательство. Пусть пара последовательностей $\{U^{(t)}\}_{t\in\mathbb{N}}$ и $\{V^{(t)}\}_{t\in\mathbb{N}}$ получены методом переменных направлений для матрицы A и начальной точки $V^{(0)} = V$. Ясно, что

$$\|A - U^{(t)}(V^{(t-1)})^T\|_C = \|A - U^{(t)}(V^{(t)})^T\|_C = \|A - U^{(t+1)}(V^{(t)})^T\|_C$$

для любых $t \in \mathbb{N}$. Тогда из Леммы 2.9 следует, что

$$S(A,U^{(1)},V^{(0)}) \supset S(A,U^{(1)},V^{(1)}) \supset S(A,U^{(2)},V^{(1)}) \supset S(A,U^{(2)},V^{(2)}) \supset \dots$$

Поскольку все множества в этой последовательности конечны и не пусты, существует $t \in \mathbb{N}$ такое, что $S(A,U^{(t+1)},V^{(t)}) = S(A,U^{(t+1)},V^{(t+1)})$. Тогда из Леммы 2.10 следует, что $(A,U^{(t+1)},V^{(t)})$ обладает двумерным альтернансом ранга r. Применяя Лемму 2.11 t раз получаем, что $(A,U^{(1)},V^{(0)}) = (A,\varphi(T,V),V)$ обладает двумерным альтернансом ранга r. □

Доказательство Теоремы 2.6. Ясно, что

$$\|A - \hat{U}\hat{V}^T\|_C \geqslant \|A - \varphi(A,\hat{V})\hat{V}^T\|_C \geqslant E(A,\hat{V}) \geqslant \|A - \hat{U}\hat{V}^T\|_C,$$

где последнее неравенство выполнено поскольку $\hat{U}$ и $\hat{V}$ являются оптимальным решением. Следовательно $\|A - \varphi(A,\hat{V})\hat{V}^T\|_C = E(A,\hat{V})$ и по Лемме 2.13

тройка $(A, \varphi(A, \hat{V}), \hat{V})$ обладает двумерным альтернансом ранга r. Заметим, что $\mathcal{I}(A, \varphi(A, \hat{V}), \hat{V}) \subset \mathcal{I}(A, \hat{U}, \hat{V})$ по определению φ. Обозначим $\tilde{U} = \varphi(A, \hat{V})$. При $i \in \mathcal{I}(A, \varphi(A, \hat{V}), \hat{V})$ получаем, что

$$\|a^i - \hat{V}\hat{u}^i\|_\infty = \|a^i - \hat{V}\tilde{u}^i\|_\infty = \min_{u \in \mathbb{R}^r} \|a^i - \hat{V}u\|_\infty.$$

Поскольку решение задачи наилучшего равномерного приближения единственно (см. Теорему 1.5), $\hat{u}^i = \tilde{u}^i$ при $i \in \mathcal{I}(A, \varphi(A, \hat{V}), \hat{V})$. Поскольку в Определении 2.5 используются только строки матрицы U такие, что $i \in \mathcal{I}(A, \varphi(A, V), V)$, получаем, что альтернанс для тройки $(A, \varphi(A, \hat{V}), \hat{V})$ является альтернансом для тройки $(A, \hat{U}, \hat{V})$. □

Доказательство Теоремы 2.7. Сразу следует из Леммы 2.12 и Леммы 2.13. □

2.5 Корректность метода переменных направлений для ранга 1

В этом и последующих двух разделах детально анализируется случай построения приближений ранга 1. В данном разделе будет обоснована корректность метода переменных направлений, а именно, будет показано, что для почти всех матриц, если начальная точка является чебышевской, то все матрицы, порождаемые методом переменных направлений, также будут чебышевскими. Введем необходимое

Определение 2.14. *Будем называть вектор $v \in \mathbb{R}^n$ пиковым, если существует единственное число $i \in \{1, \ldots, n\}$ такое, что $|v_i| = \|v\|_\infty$.*

Пусть a,$v \in \mathbb{R}^n$ и вектор v является чебышевским. Будем обозначать через $\gamma(a,v)$ решение задачи наилучшего равномерного приближения для векторов a и v, а именно,

$$\|a - \gamma(a,v)v\|_\infty = \inf_{u \in \mathbb{R}} \|a - uv\|_\infty.$$

Если вектор a является пиковым, введем также обозначения $\zeta(a) \in \{1, \ldots, n\}$ и $\delta(a) \in \mathbb{R}$ такие, что

$$|a_{\zeta(a)}| = \|a\|_\infty \quad \text{и} \quad \delta(a) = \|a\|_\infty - \max_{j \neq \zeta(a)} |a_j|.$$

Теорема 2.15. *Пусть $a \in \mathbb{R}^n$. Тогда выполнены следующие утверждения.*

(i) $\gamma(a,v) \neq 0$ *для любого чебышевского вектора* $v \in \mathbb{R}^n$ *тогда и только тогда, когда* a *является пиковым.*

(ii) Пусть вектор a *является пиковым, а* $v \in \mathbb{R}^n$ *чебышевским. Тогда* $\operatorname{sign}(\gamma(a,v)) = \operatorname{sign}(a_{\zeta(a)}v_{\zeta(a)})$ *и выполнено неравенство*

$$\frac{\delta(a)}{2\|v\|_\infty} \leqslant |\gamma(a,v)| \leqslant \frac{2\|a\|_\infty}{\|v\|_\infty}. \tag{2.6}$$

Доказательство. Пусть a является пиковым. Тогда из Следствия 1.25 получаем, что $a - \gamma(a,v)v$ не является пиковым для всех чебышевских векторов v и, следовательно, $\gamma(a,v) \neq 0$. Предположим, что a не является пиковым и пусть $i,j \in \{1,\dots,n\}$ являются различными парами индексов такими, что

$$|a_i| = |a_j| = \|a\|_\infty.$$

Пусть $v \in \mathbb{R}^n$ является произвольным чебышевским вектором таким, что $\operatorname{sign}(v_i a_i) = -\operatorname{sign}(v_j a_j)$. Тогда из Следствия 1.25 получаем, что $\gamma(a,v) = 0$. Таким образом, утверждение (i) доказано.

Докажем утверждение (ii). Поскольку a является пиковым вектором, получаем, что $\gamma(a,v) \neq 0$ и, следовательно

$$\|a - \gamma(a,v)v\|_\infty < \|a\|_\infty.$$

Таким образом,

$$|a_{\zeta(a)} - \gamma(a,v)v_{\zeta(a)}| < |a_{\zeta(a)}|$$

что возможно только если $\operatorname{sign}(\gamma(a,v)) = \operatorname{sign}(a_{\zeta(a)}v_{\zeta(a)})$. Остается доказать (2.6). Поскольку

$$\|a - \gamma(a,v)v\|_\infty \leqslant \|a\|_\infty,$$

получаем, что

$$|\gamma(a,v)|\|v\|_\infty - \|a\|_\infty \leqslant \|a\|_\infty,$$

откуда сразу следует второе неравенство в (2.6). Для того чтобы доказать оставшееся неравенство заметим, что из Следствия 1.25 получаем, что

$$|a_i - \gamma(a,v)v_i| \geqslant |a_{\zeta(a)} - \gamma(a,v)v_{\zeta(a)}|$$

для некоторого $i \neq \zeta(a)$. Таким образом,

$$|a_i| + |\gamma(a,v)||v_i| \geqslant |a_{\zeta(a)}| - |\gamma(a,v)||v_{\zeta(a)}|$$

и следовательно

$$|\gamma(a,v)|(|v_i| + |v_{\zeta(a)}|) \geqslant |a_{\zeta(a)}| - |a_i| \geqslant \delta(a).$$

□

Определение 2.16. *Будем говорить, что матрица $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы, если для любых чебышевских векторов $u \in \mathbb{R}^m$ и $v \in \mathbb{R}^n$, векторы $\varphi(A, v)$ и $\psi(A, u)$ также являются чебышевскими.*

Из последнего определения и Следствия 1.25 ясно, что если $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы и $v^{(0)} \in \mathbb{R}^n$ является чебышевским, то метод переменных направлений определен корректно и существует единственная пара последовательностей $\{u^{(k)}\}_{k\in\mathbb{N}}$ и $\{v^{(k)}\}_{k\in\mathbb{N}}$, полученная методом переменных направлений, для матрицы A и начальной точки $v^{(0)}$. Следующая лемма дает простую характеристику матриц, сохраняющих чебышевские системы, а также демонстрирует, что таких матриц достаточно много.

Лемма 2.17. *Матрица $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы тогда и только тогда, когда все строки и столбцы матрицы A являются пиковыми. Множество матриц, сохраняющих чебышевские системы, является открытым и плотным в $\mathbb{R}^{m\times n}$, а его дополнение в $\mathbb{R}^{m\times n}$ имеет лебегову меру нуль.*

Доказательство. Действительно, A сохраняет чебышевские системы тогда и только тогда, когда $\gamma(a^i, v)$ и $\gamma(a_j, u)$ не равны нулю для всех i,j и всех чебышевских векторов $v \in \mathbb{R}^n$ и $u \in \mathbb{R}^m$. Из Теоремы 2.15 (i) следует, что это имеет место тогда и только тогда, когда a^i и a_j являются пиковыми для всех $i = 1, \ldots, m$ и $j = 1, \ldots, n$. Отсюда ясно, что множество матриц, сохраняющих чебышевские системы, является открытым в $\mathbb{R}^{m\times n}$, а его дополнение в $\mathbb{R}^{m\times n}$ содержится в объединении конечного числа гиперплоскостей в $\mathbb{R}^{m\times n}$. Поэтому дополнение имеет лебегову меру нуль и пустую внутренность. □

Для того чтобы сформулировать следующую лемму, нам понадобятся дополнительные обозначения. Пусть $v \in \mathbb{R}^n$ является чебышевским вектором. Тогда будем называть $\|v\|_\infty / \min\limits_{i=1,\ldots,n} |v_i|$ *амплитудой* вектора v и обозначать ее через $\mathrm{am}(v)$.

Лемма 2.18. *Пусть матрица $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы и вектор $v^{(0)} \in \mathbb{R}^n$ является чебышевским. Пусть пара последовательностей*

$\{u^{(k)}\}_{k\in\mathbb{N}}$ и $\{v^{(k)}\}_{k\in\mathbb{N}}$ получена методом переменных направлений для матрицы A и начальной точки $v^{(0)}$. Пусть $\delta_r = \min\limits_{i=1,\dots,m} \delta(a^i)$ и $\delta_c = \min\limits_{j=1,\dots,n} \delta(a_j)$. Тогда

$$\|u^{(k)}\|_\infty \|v^{(k-1)}\|_\infty \leqslant 2\|A\|_C,\ \ \|u^{(k)}\|_\infty \|v^{(k)}\|_\infty \leqslant 2\|A\|_C,$$

$$\operatorname{am}(u^{(k)}) \leqslant 4\|A\|_C/\delta_r,\ \ \operatorname{am}(v^{(k)}) \leqslant 4\|A\|_C/\delta_c$$

для всех $k \in \mathbb{N}$.

Доказательство. Сразу следует из Теоремы 2.15 (ii) и Леммы 2.3 (i). □

2.6 Анализ поведения знаков для ранга 1

В данном разделе проводится анализ поведения знаков компонент векторов $u^{(k)}$ и $v^{(k)}$, полученных методом переменных направлений. А именно, доказывается, что знаки компонент полностью определяются приближаемой матрицей и начальным вектором и, более того, знаки стабилизируются при достаточно большом k (мы покажем, что при k большем $\min(m,n)$).

Пусть вектор $v \in \mathbb{R}^n$ является чебышевским. Обозначим через $\mathcal{S}(v)$ вектор с компонентами $\mathcal{S}(v)_i = \operatorname{sign}(v_i)$, $i = 1, \dots, n$.

Теорема 2.19. *Пусть матрица $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы. Если векторы $v_1, v_2 \in \mathbb{R}^n$ являются чебышевскими и $\mathcal{S}(v_1) = \mathcal{S}(v_2)$, то*

$$\mathcal{S}(\varphi(A, v_1)) = \mathcal{S}(\varphi(A, v_2)).$$

Аналогично, если векторы $u_1, u_2 \in \mathbb{R}^m$ являются чебышевскими и $\mathcal{S}(u_1) = \mathcal{S}(u_2)$, то

$$\mathcal{S}(\psi(A, u_1)) = \mathcal{S}(\psi(A, u_2)).$$

Доказательство. Пусть O является множеством чебышевских векторов $v \in \mathbb{R}^n$ таких, что $\mathcal{S}(v) = \mathcal{S}(v_1)$. Ясно, что множество O является выпуклым, а следовательно связным. Функция $s(v) = \mathcal{S}(\varphi(A,v))$ непрерывна на множестве чебышевских векторов в $\mathbb{R}^n$, поскольку $\varphi(A,v)$ непрерывна по v, $\mathcal{S}$ локально-постоянна и следовательно непрерывна. Поскольку образ s дискретен, получаем, что s постоянна на O. Таким образом, $s(v_1) = s(v_2)$, поскольку $v_1, v_2 \in O$. Второе утверждение может быть доказано аналогично. □

Из Теоремы 2.19 следует, что знаки компонент вектора на следующем шаге метода переменных направлений зависят только от знаков компонент вектора на предыдущем шаге. Таким образом, для пары начальных точек с одинаковыми знаками компонент, метод переменных направлений порождает последовательности векторов с одинаковыми знаками.

Проанализируем поведение знаков более детально. Пусть матрица $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы, вектор $v^{(0)} \in \mathbb{R}^n$ является чебышевским, и метод переменных направлений порождает пару последовательностей $\{u^{(k)}\}_{k\in\mathbb{N}}$ и $\{v^{(k)}\}_{k\in\mathbb{N}}$ для матрицы A и начальной точки $v^{(0)}$. Введем пару отображений $\mathcal{R} : \{-1,1\}^m \to \{-1,1\}^n$ и $\mathcal{T} : \{-1,1\}^n \to \{-1,1\}^m$, определенных по формулам

$$\mathcal{R}(p) = \mathcal{S}(\psi(A,p)),\ \ \mathcal{T}(q) = \mathcal{S}(\varphi(A,q)),$$

где $p \in \{-1,1\}^m$ и $q \in \{-1,1\}^n$. Наконец, пусть $\mathcal{V} = \mathcal{R}\circ\mathcal{T}$. Отображение $\mathcal{V}$ определяет как знаки вектора $v^{(k+1)} = \psi(A,\varphi(A,v^{(k)}))$ зависят от знаков вектора $v^{(k)}$ в процессе работы метода переменных направлений.

При помощи отображения $\mathcal{V}$ определим *граф перехода знаков* G_A матрицы A. Вершинами графа являются являются элементы множества $\{-1,1\}^n$, и между вершинами t_1 и t_2 есть ребро тогда и только тогда, когда $\mathcal{V}(t_1) = t_2$. Покажем, что граф G_A представляет собой набор изоморфных деревьев, каждое из которых содержит ровно одну вершину t такую, что $\mathcal{V}(t) = t$.

Пусть $\mathfrak{j}(i) = \zeta(a^i)$ и $\mathfrak{i}(j) = \zeta(a_j)$ (напомним, что через $\zeta(a)$ мы обозначаем позицию максимального по модулю элемента для пикового вектора a). Заметим, что по предположению на матрицу A выполнено $|a_{i,\mathfrak{j}(i)}| > 0$ и $|a_{\mathfrak{i}(j),j}| > 0$ для любых i,j. Поскольку $u^{(k+1)} = \varphi(A,v^{(k)})$, из Теоремы 2.15 (ii) следует, что

$$\operatorname{sign} u_i^{(k+1)} = \operatorname{sign} a_{i,\mathfrak{j}(i)} \operatorname{sign} v_{\mathfrak{j}(i)}^{(k)}.$$

Аналогично можно получить, что

$$\operatorname{sign} v_j^{(k+1)} = \operatorname{sign} a_{\mathfrak{i}(j),j} \operatorname{sign} u_{\mathfrak{i}(j)}^{(k+1)}.$$

Таким образом, из последних двух равенств мы имеем, что

$$\operatorname{sign} v_j^{(k+1)} = \operatorname{sign} a_{\mathfrak{i}(j),j} \cdot \operatorname{sign} a_{\mathfrak{i}(j),\mathfrak{j}(\mathfrak{i}(j))} \cdot \operatorname{sign} v_{\mathfrak{j}(\mathfrak{i}(j))}^{(k)}. \tag{2.7}$$

Равенство (2.7) выражает знаки вектора $v^{(k+1)}$ через знаки вектора $v^{(k)}$, то есть определяет отображение $\mathcal{V}$. Введем *граф зависимости знаков* G_A^{sd} матрицы A.

Множеством вершин графа G_A^{sd} является $\{1, \dots, n\}$, и в графе есть ребро из вершины k в вершину l тогда и только тогда, когда $k = \mathfrak{j}(\mathfrak{i}(l))$.

Графы G_A и G_A^{sd} являются *ориентированными*. Под *ациклическим* графом будем понимать граф, который не содержит ориентированных циклов за исключением петель. Вершину t графа G такую, что существует петля $t \to t$, будем называть *петлевой вершиной*. Наконец, *глубиной* графа G будем называть наибольшее p такое, что существует последовательность различных вершин $t_1, \dots, t_p$ таких, что существует ребро из вершины t_k в вершину t_{k+1} при всех $k = 1, \dots, p-1$.

Лемма 2.20. *Граф G_A^{sd} удовлетворяет следующим свойствам.*

(i) Для каждой вершины существует ровно одно ребро, входящее в нее.

(ii) В графе всегда существует по крайней мере одна петля. Если вершина j является петлевой, то $\operatorname{sign}(v_j^{(k)}) = \operatorname{sign}(v_j^{(l)})$ при всех k,$l \in \mathbb{N}$.

(iii) Граф G_A^{sd} является ациклическим.

(iv) Все вершины графа G_A^{sd} достижимы из петлевых вершин (то есть для любой вершины j графа G_A^{sd} существует последовательность $j_1, \dots, j_k$ такая, что $j_k = j$, j_1 является петлевой вершиной, и существует ребро из j_s в j_{s+1} для всех $s = 1, \dots, k-1$).

(v) Рассмотрим следующий процесс для матрицы $|A|$ (все элементы матрицы взяты по модулю). Возьмем произвольный столбец, найдем в нем максимальный элемент, найдем максимальный элемент в соответствующей строке, затем снова максимальный элемент в соответствующем столбце и так далее. Максимально возможное количество столбцов в описанном процессе равно глубине графа G_A^{sd}.

Доказательство. Утверждение (i) сразу следует из определения графа G_A^{sd}. Для того чтобы доказать (ii) заметим, что вершина j является петлевой тогда и только тогда, когда $\mathfrak{j}(\mathfrak{i}(j)) = j$. Таким образом, петлевые вершины в точности соответствуют индексам столбцов матрицы A, для которых максимальный по модулю элемент является также максимальным по модулю в своей строке. Нетрудно видеть, что такой элемент в матрице всегда существует, поэтому граф G_A^{sd} содержит по крайней мере одну петлю. Кроме того, если j является петлевой вершиной, то равенство (2.7) принимает вид

$$\operatorname{sign} v_j^{(k+1)} = \operatorname{sign} v_j^{(k)}.$$

Докажем (iii). Очевидно, что $\|a_j\|_\infty \leqslant \|a^{\mathfrak{i}(j)}\|_\infty$ и $\|a^i\|_\infty \leqslant \|a_{\mathfrak{j}(i)}\|_\infty$, откуда $\|a_j\|_\infty \leqslant \|a_{\mathfrak{j}(\mathfrak{i}(j))}\|_\infty$. Более того, согласно предположению на матрицу A, если $\mathfrak{j}(\mathfrak{i}(j)) \neq j$, то

$$\|a_j\|_\infty < \|a_{\mathfrak{j}(\mathfrak{i}(j))}\|_\infty.$$

Таким образом, ориентированный цикл в G_A^{sd} является петлей.

Наконец, (iv) сразу следует из (i) и (iii), а (v) ясно из доказательства утверждения (iii). □

Лемма 2.21. *Граф G_A удовлетворяет следующим свойствам.*

(i) Для каждой вершины в графе G_A существует ровно одно ребро, входящее в нее.

(ii) Любая последовательность вершин $t_1, t_2, \ldots$ такая, что существует ребро $t_k \to t_{k+1}$ при всех k, стабилизируется.

(iii) Граф G_A является ациклическим и его глубина не превосходит глубины графа G_A^{sd}.

Доказательство. Утверждение (i) сразу следует из определения G_A. Для того чтобы доказать (ii) не ограничивая общности предположим, что $\mathcal{S}(v^{(0)}) = t_1$. Заметим, что $\mathcal{S}(v^{(k)}) = t_k$ при всех $k \in \mathbb{N}$. Из равенства (2.7) следует, что $\mathcal{S}(v^{(k+1)})$ зависит только от компонент вектора $\mathcal{S}(v^{(k)})$ с индексами в множестве $\{\mathfrak{j}(\mathfrak{i}(j)) : j = 1, \ldots, n\}$. Таким образом, получаем, что $t_p = t_{p+1} = \ldots$, где p обозначает глубину графа G_A^{sd}. А именно, вектор $\mathcal{S}(v^{(k+p-1)})$ зависит только от компонент вектора $\mathcal{S}(v^{(k)})$ на множестве

$$F = \{(\mathfrak{j} \circ \mathfrak{i})^{p-1}(j) : j = 1, \ldots, n\},$$

что соответствует петлевым вершинами графа G_A^{sd} по Лемме 2.20 (v). Заметим, что из Леммы 2.20 (ii) следует, что компоненты векторов $\mathcal{S}(v^{(k)})$ на F не зависят от k, откуда получаем, что $t_k = t_p$ для всех $k \geqslant p$. Следовательно, (ii) доказано. Кроме того ясно, что глубина G_A не превосходит p. Ацикличность G_A следует из (i) и (ii). □

Из Леммы 2.21 следует, что для любой вершины $t \in \{-1, 1\}^n$ существует единственная последовательность $t_1, t_2, \ldots$ такая, что $t = t_1$ и для всех $k \in \mathbb{N}$ существует ребро $t_k \to t_{k+1}$. Поскольку последовательность стабилизируется, можно определить функцию $f(t)$ как вектор, который равен t_k для сколь угодно большого k. По определению для всех t существует путь в графе G_A из вершины

t в вершину $f(t)$, причем $f(t)$ является петлевой вершиной. Далее под *компонентной связности* ориентированного графа будем понимать компоненты слабой связности (т.е. мы соединяем вершины путями независимо от направления ребер).

Лемма 2.22. *Выполнены следующие утверждения.*

(i) Пара вершин t_1 и t_2 в графе G_A принадлежит одной компоненте связности тогда и только тогда, когда $f(t_1) = f(t_2)$.

(ii) Если t_1 и t_2 являются петлевыми вершинами в G_A и $(t_1)_l = (t_2)_l$ для всех петлевых вершин l в G_A^{sd}, то $t_1 = t_2$.

(iii) Пусть t_1 и t_2 являются петлевыми вершинами в G_A. Пусть l и j являются такими вершинами в G_A^{sd}, что l является петлевой вершиной, а j достижима из l (см. Лемму 2.20 (iii)). Тогда

$$(t_1)_l/(t_1)_j = (t_2)_l/(t_2)_j.$$

Доказательство. Для доказательства (i) предположим, что $f(t_1) = f(t_2)$. Тогда t_1 и t_2 принадлежат одной компоненте связности, поскольку t_1 и t_2 могут быть соединены путем с $f(t_1) = f(t_2)$. Обратное утверждение следует из следующего наблюдения: если существует ребро $t_1 \to t_2$ или $t_2 \to t_1$, то $f(t_1) = f(t_2)$.

Пусть t_1 и t_2 удовлетворяют условию (ii). Обозначим через $F_k \subset \{1, \dots, n\}$ множество вершин l графа G_A^{sd} таких, что существует направленный путь из петлевой вершины в l длиной не более k. Тогда F_0 состоит из всех петлевых вершин графа G_A^{sd}. Из Леммы 2.20 (iv) следует, что

$$\bigcup_{k \geqslant 0} F_k = \{1, \dots, n\}.$$

По предположению мы имеем, что $(t_1)_j = (t_2)_j$ для всех $j \in F_0$. Покажем, что если выполнено равенство $(t_1)_j = (t_2)_j$ при всех $j \in F_k$, то $(t_1)_j = (t_2)_j$ при всех $j \in F_{k+1}$. Действительно, если $j \in F_{k+1}$, то по определению существует $s \in F_k$ такое, что $\mathfrak{j}(\mathfrak{i}(j)) = s$. Поскольку t_1 является петлевой вершиной, то из (2.7) следует, что $(t_1)_j = q(t_1)_s$, где $q = \operatorname{sign} a_{\mathfrak{i}(j),j} \cdot \operatorname{sign} a_{\mathfrak{i}(j),s}$. Рассуждая аналогично получаем, что $(t_2)_j = q(t_2)_s$. Поскольку $s \in F_k$, получаем, что $(t_1)_s = (t_2)_s$ и, следовательно, $(t_1)_j = (t_2)_j$.

Наконец, докажем (iii). Пусть $j_1, \dots, j_k$ выбраны таким образом, что $j_1 = l$, $j_k = j$ и для всех $s = 1, \dots, k-1$ существует ребро из j_s в j_{s+1}. Ясно, что

$$(t_1)_l/(t_1)_{j_1} = (t_2)_l/(t_2)_{j_1}.$$

Докажем, что из равенства

$$(t_1)_l/(t_1)_{j_s} = (t_2)_l/(t_2)_{j_s}$$

следует, что

$$(t_1)_l/(t_1)_{j_{s+1}} = (t_2)_l/(t_2)_{j_{s+1}}$$

(по индукции отсюда следует (iii)). Действительно, из (2.7) следует, что

$$(t_1)_{j_{s+1}} = \operatorname{sign} a_{\mathfrak{i}(j_{s+1}),j_{s+1}} \cdot \operatorname{sign} a_{\mathfrak{i}(j_{s+1}),j_s} \cdot (t_1)_{j_s}.$$

Отсюда получаем, что

$$(t_1)_{j_s}/(t_1)_{j_{s+1}} = \operatorname{sign} a_{\mathfrak{i}(j_{s+1}),j_{s+1}} \cdot \operatorname{sign} a_{\mathfrak{i}(j_{s+1}),j_s}.$$

Очевидно, что для t_2 выполнено то же самое равенство:

$$(t_2)_{j_s}/(t_2)_{j_{s+1}} = \operatorname{sign} a_{\mathfrak{i}(j_{s+1}),j_{s+1}} \cdot \operatorname{sign} a_{\mathfrak{i}(j_{s+1}),j_s}.$$

Комбинируя равенства

$$(t_1)_l/(t_1)_{j_s} = (t_2)_l/(t_2)_{j_s} \quad \text{и} \quad (t_1)_{j_s}/(t_1)_{j_{s+1}} = (t_2)_{j_s}/(t_2)_{j_{s+1}}$$

получаем $(t_1)_l/(t_1)_{j_{s+1}} = (t_2)_l/(t_2)_{j_{s+1}}$. □

Перед тем как сформулировать заключительную теорему о структуре графа G_A введем вспомогательную функцию $d : \{-1,1\}^n \to \{-1,1\}^n$, определяемую равенством

$$d(t)_i = t_i/f(t)_i.$$

Таким образом, $d(t)_i$ равно 1, если t_i совпадает с $f(t)_i$ и -1 в противном случае. Заметим, что мы не интерпретируем значение $d(t)$ как вершину графа G_A. Из равенства (2.7), примененного к t и $f(t)$ (в виду того факта, что $\mathcal{V}(f(t)) = f(t)$), следует, что

$$d(\mathcal{V}(t))_j = d(t)_{\mathfrak{j}(\mathfrak{i}(j))} \tag{2.8}$$

для всех $j \in \{1,\dots,n\}$. Кроме того, из определения d следует, что $d(t)_j = 1$ для всех петлевых вершин j графа G_A^{sd}. Наконец $t \in \{-1,1\}^n$ является петлевой вершиной графа G_A тогда и только тогда, когда $t = f(t)$, что эквивалентно $d(t)_i = 1$ для всех i.

Теорема 2.23. *Пусть матрица $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы. Обозначим через F множество всех чисел $j \in \{1,\dots,n\}$ таких, что максимальное по модулю значение в столбце a_j также является максимальным по модулю значением в своей строке. Пусть $k = |F|$. Тогда выполнены следующие утверждения.*

(i) Пара вершин $t_1, t_2 \in \{-1,1\}^n$ графа G_A принадлежит одной и той же компоненте связности тогда и только тогда, когда $(t_1)_j = (t_2)_j$ при всех $j \in F$.

(ii) Все компоненты связности графа G_A имеют 2^{n-k} вершин (в графе G_A 2^k компонент связности). Кроме того, все компоненты связности являются деревьями и содержат ровно одну петлевую вершину.

(iii) Все компоненты связности графа G_A изоморфны. Более точно, пусть C_1 и C_2 являются компонентами связности. Тогда для любой вершины t_1 компоненты C_1 существует ровно одна вершина t_2 компоненты C_2 такая, что $d(t_2) = d(t_1)$. Более того, существует отображение g, которое отображает вершины C_1 в вершины C_2, удовлетворяет $d(g(t)) = d(t)$ при всех t из C_1 и осуществляет изоморфизм между C_1 и C_2.

(iv) Глубина графа G_A равна глубине произвольной компоненты связности G_A и совпадает с глубиной графа G_A^{sd} (которая была вычислена в утверждении (v) Леммы 2.20).

Доказательство. В виду Леммы 2.22, утверждение (i) эквивалентно следующему утверждению: $f(t_1) = f(t_2)$ тогда и только тогда, когда $(t_1)_j = (t_2)_j$ при всех $j \in F$. Предположим, что $f(t_1) = f(t_2)$. Из Леммы 2.20 (ii) следует, что $(t_1)_j = f(t_1)_j$ и $(t_2)_j = f(t_2)_j$ для всех $j \in F$, поскольку F является множеством всех петлевых вершин в графе G_A^{sd}. Таким образом, $(t_1)_j = (t_2)_j$ при всех $j \in F$. Предположим, что $(t_1)_j = (t_2)_j$ при всех $j \in F$. Рассуждая аналогично описанному выше получаем, что

$$f(t_1)_j = (t_1)_j = (t_2)_j = f(t_2)_j$$

при всех $j \in F$. Таким образом, из Леммы 2.22 (ii) следует, что $f(t_1) = f(t_2)$, поскольку $f(t_1)$ и $f(t_2)$ являются петлевыми вершинами графа G_A.

Количество элементов в компоненте связности равно количеству отображений из $\{1,\dots,n\} \setminus F$ в $\{-1,1\}$, то есть 2^{n-k}. Кроме того, из Леммы 2.21 (i) и (iii) следует, что все компоненты связности G_A являются деревьями. Наконец, легко

видеть, что каждая компонента связности содержит по крайней мере одну петлевую вершину (возьмем произвольную вершину t и заметим, что $f(t)$ является петлевой вершиной в той же компоненте связности). Если t_1 и t_2 являются петлевыми вершинами, принадлежащими одной компоненте связности, то

$$t_1 = f(t_1) = f(t_2) = t_2.$$

Таким образом, (ii) доказано.

Перед тем как доказывать (iii) заметим, что образ d содержится в множестве

$$D = \{t \in \{-1, 1\}^n : t_l = 1 \; \forall l \in F\},$$

которое содержит в точности 2^{n-k} элементов. По определению также ясно, что d является инъективным на каждой компоненте связности графа G_A, которая имеет такое же количество элементов. Следовательно d отображает каждую компоненту связности на D биективно. Тогда если C_1 и C_2 являются компонентами связности графа G_A, то существует единственное g, которое отображает вершины C_1 в вершины C_2 и удовлетворяет $d(g(t)) = d(t)$ для всех t из C_1. Кроме того ясно, что g является биективным. Следовательно остается показать, что g осуществляет изоморфизм графов, а именно, если существует ребро из t_1 в t_2, то существует ребро из $g(t_1)$ в $g(t_2)$ (заметим, что обратное не требуется доказывать, потому что C_1 и C_2 взаимозаменяемы). Заметим, что из (2.8) следует, что $d(\mathcal{V}(t_1)) = d(\mathcal{V}(t_2))$ при $d(t_1) = d(t_2)$. Наконец, предположим, что существует ребро из t_1 в t_2, где t_1 и t_2 являются вершинами в C_1. Тогда $\mathcal{V}(t_1) = t_2$ и, следовательно,

$$d(\mathcal{V}(g(t_1))) = d(\mathcal{V}(t_1)) = d(t_2) = d(g(t_2)).$$

Таким образом, $\mathcal{V}(g(t_1)) = g(t_2)$, откуда следует, что $g(t_1)$ и $g(t_2)$ соединены ребром в графе.

Поскольку все компоненты связности G_A имеют одинаковую глубину, остается доказать, что глубина графа G_A равна глубине графа G_A^{sd}. Пусть p обозначает глубину графа G_A^{sd}. В утверждении (iii) Леммы 2.21 было показано, что глубина графа G_A не превосходит p. Пусть $j_1, \ldots, j_p$ — множество различных элементов из $\{1, \ldots, n\}$ таких, что $\mathfrak{j}(\mathfrak{i}(j_k)) = j_{k-1}$ при всех $k = 2, \ldots, p$. То есть $j_1, \ldots, j_p$ является одним из самых длинных путей в G_A^{sd} (в этом случае j_1 является петлевой вершиной). Рассмотрим произвольное $t \in \{-1, 1\}^n$ такое, что $d(t)_{j_2} = -1$. Применяя равенство (2.8) k раз получаем, что $d(\mathcal{V}^k(t))_{j_{k+2}} = -1$ при $k = 0, 1, \ldots, p-2$. Следовательно $\mathcal{V}^k(t)$ не является петлевой вершиной при $k = 0,1, \ldots, p-2$. Тогда

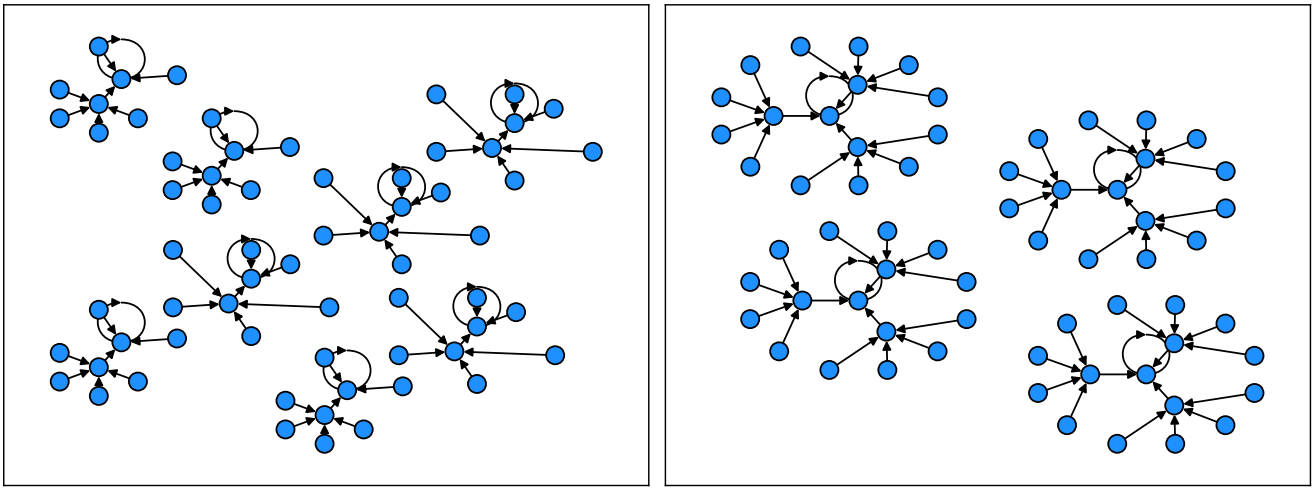


Рисунок 2.1 — Примеры графов перехода знаков для случайных матриц.

из ацикличности графа G_A следует, что величины $t, \mathcal{V}(t), \ldots, \mathcal{V}^{p-1}(t)$ различны. Таким образом, глубина графа G_A больше либо равна p. □

Примеры графов перехода знаков приведены на Рис. 2.1.

Замечание. *1. Компоненты связности множества матриц размера $m \times n$, сохраняющих чебышевские системы, легко описать. А именно, каждая компонента связности является выпуклой и матрицы $A, B \in \mathbb{R}^{m\times n}$, сохраняющие чебышевские системы, принадлежат одной компоненте тогда и только тогда, когда $\zeta(a^i) = \zeta(b^i)$ и*

$$\operatorname{sign}(a_{i,\zeta(a^i)}) = \operatorname{sign}(b_{i,\zeta(b^i)})$$

при всех $i = 1, \ldots, m$ и $\zeta(a_j) = \zeta(b_j)$ и

$$\operatorname{sign}(a_{\zeta(a_j),j}) = \operatorname{sign}(b_{\zeta(b_j),j})$$

при всех $j = 1, \ldots, n$. Ясно, что для матриц $A, B \in \mathbb{R}^{m\times n}$, сохраняющих чебышевские системы, которые принадлежат одной компоненте связности, графы G_A и G_B совпадают, поэтому поведение знаков в процессе метода переменных направлений будет для них одинаковым.

2. Компоненты связности матриц, сохраняющих чебышевские системы и содержащие матрицы ранга 1 особенно легко описать. А именно, если $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы, то компонента связности A содержит матрицу ранга 1 тогда и только тогда, когда $\zeta(a^{i_1}) = \zeta(a^{i_2})$ и $\zeta(a_{j_1}) = \zeta(a_{j_2})$ при всех $i_1, i_2 = 1, \ldots, m$ и $j_1, j_2 = 1, \ldots, n$. То есть для таких матриц графы G_A^{sd} и G_A имеют глубину 2, а G_A имеет две петлевые вершины. Более того, после первой итерации метода переменных направлений знаки векторов стабилизируются.

2.7 Построение оптимального приближения для ранга 1

Основным результатом данного раздела является следующее утверждение.

Теорема 2.24. *Пусть $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы и векторы $v,\tilde{v} \in \mathbb{R}^n$ являются чебышевскими. Пусть $\mathcal{S}(v) = \mathcal{S}(\tilde{v})$. Тогда $E(A,v) = E(A,\tilde{v})$.*

При помощи Теоремы 2.24 далее будет построен алгоритм, позволяющий строить оптимальные чебышевские приближения ранга 1.

Для доказательства Теоремы 2.24 нам потребуется следующий результат, который впервые был опубликован в [20]. Для полноты изложения мы приведем его с доказательством.

Теорема 2.25. *Пусть $A \in \mathbb{R}^{m\times n}$, векторы $v \in \mathbb{R}^n$, и $u \in \mathbb{R}^m$ являются чебышевскими. Пусть множество $\mathcal{A} = \{(i_1, j_1), (i_1, j_2), (i_2, j_2), \ldots, (i_k, j_k), (i_k, j_1)\}$ является двумерным альтернансом ранга 1 для тройки (A,u,v). Тогда если $\tilde{v} \in \mathbb{R}^n$ и $\tilde{u} \in \mathbb{R}^m$ и либо*

$$\operatorname{sign}(u_i) = \operatorname{sign}(\tilde{u}_i), \quad i \in \{i_1, i_2, \ldots, i_k\},$$

либо

$$\operatorname{sign}(v_j) = \operatorname{sign}(\tilde{v}_j), \quad j \in \{j_1, j_2, \ldots, j_k\},$$

то $\|A - \tilde{u}\tilde{v}^T\|_C \geqslant \|A - uv^T\|_C$.

Доказательство. Обозначим $g = \tilde{u} - u$ и $h = \tilde{v} - v$. Рассмотрим невязку $\tilde{R} = A - \tilde{u}\tilde{v}^T$.

$$\tilde{r}_{ij} = a_{ij} - \tilde{u}_i\tilde{v}^j = a_{ij} - (u_i + g_i)(v_j + h_j) = a_{ij} - u_iv_j - (u_ih_j + g_iv_j + g_ih_j).$$

Обозначим $R = A - uv^T$, $\beta_{ij} = \operatorname{sign}(r_{ij})(u_ih_j + g_iv_j + g_ih_j)$. В этих обозначениях

$$\tilde{r}_{ij} = \operatorname{sign}(r_{ij})|r_{ij}| - \operatorname{sign}(r_{ij})\beta_{ij}.$$

Тогда $|\tilde{r}_{ij}| = ||r_{ij}| - \beta_{ij}|$. Далее мы покажем, что при наличии двумерного альтернанса ранга 1 найдется такая точка $(i,j) \in \mathcal{A}$, что $\beta_{ij} \leqslant 0$, откуда следует утверждение теоремы.

Рассмотрим систему уравнений относительно неизвестных ζ и ξ.

$$\begin{cases} \operatorname{sign}(r_{i_tj_t})\tilde{u}_{i_t}\zeta_{j_t} + \operatorname{sign}(r_{i_tj_t})v_{j_t}\xi_{i_t} = \alpha_{2t-1} \\ \operatorname{sign}(r_{i_tj_{t+1}})\tilde{u}_{i_t}\zeta_{j_{t+1}} + \operatorname{sign}(r_{i_tj_{t+1}})v_{j_{t+1}}\xi_{i_t} = \alpha_{2t} \end{cases} \tag{2.9}$$

где t принимает значения от 1 до k. Таким образом, система состоит из $2k$ уравнений с $2k$ неизвестными, причем каждое уравнение соответствует элементу множества $\mathcal{A}$. Выберем $\zeta_{j_t} = h_{j_t}$ и $\xi_{i_t} = g_{i_t}$. Тогда первое уравнение системы принимает вид

$$\operatorname{sign}(r_{i_t j_t})\tilde{u}_{i_t} h_{j_t} + \operatorname{sign}(r_{i_t j_t}) v_{j_t} g_{i_t} = \alpha_{2t-1}.$$

Подставляя в это уравнение $\tilde{u}_{i_t} = u_{i_t} + g_{i_t}$, получаем

$$\operatorname{sign}(r_{i_t j_t})(u_{i_t} h_{j_t} + v_{j_t} g_{i_t} + g_{i_t} h_{j_t}) = \alpha_{2t-1},$$

где с левой стороны стоит в точности $\beta_{i_t j_t}$. Аналогично можно показать, что при выборе $\zeta_{j_t} = h_{j_t}$ и $\xi_{i_t} = g_{i_t}$, мы имеем $\alpha_{2t} = \beta_{i_t j_{t+1}}$, откуда следует, что образ матрицы системы (2.9) содержит всевозможные наборы β_{ij} при $(i,j) \in \mathcal{A}$. Далее мы покажем, что все векторы из образа матрицы системы (2.9) либо равны 0, либо имеют компоненты разных знаков, откуда следует, что найдется $(i,j) \in \mathcal{A}$ такой, что $\beta_{ij} \leqslant 0$.

Занумеруем переменные в следующем порядке: $\zeta_{i_1}, \xi_{i_1}, \zeta_{i_2}, \xi_{i_2}, \ldots, \zeta_{i_t}, \xi_{i_t}$ и рассмотрим матрицу системы (2.9):

$$M = \begin{bmatrix} \operatorname{sign}(r_{i_1 j_1})\tilde{u}_{i_1} & \operatorname{sign}(r_{i_1 j_1}) v_{j_1} & 0 & \ldots & 0 \\ 0 & \operatorname{sign}(r_{i_1 j_2}) v_{j_2} & \operatorname{sign}(r_{i_1 j_2})\tilde{u}_{i_1} & \ldots & 0 \\ 0 & 0 & \operatorname{sign}(r_{i_2 j_2})\tilde{u}_{i_2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \operatorname{sign}(r_{i_k j_k}) v_{j_k} \\ \operatorname{sign}(r_{i_k j_1})\tilde{u}_{i_k} & 0 & 0 & \ldots & \operatorname{sign}(r_{i_k j_1}) v_{j_1} \end{bmatrix}.$$

Покажем, что система является вырожденной. Разложим ее по первому столбцу. Тогда

$$\det M = (-1)^2 \operatorname{sign}(r_{i_1 j_1})\tilde{u}_{i_1} \det \begin{bmatrix} \operatorname{sign}(r_{i_1 j_2}) v_{j_2} & \operatorname{sign}(r_{i_1 j_2})\tilde{u}_{i_1} & \ldots & 0 \\ 0 & \operatorname{sign}(r_{i_2 j_2})\tilde{u}_{i_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \operatorname{sign}(r_{i_k j_1}) v_{j_1} \end{bmatrix} +$$

$$(-1)^{2k+1} \operatorname{sign}(r_{i_k j_1})\tilde{u}_{i_k} \det \begin{bmatrix} \operatorname{sign}(r_{i_1 j_1}) v_{j_1} & 0 & \ldots & 0 \\ \operatorname{sign}(r_{i_1 j_2}) v_{j_2} & \operatorname{sign}(r_{i_1 j_2})\tilde{u}_{i_1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \operatorname{sign}(r_{i_k j_k}) v_{j_k} \end{bmatrix},$$

откуда $\det M = 0$.

Пусть $\lambda_1, \lambda_2, \ldots, \lambda_{2k}$ — коэффициенты линейной зависимости строк матрицы M. Рассмотрим первую координату этой линейной комбинации.

$$\lambda_1 \operatorname{sign}(r_{i_1 j_1})\tilde{u}_{i_1} + \lambda_{2k} \operatorname{sign}(r_{i_k j_1})\tilde{u}_{i_k} = 0.$$

Домножим уравнение на v_{j_1}:

$$\lambda_1 \operatorname{sign}(r_{i_1 j_1})\tilde{u}_{i_1} v_{j_1} + \lambda_{2k} \operatorname{sign}(r_{i_k j_1})\tilde{u}_{i_k} v_{j_1} = 0. \tag{2.10}$$

Предположим, что $\operatorname{sign} \tilde{u}_{i_1} = \operatorname{sign} u_{i_1}$ и $\operatorname{sign} \tilde{u}_{i_k} = \operatorname{sign} u_{i_k}$. Тогда знаки множителей при λ_1 и λ_{2k} совпадают со знаками $r_{i_1 j_1} u_{i_1} v_{j_1}$ и $r_{i_k j_1} u_{i_k} v_{j_1}$ соответственно. Согласно определению двумерного альтернанса (см. Определение 2.5),

$$\operatorname{sign}(r_{i_1 j_1} u_{i_1} v_{j_1}) = -\operatorname{sign}(r_{i_k j_1} u_{i_k} v_{j_1}),$$

откуда следует, что (2.10) может быть выполнено только если $\lambda_1 = \lambda_{2k} = 0$ или $\lambda_1 \lambda_{2k} > 0$. Выписывая уравнения для всех координат линейной комбинации строк матрицы M, получаем аналогичные соотношения для пар $(\lambda_1, \lambda_2), (\lambda_2, \lambda_3)$ и т.д. Но поскольку все λ_j не могут быть равны нулю одновременно, получаем, что все $\lambda_j \neq 0$ и одного знака. Тогда в образе матрицы M все векторы либо имеют компоненты разных знаков, либо равны нулю, тем сам утверждение теоремы доказано. Аналогичное утверждение при $\operatorname{sign}(\tilde{v}_{j_t}) = \operatorname{sign}(v_{j_t})$ получается переходом к транспонированной матрице. □

Из доказанной теоремы следует, что если тройка (A, u, v) обладает двумерным альтернансом ранга 1, то точка (u, v) является локальным минимумом функционала $s(u, v) = \|A - uv^T\|_C$, причем в достаточно большой области.

Лемма 2.26. *Пусть матрица $A \in \mathbb{R}^{m \times n}$ сохраняет чебышевские системы, и вектор $v \in \mathbb{R}^n$ является чебышевским. Пусть последовательности $\{u^{(k)}\}_{k \in \mathbb{N}}$ и $\{v^{(k)}\}_{k \in \mathbb{N}}$ получены методом переменных направлений для матрицы A и начальной точки $v^{(0)} = v$. Тогда произвольная предельная точка w последовательности $w_k = v^{(k)} / \|v^{(k)}\|_\infty$ является чебышевским вектором.*

Доказательство. Пусть w является пределом подпоследовательности w_{l_k} последовательности w_k. Поскольку амплитуда чебышевского вектора не изменяется при умножении на ненулевую константу, из Леммы 2.18 следует, что $\operatorname{am}(w_k) \leqslant C$, где $C > 0$ зависит только от матрицы A. Очевидно, что сходящаяся последовательность чебышевских векторов с ограниченной амплитудой сходится либо к

нулевому вектору, либо к чебышевскому вектору. Поскольку $\|w_k\|_\infty = 1$, отсюда следует, что вектор w является чебышевским. □

Доказательство Теоремы 2.24. Пусть $v, \tilde{v} \in \mathbb{R}^n$ являются чебышевскими и $\mathcal{S}(v) = \mathcal{S}(\tilde{v})$. Пусть $\{u^{(k)}\}_{k\in\mathbb{N}}$ и $\{v^{(k)}\}_{k\in\mathbb{N}}$ (соответственно $\{\tilde{u}^{(k)}\}_{k\in\mathbb{N}}$ и $\{\tilde{v}^{(k)}\}_{k\in\mathbb{N}}$) получены методом переменных направлений для матрицы A и начальной точки $v^{(0)} = v$ (соответственно $\tilde{v}^{(0)} = \tilde{v}$). Пусть w и $\tilde{w}$ являются предельными точками последовательностей $w_k = v^{(k)}/\|v^{(k)}\|_\infty$ и $\tilde{w}_k = \tilde{v}^{(k)}/\|\tilde{v}^{(k)}\|_\infty$ соответственно. По Лемме 2.26 векторы w и $\tilde{w}$ являются чебышевскими, а по Лемме 2.12 имеем

$$E(A,v) = \|A - \varphi(A,w)w^T\|_C, \quad E(A,\tilde{v}) = \|A - \varphi(A,\tilde{w})\tilde{w}^T\|_C,$$

откуда по Лемме 2.13 тройки $(A,\varphi(A,w),w)$ и $(A,\varphi(A,\tilde{w}),\tilde{w})$ обладают двумерным альтернансом ранга 1. Заметим также, что $\mathcal{S}(w) = \mathcal{S}(\tilde{w})$. Действительно, по Теореме 2.19 имеем $\mathcal{S}(w_k) = \mathcal{S}(\tilde{w}_k)$ при всех $k \in \mathbb{N}$. Кроме того, из Леммы 2.21 (ii) следует, что $\mathcal{S}(w_k) = \mathcal{S}(w_{k+1})$ и $\mathcal{S}(\tilde{w}_k) = \mathcal{S}(\tilde{w}_{k+1})$ при достаточно больших k. Таким образом, $\mathcal{S}(w)$ совпадает с $\mathcal{S}(\tilde{w})$, который в свою очередь совпадает с $\mathcal{S}(w_k)$ при достаточно больших k. Наконец, из Теоремы 2.25 следует, что

$$\|A - \varphi(A,w)w^T\|_C \leqslant \|A - \varphi(A,\tilde{w})\tilde{w}^T\|_C$$

и

$$\|A - \varphi(A,w)w^T\|_C \geqslant \|A - \varphi(A,\tilde{w})\tilde{w}^T\|_C,$$

поскольку тройки $(A,\varphi(A,w),w)$ и $(A,\varphi(A,\tilde{w}),\tilde{w})$ обладают двумерным альтернансом ранга 1 и знаки w и $\tilde{w}$ совпадают. Тогда $E(A,v) = E(A,\tilde{v})$. □

Используя Теорему 2.24, докажем, что расстояние от матрицы A до множества матриц ранга 1 может быть вычислено за конечное число запусков метода переменных направлений.

Теорема 2.27. *Пусть $A \in \mathbb{R}^{m\times n}$ сохраняет чебышевские системы и G_A является графом перехода знаков для A. Пусть $L \subset \{-1,1\}^n$ является множеством петлевых вершин графа G_A. Тогда*

$$\inf\{\|A - uv^T\|_C : u \in \mathbb{R}^m, v \in \mathbb{R}^n\} = \min_{t\in L} E(A,t).$$

Доказательство. Пусть

$$d = \inf\{\|A - uv^T\|_C : u \in \mathbb{R}^m, v \in \mathbb{R}^n\}.$$

Из определения $E(A,t)$ ясно, что $E(A,t) \geqslant d$ для всех чебышевских $t \in \mathbb{R}^n$. Тогда $\min_{t\in L} E(A,t) \geqslant d$. Для доказательства обратного неравенства заметим, что

$$d = \inf\{\|A - uv^T\|_C : u \in \mathbb{R}^m, v \in \mathrm{R}^n, v \text{ является чебышевским}\},$$

поскольку чебышевские векторы плотны $\mathbb{R}^n$. Более того, если $v \in \mathbb{R}^n$ является чебышевским, то

$$E(A,v) \leqslant \|A - \varphi(A,v)v^T\|_C = \inf\{\|A - uv^T\|_C : u \in \mathbb{R}^m\}.$$

Следовательно

$$d \geqslant \inf\{E(A,v) : v \in \mathbb{R}^n \text{ является чебышевским}\}.$$

Из Теоремы 2.24 следует, что

$$\inf\{E(A,v) : v \in \mathbb{R}^n \text{ является чебышевским}\} = \min\{E(A,t) : t \in \{-1,1\}^n\}.$$

Наконец поскольку

$$E(A,v) = E(A,\psi(A,\varphi(A,v))) = E(A,\mathcal{S}(\psi(A,\varphi(A,v))))$$

для всех чебышевских $v \in \mathbb{R}^n$ и итерации отображения $S(\psi(A,\varphi(A,v)))$ в конце концов отображают все $t \in \{-1,1\}^n$ в L, по Лемме 2.21 (ii) получаем, что

$$\inf\{E(A,v) : v \in \mathbb{R}^n \text{ является чебышевским}\} = \min_{t\in L} E(A,t).$$

□

Замечание. *1. Формула для расстояния*

$$d = \inf\{\|A - uv^T\|_C : u \in \mathbb{R}^m, v \in \mathbb{R}^n\}$$

из Теоремы 2.27 может быть немного улучшена в виду того факта, что $E(A,v) = E(A,-v)$ *(см. Лемму 2.4 (i)). Поэтому в обозначениях Теоремы 2.27 величина* d *может быть найдена как* $\min_{t\in\tilde{L}} E(A,t)$*, где* $\tilde{L} \subset L$ *является таким подмножеством, что для всех* $t \in L$ *либо* $t \in \tilde{L}$*, либо* $-t \in \tilde{L}$*. Таким образом, количество запусков метода переменных направлений может быть уменьшено в два раза, поскольку* $L = -L$ *по Лемме 2.3 (ii).*

2. Для матриц $A \in \mathbb{R}^{m\times n}$, сохраняющих чебышевские системы, которые принадлежат компоненте связности в множестве матриц, сохраняющих чебышевские системы, содержащей матрицу ранга 1, результат может быть значительно улучшен. А именно,

$$\inf\{\|A - uv^T\|_C : u \in \mathbb{R}^m, v \in \mathbb{R}^n\} = E(A,v)$$

для произвольного чебышевского вектора v. Действительно, равенство следует из Теоремы 2.27 и замечания в Разделе 2.6, поскольку в этом случае в графе G_A существует ровно два петлевые вершины t_1 и t_2, причем $t_1 = -t_2$.

2.8 Численные эксперименты

В данном разделе численно исследуется эффективность и свойства предложенного метода переменных направлений. Реализация алгоритма доступна онлайн[1]. В данном разделе результаты работы метода переменных направлений также сравниваются с *методом переменных проекций* [34], альтернативным методом для решения задач малорангового приближения матриц в чебышевской норме. Метод переменных проекций начинает свои итерации со случайной матрицы ранга r и попеременно проектирует матрицу на ε–шар с центром в приближаемой матрице и на множество матриц малого ранга при помощи SVD. Таким образом метод пытается найти общую точку двух упомянутых множеств. Для нахождения величины ε используется бинарный поиск. Для сравнения использовались авторская реализация и параметры метода.

2.8.1 Двумерный альтернанс ранга r

Для иллюстрации результатов Теоремы 2.7 метод переменных направлений был применен к черно-белой картинке размера 64×64. Черно-белая картинка была представлена как матрица, элементами которой являются вещественные числа от 0 до 1. На Рис. 2.2а изображена исходная картинка. Пусть U и V являются

[1] `https://github.com/stanis-morozov/cheburaxa`

а) Исходное изображение б) Приближение ранга 8 в) Альтернанс

Рисунок 2.2 — Пример малорангового приближения черно-белого изображения. Левая картинка содержит исходное изображение размера 64×64, средняя — аппроксимацию ранга 8, правая соответствует двумерному альтернансу ранга 8. Синие пиксели соответствуют позициям, где максимальное по модулю значение в невязке не достигается, желтые где достигаются с положительным знаком и красные с отрицательным.

матрицами, построенными методом переменных направлений для ранга 8. На Рис. 2.2б изображена матрица UV^T, соответствующая полученной аппроксимации. Напомним, что в Разделе 2.4 были введены обозначения $G = A - UV^T$ и $S(A, U, V) = \{(i, j) : |g_{ij}| = \|G\|_C\}$. На Рис. 2.2в изображено множество $S(A, U, V)$, а именно, пиксель в позиции (i, j) является желтым, если $g_{ij} = \|G\|_C$, красным, если $g_{ij} = -\|G\|_C$ и синим, если $(i, j) \notin S(A, U, V)$. Множество $S(A, U, V)$ соответствует двумерному альтернансу ранга 8. Заметим, что любая строка и любой столбец матрицы на Рис. 2.2в либо не содержат ни одного элемента из множества $S(A, U, V)$, либо содержат по крайней мере 9 элементов. Знаки в строках и столбцах не чередуются поскольку на картинке изображены знаки только g_{ij} без определителей (см. Определение 2.5). Заметим, однако, что некоторые строки и столбцы содержат более 9 элементов из множества $S(A, U, V)$, поэтому мы не можем выбрать матрицы $\mathcal{U}$ и $\mathcal{V}$ из Определения 2.5 таким образом, чтобы обеспечить чередование знаков во всех строках и столбцах. Однако соответствующие матрицы $\mathcal{U}$ и $\mathcal{V}$ могут быть выбраны для любого подмножества элементов в строке или столбце размера 9.

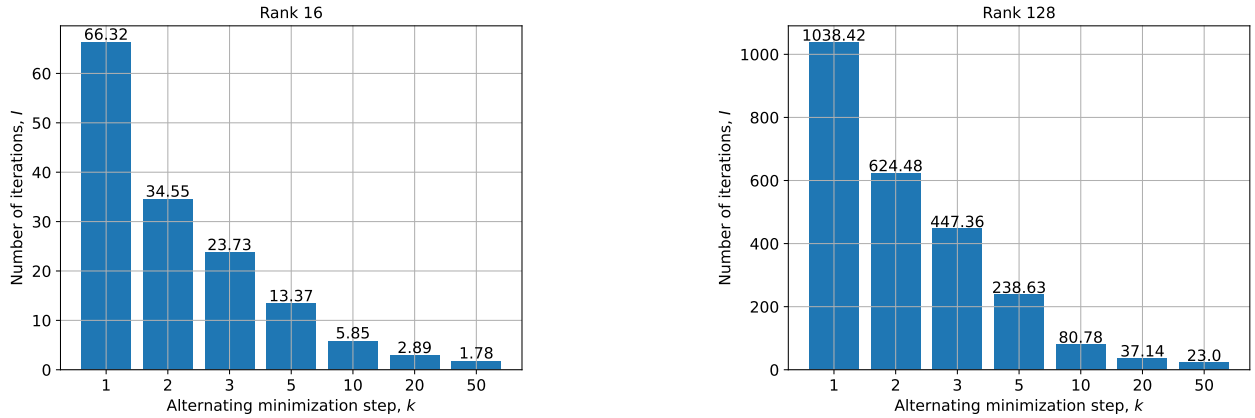


Рисунок 2.3 — Среднее число итераций для матрицы размера 8192 на различных шагах метода переменных направлений.

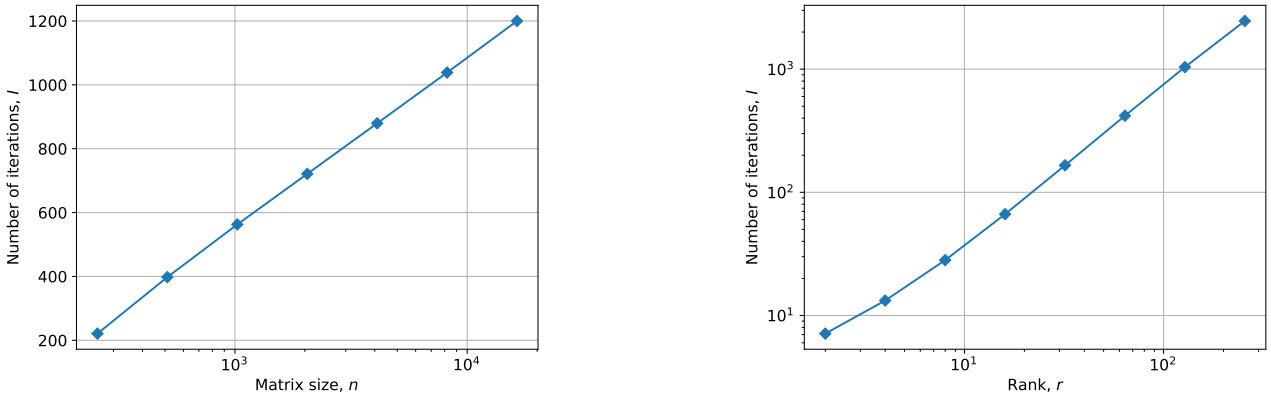


Рисунок 2.4 — Среднее число итераций на первом шаге метода переменных направлений. Левый график соответствует рангу 128 и различным размерам матрицы, а правый размеру матрицы 8192 и различным рангам. Число итераций логарифмически зависит от размера матрицы и степенным образом от ранга.

2.8.2 Сложность метода переменных направлений

Основной составляющей метода переменных направлений является алгоритм решения задачи наилучшего равномерного приближения (см. Раздел 2.3 с описанием метода переменных направлений и его сведением к задаче наилучшего равномерного приближения)

$$\|a - Vu\| \to \min_{u \in \mathbb{R}^r}, \tag{2.11}$$

где $V \in \mathbb{R}^{n \times r}$ является чебышевской матрицей и $a \in \mathbb{R}^n$. Алгоритм решения задачи (2.11) состоит в поиске характеристического множества задачи (см. Раздел 1.4)

и требует начального множества для своей работы (см. Раздел 1.10 с описанием алгоритма). Поскольку на практике метод переменных направлений сходится, начиная с некоторой итерации матрица V и вектор a в задаче (2.11) меняются не сильно. В этом случае на практике характеристические множества также начинают меняться слабо и хорошая инициализация метода решения задачи (2.11) может существенно ускорить его работу. В нашей реализации в качестве инициализации каждый раз использовалось характеристическое множество для соответствующего столбца или строки на предыдущем шаге метода переменных направлений.

Сложность алгоритма решения (2.11), описанного в Разделе 1.10, составляет $O(r^3 + Inr)$, где I — число итераций метода. В Теореме 1.30 показано, что скорость сходимости метода является геометрической, однако на практике это не позволяет оценить число итераций. Для того чтобы показать какое количество итераций совершает метод был проведен следующий эксперимент. Была сгенерирована матрица из стандартного нормального распределения размера $n \times n$ и для нее был запущен метод переменных направлений для построения малорангового приближения ранга r. На каждой итерации задача

$$\|A - UV^T\|_C \to \min_{U \in \mathbb{R}^{n \times r}}$$

(или соответствующая задача для матрицы V) разбивалась на n задач наилучшего равномерного приближения вида (2.11) для строк матрицы U и число итераций, требуемых для решения (2.11) усреднялось по всем строкам. Кроме того, результаты были усреднены по 20 случайным начальным матрицам. На Рис. 2.3 изображено среднее число итераций I на k–ом шаге метода переменных направлений для различных значений k и $n = 8192$. Нетрудно видеть, что количество итераций быстро падает с ростом номера шага метода переменных направлений. На Рис. 2.4 изображено среднее число итераций итераций на первом шаге метода переменных направлений для различных значений n и r соответственно. Легко видеть, что число итераций логарифмически зависит от размера n и степенным образом от ранга r. Мы предполагаем, что число итераций на первом шаге метода переменных направлений составляет $O(r^{1.5} \log n)$.

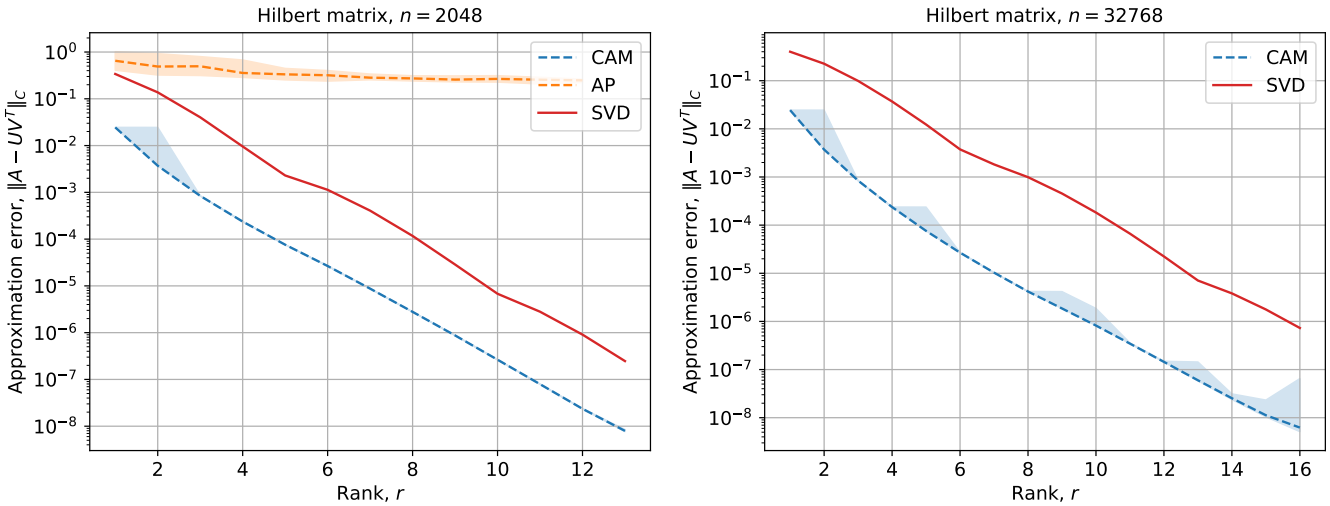


Рисунок 2.5 — Ошибка малоранговой аппроксимации для матрицы Гильбера при помощи метода переменных направлений (CAM), переменных проекций (AP) и SVD. На левом графике приведены результаты для матрицы размера $n = 2{,}048$, а на правом $n = 32{,}768$. Итерационная процедура была запущена из 20 случайных начальных точек. Закрашенная область соответствует минимальным и максимальным значениям для различных инициализаций, а пунктирная кривая медианным значениям.

2.8.3 Матрица Гильберта

Известно, что сингулярные числа матрицы $H = \left[\dfrac{1}{i+j}\right]_{i,j=1}^{n}$ убывают экспоненциально. Это означает, что она может быть хорошо приближена матрицами малого ранга в унитарно-инвариантных нормах при помощи SVD. На Рис. 2.5 приведены графики точности приближения для матриц размера $n = 2{,}048$ и $n = 32{,}768$ и различных рангов. В данном эксперименте сравниваются результаты работы метода переменных направлений (CAM, Chebyshev Alternating Minimization), SVD и метода переменных проекций (AP, Alternating Projections). Для методов переменных направлений и переменных проекций итерационные процедуры были запущены из 20 случайных точек. Для исследования устойчивости методов к выбору начальной точки на графиках приведены минимальные и максимальные значения ошибки среди 20 начальных точек, а область между ними закрашена. Пунктирная линия соответствует медианным значениям. Из графиков можно видеть, что метод переменных направлений существенно превосходит остальные методы по точности приближения и устойчивости. Дан-

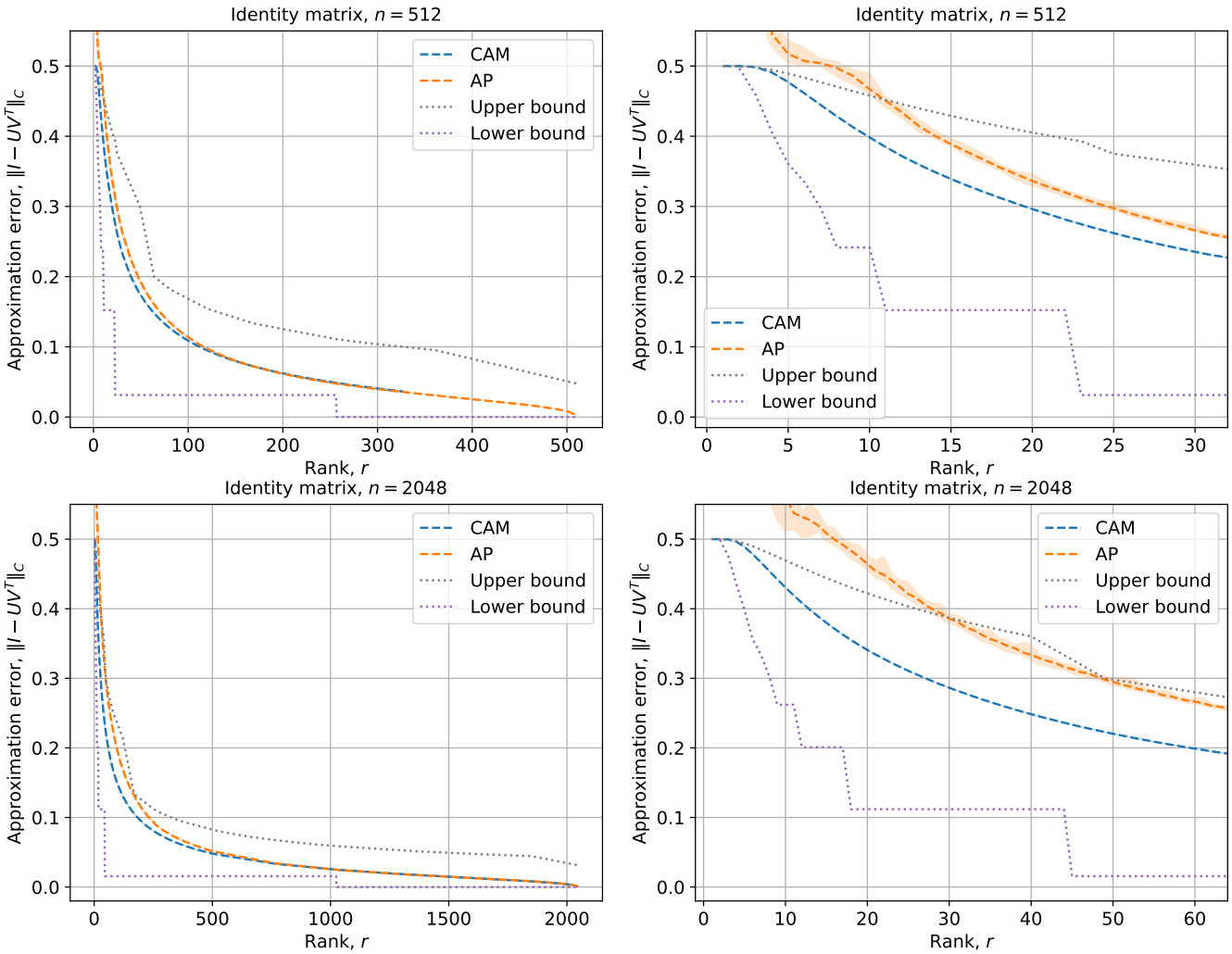


Рисунок 2.6 — Ошибка малоранговой аппроксимации единичной матрицы при помощи метода переменных направлений (CAM) и метода переменных проекций (AP). Верхние графики соответствуют матрице размера $n = 512$, а нижние $n = 2{,}048$. Правые графики содержат результаты для малых рангов в более крупном масштабе. Итерационные процедуры запущены из 20 случайных начальных точек. Закрашенные области соответствуют минимальным и максимальным значениям, а пунктирные кривые медианным значениям. Графики также содержат известные в литературе верхние и нижние оценки.

ный эксперимент демонстрирует, что даже для матриц с быстрым убыванием сингулярных чисел могут существовать приближения в чебышевской норме, значительно превосходящие по точности аппроксимации, полученные с помощью SVD. Для матрицы размера $n = 32{,}768$ не приведены результаты работы метода переменных проекций, поскольку в этом случае метод работает неприемлемо долго.

Точность, ε	128	256	512	1,024	2,048	4,096	8,192	16,384
0.45	6	6	7	8	9	10	11	12
0.4	8	9	10	12	13	15	17	18
0.25	17	22	27	33	40	47	54	62
0.1	60	84	112	145	184	228	278	333

Таблица 1 — Минимальный ранг, требуемый методу переменных направлений для достижения точности ε для единичной матрицы различных размеров.

2.8.4 Единичная матрица

В данном разделе приведены эксперименты по приближению единичной матрицы в чебышевской норме. На Рис. 2.6 приведены результаты для матриц размера $n = 512$ и $n = 2{,}048$. Как и в предыдущем эксперименте, метод был запущен из 20 случайных начальных точек и на графике приведены максимальные, минимальные и медианные значения. Также на графике приведены известные в литературе верхние и нижние оценки на чебышевские приближения единичной матрицы (см. [7; 8; 35] для верхних оценок и [9; 10] для нижних). Легко видеть, результаты работы метода переменных направлений лежат в точности между верхними и нижними оценками для всех рангов. Кроме того, результаты работы метода переменных направлений превосходят результаты для метода переменных проекций, особенно для маленьких рангов.

В Таблице 1 приведены минимальные значения рангов, требуемые для достижения точности ε при помощи метода переменных направлений при $\varepsilon \in \{0.45, 0.4, 0.25, 0.1\}$ и различных размеров единичной матрицы. Эксперимент показывает как ведет себя точность приближения единичной матрицы для различных размеров и демонстрирует масштабируемость алгоритма.

2.8.5 Функционально порожденные матрицы

В [5] и [6] показано, что некоторые классы функционально порожденных матриц могут быть хорошо приближены матрицами малого ранга в чебышевской норме. В частности, было показано, что если $\{x_j\}_{j=1}^n$ равномерно распределены

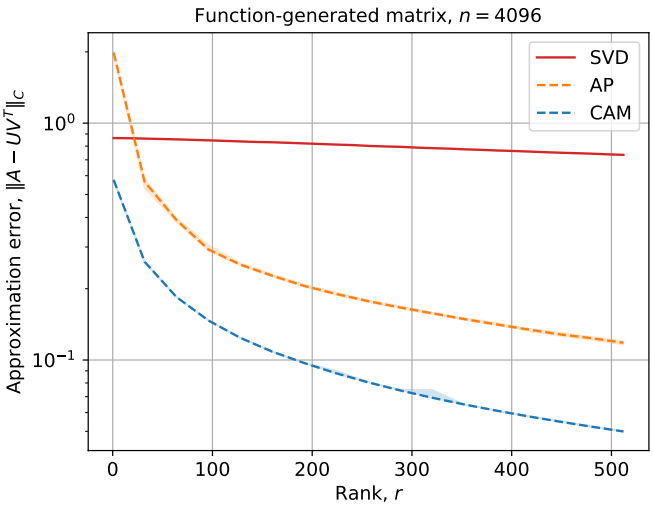


Рисунок 2.7 — Ошибка малоранговой аппроксимации для функционально порожденных матриц при помощи метода переменных направлений (CAM), метода переменных проекций (AP) и SVD. Размер матрицы $n = 4{,}096$. Итерационные процедуры были запущены из 20 случайных начальных точек. Закрашенная область соответствует минимальным и максимальным значениям, а пунктирная кривая медианным значениям.

на d–мерной сфере, то матрица $A \in \mathbb{R}^{n \times n}$, где

$$a_{ij} = \exp\left(-\|x_i - x_j\|_2^2\right), \tag{2.12}$$

допускает малоранговое приближение в чебышевской норме. В данном эксперименте было сгенерировано $n = 4{,}096$ случайных точек из равномерного распределения на d–мерной сфере, где $d = 8{,}192$, и вычислена матрица A по формулам (2.12). На Рис. 2.7 приведены результаты работы метода переменных направлений (CAM), SVD и метода переменных проекций (AP) для построения малоранговых приближений матрицы A. Для методов переменных направлений и переменных проекций итерации снова были запущены из 20 случайных начальных точек и на графиках изображены максимальные, минимальные и медианные значения. Матрица A была использована одна и та же для всех методов. На Рис. 2.7 можно видеть, что методом переменных направлений снова превосходит остальные методы по точности приближения и устойчивости.

Рисунок 2.8 — Аппроксимация черно-белых изображений при помощи метода переменных направлений и SVD.

2.8.6 Черно-белые изображения

Для визуализации разницы между малоранговыми приближениями матриц в чебышевской и фробениусовой нормах, был проведен эксперимент по аппроксимациям черно-белых изображений. Черно-белые изображения могут быть представлены как матрицы $T \in \mathbb{R}^{w \times h}$, где w и h соответствуют геометрическим размерам картинки. Значения матриц являются вещественными числами от 0 до 1. В данном эксперименте были использованы картинки с фиксированными

геометрическими размерами 512×512. На Рис. 2.8 приведены результаты аппроксимации картинок с различными рангами при помощи метода переменных направлений и SVD. Можно заметить, что приближения при помощи SVD являются более размытыми, в то время как аппроксимации, полученные с помощью метода переменных направлений, более резкие, но имеют дрожащую структуру. Дрожащая структура может быть объяснена наличием двумерного альтернанса (см. Раздел 2.4).

Глава 3. Построение малоранговых приближений тензоров в чебышевской норме

3.1 Постановка задачи

В данной главе рассматривается задача построения малоранговых приближений тензоров в каноническом формате. Пусть $T \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. Каноническим тензорным разложением тензора T называется его представление в виде

$$T = \sum_{t=1}^{r} u_t^{(1)} \otimes \cdots \otimes u_t^{(d)}, \tag{3.1}$$

где $u_t^{(j)} \in \mathbb{R}^{n_j}$ и $\otimes$ обозначает тензорное произведение. Каждое слагаемое вида $u^{(1)} \otimes \cdots \otimes u^{(d)}$ называется *тензором ранга 1*. Когда r минимально в (3.1), число r называется *каноническим тензорным рангом* тензора T.

Каноническое тензорное разложение, по сравнению с другими малопараметрическими тензорными форматами, обычно требует меньше параметров для достижения заданной точности аппроксимации [36] и естественным образом возникает во многих приложениях, например, для быстрого умножения матриц [37], в теории обработки сигналов [38] и машинном обучении [39]. Однако задача построения канонического тензорного разложения является NP–полной [40], что делает сложным построение приближений в этом формате. Обычно для построения канонического разложения используют итерационные процедуры, такие как метод переменных наименьших квадратов [33]. На сегодняшний день большинство алгоритмов строят приближения тензоров в норме Фробениуса, однако в некоторых приближениях требуется аппроксимировать тензоры поэлементно, то есть таким образом, что ошибка приближения для каждого элемента ограничена и мала.

В данной главе решается задача построения малорангового приближения тензоров в каноническом формате в чебышевской норме. А именно, пусть $T \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. Требуется решить задачу

$$\left\| T - \sum_{t=1}^{r} u_t^{(1)} \otimes \cdots \otimes u_t^{(d)} \right\|_C \to \min_{u_{(1)},\ldots,u_{(d)}}, \tag{3.2}$$

где чебышевская норма тензора $X \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ определяется как

$$\|X\|_C = \max_{i_1,\ldots,i_d} |x_{i_1,\ldots,i_d}|.$$

Для решения задачи (3.2) предлагается *метод переменных направлений*, который, также как алгоритм ALS, фиксирует все факторы разложения, кроме одного и находит значение незафиксированного фактора, минимизирующее чебышевскую норму ошибки. Все представленные в данной главе результаты сформулированы для трехмерных тензоров для простоты изложения, однако они могут быть естественным образом перенесены на случай d–мерных тензоров.

3.2 Метод переменных направлений

В данном разделе приводится описание метода переменных направлений для решения задачи (3.2) и формулируются его базовые свойства. Пусть $T \in \mathbb{R}^{m \times n \times k}$. Здесь и далее будем считать, что размеры m, n и k строго больше 1. Рассмотрим задачу построения малорангового поэлементного приближения тензора с каноническим рангом r, а именно,

$$T \approx \sum_{t=1}^{r} u_t \otimes v_t \otimes w_t. \tag{3.3}$$

Обозначим через $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ и $W \in \mathbb{R}^{k \times r}$ матрицы, составленные из векторов u_t, v_t и w_t соответственно. Будем сокращенно записывать (3.3) как

$$T \approx U \otimes V \otimes W.$$

Пусть матрицы U и V известны и построим оптимальную матрицу W, минимизирующую чебышевскую норму ошибки:

$$W = \arg\min_{X \in \mathbb{R}^{k \times r}} \|T - U \otimes V \otimes X\|_C. \tag{3.4}$$

Уравнение (3.4) может быть записано как

$$W = \arg\min_{X \in \mathbb{R}^{k \times r}} \|(U \odot V)X^T - T^{(1,2)}\|_C, \tag{3.5}$$

где через $U \odot V = \begin{bmatrix} u_1 \otimes v_1 & \dots & u_r \otimes v_r \end{bmatrix} \in \mathbb{R}^{mn \times r}$ обозначено произведение Хатри-Рао матриц U и V, а матрица $T^{(1,2)} \in \mathbb{R}^{mn \times k}$ получена из тензора T склеиванием первых двух размерностей. Обозначим через $T[:,:,l] \in \mathbb{R}^{m \times n}$ l–ую срезку тензора T вдоль третьей координаты. Нетрудно заметить, что задача (3.5) может быть разбита на набор независимых подзадач

$$w^l = \arg\min_{x \in \mathbb{R}^r} \|(U \odot V)x - \operatorname{vec}(T[:,:,l])\|_\infty, \quad l = 1, 2, \dots, k, \tag{3.6}$$

где через w^l обозначена l–ая строка матрицы W, а $\operatorname{vec}(\cdot)$ обозначает операцию векторизации матрицы. Из Теоремы 1.3, Теоремы 1.5 и Теоремы 1.7 следует, что если матрица $U \odot V$ является чебышевской, то решение задачи (3.6) существует, единственно и непрерывно зависит от матрицы $U \odot V$ и правой части $\operatorname{vec}(T[:,:,l])$. Таким образом, если матрица $U \odot V$ является чебышевской, можно определить функцию $\chi : \mathbb{R}^{m \times n \times k} \times \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \to \mathbb{R}^{k \times r}$ такую, что l–ая строка $\chi(T, U, V)$ определяется как

$$\chi(T, U, V)^l = \arg\min_{x \in \mathbb{R}^r} \|(U \odot V)x - \operatorname{vec}(T[:,:,l])\|_\infty, \quad l = 1, 2, \dots, k.$$

Функция χ определяет оптимальное значение матрицы W при известных U и V. Аналогично обозначим через $T[i,:,:] \in \mathbb{R}^{n \times k}$ i–ую срезку вдоль первой координаты, а через $T[:,j,:] \in \mathbb{R}^{m \times k}$ j–ую срезку вдоль второй координаты. Если матрица $U \odot W$ является чебышевской, можно определить функцию $\psi : \mathbb{R}^{m \times n \times k} \times \mathbb{R}^{m \times r} \times \mathbb{R}^{k \times r} \to \mathbb{R}^{n \times r}$ такую, что j–ая строка $\psi(T, U, W)$ определяется как

$$\psi(T, U, W)^j = \arg\min_{x \in \mathbb{R}^r} \|(U \odot W)x - \operatorname{vec}(T[:,j,:])\|_\infty, \quad j = 1, 2, \dots, n.$$

Наконец, если матрица $V \odot W$ является чебышевской, можно определить функцию $\varphi : \mathbb{R}^{m \times n \times k} \times \mathbb{R}^{n \times r} \times \mathbb{R}^{k \times r} \to \mathbb{R}^{m \times r}$ такую, что i–ая строка $\varphi(T, V, W)$ определяется как

$$\varphi(T, V, W)^i = \arg\min_{x \in \mathbb{R}^r} \|(V \odot W)x - \operatorname{vec}(T[i,:,:])\|_\infty, \quad i = 1, 2, \dots, m.$$

Заметим, что отображения φ, ψ и χ непрерывны на множестве точек, где $V \odot W$, $U \odot W$ и $U \odot V$ соответственно являются чебышевскими.

Определение 3.1. *Пусть $T \in \mathbb{R}^{m \times n \times k}$. Будем говорить, что тройка последовательностей чебышевских матриц $\{U^{(t)} \in \mathbb{R}^{m \times r}\}_{t \in \mathbb{N}}$, $\{V^{(t)} \in \mathbb{R}^{n \times r}\}_{t \in \mathbb{N}}$ и*

$\{W^{(t)} \in \mathbb{R}^{k\times r}\}_{t\in\mathbb{N}}$ получена методом переменных направлений для тензора T и пары начальных точек $(V^{(0)}, W^{(0)})$, где $V^{(0)} \in \mathbb{R}^{n\times r}$ и $W^{(0)} \in \mathbb{R}^{k\times r}$ являются чебышевскими матрицами, если

$$\begin{cases} U^{(t)} = \varphi(T, V^{(t-1)}, W^{(t-1)}), \\ V^{(t)} = \psi(T, U^{(t)}, W^{(t-1)}), \\ W^{(t)} = \chi(T, U^{(t)}, V^{(t)}) \end{cases}$$

при всех $t \in \mathbb{N}$.

Заметим, что если матрица $U \odot V$ является чебышевской, то это не значит, что $\chi(T, U, V)$ также является чебышевской. В Разделе 3.3 будет показано, что при построении приближений ранга 1 для почти всех тензоров, если U и V являются чебышевскими, то $\chi(T, U, V)$ также является чебышевской. Однако это не верно при аппроксимациях произвольного ранга. Более того, мы предполагаем, что при $r \geqslant 2$ для почти всех тензоров T существуют чебышевские матрицы $U \in \mathbb{R}^{m\times r}$ и $V \in \mathbb{R}^{n\times r}$ такие, что $\chi(T, U, V)$ не является чебышевской. Однако проведенные численные эксперименты показывают, что такие ситуации редки на практике. Тем не менее, при проведении теоретических исследований нам будет требоваться явно предполагать, что получаемые в результате работы метода переменных направлений матрицы являются чебышевскими.

Сформулируем базовые свойства метода переменных направлений для тензоров.

Лемма 3.2. *Пусть $T \in \mathbb{R}^{m\times n\times k}$ и матрицы $V^{(0)} \in \mathbb{R}^{n\times r}$ и $W^{(0)} \in \mathbb{R}^{k}$ являются чебышевскими. Пусть тройка последовательностей $\{U^{(t)} \in \mathbb{R}^{m\times r}\}_{t\in\mathbb{N}}$, $\{V^{(t)} \in \mathbb{R}^{n\times r}\}_{t\in\mathbb{N}}$ и $\{W^{(t)} \in \mathbb{R}^{k\times r}\}_{t\in\mathbb{N}}$ порождена методом переменных направлений для тензора T и пары начальных точек $(V^{(0)}, W^{(0)})$. Тогда выполнены следующие утверждения.*

(i) Выполнены неравенства

$$\|T - U^{(t)} \otimes V^{(t-1)} \otimes W^{(t-1)}\|_C \geqslant \|T - U^{(t)} \otimes V^{(t)} \otimes W^{(t-1)}\|_C \geqslant \|T - U^{(t)} \otimes V^{(t)} \otimes W^{(t)}\|_C \geqslant \|T - U^{(t+1)} \otimes V^{(t)} \otimes W^{(t)}\|_C$$

при всех $t \in \mathbb{N}$.

(ii) Если тройка последовательностей $\{\tilde{U}^{(t)}\}_{k\in\mathbb{N}}$, $\{\tilde{V}^{(t)}\}_{k\in\mathbb{N}}$ *и* $\{\tilde{W}^{(t)}\}_{k\in\mathbb{N}}$ *получена методом переменных направлений для тензора* T *и пары начальных точек* $(\alpha V^{(0)}, \beta W^{(0)})$*, где* $\alpha, \beta \neq 0$*, то* $\tilde{U}^{(t)} = 1/(\alpha\beta)\ U^{(t)}$*,* $\tilde{V}^{(t)} = \alpha V^{(t)}$*,* $\tilde{W}^{(t)} = \beta W^{(t)}$*.*

Доказательство. По построению φ,

$$\inf_{U\in\mathbb{R}^{m\times r}} \|T - U\otimes V^{(t)}\otimes W^{(t)}\|_C = \|T - \varphi(T,V^{(t)},W^{(t)})\otimes V^{(t)}\otimes W^{(t)}\|_C,$$

откуда имеем

$$\|T - U^{(t)}\otimes V^{(t)}\otimes W^{(t)}\|_C \geqslant \|T - U^{(t+1)}\otimes V^{(t)}\otimes W^{(t)}\|_C$$

поскольку $U^{(t+1)} = \varphi(T,V^{(t)},W^{(t)})$. Оставшиеся неравенства в (i) могут быть доказаны аналогично.

Утверждение (ii) следует из единственности решения задачи (3.6). □

Пусть $V \in \mathbb{R}^{n\times r}$ и $W \in \mathbb{R}^{k\times r}$ являются чебышевскими матрицами и тройка последовательностей чебышевских матриц $\{U^{(t)}\}_{t\in\mathbb{N}}$, $\{V^{(t)}\}_{t\in\mathbb{N}}$ и $\{W^{(t)}\}_{t\in\mathbb{N}}$ получена методом переменных направлений для тензора T и пары начальных точек $V^{(0)} = V$ и $W^{(0)} = W$. Из Леммы 3.2 (i) следует, что последовательность $\|T - U^{(t)}\otimes V^{(t)}\otimes W^{(t)}\|_C$ не возрастает и, поскольку состоит из неотрицательных элементов, сходится. Обозначим предел этой последовательности через $E(T,U,V)$. Следующая лемма содержит базовые свойства функции E.

Лемма 3.3. *Пусть* $T \in \mathbb{R}^{m\times n\times k}$ *и матрицы* $V \in \mathbb{R}^{n\times r}$ *и* $W \in \mathbb{R}^{k\times r}$ *являются чебышевскими. Пусть метод переменных направлений для тензора* T *и пары начальных точек* $V^{(0)} = V$ *и* $W^{(0)} = W$ *является корректным (то есть все порожденные методом матрицы являются чебышевскими). Тогда выполнены следующие утверждения.*

(i) $E(T,V,W) \geqslant 0$ *и* $E(T,V,W) = E(T,\alpha V,\beta W)$ *при* $\alpha, \beta \neq 0$*.*

(ii) $E(T,V,W) = E(T,\tilde{V},W) = E(T,\tilde{V},\tilde{W})$*, где* $\tilde{V} = \psi(T,\varphi(T,V,W),W)$ *и* $\tilde{W} = \chi(T,\varphi(T,V,W),\tilde{V})$*.*

(iii) Функция $E(T,V,W)$ *является полунепрерывной сверху по паре* (V,W)*.*

Доказательство. Утверждения (i) и (ii) следуют из определения функции $E(T,V,W)$ и Леммы 3.2. Полунепрерывность сверху следует из того, что $E(T,U,V)$ является пределом убывающей последовательности непрерывных функций. □

Итоговая процедура метода переменных направлений приведена в Алгоритме 6. Заметим, что в алгоритме применяются перенормировки на каждом шаге, поскольку по Лемме 3.3 они не влияют на решение, но позволяют улучшить численную устойчивость.

Входные данные: Тензор $T \in \mathbb{R}^{m\times n\times k}$, ранг $r \geqslant 1$, начальные матрицы $V^{(0)} \in \mathbb{R}^{n\times r}$, $W^{(0)} \in \mathbb{R}^{k\times r}$.

Результат: Факторы канонического разложения ранга r: $\hat{U} \in \mathbb{R}^{m\times r}$, $\hat{V} \in \mathbb{R}^{n\times r}$, $\hat{W} \in \mathbb{R}^{k\times r}$.

$t = 1$;

repeat

$U^{(t)} = \varphi(T, V^{(t-1)}, W^{(t-1)})$;

$V^{(t)} = \psi(T, U^{(t)}, W^{(t-1)})$;

$W^{(t)} = \chi(T, U^{(t)}, V^{(t)})$;

$C = \|U^{(t)}\|_C \|V^{(t)}\|_C \|W^{(t)}\|_C$;

$U^{(t)} = U^{(t)} / \|U^{(t)}\|_C \cdot C^{1/3}$;

$V^{(t)} = V^{(t)} / \|V^{(t)}\|_C \cdot C^{1/3}$;

$W^{(t)} = W^{(t)} / \|W^{(t)}\|_C \cdot C^{1/3}$;

$t = t + 1$

until *сходимость*;

$\hat{U} = U^{(t-1)}, \quad \hat{V} = V^{(t-1)}, \quad \hat{W} = W^{(t-1)}$;

Алгоритм 6: Метод переменных направлений для тензоров.

3.3 Корректность метода переменных направлений для ранга 1

В Разделе 3.2 было введено понятие метода переменных направлений, однако Определение 3.1 требует явно предполагать, что все получаемые в результате работы метода матрицы являются чебышевскими. В этом разделе будет показано, что в случае приближений ранга 1 для почти всех тензоров получаемые векторы будут чебышевскими.

Определение 3.4. *Будем говорить, что тензор $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы, если для любых чебышевских векторов $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^n$, векторы $\varphi(T, v, w)$, $\psi(T, u, w)$ и $\chi(T, u, v)$ также являются чебышевскими.*

Из Определения 3.4 и Следствия 1.25 ясно, что если $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы и $v^{(0)} \in \mathbb{R}^n$ и $w^{(0)} \in \mathbb{R}^k$ являются чебышевскими, то существует единственная тройка последовательностей $\{u^{(t)}\}_{t\in\mathbb{N}}$, $\{v^{(t)}\}_{t\in\mathbb{N}}$ и $\{w^{(t)}\}_{t\in\mathbb{N}}$, полученная методом переменных направлений для тензора T и пары начальных точек $(v^{(0)}, w^{(0)})$. Следующая лемма дает описание множества тензоров, сохраняющих чебышевские системы.

Лемма 3.5. *Тензор $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы тогда и только тогда, когда любая его двумерная срезка имеет единственный максимальный по модулю элемент, то есть векторы $\operatorname{vec}(T[i,:,:])$, $\operatorname{vec}(T[:,j,:])$ и $\operatorname{vec}(T[:,:,l])$ являются пиковыми для любых i, j и l. Множество всех тензоров, сохраняющих чебышевские системы является открытым и плотным $\mathbb{R}^{m\times n\times k}$, а его дополнение в $\mathbb{R}^{m\times n\times k}$ имеет лебегову меру нуль.*

Доказательство. Действительно, T сохраняет чебышевские системы тогда и только тогда, когда $\mu(\operatorname{vec}(T[i,:,:]), v\otimes w)$, $\mu(\operatorname{vec}(T[:,j,:]), u\otimes w)$ и $\mu(\operatorname{vec}(T[:,:,l]), u\otimes v)$ не равны нулю при всех i,j и l и всех чебышевских векторов $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^k$. Из Теоремы 2.15 (i) следует, что это имеет место тогда и только тогда, когда $\operatorname{vec}(T[i,:,:])$, $\operatorname{vec}(T[:,j,:])$ и $\operatorname{vec}(T[:,:,l])$ являются пиковыми при всех $i = 1,\dots,m$, $j = 1,\dots,n$ и $l = 1,\dots,k$. Из этого описания ясно, что множество тензоров, сохраняющих чебышевские системы является открытым в $\mathbb{R}^{m\times n\times k}$ и его дополнение содержится в объединении конечного числа гиперплоскостей в $\mathbb{R}^{m\times n\times k}$. Поэтому дополнение имеет меру нуль и пустую внутренность. □

Напомним, что *амплитудой* чебышевского вектора $v \in \mathbb{R}^n$ мы называем $\operatorname{am}(v) = \|v\|_\infty / \min\limits_{i=1,\dots,n} |v_i|$.

Лемма 3.6. *Пусть $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы и $v^{(0)} \in \mathbb{R}^n$ и $w^{(0)} \in \mathbb{R}^k$ являются чебышевскими векторами. Пусть тройка последовательностей $\{u^{(t)}\}_{t\in\mathbb{N}}$, $\{v^{(t)}\}_{t\in\mathbb{N}}$ и $\{w^{(t)}\}_{t\in\mathbb{N}}$ получена методом переменных направлений для тензора T и пары начальных точек $(v^{(0)}, w^{(0)})$. Пусть $\delta_1 = \min\limits_{i=1,\dots,m} \delta(\operatorname{vec}(T[i,:,:]))$, $\delta_2 = \min\limits_{j=1,\dots,n} \delta(\operatorname{vec}(T[:,j,:]))$ и $\delta_3 = \min\limits_{l=1,\dots,k} \delta(\operatorname{vec}(T[:,:,l]))$. Тогда*

$$\|u^{(t)}\|_\infty \|v^{(t-1)}\|_\infty \|w^{(t-1)}\|_\infty \leqslant 2\|T\|_C,$$

$$\|u^{(t)}\|_\infty \|v^{(t)}\|_\infty \|w^{(t-1)}\|_\infty \leqslant 2\|T\|_C,$$

$$\|u^{(t)}\|_\infty \|v^{(t)}\|_\infty \|w^{(t)}\|_\infty \leqslant 2\|T\|_C,$$
$$\operatorname{am}(u^{(t)}) \leqslant 4\|T\|_C/\delta_1,\ \operatorname{am}(v^{(t)}) \leqslant 4\|T\|_C/\delta_2,\ \operatorname{am}(w^{(t)}) \leqslant 4\|T\|_C/\delta_3$$

при всех $t \in \mathbb{N}$.

Доказательство. По определению функции φ и Теореме 2.15 (ii) имеем

$$\frac{\delta(\operatorname{vec}(T[i,:,:]))}{2\|v \otimes w\|_\infty} \leqslant |\varphi(T,v,w)_i| \leqslant \frac{2\|\operatorname{vec}(T[i,:,:])\|_\infty}{\|v \otimes w\|_\infty},$$

следовательно

$$\frac{\delta_1}{2\|v\|_\infty \|w\|_\infty} \leqslant \|\varphi(T,v,w)\|_\infty \leqslant \frac{2\|T\|_C}{\|v\|_\infty \|w\|_\infty},$$

откуда следует утверждение леммы. □

3.4 Анализ поведения знаков для ранга 1

В данном разделе анализируется поведение знаков компонент векторов $u^{(t)}$, $v^{(t)}$ и $w^{(t)}$, полученных при помощи метода переменных направлений. А именно, будет показано, что знаки векторов $u^{(t)}$, $v^{(t)}$ и $w^{(t)}$ полностью определяются приближаемым тензором и знаками начальной точки $(v^{(0)}, w^{(0)})$, а также, что знаки стабилизируются для достаточно больших t.

Пусть $v \in \mathbb{R}^n$ являются чебышевским вектором. Напомним, что через $\mathcal{S}(v)$ мы обозначаем вектор с компонентами $\mathcal{S}(v)_i = \operatorname{sign}(v_i)$, $i = 1, \ldots, n$.

Теорема 3.7. *Пусть* $T \in \mathbb{R}^{m\times n\times k}$ *сохраняет чебышевские системы. Если* $v_1, v_2 \in \mathbb{R}^n$ *и* $w_1, w_2 \in \mathbb{R}^k$ *являются чебышевскими,* $\mathcal{S}(v_1) = \mathcal{S}(v_2)$ *и* $\mathcal{S}(w_1) = \mathcal{S}(w_2)$, *то*

$$\mathcal{S}(\varphi(T, v_1, w_1)) = \mathcal{S}(\varphi(T, v_2, w_2)).$$

Аналогичное утверждение верно для ψ *и* χ.

Доказательство. Пусть O_v является множеством всех чебышевских векторов $v \in \mathbb{R}^n$ таких, что $\mathcal{S}(v) = \mathcal{S}(v_1)$, а O_w является множеством всех чебышевских векторов $w \in \mathbb{R}^k$ таких, что $\mathcal{S}(w) = \mathcal{S}(w_1)$. Обозначим $O = O_v \times O_w$. Ясно, что множества O_v и O_w являются выпуклыми, а поэтому связным, следовательно O

также является выпуклым и связным. Функция $s(v,w) = \mathcal{S}(\varphi(T,v,w))$ является непрерывной на множестве чебышевских пар векторов, поскольку $\varphi(T,v,w)$ непрерывна, а $\mathcal{S}$ локально постоянна, и, следовательно, тоже непрерывна. Поскольку образ s дискретен, отсюда следует, что s постоянна на O. Тогда $s(v_1,w_1) = s(v_2,w_2)$, поскольку $(v_1,w_1),(v_2,w_2) \in O$. Утверждения для ψ и χ могут быть доказаны аналогично. □

Из Теоремы 3.7 следует, что знаки векторов на следующем шаге метода переменных направлений зависят только от знаков векторов на предыдущем шаге. Таким образом, для пары начальных точек с одинаковыми знаками метод переменных направлений порождает последовательности векторов с совпадающими знаками.

Поведение знаков может быть легко описано и без явного применения метода переменных направлений. Напомним, что $\zeta(a)$ для пикового вектора a обозначает позицию максимального по модулю элемента. Будем использовать то же самое обозначение для функции $\zeta : \mathbb{R}^{m\times n} \to \{1,\dots,m\} \times \{1,\dots,n\}$ которая ставит в соответствие матрице $A \in \mathbb{R}^{m\times n}$ такой, что вектор $\mathrm{vec}\,(A)$ является пиковым, позицию максимального по модулю элемента в матрице. Обозначим также через $\zeta_1 : \mathbb{R}^{m\times n} \to \{1,\dots,m\}$ функцию, которая ставит в соответствие матрице $A \in \mathbb{R}^{m\times n}$ строку, содержащую максимальный по модулю элемент, а через $\zeta_2 : \mathbb{R}^{m\times n} \to \{1,\dots,n\}$ отображение в номер столбца, содержащего максимальный по модулю элемент.

Теорема 3.8. *Пусть $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы. Тогда выполнены следующие утверждения.*

(i) Пусть $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^k$ являются чебышевскими. Тогда

$$\operatorname{sign}\left(\varphi(T,v,w)_i\right) = \operatorname{sign}\left(T[i,:,:]_{\zeta(T[i,:,:])} v_{\zeta_1(T[i,:,:])} w_{\zeta_2(T[i,:,:])}\right),$$

$$\operatorname{sign}\left(\psi(T,u,w)_j\right) = \operatorname{sign}\left(T[:,j,:]_{\zeta(T[:,j,:])} u_{\zeta_1(T[:,j,:])} w_{\zeta_2(T[:,j,:])}\right),$$

$$\operatorname{sign}\left(\chi(T,u,v)_l\right) = \operatorname{sign}\left(T[:,:,l]_{\zeta(T[:,:,l])} u_{\zeta_1(T[:,:,l])} v_{\zeta_2(T[:,:,l])}\right).$$

(ii) Пусть тройка последовательностей $\{u^{(t)}\}_{t\in\mathbb{N}}$, $\{v^{(t)}\}_{t\in\mathbb{N}}$ и $\{w^{(t)}\}_{t\in\mathbb{N}}$ получена методом переменных направлений для тензора T и чебышевской пары начальных точек. Тогда знаки векторов $u^{(t)}$, $v^{(t)}$ и $w^{(t)}$ стабилизируются для достаточно большого t.

Доказательство. Из Определения 3.1 имеем

$$\varphi(T, v, w)_i = \mu(\mathrm{vec}\,(T[i, :, :]), v \otimes w).$$

Тогда первое утверждение следует из Теоремы 2.15 (ii).

Обозначим через $\mathcal{M}_0$ множество троек (i, j, l) таких, что элементы в позиции (i, j, l) для тензора T соответствуют максимальным по модулю значениям в срезках $T[i, :, :]$, $T[:, j, :]$ и $T[:, :, l]$ одновременно. Очевидно, что $\mathcal{M}_0$ не пусто, так как содержит по крайней мере позицию максимального по модулю элемента в тензоре T. Обозначим также

$$\mathcal{M}_0^{(1)} = \{i : \exists (j, l) \text{ такие, что } (i, j, l) \in \mathcal{M}_0\},$$
$$\mathcal{M}_0^{(2)} = \{j : \exists (i, l) \text{ такие, что } (i, j, l) \in \mathcal{M}_0\},$$
$$\mathcal{M}_0^{(3)} = \{l : \exists (i, j) \text{ такие, что } (i, j, l) \in \mathcal{M}_0\}.$$

Заметим, что при $(i, j, l) \in \mathcal{M}_0$ мы имеем

$$\mathrm{sign}\left(T[i, :, :]_{\zeta(T[i,:,:])}\right) = \mathrm{sign}\left(T[:, j, :]_{\zeta(T[:,j,:])}\right) = \mathrm{sign}\left(T[:, :, l]_{\zeta(T[:,:,l])}\right) = c,$$

где $c = \pm 1$. Тогда при $(i, j, l) \in \mathcal{M}_0$ утверждение (i) сводится к

$$c \cdot \mathrm{sign}\,(u_i v_j w_l) = 1, \quad (i, j, l) \in \mathcal{M}_0,$$

что означает, что знаки вектора $u^{(t)}$ на позициях из $\mathcal{M}_0^{(1)}$, знаки вектора $v^{(t)}$ на позициях из $\mathcal{M}_0^{(2)}$ и знаки вектора $w^{(t)}$ на позициях из $\mathcal{M}_0^{(3)}$ не изменяются в процессе метода переменных направлений.

Обозначим

$$\mathcal{M}_p^{(1)} = \{i : \zeta_1(T[i, :, :]) \in \mathcal{M}_{p-1}^{(2)} \text{ and } \zeta_2(T[i, :, :]) \in \mathcal{M}_{p-1}^{(3)}\},$$
$$\mathcal{M}_p^{(2)} = \{j : \zeta_1(T[:, j, :]) \in \mathcal{M}_{p-1}^{(1)} \text{ and } \zeta_2(T[:, j, :]) \in \mathcal{M}_{p-1}^{(3)}\},$$
$$\mathcal{M}_p^{(3)} = \{l : \zeta_1(T[:, :, l]) \in \mathcal{M}_{p-1}^{(1)} \text{ and } \zeta_2(T[:, :, l]) \in \mathcal{M}_{p-1}^{(2)}\}.$$

Поскольку знаки векторов $u^{(t)}$, $v^{(t)}$ и $w^{(t)}$ на позициях $\mathcal{M}_0^{(1)}$, $\mathcal{M}_0^{(2)}$ и $\mathcal{M}_0^{(3)}$ соответственно не изменяются, из (i) следует, что знаки векторов $u^{(t)}$, $v^{(t)}$ и $w^{(t)}$ на позициях из $\mathcal{M}_1^{(1)}$, $\mathcal{M}_1^{(2)}$ и $\mathcal{M}_1^{(3)}$ не изменяются после одного шага метода переменных направлений. Аналогично, знаки векторов $u^{(t)}$, $v^{(t)}$ и $w^{(t)}$ на позициях $\mathcal{M}_p^{(1)}$, $\mathcal{M}_p^{(2)}$ и $\mathcal{M}_p^{(3)}$ соответственно не меняются после p шагов метода переменных направлений. Остается заметить, что $\mathcal{M}_{p-1}^{(j)} \subset \mathcal{M}_p^{(j)}$, $j \in \{1, 2, 3\}$ и для достаточно больших p мы имеем $\mathcal{M}_p^{(1)} = \{1, \ldots, m\}$, $\mathcal{M}_p^{(2)} = \{1, \ldots, n\}$, $\mathcal{M}_p^{(3)} = \{1, \ldots, k\}$. □

3.5 Теорема об альтернансе

В данном разделе вводится понятие трехмерного альтернанса и доказывается, что оптимальное приближение ранга 1 в чебышевской норме обладает введенной структурой альтернанса (см. Теорему 3.15). Более того, в данном разделе будет показано (см. Теорему 3.16), что все предельные точки метода переменных направлений также обладают трехмерным альтернансом.

Введем необходимые обозначения. Пусть $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы и векторы $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^k$ являются чебышевскими. Обозначим $G = T - u\otimes v\otimes w$, $\sigma_{ijl}(T,u,v,w) = \operatorname{sign}(g_{ijl}u_iv_jw_l)$. Обозначим также

$$S(T,u,v,w) = \{(i,j,l) : |g_{ijl}| = \|G\|_C\},$$

$$\mathcal{I}(T,u,v,w) = \{i : \exists (j,l) \text{ такие, что } (i,j,l) \in S(T,u,v,w)\},$$
$$\mathcal{J}(T,u,v,w) = \{j : \exists (i,l) \text{ такие, что } (i,j,l) \in S(T,u,v,w)\},$$
$$\mathcal{L}(T,u,v,w) = \{l : \exists (i,j) \text{ такие, что } (i,j,l) \in S(T,u,v,w)\}.$$

Определение 3.9. *Пусть $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы и векторы $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^k$ являются чебышевскими. Будем говорить, что четверка (T,u,v,w) обладает трехмерным альтернансом, если существует непустое множество $\mathcal{A} \subset \{1,\dots,m\}\times\{1,\dots,n\}\times\{1,\dots,k\}$ такое, что $\mathcal{A} \subset S(T,u,v,w)$ и если $(i,j,l) \in \mathcal{A}$, то*

1. *существует пара $(i_1,j_1) \neq (i,j)$ такая, что $(i_1,j_1,l) \in \mathcal{A}$ и $\sigma_{ijl}(T,u,v,w) = -\sigma_{i_1j_1l}(T,u,v,w)$.*
2. *существует пара $(i_2,l_2) \neq (i,l)$ такая, что $(i_2,j,l_2) \in \mathcal{A}$ и $\sigma_{ijl}(T,u,v,w) = -\sigma_{i_2jl_2}(T,u,v,w)$.*
3. *существует пара $(j_3,l_3) \neq (j,l)$ такая, что $(i,j_3,l_3) \in \mathcal{A}$ и $\sigma_{ijl}(T,u,v,w) = -\sigma_{ij_3l_3}(T,u,v,w)$.*

Лемма 3.10. *Пусть $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы и векторы $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^k$ являются чебышевскими. Пусть $u = \varphi(T,v,w)$, $\tilde{v} = \psi(T,u,w)$, $\tilde{w} = \chi(T,u,\tilde{v})$. Тогда для любого $i \in \mathcal{I}(T,u,v,w)$ существуют пары (j_1,l_1) и (j_2,l_2) такие, что*

$$(i,j_1,l_1),(i,j_2,l_2) \in S(T,u,v,w), \quad \sigma_{ij_1l_1}(T,u,v,w) = -\sigma_{ij_2l_2}(T,u,v,w).$$

Аналогичные утверждения верны для любых $j \in \mathcal{J}(T, u, \tilde{v}, w)$ *и любых* $l \in \mathcal{L}(T, u, \tilde{v}, \tilde{w})$.

Доказательство. Пусть $i \in \mathcal{I}(T, u, v, w)$. Поскольку

$$u_i = \varphi(T,v,w)_i = \mu(\mathrm{vec}\,(T[i,:,:]), v \otimes w), \quad i = 1, \ldots, m,$$

из Следствия 1.25 получаем, что существуют такие пары (j_1, l_1) и (j_2, l_2), что

$$|t_{ij_1l_1} - u_i(v \otimes w)_{(j_1,l_1)}| = |t_{ij_2l_2} - u_i(v \otimes w)_{(j_2,l_2)}| = \\ \|\,\mathrm{vec}\,(T[i,:,:]) - u_i(v \otimes w)\|_\infty = \|G\|_C,$$

$$\mathrm{sign}((v\otimes w)_{(j_1,l_1)}(t_{ij_1l_1} - u_i(v\otimes w)_{(j_1,l_1)})) = -\mathrm{sign}((v\otimes w)_{(j_2,l_2)}(t_{ij_2l_2} - u_i(v\otimes w)_{(j_2,l_2)})),$$

откуда следует утверждение леммы. Оставшиеся утверждения могут быть доказаны аналогично. □

Лемма 3.11. *Пусть* $T \in \mathbb{R}^{m\times n\times k}$ *сохраняет чебышевские системы и векторы* $v \in \mathbb{R}^n$ *и* $w \in \mathbb{R}^k$ *являются чебышевскими. Пусть* $u = \varphi(T, v, w)$, $\tilde{v} = \psi(T, u, w)$, $\tilde{w} = \chi(T, u, \tilde{v})$, $\tilde{u} = \varphi(T, \tilde{v}, \tilde{w})$. *Тогда выполнены следующие утверждения.*

1. *Пусть* $\|T - u \otimes v \otimes w\|_C = \|T - u \otimes \tilde{v} \otimes w\|_C$. *Тогда* $S(T, u, \tilde{v}, w) \subset S(T, u, v, w)$. *Более того,* $j \in \mathcal{J}(u, \tilde{v}, w)$ *тогда и только тогда, когда* $j \in \mathcal{J}(u, v, w)$ *и* $v_j = \tilde{v}_j$, *что имеет место тогда и только тогда, когда существуют две различные пары* (i_1, l_1) *и* (i_2, l_2) *такие, что* $(i_1, j, l_1), (i_2, j, l_2) \in S(T, u, v, w)$ *и* $\sigma_{i_1jl_1}(T, u, v, w) = -\sigma_{i_2jl_2}(T, u, v, w)$.
2. *Пусть* $\|T - u \otimes \tilde{v} \otimes w\|_C = \|T - u \otimes \tilde{v} \otimes \tilde{w}\|_C$. *Тогда* $S(T, u, \tilde{v}, \tilde{w}) \subset S(T, u, \tilde{v}, w)$. *Более того,* $l \in \mathcal{L}(u, \tilde{v}, \tilde{w})$ *тогда и только тогда, когда* $l \in \mathcal{L}(u, \tilde{v}, w)$ *и* $w_l = \tilde{w}_l$, *что имеет место тогда и только тогда, когда существуют две различные пары* (i_1, j_1) *и* (i_2, j_2) *такие, что* $(i_1, j_1, l), (i_2, j_2, l) \in S(T, u, \tilde{v}, w)$ *и* $\sigma_{i_1j_1l}(T, u, \tilde{v}, w) = -\sigma_{i_2j_2l}(T, u, \tilde{v}, w)$.
3. *Пусть* $\|T - u \otimes \tilde{v} \otimes \tilde{w}\|_C = \|T - \tilde{u} \otimes \tilde{v} \otimes \tilde{w}\|_C$. *Тогда* $S(T, \tilde{u}, \tilde{v}, \tilde{w}) \subset S(T, u, \tilde{v}, \tilde{w})$. *Более того,* $i \in \mathcal{I}(\tilde{u}, \tilde{v}, \tilde{w})$ *тогда и только тогда, когда* $i \in \mathcal{I}(u, \tilde{v}, \tilde{w})$ *и* $u_i = \tilde{u}_i$, *что имеет место тогда и только тогда, когда существуют две различные пары* (j_1, l_1) *и* (j_2, l_2) *такие, что* $(i, j_1, l_1), (i, j_2, l_2) \in S(T, u, \tilde{v}, \tilde{w})$ *и* $\sigma_{ij_1l_1}(T, u, \tilde{v}, \tilde{w}) = -\sigma_{ij_2l_2}(T, u, \tilde{v}, \tilde{w})$.

Доказательство. Рассмотрим произвольный индекс j. Поскольку $\tilde{v} = \psi(T, u, w)$,

$$\|\,\mathrm{vec}\,(T[:,j,:]) - v_j(u \otimes w)\|_\infty \geqslant \|\,\mathrm{vec}\,(T[:,j,:]) - \tilde{v}_j(u \otimes w)\|_\infty.$$

Пусть $j \notin \mathcal{J}(T, u, v, w)$. Тогда

$$\| \operatorname{vec}(T[:, j, :]) - v_j(u \otimes w)\|_\infty < \|T - u \otimes v \otimes w\|_C,$$

откуда

$$\| \operatorname{vec}(T[:, j, :]) - \tilde{v}_j(u \otimes w)\|_\infty < \|T - u \otimes v \otimes w\|_C = \|T - u \otimes \tilde{v} \otimes w\|_C,$$

следовательно $j \notin \mathcal{J}(T, u, \tilde{v}, w)$.

Пусть теперь $\tilde{v}_j \neq v_j$. Поскольку решение задачи о наилучшем равномерном приближении единственно (см. Теорему 1.5),

$$\| \operatorname{vec}(T[:, j, :]) - \tilde{v}_j(u \otimes w)\|_\infty < \| \operatorname{vec}(T[:, j, :]) - v_j(u \otimes w)\|_\infty \leqslant \|T - u \otimes v \otimes w\|_C = \|T - u \otimes \tilde{v} \otimes w\|_C,$$

откуда следует, что $j \notin \mathcal{J}(T, u, \tilde{v}, w)$. Таким образом, было доказано, что если $j \in \mathcal{J}(T, u, \tilde{v}, w)$, то $j \in \mathcal{J}(T, u, v, w)$ и $\tilde{v}_j = v_j$.

Докажем обратное, пусть $j \in \mathcal{J}(T, u, v, w)$ и $\tilde{v}_j = v_j$. Поскольку $\tilde{v}_j = v_j$, имеем

$$\| \operatorname{vec}(T[:, j, :]) - v_j(u \otimes w)\|_\infty = \| \operatorname{vec}(T[:, j, :]) - \tilde{v}_j(u \otimes w)\|_\infty,$$

а из того, что $j \in \mathcal{J}(T, u, v, w)$ получаем

$$\| \operatorname{vec}(T[:, j, :]) - v_j(u \otimes w)\|_\infty = \|T - u \otimes v \otimes w\|_C = \|T - u \otimes \tilde{v} \otimes w\|_C.$$

Следовательно $j \in \mathcal{J}(T, u, \tilde{v}, w)$.

Из Следствия 1.25 имеем, что $\tilde{v} = \psi(T, u, w)$ эквивалентно существованию различных пар (i_1, l_1) и (i_2, l_2) таких, что

$$|t_{i_1 j l_1} - \tilde{v}_j(u \otimes w)_{i_1, l_1}| = |t_{i_2 j l_2} - \tilde{v}_j(u \otimes w)_{i_2, l_2}| = \| \operatorname{vec}(T[:, j, :]) - \tilde{v}_j(u \otimes w)\|_\infty$$

и

$$\operatorname{sign}((u \otimes w)_{i_1, l_1}(t_{i_1 j l_1} - \tilde{v}_j(u \otimes w)_{i_1, l_1})) = -\operatorname{sign}((u \otimes w)_{i_2, l_2}(t_{i_2 j l_2} - \tilde{v}_j(u \otimes w)_{i_2, l_2})).$$

Пусть $j \in \mathcal{J}(T, u, v, w)$ и $\tilde{v}_j = v_j$, тогда

$$|t_{i_1 j l_1} - v_j(u \otimes w)_{i_1, l_1}| = |t_{i_2 j l_2} - v_j(u \otimes w)_{i_2, l_2}| = \| \operatorname{vec}(T[:, j, :]) - \tilde{v}_j(u \otimes w)\|_\infty = \|T - u \otimes \tilde{v} \otimes w\|_C = \|T - u \otimes v \otimes w\|_C$$

и

$$\operatorname{sign}((u\otimes w)_{i_1,l_1}(t_{i_1jl_1}-v_j(u\otimes w)_{i_1,l_1}))=-\operatorname{sign}((u\otimes w)_{i_2,l_2}(t_{i_2jl_2}-v_j(u\otimes w)_{i_2,l_2})),$$

откуда следует, что $(i_1,j,l_1),(i_2,j,l_2)\in S(T,u,v,w)$ и $\sigma_{i_1jl_1}(T,u,v,w)=-\sigma_{i_2jl_2}(T,u,v,w)$.

Обратно, пусть $(i_1,j,l_1),(i_2,j,l_2)\in S(T,u,v,w)$ и $\sigma_{i_1jl_1}(T,u,v,w)=-\sigma_{i_2jl_2}(T,u,v,w)$. Тогда из Следствия 1.25 получаем, что v_j является решением задачи

$$\inf_{t\in\mathbb{R}}\|\operatorname{vec}(T[:,j,:])-t(u\otimes w)\|_\infty,$$

откуда в силу единственности решения следует, что $v_j=\tilde{v}_j$. Более того, поскольку $(i_1,j,l_1),(i_2,j,l_2)\in S(T,u,v,w)$, имеем $j\in\mathcal{J}(T,u,v,w)$.

Докажем наконец, что $S(T,u,\tilde{v},w)\subset S(T,u,v,w)$. Пусть $(i,j,l)\in S(T,u,\tilde{v},w)$. Тогда $j\in\mathcal{J}(T,u,\tilde{v},w)$, откуда $\tilde{v}_j=v_j$. Следовательно

$$t_{ijl}-v_j(u\otimes w)_{i,l}=t_{ijl}-\tilde{v}_j(u\otimes w)_{i,l}.$$

Однако

$$|t_{ijl}-\tilde{v}_j(u\otimes w)_{i,l}|=\|T-u\otimes\tilde{v}\otimes w\|_C=\|T-u\otimes v\otimes w\|_C,$$

откуда следует, что $(i,j,l)\in S(T,u,v,w)$. Оставшиеся утверждения леммы могут быть доказаны аналогично. □

Лемма 3.12. *Пусть $T\in\mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы и векторы $v\in\mathbb{R}^n$ и $w\in\mathbb{R}^k$ являются чебышевскими. Пусть $u=\varphi(T,v,w)$, $\tilde{v}=\psi(T,u,w)$, $\tilde{w}=\chi(T,u,\tilde{v})$. Пусть также*

$$\|T-u\otimes v\otimes w\|_C=\|T-u\otimes\tilde{v}\otimes w\|_C=\|T-u\otimes\tilde{v}\otimes\tilde{w}\|_C$$

и

$$S(T,u,v,w)=S(T,u,\tilde{v},w)=S(T,u,\tilde{v},\tilde{w}).$$

Тогда (T,u,v,w) обладает трехмерным альтернансом.

Доказательство. Покажем, что (T,u,v,w) обладает трехмерным альтернансом с множеством индексов $\mathcal{A}=S(T,u,v,w)$. Пусть $(i,j,l)\in\mathcal{A}$, тогда $i\in\mathcal{I}(T,u,v,w)$ и по Лемме 3.10 существуют пары (j_1,l_1) и (j_2,l_2) такие, что

$$(i,j_1,l_1),(i,j_2,l_2)\in S(T,u,v,w),\quad \sigma_{ij_1l_1}(T,u,v,w)=-\sigma_{ij_2l_2}(T,u,v,w).$$

Тогда по крайней мере для одной из пар $(\tilde{j}_1, \tilde{l}_1)$ (либо $(\tilde{j}_1, \tilde{l}_1) = (j_1, l_1)$, либо $(\tilde{j}_1, \tilde{l}_1) = (j_2, l_2)$) имеем

$$\sigma_{i,\tilde{j}_1,\tilde{l}_1}(T, u, v, w) = -\sigma_{i,j,l}(T, u, v, w),$$

$$(i, \tilde{j}_1, \tilde{l}_1) \in S(T, u, v, w) = \mathcal{A}.$$

Таким образом, выполнено первое свойство из определения трехмерного альтернанса.

Пусть $(i, j, l) \in S(T, u, \tilde{v}, w) = S(T, u, v, w) = \mathcal{A}$. Тогда $j \in \mathcal{J}(T, u, \tilde{v}, w)$, откуда по Лемме 3.11 существуют различные пары (i_1, l_1) и (i_2, l_2) такие, что $(i_1, j, l_1), (i_2, j, l_2) \in S(T, u, v, w)$ и $\sigma_{i_1 j l_1}(T, u, v, w) = -\sigma_{i_2 j l_2}(T, u, v, w)$. Тогда по крайней мере для одной пары $(\tilde{i}_2, \tilde{l}_2)$ (либо $(\tilde{i}_2, \tilde{l}_2) = (i_1, l_1)$, либо $(\tilde{i}_2, \tilde{l}_2) = (i_2, l_2)$) имеем

$$\sigma_{\tilde{i}_2,j,\tilde{l}_2}(T, u, v, w) = -\sigma_{i,j,l}(T, u, v, w),$$

$$(\tilde{i}_2, j, \tilde{l}_2) \in S(T, u, v, w) = \mathcal{A}.$$

Таким образом, выполнено второе свойство из определения трехмерного альтернанса.

Пусть $(i, j, l) \in S(T, u, \tilde{v}, \tilde{w}) = S(T, u, v, w) = \mathcal{A}$. Тогда $l \in \mathcal{L}(T, u, \tilde{v}, \tilde{w})$, откуда по Лемме 3.11 существуют различные пары (i_1, j_1) и (i_2, j_2) такие, что $(i_1, j_1, l), (i_2, j_2, l) \in S(T, u, \tilde{v}, w) = S(T, u, v, w)$ и $\sigma_{i_1 j_1 l}(T, u, \tilde{v}, w) = -\sigma_{i_2 j_2 l}(T, u, \tilde{v}, w)$. Поскольку $(i_1, j_1, l) \in S(T, u, \tilde{v}, w)$, получаем $j_1 \in \mathcal{J}(u, \tilde{v}, w)$, откуда из Леммы 3.11 имеем $v_{j_1} = \tilde{v}_{j_1}$. Аналогично $v_{j_2} = \tilde{v}_{j_2}$. Тогда

$$\sigma_{i_1 j_1 l}(T, u, v, w) = -\sigma_{i_2 j_2 l}(T, u, v, w).$$

Тогда по крайней мере для одной из пар $(\tilde{i}_3, \tilde{j}_3)$ (либо $(\tilde{i}_3, \tilde{j}_3) = (i_1, j_1)$, либо $(\tilde{i}_3, \tilde{j}_3) = (i_2, j_2)$) имеем

$$\sigma_{\tilde{i}_3,\tilde{j}_3,l}(T, u, v, w) = -\sigma_{i,j,l}(T, u, v, w),$$

$$(\tilde{i}_3, \tilde{j}_3, l) \in S(T, u, v, w) = \mathcal{A}.$$

Таким образом, выполнены все свойства из определения трехмерного альтернанса. □

Лемма 3.13. *Пусть $T \in \mathbb{R}^{m \times n \times k}$ сохраняет чебышевские системы и векторы $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^k$ являются чебышевскими. Пусть $u = \varphi(T, v, w)$, $\tilde{v} = \psi(T, u, w)$,*

$\tilde{w} = \chi(T, u, \tilde{v})$, $\tilde{u} = \varphi(T, \tilde{v}, \tilde{w})$. *Пусть также* $\|T - u \otimes v \otimes w\|_C = \|T - \tilde{u} \otimes \tilde{v} \otimes \tilde{w}\|_C$. *Тогда если* $(T, \tilde{u}, \tilde{v}, \tilde{w})$ *обладает трехмерным альтернансом, то* (T, u, v, w) *также обладает трехмерным альтернансом.*

Доказательство. По Лемме 3.2 (i)

$$\|T - u \otimes v \otimes w\|_C \geqslant \|T - u \otimes \tilde{v} \otimes w\|_C \geqslant \|T - u \otimes \tilde{v} \otimes \tilde{w}\|_C \geqslant \|T - \tilde{u} \otimes \tilde{v} \otimes \tilde{w}\|_C,$$

однако первый и последний члены неравенства равны, поэтому

$$\|T - u \otimes v \otimes w\|_C = \|T - u \otimes \tilde{v} \otimes w\|_C = \|T - u \otimes \tilde{v} \otimes \tilde{w}\|_C = \|T - \tilde{u} \otimes \tilde{v} \otimes \tilde{w}\|_C.$$

Применяя Лемму 3.11 к каждому из равенств получаем $S(T, \tilde{u}, \tilde{v}, \tilde{w}) \subset S(T, u, v, w)$. Кроме того, из Леммы 3.11 следует, что

1. $u_i = \tilde{u}_i$ при $i \in \mathcal{I}(T, \tilde{u}, \tilde{v}, \tilde{w})$;
2. $v_j = \tilde{v}_j$ при $j \in \mathcal{J}(T, \tilde{u}, \tilde{v}, \tilde{w})$;
3. $w_l = \tilde{w}_l$ при $k \in \mathcal{L}(T, \tilde{u}, \tilde{v}, \tilde{w})$.

Тогда $\sigma_{ijl}(T, \tilde{u}, \tilde{v}, \tilde{w}) = \sigma_{ijl}(T, u, v, w)$ при всех $(i,j,l) \in S(T, \tilde{u}, \tilde{v}, \tilde{w})$. Следовательно трехмерный альтернанс для $(T, \tilde{u}, \tilde{v}, \tilde{w})$ является трехмерным альтернансом для (T, u, v, w). □

Лемма 3.14. *Пусть* $T \in \mathbb{R}^{m \times n \times k}$ *сохраняет чебышевские системы и векторы* $v \in \mathbb{R}^n$ *и* $w \in \mathbb{R}^k$ *являются чебышевскими. Тогда выполнены следующие утверждения.*

(i) Пусть последовательности $\{u^{(t)}\}_{t \in \mathbb{N}}$, $\{v^{(t)}\}_{t \in \mathbb{N}}$ *и* $\{w^{(t)}\}_{t \in \mathbb{N}}$ *получены методом переменных направлений для тензора* T *и пары начальных точек* $v^{(0)} = v$, $w^{(0)} = w$. *Тогда произвольная предельная точка* (ξ, η) *последовательности пар* (ξ_t, η_t), *где* $\xi_t = v^{(t)}/\|v^{(t)}\|_\infty$ $\eta_t = w^{(t)}/\|w^{(t)}\|_\infty$ *такова, что векторы* ξ *и* η *являются чебышевскими и*

$$E(A,\xi,\eta) = \|T - \varphi(T,\xi,\eta) \otimes \xi \otimes \eta\|_C = E(T,v,w).$$

(ii) Если $\|T - \varphi(T,v,w) \otimes v \otimes w\|_C = E(T,v,w)$, *то* $(T,\varphi(T,v,w), v, w)$ *обладает трехмерным альтернансом.*

Доказательство. Пусть ξ является пределом последовательности ξ_{l_t}, являющейся подпоследовательностью ξ_t, а η является пределом η_{l_t}, подпоследовательности η_t. Поскольку амплитуды чебышевских векторов не меняются при умножении на ненулевую константу, из Леммы 3.6 следует, что $\mathrm{am}(\xi_t) \leqslant C$, $\mathrm{am}(\eta_t) \leqslant C$, где

$C > 0$ зависит только от тензора T. Нетрудно понять, что сходящаяся последовательность чебышевских векторов с ограниченной амплитудой сходится либо к нулевому вектору, либо к чебышевскому. Поскольку $\|\xi_t\|_\infty = 1$ и $\|\eta_t\|_\infty = 1$, получаем, что ξ и η являются чебышевскими векторами.

Из Леммы 3.3 (ii) следует, что $E(T,\xi_t,\eta_t) = E(T,v,w)$ при всех $t \in \mathbb{N}$. Из полунепрерывности сверху функции E (см. Лемму 3.3 (iii)) получаем, что $E(T,\xi,\eta) \geqslant E(A,v,w)$. Более того, $\|T - \varphi(T,\xi,\eta) \otimes \xi \otimes \eta\|_C \geqslant E(T,\xi,\eta)$.

$$\|T - \varphi(T,\xi,\eta) \otimes \xi \otimes \eta\|_C = \lim_{t\to\infty} \|T - \varphi(T,\xi_{l_t},\eta_{l_t}) \otimes \xi_{l_t} \otimes \eta_{l_t}\|_C = \lim_{t\to\infty} \|T - \varphi(T, v^{(l_t)}, w^{(l_t)}) \otimes v^{(l_t)} \otimes w^{(l_t)}\|_C = E(T,v,w).$$

Следовательно,

$$E(T,v,w) = \|T - \varphi(T,\xi,\eta) \otimes \xi \otimes \eta\|_C \geqslant E(T,\xi,\eta) \geqslant E(T,v,w)$$

и (i) доказано.

Пусть $\|T - \varphi(T,v,w) \otimes v \otimes w\|_C = E(T,v,w)$ и пусть $\{u^{(t)}\}_{t\in\mathbb{N}}$, $\{v^{(t)}\}_{t\in\mathbb{N}}$ и $\{w^{(t)}\}_{t\in\mathbb{N}}$ получены методом переменных направлений для T и пары начальных точек $v^{(0)} = v$ и $w^{(0)} = w$. Ясно, что

$$\|T - u^{(t)} \otimes v^{(t-1)} \otimes w^{(t-1)}\|_C = \|T - u^{(t)} \otimes v^{(t)} \otimes w^{(t-1)}\|_C = \|T - u^{(t)} \otimes v^{(t)} \otimes w^{(t)}\|_C = \|T - u^{(t+1)} \otimes v^{(t)} \otimes w^{(t)}\|_C$$

при всех $t \in \mathbb{N}$. Тогда из Леммы 3.11 следует, что

$$S(T, u^{(1)}, v^{(0)}, v^{(0)}) \supset S(T, u^{(1)}, v^{(1)}, v^{(0)}) \supset S(T, u^{(1)}, v^{(1)}, v^{(1)}) \supset S(T, u^{(2)}, v^{(1)}, v^{(1)}) \supset S(T, u^{(2)}, v^{(2)}, v^{(1)}) \supset \dots.$$

Поскольку все множества в этой последовательности не пусты, существует $t \in \mathbb{N}$ такое, что

$$S(T, u^{(t+1)}, v^{(t)}, w^{(t)}) = S(T, u^{(t+1)}, v^{(t+1)}, w^{(t)}) = S(T, u^{(t+1)}, v^{(t+1)}, w^{(t+1)}).$$

Тогда по Лемме 3.12 получаем, что $(T, u^{(t+1)}, v^{(t)}, w^{(t)})$ обладает трехмерным альтернансом. Применяя Лемму 3.13 t раз получаем, что $(T, u^{(1)}, v^{(0)}, w^{(0)}) = (T, \varphi(T,v,w), v, w)$ также обладает трехмерным альтернансом. □

Основные результаты раздела сразу следуют из доказанных лемм.

Теорема 3.15. *Пусть тензор $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы. Пусть $\hat{u} \in \mathbb{R}^m$, $\hat{v} \in \mathbb{R}^n$ и $\hat{w} \in \mathbb{R}^k$ являются решением задачи*

$$\|T - u \otimes v \otimes k\|_C \to \min_{u\in\mathbb{R}^m, v\in\mathbb{R}^n, w\in\mathbb{R}^k}.$$

Пусть $\hat{v}$ и $\hat{w}$ являются чебышевскими. Тогда $(T, \hat{u}, \hat{v}, \hat{w})$ обладает трехмерным альтернансом.

Доказательство. Ясно, что

$$\|T - \hat{u}\otimes\hat{v}\otimes\hat{w}\|_C \geqslant \|T - \varphi(T,\hat{v},\hat{w})\otimes\hat{v}\otimes\hat{w}\|_C \geqslant E(T,\hat{v},\hat{w}) \geqslant \|T - \hat{u}\otimes\hat{v}\otimes\hat{w}\|_C,$$

где последнее неравенство следует из того, что $\hat{u}$, $\hat{v}$ и $\hat{w}$ являются оптимальными решениями. Следовательно $\|T - \hat{u}\otimes\hat{v}\otimes\hat{w}\|_C = E(T,\hat{v},\hat{w})$ и по Лемме 3.14 (ii) четверка $(T, \varphi(T,\hat{v},\hat{w}),\hat{v},\hat{w})$ обладает трехмерным альтернансом. Заметим, что $\mathcal{I}(T,\varphi(T,\hat{v},\hat{w}),\hat{v},\hat{w}) \subset \mathcal{I}(T,\hat{u},\hat{v},\hat{w})$ по определению φ. Обозначим $\tilde{u} = \varphi(T,\hat{v},\hat{w})$. При $i \in \mathcal{I}(T,\varphi(T,\hat{v},\hat{w}),\hat{v},\hat{w})$ имеем

$$\|\operatorname{vec}(T[i,:,:]) - (\hat{v}\otimes\hat{w})\hat{u}_i\|_\infty = \|\operatorname{vec}(T[i,:,:]) - (\hat{v}\otimes\hat{w})\tilde{u}_i\|_\infty = \min_{u\in\mathbb{R}}\|\operatorname{vec}(T[i,:,:]) - (\hat{v}\otimes\hat{w})u\|_\infty.$$

Благодаря единственности решения, $\hat{u}_i = \tilde{u}_i$ при $i \in \mathcal{I}(T,\varphi(T,\hat{v},\hat{w}),\hat{v},\hat{w})$. Поскольку в Определении 3.9 используются только позиции вектора $\hat{u}$ такие, что $i \in \mathcal{I}(T,\varphi(T,\hat{v},\hat{w}),\hat{v},\hat{w})$, получаем, что альтернанс для четверки $(T,\varphi(T,\hat{v},\hat{w}),\hat{v},\hat{w})$ является альтернансом для четверки $(T,\hat{u},\hat{v},\hat{w})$. □

Теорема 3.16. *Пусть $T \in \mathbb{R}^{m\times n\times k}$ сохраняет чебышевские системы и векторы $v \in \mathbb{R}^n$ и $w \in \mathbb{R}^k$ являются чебышевскими. Пусть последовательности $\{u^{(t)}\}_{t\in\mathbb{N}}$, $\{v^{(t)}\}_{t\in\mathbb{N}}$ и $\{w^{(t)}\}_{t\in\mathbb{N}}$ получены методом переменных направлений для тензора T и пары начальных точек $v^{(0)} = v$, $w^{(0)} = w$. Тогда произвольная предельная точка (ξ,η) последовательности пар (ξ_t,η_t), где $\xi_t = v^{(t)}/\|v^{(t)}\|_\infty$ $\eta_t = w^{(t)}/\|w^{(t)}\|_\infty$ такова, что ξ и η являются чебышевскими и $(T,\varphi(T,\xi,\eta),\xi,\eta)$ обладает трехмерным альтернансом.*

Доказательство. Сразу следует из Леммы 3.14. □

3.6 О сходимости к локальному минимуму

В Разделе 2.7 структура двумерного альтернанса для матриц была использована для построения оптимального чебышевского приближения ранга 1. Одним из ключевых утверждений в случае приближения матриц является Теорема 2.25 [20], гарантирующая, что если тройка $(A, \hat{u}, \hat{v})$ обладает двумерным альтернансом, где $A \in \mathbb{R}^{m\times n}$ и $\hat{u} \in \mathbb{R}^m$, $\hat{v} \in \mathbb{R}^n$, то точка $(\hat{u}, \hat{v})$ является локальным минимумом функционала $c(u, v) = \|A - uv^T\|_C$, причем в достаточно большой области. Естественный вопрос — обеспечивает ли наличие трехмерного альтернанса для тензоров аналогичные свойства решения?

Рассмотрим пример. Пусть тензор $T \in \mathbb{R}^{3\times3\times3}$ имеет следующие срезки:

$$T[1,:,:] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 2 \\ 1 & 0 & 0 \end{bmatrix}, \quad T[2,:,:] = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad T[3,:,:] = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 2 \\ 0 & 0 & 2 \end{bmatrix}.$$

Пусть векторы $\hat{u} = \hat{v} = \hat{w} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$. Легко видеть, что четверка $(T, \hat{u}, \hat{v}, \hat{w})$ обладает трехмерным альтернансом и

$$S(T, \hat{u}, \hat{v}, \hat{w}) = \{(1,2,1), (1,2,2), (1,2,3), (1,3,2), (1,3,3), (2,2,1), \\ (2,3,1), (2,3,3), (3,2,1), (3,2,3), (3,3,1), (3,3,2), (3,3,3)\}.$$

В этом случае тройка $(\hat{u}, \hat{v}, \hat{w})$ является неподвижной точкой метода переменных направлений и $\|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|_C = 1$. Рассмотрим также векторы

$$u(\varepsilon) = \begin{bmatrix} 1-\varepsilon & 1+\varepsilon & 1+3\varepsilon \end{bmatrix}^T, \\ v(\varepsilon) = \begin{bmatrix} 1 & 1+\varepsilon & 1-5\varepsilon \end{bmatrix}^T, \\ w(\varepsilon) = \begin{bmatrix} 1-\varepsilon & 1+\varepsilon & 1+3\varepsilon \end{bmatrix}^T.$$

В этом случае $G(\varepsilon) = T - u(\varepsilon) \otimes v(\varepsilon) \otimes w(\varepsilon)$ имеет следующие срезки:

$$G(\varepsilon)[1,:,:] = \begin{bmatrix} 2\varepsilon - \varepsilon^2 & \varepsilon^2 & -2\varepsilon + 3\varepsilon^2 \\ -1 + \varepsilon + \varepsilon^2 - \varepsilon^3 & 1 - \varepsilon + \varepsilon^2 + \varepsilon^3 & 1 - 3\varepsilon + \varepsilon^2 + 3\varepsilon^3 \\ 7\varepsilon - 11\varepsilon^2 + 5\varepsilon^3 & -1 + 5\varepsilon + \varepsilon^2 - 5\varepsilon^3 & -1 + 3\varepsilon + 13\varepsilon^2 - 15\varepsilon^3 \end{bmatrix},$$

$$G(\varepsilon)[2,:,:] = \begin{bmatrix} \varepsilon^2 & -2\varepsilon - \varepsilon^2 & -4\varepsilon - 3\varepsilon^2 \\ 1 - \varepsilon + \varepsilon^2 + \varepsilon^3 & -3\varepsilon - 3\varepsilon^2 - \varepsilon^3 & -5\varepsilon - 7\varepsilon^2 - 3\varepsilon^3 \\ -1 + 5\varepsilon + \varepsilon^2 - 5\varepsilon^3 & 3\varepsilon + 9\varepsilon^2 + 5\varepsilon^3 & -1 + \varepsilon + 17\varepsilon^2 + 15\varepsilon^3 \end{bmatrix},$$

$$G(\varepsilon)[3,:,:] = \begin{bmatrix} -2\varepsilon + 3\varepsilon^2 & -4\varepsilon - 3\varepsilon^2 & -6\varepsilon - 9\varepsilon^2 \\ 1 - 3\varepsilon + \varepsilon^2 + 3\varepsilon^3 & -5\varepsilon - 7\varepsilon^2 - 3\varepsilon^3 & 1 - 7\varepsilon - 15\varepsilon^2 - 9\varepsilon^3 \\ -1 + 3\varepsilon + 13\varepsilon^2 - 15\varepsilon^3 & -1 + \varepsilon + 17\varepsilon^2 + 15\varepsilon^3 & 1 - \varepsilon + 21\varepsilon^2 + 45\varepsilon^3 \end{bmatrix}.$$

Из формул выше ясно, что $\|G(\varepsilon)\|_C < 1$ при любом достаточно малом ε, что означает, что несмотря на то, что четверка $(T, \hat{u}, \hat{v}, \hat{w})$ обладает трехмерным альтернансом, тройка $(\hat{u}, \hat{v}, \hat{w})$ не является локальным минимумом функционала $c(u, v, w) = \|T - u \otimes v \otimes w\|_C$. Однако в этом случае можно предложить процедуру, позволяющую строить направление, вдоль которого значение функционала $c(u, v, w)$ строго убывает.

Пусть $T \in \mathbb{R}^{m \times n \times k}$ и $\hat{u} \in \mathbb{R}^m$, $\hat{v} \in \mathbb{R}^n$, $\hat{w} \in \mathbb{R}^k$. Пусть $q \in \mathbb{R}^m$, $h \in \mathbb{R}^n$ и $f \in \mathbb{R}^k$ являются поправками к векторам $\hat{u}$, $\hat{v}$ и $\hat{w}$ соответственно. Обозначим

$$u = \hat{u} + q \quad v = \hat{v} + h \quad w = \hat{w} + f.$$

Обозначим также $G = T - u \otimes v \otimes w$, $\hat{G} = T - \hat{u} \otimes \hat{v} \otimes \hat{w}$. Тогда

$$g_{ijl} = t_{ijl} - u_i v_j w_l = t_{ijl} - (\hat{u}_i + q_i)(\hat{v}_j + h_j)(\hat{w}_l + f_l) =$$
$$t_{ijl} - (\hat{u}_i \hat{v}_j \hat{w}_l + \hat{u}_i \hat{v}_j f_l + \hat{u}_i h_j \hat{w}_l + \hat{u}_i h_j f_l + q_i \hat{v}_j \hat{w}_l + q_i \hat{v}_j f_l + q_i h_j \hat{w}_l + q_i h_j f_l).$$

Обозначим

$$\beta_{ijl} = \operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i \hat{v}_j f_l + \hat{u}_i h_j \hat{w}_l + \hat{u}_i h_j f_l + q_i \hat{v}_j \hat{w}_l + q_i \hat{v}_j f_l + q_i h_j \hat{w}_l + q_i h_j f_l).$$

Тогда

$$g_{ijl} = \operatorname{sign}(\hat{g}_{ijl})|\hat{g}_{ijl}| - \operatorname{sign}(\hat{g}_{ijl})\beta_{ijl},$$

откуда

$$|g_{ijl}| = ||\hat{g}_{ijl}| - \beta_{ijl}|. \tag{3.7}$$

Нетрудно доказать следующий результат.

Лемма 3.17. *Точка $(\hat{u}, \hat{v}, \hat{w})$ является локальным минимумом функционала $c(u, v, w) = \|T - u \otimes v \otimes w\|_C$ тогда и только тогда, когда для любых векторов q, h и f в некоторой окрестности нуля выполнено $\beta_{ijl} \leqslant 0$ по крайней мере для одной тройки $(i, j, l) \in S(T, \hat{u}, \hat{v}, \hat{w})$.*

Доказательство. Действительно, пусть $(\hat{u}, \hat{v}, \hat{w})$ является локальным минимумом $c(u, v, w)$. Тогда при всех (u, v, w) в некоторой окрестности $(\hat{u}, \hat{v}, \hat{w})$ выполнено $\|T - u \otimes v \otimes w\|_C \geqslant \|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|_C$. Если окрестность достаточно мала, то поскольку при любых $(i, j, l) \notin S(T, \hat{u}, \hat{v}, \hat{w})$

$$|\hat{g}_{ijl}| = |t_{ijl} - \hat{u}_i \hat{v}_j \hat{w}_l| < \|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|_C,$$

то из непрерывности получаем, что в этой окрестности также выполнено

$$|g_{ijl}| = |t_{ijl} - u_i v_j w_l| < \|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|_C \leqslant \|T - u \otimes v \otimes w\|_C,$$

то есть максимум не достигается на позициях $(i,j,l) \notin S(T,\hat{u},\hat{v},\hat{w})$. Следовательно он достигается при $(i,j,l) \in S(T,\hat{u},\hat{v},\hat{w})$, что возможно только если $\beta_{ijl} \leqslant 0$ при некотором $(i,j,l) \in S(T,\hat{u},\hat{v},\hat{w})$ в силу (3.7) и $\|T - u \otimes v \otimes w\|_C \geqslant \|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|_C$.

Обратно, если выполнено, что $\beta_{ijl} \leqslant 0$ при некотором $(i,j,l) \in S(T,\hat{u},\hat{v},\hat{w})$, то из (3.7) получаем, что $\|T - u \otimes v \otimes w\|_C \geqslant \|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|_C$. □

Заметим, что

$$\beta_{ijl} = \operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i\hat{v}_j f_l + \hat{u}_i(h_j\hat{w}_l + h_j f_l) + q_i(\hat{v}_j\hat{w}_l + \hat{v}_j f_l + h_j\hat{w}_l + h_j f_l)) = \operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i\hat{v}_j f_l + \hat{u}_i h_j w_l + q_i v_j w_l),$$

где

$$u = \hat{u} + q \quad v = \hat{v} + h \quad w = \hat{w} + f.$$

Рассмотрим систему (нелинейную) уравнений от переменных ξ, ζ и η:

$$\operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i\hat{v}_j\eta_l + \hat{u}_i\zeta_j w_l(\eta_l) + \xi_i v_j(\zeta_j) w_l(\eta_l)) = b_{ijk} \tag{3.8}$$

при $(i,j,l) \in S(T,\hat{u},\hat{v},\hat{w})$, где $v_j(\zeta_j) = \hat{v}_l + \zeta_j$, $w_l(\eta_l) = \hat{w}_l + \eta_l$. Образом этой системы являются всевозможные значения β_{ijl}. Система (3.8) является системой полиномиальных уравнений и анализировать ее свойства может быть достаточно трудно, поэтому рассмотрим другую систему, являющуюся ее линеаризацией (3.8):

$$\operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i\hat{v}_j\eta_l + \hat{u}_i\hat{w}_l\zeta_j + \hat{v}_j\hat{w}_l\xi_i) = b_{ijk}, \quad (i,j,l) \in S(T,\hat{u},\hat{v},\hat{w}). \tag{3.9}$$

Обозначим матрицу системы (3.9) через L. Будем также обозначать правую часть (3.9), соответствующую $\xi \in \mathbb{R}^{|R(T,\hat{u},\hat{v},\hat{w})|}$, $\zeta \in \mathbb{R}^{|C(T,\hat{u},\hat{v},\hat{w})|}$, $\eta \in \mathbb{R}^{|D(T,\hat{u},\hat{v},\hat{w})|}$, через $L(\xi,\zeta,\eta) \in \mathbb{R}^{|S(T,\hat{u},\hat{v},\hat{w})|}$.

Лемма 3.18. *Пусть образ матрицы L содержит по крайней мере один вектор с компонентами одного знака (ненулевыми). Тогда в любой окрестности нуля найдутся векторы ξ, ζ и η такие, что $b_{ijk} > 0$ для всех уравнений системы (3.8).*

Доказательство. Пусть при некоторых значениях $(\hat{\xi}, \hat{\zeta}, \hat{\eta})$ выполнено

$$\operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i\hat{v}_j\hat{\eta}_l + \hat{u}_i\hat{w}_l\hat{\zeta}_j + \hat{v}_j\hat{w}_l\hat{\xi}_i) = \alpha_{ijk}, \quad (i,j,l) \in S(T,\hat{u},\hat{v},\hat{w}).$$

Пусть $p > 0$ и возьмем

$$q = p\hat{\xi}, \quad h = p\hat{\zeta}, \quad f = p\hat{\eta}.$$

Тогда

$$\beta_{ijl} = \operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i\hat{v}_j\hat{\eta}_l p + \hat{u}_i\hat{\zeta}_j\hat{w}_l p + \hat{u}_i\hat{\zeta}_j\hat{\eta}_l p^2 + \hat{\xi}_i\hat{v}_j\hat{w}_l p + \hat{\xi}_i\hat{v}_j\hat{\eta}_l p^2 + \hat{\xi}_i\hat{\zeta}_j\hat{w}_l p^2 + \hat{\xi}_i\hat{\zeta}_j\hat{\eta}_l p^3) = \alpha_{ijl}p + c^{(1)}_{ijl}p^2 + c^{(2)}_{ijl}p^3, \quad (3.10)$$

где $c^{(1)}_{ijl}$ и $c^{(2)}_{ijl}$ обозначают суммы коэффициентов перед p^2 и p^3 соответственно. Пусть образ системы (3.9) содержит вектор с компонентами одного знака. Тогда существует тройка $(\hat{\xi}, \hat{\zeta}, \hat{\eta})$ такая, что $\alpha_{ijl} > 0$ при всех $(i,j,l) \in S(T,\hat{u},\hat{v},\hat{w})$. Тогда из (3.10) следует, что для любого достаточно малого p выполнено $\beta_{ijl} > 0$ для любых $(i,j,l) \in S(T,\hat{u},\hat{v},\hat{w})$. □

Комбинируя Лемму 3.17 и Лемму 3.18, получаем следующий важный результат.

Лемма 3.19. *Пусть образ матрицы L содержит по крайней мере один вектор с компонентами одного знака (ненулевыми). Тогда $(\hat{u},\hat{v},\hat{w})$ не является локальным минимумом функционала $c(u,v,w) = \|T - u\otimes v\otimes w\|_C$. Более того, если для тройки (ξ,ζ,η) выполнено*

$$\operatorname{sign}(\hat{g}_{ijl})(\hat{u}_i\hat{v}_j\eta_l + \hat{u}_i\hat{w}_l\zeta_j + \hat{v}_j\hat{w}_l\xi_i) > 0$$

при всех $(i,j,l) \in S(T,\hat{u},\hat{v},\hat{w})$, то существуют $\tilde{\xi} \in \mathbb{R}^m$, $\tilde{\zeta} \in \mathbb{R}^n$, $\tilde{\eta} \in \mathbb{R}^k$ и $p_0 > 0$ такие, что для любого $p < p_0$ и

$$u = \hat{u} + p\tilde{\xi} \quad v = \hat{v} + p\tilde{\zeta} \quad w = \hat{w} + p\tilde{\eta}$$

выполнено $\|T - u\otimes v\otimes w\|_C < \|T - \hat{u}\otimes\hat{v}\otimes\hat{w}\|_C$.

Доказательство. Достаточно взять ξ, ζ и η из Леммы 3.18 и $\tilde{\xi}$, $\tilde{\zeta}$ и $\tilde{\eta}$ такими, что они равны ξ, ζ и η на позициях из $\mathcal{I}(T,\hat{u},\hat{v},\hat{w})$, $\mathcal{J}(T,\hat{u},\hat{v},\hat{w})$ и $\mathcal{L}(T,\hat{u},\hat{v},\hat{w})$ соответственно, и нулевые на остальных позициях. □

Для нахождения значений (ξ, ζ, η) из Леммы 3.19 можно использовать методы линейного программирования. Обозначим размеры матрицы L через $M = |S(T, \hat{u}, \hat{v}, \hat{w})|$ и $N = |\mathcal{I}(T, \hat{u}, \hat{v}, \hat{w})| + |\mathcal{J}(T, \hat{u}, \hat{v}, \hat{w})| + |\mathcal{L}(T, \hat{u}, \hat{v}, \hat{w})|$. Требуется узнать существует ли такой вектор $x \in \mathbb{R}^N$, что $Lx > 0$ (все компоненты строго больше нуля). Более того, для численной устойчивости лучше, чтобы компоненты вектора Lx были сбалансированы, то есть

$$\frac{\max\limits_i |(Lx)_i|}{\min\limits_i |(Lx)_i|} \leqslant C.$$

Поскольку $(Lx)_i$ при всех i, это эквивалентно

$$(Lx)_i \leqslant C(Lx)_j, \quad \text{для всех } i \text{ и } j.$$

Тогда задача нахождения векторов (ξ, ζ, η) (или получения результата, что таких векторов не существует) может быть записана как

$$\begin{cases} \langle 0, x \rangle \to \min\limits_x, \\ Lx \geqslant \varepsilon e, \\ (Lx)_i \leqslant C(Lx)_j, \quad \text{для всех } i \text{ и } j, \end{cases} \tag{3.11}$$

где $e = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$, $\varepsilon > 0$ является достаточно маленькой константой (например, 10^{-6}), а C может быть найдено при помощи бинарного поиска. Ясно, что (3.11) является канонической задачей линейного программирования и может быть решена, например, при помощи симплекс-метода.

Существуют различные способы нахождения оптимального значения p. Заметим, что функция

$$z(p) = \|T - (\hat{u} + p\tilde{\xi}) \otimes (\hat{v} + p\tilde{\zeta}) \otimes (\hat{w} + p\tilde{\eta})\|_C$$

может быть записана как

$$z(p) = \max\{|c_0^{(1)} + c_1^{(1)}p + c_2^{(1)}p^2 + c_3^{(1)}p^3|, \dots |c_0^{(N)} + c_1^{(mnk)}p + c_2^{(mnk)}p^2 + c_3^{(mnk)}p^3|\}.$$

Функция $z(p)$ является кусочно-полиномиальной и поэтому может принимать свое минимальное значение только либо в точках экстремума полиномов, либо в точках излома, когда один из полиномов равен нулю или выполнено

$$|c_0^{(i)} + c_1^{(i)}p + c_2^{(i)}p^2 + c_3^{(i)}p^3| = |c_0^{(j)} + c_1^{(j)}p + c_2^{(j)}p^2 + c_3^{(j)}p^3|$$

при $i \neq j$. Заметим, что таких точек существует конечное число и все они могут быть легко вычислены. Таким образом может быть найдено оптимальное значение p, минимизирующее функционал $z(p)$. Однако поскольку по Лемме 3.19 существует такое $p_0 > 0$, что для любого $p < p_0$ выполнено $z(p) < z(0)$, значение p, строго уменьшающее величину $z(p)$, может быть найдено при помощи деления пополам.

Будем называть метод, основанный на применении Леммы 3.19, *методом переменных направлений с коррекциями*. Общая схема метода приведена в Алгоритме 7.

Входные данные: Тензор $T \in \mathbb{R}^{m \times n \times k}$,начальные векторы $v^{(0)} \in \mathbb{R}^n$, $w^{(0)} \in \mathbb{R}^k$, $\varepsilon > 0$.
Результат: Факторы канонического разложения $\hat{u} \in \mathbb{R}^m$, $\hat{v} \in \mathbb{R}^n$, $\hat{w} \in \mathbb{R}^k$.
$t = 1$;
$\hat{u}^{(t)}, \hat{v}^{(t)}, \hat{w}^{(t)} = \texttt{alternating\_minimization}(T, v^{(t-1)}, w^{(t-1)})$;
Построим матрицу $L^{(t)}$, соответствующую $\hat{u}^{(t)}, \hat{v}^{(t)}, \hat{w}^{(t)}$;
Решим задачу линейного программирования для нахождения $(\xi^{(t)}, \zeta^{(t)}, \eta^{(t)})$ таких, что $L^{(t)}(\xi^{(t)}, \eta^{(t)}, \zeta^{(t)}) > 0$;
while *существуют* $(\xi^{(t)}, \zeta^{(t)}, \eta^{(t)})$ *такие, что* $L^{(t)}(\xi^{(t)}, \eta^{(t)}, \zeta^{(t)}) > 0$ **do**
  Продолжим $\xi^{(t)}$, $\zeta^{(t)}$, $\eta^{(t)}$ нулями до $\tilde{\xi}^{(t)}$, $\tilde{\zeta}^{(t)}$, $\tilde{\eta}^{(t)}$;
  Найдем $p^{(t)} > 0$, минимизирующее

$$z^{(t)}(p) = \|T - (\hat{u}^{(t)} + p\tilde{\xi}^{(t)}) \otimes (\hat{v}^{(t)} + p\tilde{\zeta}^{(t)}) \otimes (\hat{w}^{(t)} + p\tilde{\eta}^{(t)})\|_C;$$

  $u^{(t)} = \hat{u}^{(t)} + p^{(t)}\tilde{\xi}^{(t)}$, $v^{(t)} = \hat{v}^{(t)} + p^{(t)}\tilde{\zeta}^{(t)}$, $w^{(t)} = \hat{w}^{(t)} + p^{(t)}\tilde{\eta}^{(t)}$;
  $t = t + 1$;
  $\hat{u}^{(t)}, \hat{v}^{(t)}, \hat{w}^{(t)} = \texttt{alternating\_minimization}(T, v^{(t-1)}, w^{(t-1)})$;
  Построим матрицу $L^{(t)}$, соответствующую $\hat{u}^{(t)}, \hat{v}^{(t)}, \hat{w}^{(t)}$;
  Решим задачу линейного программирования для нахождения $(\xi^{(t)}, \zeta^{(t)}, \eta^{(t)})$ таких, что $L^{(t)}(\xi^{(t)}, \eta^{(t)}, \zeta^{(t)}) > 0$;
end
$\hat{u} = \hat{u}^{(t)}$, $\hat{v} = \hat{v}^{(t)}$, $\hat{w} = \hat{w}^{(t)}$;

Алгоритм 7: Метод переменных направлений с коррекциями.

Для оценки качества работы метода переменных направлений с коррекциями был проведен следующий численный эксперимент. Рассмотрим случайный

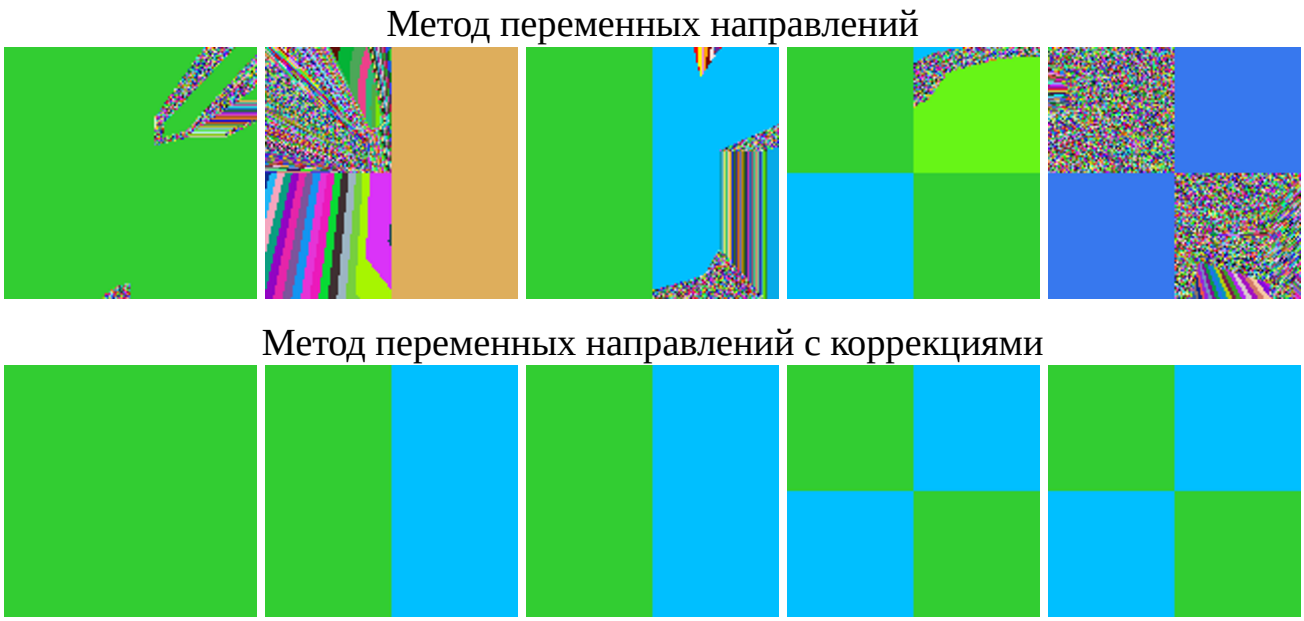


Рисунок 3.1 — Сходимость метода переменных направлений и метод переменных направлений с коррекциями для одинаковых начальных точек и случайных тензоров размера $3 \times 3 \times 3$.

тензор $T \in \mathbb{R}^{3\times3\times3}$. Заметим, что начальная чебышевская точка для метода переменных направлений может быть параметризована как

$$v^{(0)} = \begin{bmatrix} 1 \\ v_1 \\ v_2 \end{bmatrix} \quad w^{(0)} = \begin{bmatrix} 1 \\ w_1 \\ w_2 \end{bmatrix}.$$

Возьмем случайные значения v_1 и v_2 и зафиксируем их. Значения w_1 и w_2 будем выбирать из равномерной сетки из 128 точек на отрезке $[-2, 2]$. Обозначим узлы сетки через $w_1^{(i)}$ и $w_2^{(j)}$, $i, j = 1, \ldots, 128$. Для каждой начальной точки был запущен метод переменных направлений и метод переменных направлений с коррекциями (см. Алгоритм 7). В результате работы эксперимента была построена картинка, где пиксель в позиции (i, j) соответствует значению $\|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|$ для начальной точки $(v^{(0)}, w^{(0)}) = (w_1^{(i)}, w_2^{(j)})$. Два пикселя имеют один и тот же цвет, если значения $\|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|$ отличаются не более чем на 10^{-6}. Метод переменных направлений был запущен до сходимости с точностью 10^{-12}. Примеры получившихся картинок приведены на Рис. 3.1. Верхний ряд картинок соответствует обычному методу переменных направлений, а нижний ряд тем же самым тензорам и начальным точкам, но для метода переменных направлений с коррекцией. Проведенный эксперимент показывает, что в общем случае наличие трехмерного альтернанса (напомним, что все предельные точки обладают

трехмерным альтернансом по Теореме 3.16) не гарантирует, что предельная точка является локальным минимумом, и точность приближения зависит не только от знаков начальной точки. Однако для метода переменных направлений с коррекциями значение функционала $\|T - \hat{u} \otimes \hat{v} \otimes \hat{w}\|_C$ зависит только от знаков начальной точки. Аналогичные эксперименты были приведены для тензоров размера $7 \times 7 \times 7$, причем варьируемые параметры выбирались как по одному из векторов $v^{(0)}$ и $w^{(0)}$, так и по два из одного вектора. Во всех проведенных экспериментах наблюдалась картина аналогичная приведенной на Рис. 3.1.

В настоящий момент нам неизвестно о каких-либо результатах, доказывающих или опровергающих, что значение ошибки для метода переменных направлений с коррекциями зависит только знаков начальных векторов. Однако, если это верно, то Алгоритм 7 позволяет строить оптимальные приближения ранга 1 для тензоров в чебышевской норме. Для этого достаточно запустить метод из конечного числа точек, соответствующих различным знакам начальных векторов.

3.7 Численные эксперименты

Предложенный в Разделе 3.2 метод переменных направлений, не имеет гарантий оптимальности. Более того, также как для известного алгоритма ALS, для метода переменных направлений не удается даже доказать сходимость итерационной процедуры. Для оценки эффективности предложенной процедуры была проведена серия численных экспериментов. Во всех проведенных экспериментах предложенная итерационная процедура сходится. Реализация метода доступна онлайн[1].

3.7.1 Тензор Гильберта

Рассмотрим тензор

$$T = \left[\frac{1}{i + j + l} \right]_{i,j,l=1}^{512}.$$

[1] `https://github.com/stanis-morozov/cheburaxa`

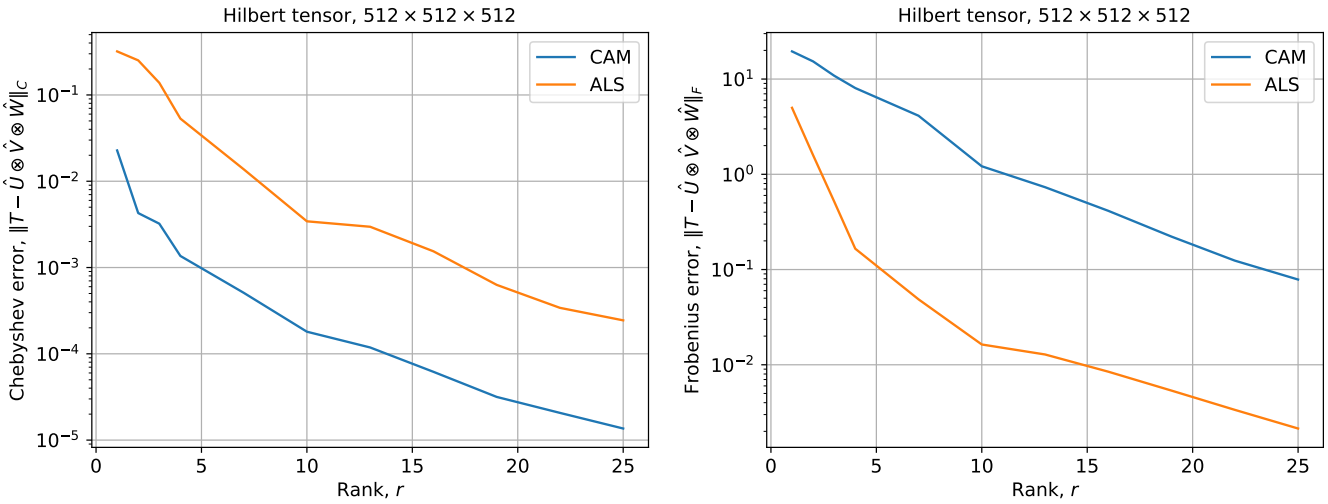


Рисунок 3.2 — Ошибка малоранговой аппроксимации для тензора Гильберта при помощи метода переменных направлений (CAM) и метода переменных наименьших квадратов (ALS). На левом графике приведена ошибка в норме Чебышева, а на правом в норме Фробениуса.

В данном эксперименте для построения малорангового приближения тензора T в каноническом формате были использованы метод переменных направлений (см. Алгоритм 6) и метод переменных наименьших квадратов (ALS). В случае алгоритма ALS итерационная процедура была запущена из 20 случайных матриц $V^{(0)}$ и $W^{(0)}$ с элементами из стандартного нормального распределения. Для того чтобы сделать время вычислений приемлемым, число итераций было ограничено $1{,}000$. В случае метода переменных направлений итерационная процедура была запущена из тех же самых точек и наилучшей точки, полученной в результате работы алгоритма ALS. Среди всех начальных точек для обоих методов была выбрана наилучшая. На Рис. 3.2 приведен график ошибок аппроксимации в чебышевской норме и норме Фробениуса. Из приведенных графиков видно, что алгоритмы качественно различны, а именно, метод переменных направлений лучше оптимизирует чебышевскую норму ошибки, а алгоритм ALS ошибку в норме Фробениуса.

3.7.2 Функционально порожденные тензоры

Рассмотрим тензор, порожденный значениями функции

$$f(x, y, z) = \cos(2\pi xyz)$$

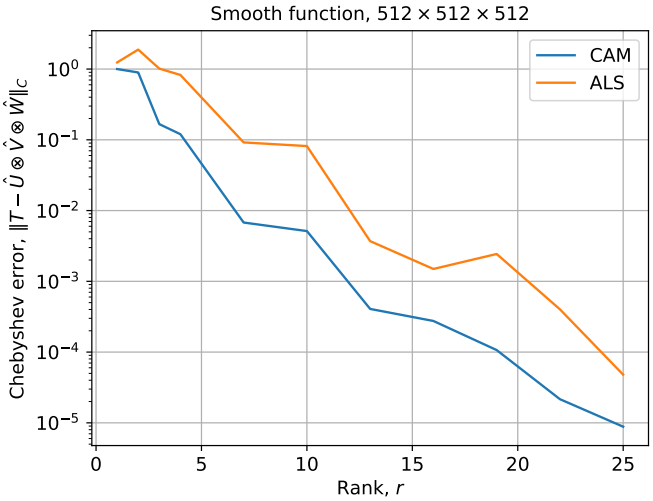


Рисунок 3.3 — Чебышевская ошибка малоранговой аппроксимации для тензора, порожденного гладкой функцией.

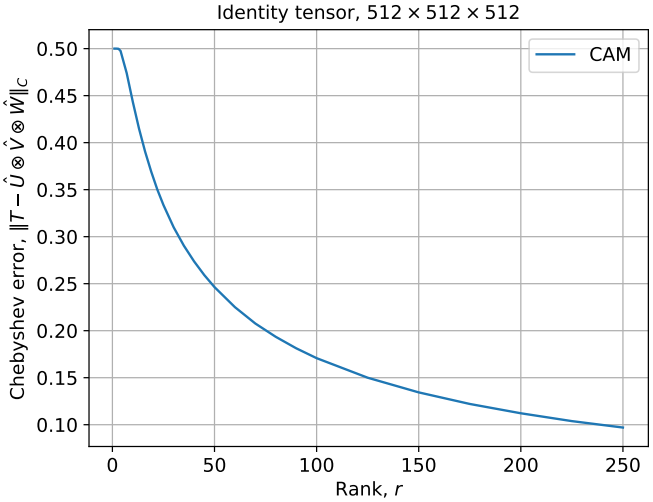


Рисунок 3.4 — Чебышевская ошибка малоранговой аппроксимации для «единичного» тензора.

на равномерной сетке в кубе $[0, 1]^3$, а именно,

$$T = \left[\cos\left(2\pi \cdot \frac{i-1}{n} \cdot \frac{j-1}{n} \cdot \frac{l-1}{n}\right)\right]_{i,j,l=1}^{n}.$$

В данном эксперименте было выбрано $n = 512$. Все остальные параметры такие же, как в предыдущем эксперименте. На Рис. 3.3 приведена ошибка в чебышевской норме для метода переменных направлений (CAM) и алгоритма ALS. Из графика легко видеть, что метод переменных направлений снова достигает существенно лучшей аппроксимации в чебышевской норме.

3.7.3 Единичный тензор

Рассмотрим тензор $T \in \mathbb{R}^{n \times n \times n}$, заданный формулой

$$t_{ijk} = \begin{cases} 1, & i = j = k, \\ 0, & \text{иначе}. \end{cases}$$

В данном эксперименте было выбрано $n = 512$. На Рис. 3.4 изображена ошибка аппроксимации в норме Чебышева для метода переменных направлений (CAM). Данный эксперимент демонстрирует качественную разницу между чебышевской и фробениусовой нормами, поскольку «единичный» тензор не имеет разумного

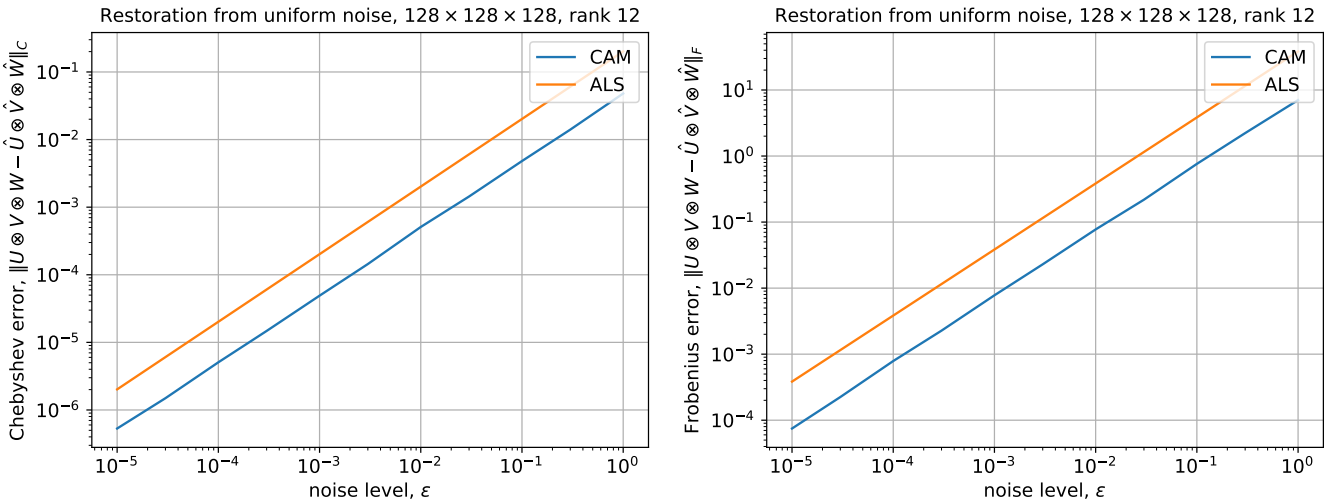


Рисунок 3.5 — Ошибка восстановления малорангового тензора из равномерного шума при помощи метода переменных направлений (CAM) и метода переменных наименьших квадратов (ALS). На левом графике приведена ошибка в норме Чебышева, а на правом в норме Фробениуса.

малорангового приближения в норме Фробениуса (ошибка равна 1), однако в норме Чебышева тензор может быть приближен достаточно хорошо.

Восстановление из равномерного шума

В данном эксперименте рассматривается задача восстановления малорангового тензора из шума. Пусть $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ и $W \in \mathbb{R}^{k \times r}$. Рассмотрим тензор

$$T = U \otimes V \otimes W + N,$$

где $N \in \mathbb{R}^{m \times n \times k}$ является случайным тензором с независимыми одинаково распределенными элементами из равномерного распределения $\mathcal{U}(-\varepsilon, \varepsilon)$. В данном эксперименте варьировался уровень шума ε и для каждого ε было сгенерировано 10 случайных матриц U, V и W с элементами из стандарного нормального распределения и тензор N с элементами из равномерного распределения. Для восстановления малорангового тензора из шума, тензор $T = U \otimes V \otimes W + N$ был приближен тензором малого ранга при помощи метода переменных направлений и алгоритма ALS. Для каждой процедуры восстановления итерационные процедуры были запущены из 10 случайных начальных точек и выбран наилучший результат. Пусть итерационная процедура построила факторы $\hat{U}$, $\hat{V}$ и $\hat{W}$. Тогда

Рисунок 3.6 — Примеры аппроксимации цветных изображений при помощи метода переменных направлений (CAM) и алгоритма ALS.

ошибка восстановления была вычислена как

$$\|U \otimes V \otimes W - \hat{U} \otimes \hat{V} \otimes \hat{W}\|$$

и результат усреднен по всем выборкам тензора T. В данном эксперименте было выбрано $m = n = k = 128$ и $r = 12$. На Рис. 3.5 приведена ошибка восстановления в нормах Чебышева и Фробениуса. Из приведенных графиков видно, что метод переменных направлений лучше восстанавливает малоранговую структуру как в норме Чебышева, так и в норме Фробениуса.

3.7.4 Цветные изображения

Для того чтобы визуализировать разницу между малоранговыми приближениями тензоров в чебышевской и фробениусовой нормах, был проведен эксперимент по построению аппроксимации для цветных изображений. Цветная картинка может быть представлена как тензор $T \in \mathbb{R}^{3\times w\times h}$, где w и h соответствуют геометрическим размерам картинки. Элементы тензора являются вещественными числами от 0 до 1. В данном эксперименте были рассмотрены картинки с фиксированными геометрическими размерами 512×512. На Рис. 3.6 показаны результаты приближения тензоров картинок в каноническом формате при помощи метода переменных направлений (CAM) и метода переменных наименьших квадратов (ALS). Каждая картинка на Рис. 3.6 также содержит ошибку приближения в чебышевской и фробениусовой нормах. Можно заметить, что аппроксимации, полученные методом ALS, более размытые, а изображения, полученные методом переменных направлений, являются более резкими, но имеют дрожащую структуру, которая может быть объяснена наличием альтернанса у предельных точек (см. Раздел 3.5).

Заключение

Основные результаты работы заключаются в следующем.

1. Предложен эффективный алгоритм решения задачи наилучшего равномерного приближения, доказаны гарантии его сходимости, приведены различные критерии оптимальности решения задачи.
2. Предложен метод переменных направлений для построения малоранговых чебышевских приближений матриц, теоретически изучены его свойства. Проведенные численные эксперименты демонстрируют эффективность и масштабируемость предложенного алгоритма.
3. Предложен алгоритм, позволяющий находить оптимальные приближения ранга 1 для матриц в чебышевской норме.
4. Предложен метод переменных направлений, позволяющий строить эффективные малоранговые приближения тензоров в каноническом формате в чебышевской норме и изучены его свойства. При помощи обширного численного исследования продемонстрирована эффективность предложенной процедуры.

Список литературы

1. *Bebendorf, M.* A means to efficiently solve elliptic boundary value problems / M. Bebendorf // Hierarchical Matrices. LNCS. — 2008. — Т. 63. — С. 49—98.
2. Data compression for the exascale computing era-survey / S. W. Son [и др.] // Supercomputing frontiers and innovations. — 2014. — Т. 1, № 2. — С. 76—88.
3. Fast matrix factorization for online recommendation with implicit feedback / X. He [и др.] // Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. — 2016. — С. 549—558.
4. OBOE: Collaborative filtering for AutoML model selection / C. Yang [и др.] // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. — 2019. — С. 1173—1183.
5. *Udell, M.* Why are big data matrices approximately low rank? / M. Udell, A. Townsend // SIAM Journal on Mathematics of Data Science. — 2019. — Т. 1, № 1. — С. 144—160.
6. *Budzinskiy, S.* When big data actually are low-rank, or entrywise approximation of certain function-generated matrices / S. Budzinskiy // arXiv preprint arXiv:2407.03250. — 2024.
7. *Pinkus, A.* N-widths in Approximation Theory / A. Pinkus. — Springer Berlin, Heidelberg, 1985. — С. 294.
8. *Кашин, Б. С.* О поперечниках октаэдров / Б. С. Кашин // Успехи математических наук. — 1975. — Т. 30, № 4. — С. 251—252.
9. *Глускин, Е. Д.* Октаэдр плохо приближается случайными подпространствами / Е. Д. Глускин // Функциональный анализ и его приложения. — 1986. — Т. 20, № 1. — С. 14—20.
10. *Гарнаев, А. Ю.* О поперечниках евклидова шара / А. Ю. Гарнаев, Е. Д. Глускин // Доклады Академии наук. Т. 277. — 1984. — С. 1048—1052.
11. *Khoromskij, B. N.* Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs / B. N. Khoromskij, C. Schwab // SIAM journal on scientific computing. — 2011. — Т. 33, № 1. — С. 364—385.

12. *Eshelby, J. D.* Energy relations and the energy-momentum tensor in continuum mechanics / J. D. Eshelby // Fundamental Contributions to the Continuum Theory of Evolving Phase Interfaces in Solids: A Collection of Reprints of 14 Seminal Papers. — Springer, 1999. — С. 82—119.

13. Tensor train versus Monte Carlo for the multicomponent Smoluchowski coagulation equation / S. A. Matveev [и др.] // Journal of Computational Physics. — 2016. — Т. 316. — С. 164—179.

14. *Lu, H.* A survey of multilinear subspace learning for tensor data / H. Lu, K. N. Plataniotis, A. N. Venetsanopoulos // Pattern Recognition. — 2011. — Т. 44, № 7. — С. 1540—1551.

15. *Hitchcock, F. L.* Multiple invariants and generalized rank of a p-way matrix or tensor / F. L. Hitchcock // Journal of Mathematics and Physics. — 1928. — Т. 7, № 1—4. — С. 39—79.

16. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis / R. A. Harshman [и др.] // UCLA working papers in phonetics. — 1970. — Т. 16, № 1. — С. 84.

17. *Tucker, L. R.* Some mathematical notes on three-mode factor analysis / L. R. Tucker // Psychometrika. — 1966. — Т. 31, № 3. — С. 279—311.

18. *De Lathauwer, L.* A multilinear singular value decomposition / L. De Lathauwer, B. De Moor, J. Vandewalle // SIAM journal on Matrix Analysis and Applications. — 2000. — Т. 21, № 4. — С. 1253—1278.

19. *Oseledets, I. V.* Tensor-train decomposition / I. V. Oseledets // SIAM Journal on Scientific Computing. — 2011. — Т. 33, № 5. — С. 2295—2317.

20. *Даугавет, В.* О равномерном приближении функции двух переменных, заданной таблично, произведением функций одной переменной / В. Даугавет // Журнал вычислительной математики и математической физики. — 1971. — Т. 11, № 2. — С. 289—303.

21. *Gillis, N.* Low-rank matrix approximation in the infinity norm / N. Gillis, Y. Shitov // Linear Algebra and its Applications. — 2019. — Т. 581. — С. 367—382.

22. *Замарашкин, Н. Л.* Об алгоритме наилучшего приближения матрицами малого ранга в норме Чебышёва / Н. Л. Замарашкин, С. В. Морозов, Е. Е. Тыртышников // Журнал вычислительной математики и математической физики. — 2022. — Т. 62, № 5. — С. 723—741.

23. *Morozov, S.* On the optimal rank-1 approximation of matrices in the Chebyshev norm / S. Morozov, M. Smirnov, N. Zamarashkin // Linear Algebra and its Applications. — 2023. — Т. 679. — С. 4—29.

24. *Морозов, С. В.* Метод переменных направлений для построения малорангового поэлементного приближения тензоров в каноническом формате / С. В. Морозов // Вычислительные методы и программирование. — 2024. — Т. 25, № 3. — С. 302—314.

25. *Morozov, S.* Refining uniform approximation algorithm for low-rank Chebyshev embeddings / S. Morozov, D. Zheltkov, A. Osinsky // Russian Journal of Numerical Analysis and Mathematical Modelling. — 2024. — Т. 39, № 5. — С. 311—328.

26. *Дзядык, В. К.* Введение в теорию равномерного приближения функций полиномами / В. К. Дзядык. — Наука, 1977.

27. *Смирнов, В. И.* Конструктивная теория функций комплексного переменного / В. И. Смирнов, Н. А. Лебедев. — М, 1964.

28. *Дзядык, В. К.* О приближении функций на множествах, состоящих из конечного числа точек / В. К. Дзядык // Сборник «Теория приближения функций и ее приложения». — 1974. — С. 69—80.

29. *Osinsky, A.* Rectangular maximum volume and projective volume search algorithms / A. Osinsky // arXiv preprint arXiv:1809.02334. — 2018.

30. *Golub, G. H.* Matrix computations / G. H. Golub, C. F. Van Loan. — JHU press, 2013.

31. *Budzinskiy, S.* On the distance to low-rank matrices in the maximum norm / S. Budzinskiy // Linear Algebra and its Applications. — 2024. — Т. 688. — С. 44—58.

32. *Pinkus, A.* Matrices and n-widths / A. Pinkus // Linear Algebra and its Applications. — 1979. — Т. 27. — С. 245—278.

33. *Mohlenkamp, M. J.* Musings on multilinear fitting / M. J. Mohlenkamp // Linear Algebra and its Applications. — 2013. — Т. 438, № 2. — С. 834—852.

34. *Budzinskiy, S.* Quasioptimal alternating projections and their use in low-rank approximation of matrices and tensors / S. Budzinskiy // arXiv preprint arXiv:2308.16097. — 2023.

35. *Shannon, C. E.* Probability of error for optimal codes in a Gaussian channel / C. E. Shannon // Bell System Technical Journal. — 1959. — Т. 38, № 3. — С. 611—656.

36. *Shi, T.* On the compressibility of tensors / T. Shi, A. Townsend // SIAM Journal on Matrix Analysis and Applications. — 2021. — Т. 42, № 1. — С. 275—298.

37. *Strassen, V.* Gaussian elimination is not optimal / V. Strassen // Numerische mathematik. — 1969. — Т. 13, № 4. — С. 354—356.

38. Tensor decompositions for signal processing applications: From two-way to multiway component analysis / A. Cichocki [и др.] // IEEE signal processing magazine. — 2015. — Т. 32, № 2. — С. 145—163.

39. Speeding-up convolutional neural networks using fine-tuned cp-decomposition / V. Lebedev [и др.] // arXiv preprint arXiv:1412.6553. — 2014.

40. *Håstad, J.* Tensor rank is NP-complete / J. Håstad // Journal of algorithms. — 1990. — Т. 11, № 4. — С. 644—654.