

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.16

АППРОКСИМАЦИЯ ДИФФЕРЕНЦИАЛЬНЫХ ОПЕРАТОРОВ С УЧЕТОМ ГРАНИЧНЫХ УСЛОВИЙ

© 2023 г. В. П. Варин^{1,*}

¹ 125047 Москва, Миусская пл., 4, Институт прикладной математики им. М.В. Келдыша РАН, Россия

*e-mail: varin@keldysh.ru

Поступила в редакцию 11.01.2023 г.
Переработанный вариант 19.02.2023 г.
Принята к публикации 30.03.2023 г.

Применение спектральных методов для решения краевых задач является весьма эффективным, но сопряженным с большими техническими трудностями, связанными с учетом граничных условий. Существует несколько способов такого учета, но все они либо весьма трудоемки, либо требуют предварительного анализа задачи и записи ее в интегральной форме. Мы предлагаем универсальный способ учета граничных условий для линейных дифференциальных операторов на конечном отрезке, который весьма прост в реализации. Применение рациональной арифметики позволяет оценить эффективность метода без учета ошибок округления. Мы применили наш подход к вычислению рациональных приближений для некоторых фундаментальных констант. Получены приближения, которые в ряде случаев лучше, чем те, которые дают обыкновенные цепные дроби этих констант. Библ. 22. Фиг. 2.

Ключевые слова: полиномы Лежандра, краевые задачи, доказательства иррациональности, голономные функции, высокоточные вычисления.

DOI: 10.31857/S004446692308015X, **EDN:** IIGDWE

1. ВВЕДЕНИЕ

Численные методы решения краевых задач на конечном интервале можно условно разделить на сеточные и спектральные. В сеточных методах дифференциальные операторы аппроксимируются с помощью конечных разностей, а функции приближаются с помощью таблиц их значений на выбранной сетке. В спектральных методах функции приближаются отрезками разложений этих функций по некоторым наборам пробных функций, в качестве которых чаще всего используются полиномы, ортогональные на данном интервале с некоторым весом.

Среди спектральных методов выделяются два, которые обладают уникальными свойствами, выделяющими их из всех прочих. Это разложение периодических функций в ряды Фурье и разложение функций, аналитических в некоторой окрестности отрезка, по (смещенным) полиномам Чебышёва (что можно рассматривать как модификацию метода Фурье).

Принципиальное отличие этих двух методов от всех прочих спектральных состоит в том, что они одновременно являются также и сеточными, что гарантируется дискретным преобразованием Фурье или Фурье–Чебышёва.

Эти преобразования, которые также можно записать в матричной форме в явном виде, всегда хорошо обусловлены, являются (почти) инволютивными, и не добавляют погрешности к результату при вычислениях с конечной разрядной сеткой (см. [1]).

Второе важное отличие этих двух методов от (почти) всех остальных спектральных — это сходимость к решению не только в среднем, что гарантировано изначально, но также и равномерно на выбранном интервале при весьма слабых ограничениях на решение задачи.

Эти свойства позволяют по таблице полученного решения сделать заключение об аналитических свойствах решения, которые редко заранее известны. Иными словами, скорость убывания коэффициентов Фурье–(Чебышёва) находится в хорошо известной зависимости от гладкости решения. Например, для аналитических функций коэффициенты Фурье–(Чебышёва) убывают экспоненциально быстро.

Для решений, сингулярных в концах интервала, коэффициенты Фурье–(Чебышёва) убывают обычно все же достаточно быстро, чтобы метод был эффективным.

Поэтому, получив таблицу коэффициентов Фурье решения, можно сделать ряд важных заключений о свойствах полученного решения, которые невозможно сделать в случае применения сеточных методов. Помимо выводов о гладкости решения, можно достоверно оценить достигнутую глобальную точность, а также оценить достаточность (или нет) выбранной размерности аппроксимации при данной разрядной сетке.

Изложенные соображения были положены К.И. Бабенко в основу т.н. *методов без насыщения* (см. [2]).

К сожалению, Бабенко не дал в своей книге простого и понятного определения этих методов, а явление насыщения излагал в рамках общей теории аппроксимаций. Это послужило поводом к большому количеству спекуляций на тему методов без насыщения в русскоязычной литературе.

Применение спектральных методов часто сопряжено с большими техническими трудностями, связанными с учетом краевых условий задачи. Например, книга [1] дает практически все о полиномах Чебышёва, что нужно при решении практических задач. Однако в примерах решения задач учет краевых условий там осуществляется путем подбора вручную уравнений для нужных коэффициентов. Похожий подход применялся в [3], р. 119.

Эта же проблема возникает в т.н. τ -методе Ланцоша (см. [4]), который также является по сути спектральным. Учет краевых условий в τ -методе проводится путем добавления уравнений для некоторых малых параметров τ_i , $i = 1, 2, \dots$, а также выбрасывания некоторых уравнений для коэффициентов разложения решения из полученной переопределенной системы, но каждый раз в индивидуальном порядке.

Наиболее общий метод учета краевых условий — это обращение дифференциального оператора и запись задачи в интегральной форме. Однако обращение дифференциального оператора с учетом граничных условий — это весьма нетривиальная и очень трудоемкая задача, которая каждый раз требует индивидуального подхода.

Другой (и весьма малоизвестный) способ учета краевых условий был предложен Бабенко (см. [2]).

Бабенко предложил способ аппроксимации некоторых типовых дифференциальных операторов с учетом граничных условий так, как если бы аппроксимировались интегральные. Далее нужный дифференциальный оператор собирается из имеющихся прототипов так, что его граничные условия достаточно хорошо аппроксимируются (подробности см. в [5]).

Однако и этот способ учета краевых условий задачи является весьма трудоемким и имеющим ограниченное применение (см. ниже).

Таким образом, проблема учета краевых условий в спектральных методах остается (с практической точки зрения) открытой и пока не имеет общего и эффективного решения.

В этой статье мы даем, в определенном смысле, полное решение этой задачи для конкретного спектрального метода — разложения функций на интервале $[0, 1]$ по смещенным полиномам Лежандра (т.е. с единичным весом).

Выбор этих полиномов обусловлен тем, что разложения по ним также сходятся равномерно (при определенных условиях), т.е. полученные аппроксимации будут давать методы без насыщения.

Вторая причина, по которой мы выбрали эти полиномы — это наличие приложений данного метода к аналитической теории чисел, что возможно при выборе рациональной арифметики в расчетах и использовании компьютерной алгебры (CAS).

Учет краевых условий в разложениях по полиномам Чебышёва, на самом деле, осуществляется таким же образом. Однако эти разложения больше подходят для численных расчетов, и им будет посвящена отдельная статья.

В следующем разделе мы приведем готовые формулы для некоторых типовых дифференциальных операторов, а также покажем, как моделируются (в принципе) любые алгебраические линейные дифференциальные операторы с любыми краевыми условиями для них. В этом смысле наш подход напоминает (чисто внешне) метод Бабенко.

В последующих разделах мы дадим ряд примеров приложений нашего метода, эффективность которого особенно видна при использовании рациональной арифметики.

2. АРИФМЕТИКА РАЗЛОЖЕНИЙ ПО ПОЛИНОМАМ ЛЕЖАНДРА

Везде в этой статье рассматриваются аналитические функции на отрезке $[0, 1]$, которые могут иметь особенности на концах интервала. Эти функции раскладываются в ряды по смещенным полиномам Лежандра, и эти разложения сходятся (вообще говоря) равномерно на любом подынтервале отрезка $[0, 1]$.

Так же как и при вычислениях с обычными числами, мы всегда имеем дело лишь с конечным представлением функций в виде отрезков рядов. Поэтому необходимо определить операции на данном множестве функций так, чтобы разложения автоматически обрывались нужным образом. Для сложения и вычитания функций это очевидно, однако умножение, дифференцирование и интегрирование функций представляют некоторую проблему, которой мы и займемся.

Пусть $\tilde{P}(n, x)$ — это обычные полиномы Лежандра, ортогональные на отрезке $[-1, 1]$ с единичным весом. Мы используем смещенные полиномы Лежандра, $P(n, x) = \tilde{P}(n, 2x - 1)$, ортогональные на отрезке $[0, 1]$,

$$P(n, x) = \frac{1}{n!} \frac{d^n}{dx^n} (x^n (x-1)^n), \quad \int_0^1 P^2(n, x) dx = \frac{1}{2n+1}, \quad n \in \mathbb{N}_0.$$

Пусть N — это размерность аппроксимации, т.е. любая функция $f(x)$ в этом разделе понимается как отрезок ее разложения по полиномам $P(n, x)$ до степени $N-1$, т.е. $f \in \mathcal{P}_N$. Это представление функций аналогично представлению чисел в арифметике с фиксированной разрядной сеткой. Ясно, что дифференцирование и интегрирование таких функций нарушают это представление и требуют переразложения результата операции по полиномам $P(n, x)$ до номера $N-1$ включительно. Это же справедливо для умножения функций. Например,

$$x \sum_{k=0}^{N-1} a_k P(k, x) = \sum_{k=0}^{N-1} b_k P(k, x),$$

где равенство понимается в проективном смысле, т.е. при известных коэффициентах a_n коэффициенты b_n необходимо определить так, чтобы разность этих функций была ортогональна полиномам $P(n, x)$, $n < N$.

Таким образом, любая функция $f(x)$ представлена N -мерным вектором $f = \langle a_0, a_1, \dots, a_{N-1} \rangle^t$, а операции умножения на x , дифференцирование и т.д. — это линейные операторы в N -мерном векторном пространстве.

Мы будем обозначать векторы обычными, а линейные операторы — большими латинскими буквами. Введем следующие обозначения:

e — единичный вектор, т.е. $e(x) = 1$, $e = \langle 1, 0, \dots, 0 \rangle^t$;

E — единичная матрица;

X — оператор умножения на x ;

$X2$ — оператор умножения на x^2 ;

x — вектор, соответствующий функции x , т.е. $x = \langle \frac{1}{2}, \frac{1}{2}, 0, \dots, 0 \rangle^t$;

D — оператор дифференцирования по x ;

I — оператор формального интегрирования по x .

Таким образом,

$$x = X.e, \quad f'(x) = D.f, \quad \int f(x) dx = I.f, \quad \text{и т.д.,}$$

где точка обозначает применение оператора к вектору либо умножение матриц, а интеграл неопределенный, т.е. константа интегрирования неизвестна и может меняться при изменении N .

Заметим также, что $X.X \neq X2$, хотя эти две матрицы отличаются только в одном элементе, $(X2 - X.X)[N, N] = N^2/(4N^2 - 1)/4$. Однако оператор $X2$ все же лучше в численных (но не символьных) расчетах, чем $X.X$, но не настолько, чтобы вводить отдельные операторы $X3$, $X4$ и т.д.

Мы приведем готовые формулы для всех перечисленных операторов. Вывод этих формул весьма громоздок, и мы его опускаем. Сами формулы легко проверяются и могут быть доказаны с помощью математической индукции.

Матрицы X и $X2$ являются соответственно трех- и пятидиагональными. Приведем для примера эти матрицы для размерности $N = 6$:

$$X = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{5} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{2} & \frac{3}{14} & 0 & 0 \\ 0 & 0 & \frac{3}{10} & \frac{1}{2} & \frac{2}{9} & 0 \\ 0 & 0 & 0 & \frac{2}{7} & \frac{1}{2} & \frac{5}{22} \\ 0 & 0 & 0 & 0 & \frac{5}{18} & \frac{1}{2} \end{bmatrix}, \quad X2 = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{30} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{2}{5} & \frac{1}{5} & \frac{3}{70} & 0 & 0 \\ \frac{1}{6} & \frac{1}{3} & \frac{8}{21} & \frac{3}{14} & \frac{1}{21} & 0 \\ 0 & \frac{1}{10} & \frac{3}{10} & \frac{17}{45} & \frac{2}{9} & \frac{5}{99} \\ 0 & 0 & \frac{3}{35} & \frac{2}{7} & \frac{29}{77} & \frac{5}{22} \\ 0 & 0 & 0 & \frac{5}{63} & \frac{5}{18} & \frac{44}{117} \end{bmatrix}.$$

Для матрицы X диагонали образуют последовательности, начиная с нижней, соответственно,

$$\frac{n}{2(2n-1)}, \frac{1}{2}, \frac{n}{2(2n+1)}, \quad n \in \mathbb{N}.$$

Для матрицы $X2$ диагонали образуют последовательности, начиная с нижней, соответственно,

$$\frac{n(n+1)}{4(2n-1)(2n+1)}, \frac{n}{2(2n-1)}, \frac{3n^2-3n-2}{2(2n+1)(2n-3)}, \frac{n}{2(2n+1)}, \frac{n(n+1)}{4(2n+3)(2n+1)}, \quad n \in \mathbb{N}.$$

Проще всего устроена матрица дифференцирования D , а сложнее всего — матрица интегрирования I . Эти матрицы для размерности $N = 6$ имеют вид:

$$D = \begin{bmatrix} 0 & 2 & 0 & 2 & 0 & 2 \\ 0 & 0 & 6 & 0 & 6 & 0 \\ 0 & 0 & 0 & 10 & 0 & 10 \\ 0 & 0 & 0 & 0 & 14 & 0 \\ 0 & 0 & 0 & 0 & 0 & 18 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad I = \begin{bmatrix} \frac{1}{2} & -\frac{1}{6} & -\frac{1}{4} & \frac{5}{16} & \frac{3}{16} & -\frac{7}{16} \\ \frac{1}{2} & 0 & -\frac{1}{10} & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & -\frac{1}{14} & 0 & 0 \\ 0 & 0 & \frac{1}{10} & 0 & -\frac{1}{18} & 0 \\ 0 & 0 & 0 & \frac{1}{14} & 0 & -\frac{1}{22} \\ 0 & 0 & 0 & 0 & \frac{1}{18} & 0 \end{bmatrix}.$$

Элемент с индексами $[m, n]$ в верхнетреугольной матрице D имеет вид

$$D[m, n] = \begin{cases} 2(2m-1), & m+n \text{ нечетно,} \\ 0, & m+n \text{ четно.} \end{cases}$$

Матрица I является трехдиагональной, за исключением первой строки. Элементы главной диагонали нулевые. Ненулевой элемент с индексами $[m, n]$ в матрице I имеет вид

$$I[m, n] = \begin{cases} c(n), & m=1, \\ \frac{1}{2(2m-3)}, & m-n=1, \\ \frac{-1}{2(2m+1)}, & n-m=1, \end{cases}$$

где функция номера столбца $c(n)$ дается формулой

$$c(n) = \begin{cases} -\frac{1}{6}, & n = 2, \\ (-1)^{\frac{n-1}{2}} \Gamma\left(\frac{n}{2}\right) / \Gamma\left(\frac{n+1}{2}\right) / 2/\sqrt{\pi}, & n \text{ нечетно}, \\ (-1)^{\frac{n}{2}} 2\Gamma\left(\frac{n+3}{2}\right) / \Gamma\left(\frac{n}{2} - 1\right) / (n-1)/n/\sqrt{\pi}, & n \text{ четно}. \end{cases}$$

Матрица X размерности N обладает важным свойством, которое нам понадобится. Справедливо

Предложение 1. Матрица $P(N, X)$ нулевая, $N = 1, 2, \dots$

Доказательство. Введем вектор $p = \langle P(0, x), P(1, x), \dots, P(N-1, x) \rangle^t$. Для трехдиагональной матрицы Якоби J размерности N , ассоциированной с системой ортогональных полиномов Лежандра $\{P(n, x), n = 0, 1, \dots\}$, справедливо тождество

$$xp = J.p + \left\langle 0, 0, \dots, \frac{N}{2(2N-1)} P(N, x) \right\rangle^t, \quad (1)$$

которое является прямым следствием трехзвенного линейного рекуррентного соотношения для этих полиномов (см. [6]).

Результат умножения любого вектора $f \in \mathcal{P}_N$ на x можно записать как скалярное произведение этого вектора с левой частью (1). Поэтому транспонируя (1), получим $X = J^t$.

Но спектр матрицы J — это корни полинома $P(N, x)$, что следует из (1). Поэтому матрица J , а значит, и X аннулируются своим характеристическим полиномом $P(N, x)$ (теорема Кэли–Гамильтона). Что и требовалось доказать.

Как следствие, получаем,

$$P(n, x) = \frac{4^n \Gamma(n+1/2)}{\sqrt{\pi} n!} \det(xE - X), \quad n \in \mathbb{N}.$$

Также при нечетном N все матрицы $P(M, X)$ для нечетных M вырождены (но ненулевые). Это связано с тем, что у всех полиномов $P(M, x)$ нечетной степени есть общий рациональный корень $x = 1/2$. Для четных N все матрицы $P(M, X)$, $M \neq N$ невырождены.

Обратим внимание, что последняя строка матрицы дифференцирования D всегда нулевая, т.е. этот оператор всегда вырожден, а его матрица имеет ранг $N-1$. Это прямое следствие того, что ОДУ $y'(x) = f(x)$ имеет одномерное многообразие решений. Иными словами, для однозначного определения решения $y(x)$ необходимо задать начальное (либо краевое) условие.

Эти условия можно учесть совершенно естественным образом в последней строке матрицы D . Поскольку

$$P(k-1, 0) = (-1)^{k+1}, \quad P(k-1, 1) = 1, \quad k \in \mathbb{N},$$

то последнюю строку матрицы D надо заполнить этими значениями для учета, соответственно, начального (т.е. при $x = 0$) или краевого (т.е. при $x = 1$) условия (или их линейной комбинации). Затем нужно заменить последний элемент вектора f в правой части на нужное начальное (краевое) условие. Тогда полученная линейная система будет невырожденной, а полученное решение $y(x)$ будет автоматически иметь нужное начальное (краевое) условие.

Последнее утверждение можно сформулировать точно. Справедливо

Предложение 2. Пусть правая часть $f(x)$ ОДУ $y'(x) = f(x)$ имеет вид

$$f(x) = \sum_{k=0}^{N-2} a_k P(k, x), \quad (2)$$

т.е. последний элемент вектора $f = f(x)$ (который заменяется на граничное условие) равен нулю. Тогда решение $y(x)$ находится точно в виде полинома описанным выше способом при любом начальном (краевом) условии или их линейной комбинации.

Доказательство. Для того, чтобы ОДУ $y'(x) = f(x)$ имело решение $y \in \mathcal{P}_N$ необходимо и достаточно, чтобы правая часть f имела вид (2). Любое линейное краевое условие учитывается подбором константы интегрирования, для чего достаточно одного линейного уравнения. Что и требовалось доказать.

Можно дать еще одну интерпретацию описанного способа решения ОДУ $y'(x) = f(x)$ с полиномиальной правой частью.

Как было отмечено, ранг матрицы аппроксимации оператора D равен $N - 1$. Поэтому если решать полученную вырожденную линейную систему методом Гаусса с выбором главного элемента по всей матрице, то на последнем шаге N получится уравнение $0 \times \tilde{y}[N] = \tilde{f}[N]$, где тильда символизирует операции, проделанные в методе Гаусса. Последнее уравнение может быть совместно, только если $\tilde{f}[N] = 0$, т.е. $f(x)$ — это полином вида (2).

Практически можно взять любое значение $\tilde{y}[N]$, а затем скорректировать начальное (краевое) условие полученного решения, как требуется. Эта операция полностью эквивалентна неопределенному интегрированию функции, т.е. третий способ решения ОДУ $y'(x) = f(x)$ — это вычисление решения $y = I.f$, где вектор f вида (2), с последующей корректировкой константы интегрирования.

В операторе $D^2 = D.D$ будут, очевидно, две нулевые последние строки, которые следует заполнить нужными краевыми условиями. Вообще, для учета краевых условий нашим методом необходимо и достаточно знать следующие векторы-строки, линейные комбинации которых заполняют последние строки нужных линейных дифференциальных операторов. Эти векторы размерности N имеют вид

$$i_m = \langle P^{(m)}(k-1, 0) \rangle, \quad b_m = \langle P^{(m)}(k-1, 1) \rangle, \quad k = 1, \dots, N, \quad m \in \mathbb{N}_0,$$

где m обозначает номер производной полинома $P(n, x)$ по x . Векторы i_m отвечают за граничные значения функции при $x = 0$, а b_m — за граничные значения при $x = 1$. Эти векторы устроены весьма просто:

$$\begin{aligned} i_0 &= \langle (-1)^{k+1} \rangle, & b_0 &= \langle 1 \rangle, \\ i_1 &= \langle (-1)^k k(k-1) \rangle, & b_1 &= \langle k(k-1) \rangle, \\ i_2 &= \left\langle \frac{(-1)^{k+1} k(k-2)(k^2-1)}{2} \right\rangle, & b_2 &= \left\langle \frac{k(k-2)(k^2-1)}{2} \right\rangle, \\ &\dots & & \end{aligned}$$

причем этот список может быть легко продолжен (с использованием CAS).

Любые начальные либо краевые линейные условия для функций $f(x)$, $f'(x)$, $f''(x)$ и т.д. и их комбинаций получаются в виде линейных комбинаций перечисленных векторов.

Аналогично можно задать условия и внутри интервала, но у нас пока нет содержательной задачи для демонстрации этого факта.

Приведем численный пример, на котором видны преимущества данного метода. Рассмотрим спектральную задачу на интервале $x \in [0, 1]$:

$$\frac{d^2 y}{dx^2} + a^2 y = 0, \quad y(0) = 0, \quad y(1) + y'(1) = 0, \quad (3)$$

с нестандартными краевыми условиями. Эта задача приводилась в [7] для демонстрации приложения τ -метода.

Спектральная задача (3) решается в явном виде. Несложные вычисления показывают, что все собственные значения являются решениями трансцендентного уравнения $x + \tan x = 0$, т.е.

$$a_1 \approx 2.028757838110, \quad a_2 \approx 4.913180439434, \quad a_3 \approx 7.978665712413, \dots$$

Если решать эту задачу методом Бабенко, то необходимо сперва сконструировать интерполяционный полином Чебышёва с указанными краевыми условиями. Затем этот полином нужно дважды продифференцировать и получить таким образом конечномерную спектральную задачу. Мы проделали эти вычисления, но они весьма громоздки, и мы их опускаем.

Согласно нашему методу, нужно взять матрицу $A = D.D$ и заменить ее две последние строки на векторы i_0 и $b_0 + b_1$ (в любом порядке). Полученная спектральная задача размерности $N = 10$ дает приближенные значения a_1, a_2, a_3 , соответственно, с шестью, тремя и одним десятичными разрядами при вычислениях в обычной арифметике, что соответствует точности, полученной для этой задачи в [7] с помощью τ -метода той же размерности.

При увеличении N количество полученных собственных значений увеличивается, а их точность возрастает, т.е. спектр стабилизируется. Никаких “паразитных” собственных значений в нашем методе не появляется, т.е. данный метод является методом без насыщения.

В модификации τ -метода, которая использовалась в [7], для учета краевых условий к разложению функции добавляются дополнительные члены, уравнения для коэффициентов которых получаются из краевых условий. Это требует весьма громоздких вычислений, предназначенных только для данного конкретного случая. В результате для системы (3) τ -метод в [7] дает спектральную задачу для пучка матриц, т.е. $Ay = \lambda By$, где матрицы A и B имеют блочную структуру, собранную из отдельных кусков.

Отметим также, что, помимо учета краевых условий, наш подход схож с операторным подходом к τ -методу, предложенному в [8]. Однако в [8] используются бесконечные матрицы, которые обрезаются нужным образом для получения конечномерных аппроксимаций операторов. Затем эти конечномерные матрицы обрамляются дополнительными блоками (примерно как в [7]) для учета краевых условий.

3. КВАДРАТУРЫ ГАУССА И ПАДЕ-АППРОКСИМАЦИИ

Вычислительные эксперименты показали (неожиданно высокую) численную эффективность предложенного алгоритма. При расчетах в обычной плавающей арифметике обнаруженные нами эффекты ненаблюдаемы, и, вероятно, поэтому они (насколько нам известно) никогда ранее не отмечались в спектральных методах.

Вычисления с расширенной разрядной сеткой, а также в рациональной арифметике, проделанные нами в CAS Maple, позволили установить связь нашего метода с квадратурами Гаусса и паде-аппроксимациями функций.

Напомним, что в классической M -точечной квадратуре Гаусса с единичным весом на интервале $[-1, 1]$ требуется вычислить M корней полинома Лежандра $\tilde{P}(M, x)$ и M весов квадратурной формулы (см. [9]). На самом деле достаточно вычислить половину из каждого набора ввиду их симметрии, но проще вычислить их все сразу. Это возможно сделать только в плавающей арифметике. Тогда полученная квадратура будет точна на полиномах степени не выше $2M - 1$.

Поскольку при нашем подходе функция приближается полиномами Лежандра степени не выше $N - 1$, то можно было ожидать, что данный метод будет соответствовать квадратуре Гаусса порядка $[N/2]$. Приведем пример, показывающий, что точность на самом деле значительно выше.

Рассмотрим задачу Коши

$$y'(x) = \frac{1}{1+x^2} = r(x), \quad x \in [0, 1], \quad y(0) = 0, \quad (4)$$

решением которой является функция $y(x) = \arctan x$.

Выберем для этого примера размерность аппроксимации $N = 10$.

Правая часть уравнения (4) может быть сразу записана в виде вектора $r = (E + X^2)^{-1}e$ (см. теорему ниже), т.е. мы получаем приближенное разложение рациональной функции $r(x)$ по сдвинутым полиномам Лежандра, просто обратив нужную матрицу. Обозначим точные коэффициенты разложения через $\tilde{r}_k$:

$$\tilde{r}_k = (2k - 1) \int_0^1 r(x) P(k - 1, x) dx, \quad k = 1, 2, \dots, N.$$

Тогда

$$|r_k - \tilde{r}_k| \approx \{3 \times 10^{-15}, 3 \times 10^{-15}, 1 \times 10^{-13}, 6 \times 10^{-13}, 6 \times 10^{-13}, 2 \times 10^{-11}, 6 \times 10^{-11}, 2 \times 10^{-10}, 2 \times 10^{-9}, 4 \times 10^{-9}\}. \quad (5)$$

Поскольку

$$\tilde{r}_1 = \int_0^1 r(x)P(0, x)dx = \int_0^1 \frac{1}{1+x^2} dx = \frac{\pi}{4}, \quad (6)$$

то мы вычислили интеграл от правой части (4), даже не решая дифференциального уравнения. Имеем

$$r_1 = \frac{281845376409124}{358856678744865}, \quad \frac{\pi}{4} - r_1 \approx 3.057054520150 \times 10^{-15}. \quad (7)$$

Уместно сравнить полученную точность с той, которую можно достичь с помощью квадратур Гаусса при вычислении интеграла (6). Обозначим соответствующие квадратуры G_M . Тогда

$$\frac{\pi}{4} - G_{10} \approx -6.3 \times 10^{-14}, \quad \frac{\pi}{4} - G_{11} \approx 3.056102 \times 10^{-15}, \quad \frac{\pi}{4} - G_{12} \approx -4.8 \times 10^{-17},$$

т.е. наш метод соответствует (в данном примере) 11-точечной квадратуре Гаусса.

Теперь вычислим решение задачи Коши (4) согласно нашему методу. Возьмем матрицу $A = D$ и заменим последнюю (нулевую) строку на вектор i_0 . Далее заменим последний (в данном случае ненулевой) элемент вектора r на ноль, согласно начальному значению решения. Получаем конечномерную аппроксимацию задачи (4), $A \cdot y = r$, откуда находим вектор $y = A^{-1} \cdot r$. Таким образом, получаем приближенное разложение арктангенса

$$y(x) = \arctan x \approx \sum_{k=1}^N y_k P(k-1, x).$$

Подставив $x = 1$ в эту формулу, получаем рациональное приближение числа $\pi/4$, которое оказывается в точности равным значению r_1 в (7), вычисленному ранее.

Вычислим паде-аппроксимации арктангенса различной размерности в рациональной арифметике, подставим туда $x = 1$ и сравним с приближением числа $\pi/4$, полученным нами. Оказалось, что

$$P[19,18](y(1)) = \frac{270807143217152}{344802363740835}, \quad P[19,18](y(1)) - \frac{\pi}{4} \approx 3.672520113 \times 10^{-15}.$$

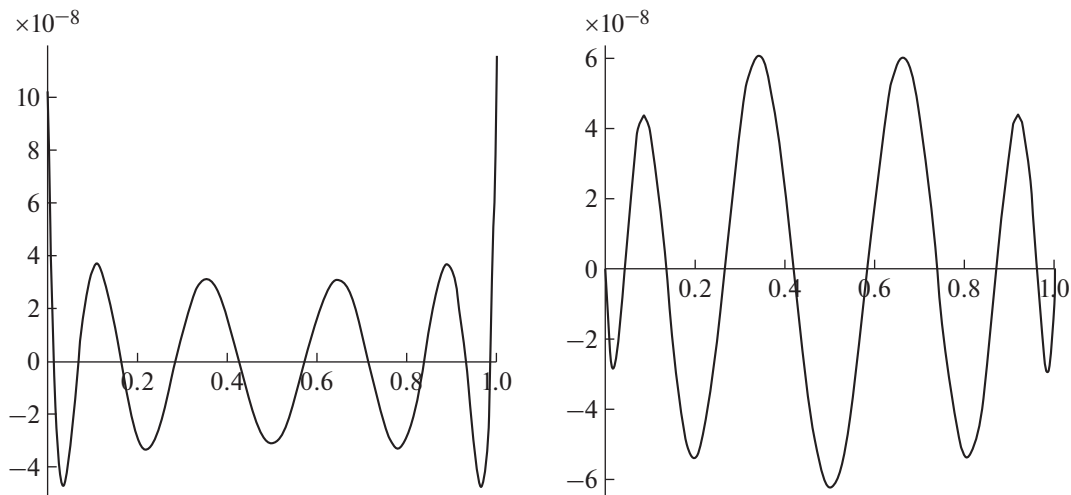
Как мы увидим, близость (и даже схожесть) полученных аппроксимаций не является случайной.

Коэффициенты y_k отличаются от точных значений разложения арктангенса примерно так же, как в (5). На фиг. 1 показаны графики разности $\arctan x$ с их аппроксимациями полиномами Лежандра для $N = 10$. На фиг. 1 слева использовались точные (иррациональные) коэффициенты разложения, а на фиг. 1 справа — полученные нами.

Обратим внимание, что глобальная (или равномерная) погрешность нашей аппроксимации примерно соответствует ожидаемой для данной размерности (и даже лучше). Однако краевое значение в задаче (4) находится с точностью, превышающей равномерную на несколько порядков.

Последнее обстоятельство также не является случайным. Чтобы это увидеть, решим задачу (4) нашим методом еще раз, но без использования матрицы $X2$. Сохраним те же обозначения, но возьмем правую часть уравнения (4) в виде вектора $r = (E + X \cdot X)^{-1} \cdot e$. Тогда

$$|r_k - \tilde{r}_k| \approx \{6 \times 10^{-14}, 4 \times 10^{-13}, 7 \times 10^{-13}, 1 \times 10^{-11}, 4 \times 10^{-11}, 2 \times 10^{-10}, 2 \times 10^{-9}, 3 \times 10^{-9}, 3 \times 10^{-8}, 2 \times 10^{-7}\},$$



Фиг. 1. Аппроксимация $\arctan x$ полиномами Лежандра ($N = 10$).

т.е. точность аппроксимации правой части (4) несколько хуже, чем ранее. Это же справедливо для нового вектора y . Однако теперь

$$r_1 = \frac{140675049238504}{179113035648015}, \quad \frac{\pi}{4} - r_1 \approx -6.326259326040 \times 10^{-14}, \quad (8)$$

что соответствует 10-точечной квадратуре Гаусса (см. выше).

Данное соответствие нашего рационального приближения интеграла его квадратуре Гаусса на самом деле является полным совпадением.

Возьмем произвольную рациональную функцию $r(x)$ с рациональными коэффициентами (можно и комплексно-рациональными или даже символьными). Рассмотрим задачу Коши

$$z'(x) = r(x) = \frac{p(x)}{q(x)}, \quad x \in [0, 1], \quad z(0) = 0. \quad (9)$$

Ее решением является функция $z(x) = \int_0^x r(t)dt$, если она существует. Но нас интересуют формальные символьные вычисления в конечномерной аппроксимации.

Для задачи (9) в случае общего положения (см. ниже) вычислим следующие массивы для размерностей аппроксимации $N = 1, 2, \dots$

$S_1[N] = r_1$, где вектор $r = \langle r_1, \dots, r_N \rangle^t$ вычисляется по формуле

$$r = p(X).q(X)^{-1}e = q(X)^{-1}p(X).e,$$

где коммутирующие матрицы $p(X)$ и $q(X)$ получаются подстановкой матрицы X вместо независимой переменной x в полиномы $p(x)$ и $q(x)$. Напомним, что X^2 мы теперь не используем.

$S_2[N] = \sum_{k=1}^N z_k$, где вектор z вычисляется так же, как вектор y для (4). То есть берем матрицу $A = D$, заменяем последнюю (нулевую) строку на вектор i_0 . Далее заменяем последний (ненулевой) элемент вектора r на ноль, т.е. берем $r[N] = 0$, согласно начальному значению решения. Затем вычисляем вектор $z = A^{-1}.r$, который дает приближенное разложение функции $z(x)$, если она существует. Иначе это просто формальный вектор. Отметим, что $S_2[1] = 0$ в силу способа вычисления.

Этот массив можно также вычислять альтернативным способом с помощью оператора формального интегрирования, как это описано в разд. 2.

$S_3[N] = G_N(r)$, т.е. $S_3[N]$ — это значение N -точечной квадратуры Гаусса функции $r(x)$ на интервале $[0, 1]$.

Напомним, что корни x_n полиномов Лежандра $P(N, x)$ могут быть вычислены (в принципе) в виде радикалов только до $N = 9$ включительно. Это же относится к весам w_n квадратурной формулы Гаусса,

$$w_n = \frac{4x_n(1-x_n)}{N^2 P^2(N-1, x_n)}, \quad n = 1, \dots, N. \quad (10)$$

Разумеется, практически делать это бессмысленно, поэтому массив $S_3[N]$ (в отличие от предыдущих) дается в плавающей арифметике, но с любым (в принципе) числом десятичных разрядов.

Наложим очевидное ограничение на полюса функции $r(x)$, т.е. если $q(x)$ делится на полином Лежандра $P(M, x)$, то в массивах S_1, S_2, S_3 следует пропустить размерность $N = M$ (см. предложение 1). А в случае если $q(x)$ имеет корень $x = 1/2$, то эти массивы вычисляются только для четных N .

Заметим, что функция $r(x)$ не может иметь полюс ($\neq 1/2$) в одном из корней полинома $P(M, x)$ и не иметь полюсов во всех остальных корнях.

С учетом этих условий справедлива

Теорема. Массивы $S_k, k = 1, 2, 3$, совпадают.

Доказательство. Напомним, что $x = Xe$ по определению, т.е. вектор $x \in \mathcal{P}_N$, соответствующий функции x , получается применением матрицы X к вектору e . Аналогично получается вектор $p = p(X)e$, соответствующий любому полиному $p(x)$. Но эти же рассуждения применимы для общей функции от матрицы, $f(X)$, т.е. вектор $f = f(X)e$ — это коэффициенты приближенного разложения функции $f(x)$ по смещенным полиномам Лежандра при условии, что матрица $f(X)$ определена.

Функция от матрицы $f(X)$ однозначно определяется значениями функции $f(x)$ на спектре матрицы X (см. [10]), т.е. значениями этой функции в корнях полинома $P(N, x)$ (так как матрица X диагонализирована, т.е. производные $f(x)$ не нужны).

Для рациональной функции $r(x)$ не требуется вычислять ее значения на спектре матрицы X , а можно формально подставить матрицу X вместо переменной x (см. [10]). Поэтому совпадение массивов S_1 и S_2 следует из совпадения двух интегралов.

Величина $S_1[N]$ — это первая компонента вектора правой части (9), т.е. нулевой коэффициент ее аппроксимации полиномами $P(n, x)$. Иными словами, это аппроксимация интеграла $r(x)$ на отрезке $[0, 1]$. Величина $S_2[N]$ получается из аппроксимации интеграла $\int_0^x r(t)dt$ как функции на отрезке $[0, 1]$ подстановкой $x = 1$ в эту функцию, т.е. $S_1 = S_2$.

Вторая часть утверждения теоремы, т.е. $S_1 = S_3$, следует из представления матричной функции $r(X)$ интерполяционным полиномом Лагранжа—Сильвестра (см. [10]). Имеем

$$r(X) = \sum_{n=1}^N r(x_n) l_{N,n}(X), \quad l_{N,n}(x) = \prod_{k \neq n} \frac{x - x_k}{x_n - x_k},$$

где $\{x_n, n = 1, \dots, N\}$ — это спектр матрицы X (корни полинома $P(N, x)$), а $l_{N,n}(x)$ — это фундаментальные полиномы лагранжевой интерполяции.

Поскольку $l_{N,n}(x) \in \mathcal{P}_N$, то

$$l_{N,n}(X).e[1] = \int_0^1 l_{N,n}(x) dx = w_n,$$

где w_n — это веса (10) квадратурной формулы Лежандра—Гаусса (см. [9]). Таким образом, вектор правой части (9) представим в виде

$$r(X).e = \sum_{n=1}^N r(x_n) l_{N,n}(X).e,$$

поэтому $S_1 = r(X).e[1]$ — это квадратура Лежандра—Гаусса функции $r(x)$. Что и требовалось доказать.

Следствие. Любая квадратура Лежандра–Гаусса на интервале $[0, 1]$ от любой рациональной функции $r(x)$ (с учетом сделанных оговорок) является рациональным числом, т.е.

$$\sum_{n=1}^N w_n r(x_n) \in \mathbb{Q}. \quad (11)$$

Этот результат может быть (относительно легко) проверен в CAS до размерности $N = 5$ и, в принципе, до размерности $N = 9$, хотя выражение для корня полинома $\tilde{P}(8, x)$ занимает страницу. Для $N > 9$ этот результат может быть проверен только численно, но зато с любой, в принципе, точностью.

Возьмем, например, $r(x) = 1/x$, для которой решение задачи Коши (9) $z(x)$ не существует. Но это не мешает нам вычислить квадратуры Гаусса, так как все корни x_n лежат внутри интервала. Получим в точности числа $2H(N)$, $N = 1, 2, \dots$, где $H(N)$ – это гармоническое число.

Возьмем теперь $r(x) = q/(1 + qx)$. Тогда $z(x) = \ln(1 + qx)$ является формальным решением задачи Коши (9). По нашему алгоритму получим для $N = 1, 2, \dots$,

$$S_1[N] = \left\{ \frac{2q}{2+q}, \frac{3q(2+q)}{(6+6q+q^2)}, \frac{q(60+60q+11q^2)}{3(2+q)(10+10q+q^2)}, \frac{5q(2+q)(42+42q+5q^2)}{6(70+140q+90q^2+20q^3+q^4)}, \dots \right\}. \quad (12)$$

Квадратуры Гаусса (11) получаются весьма громоздкими и в плавающей арифметике, но при подстановке численных значений параметра q совпадают с рациональными функциями (12).

Формула (12) также показывает, почему полюс $x = 1/2$ при $q = -2$ неприемлем. Он соответствует корню $x = 0$ полиномов Лежандра нечетной степени. Других рациональных корней у полиномов Лежандра не существует.

Предложение 3. Последовательность рациональных приближений (12), т.е. $S_1[N]$, $N = 1, 2, \dots$, совпадает с диагональными паде-аппроксимациями $P[N, N](z(x))|_{x=1} = P[N, N](\ln(1 + q))$.

Доказательство. Докажем другое утверждение, которое равносильно этому согласно доказанной теореме. Последовательность квадратур Гаусса функции $r(x) = q/(1 + qx)$, т.е. $S_3[N]$, $N = 1, 2, \dots$, совпадает с этими паде-аппроксимациями. В самом деле,

$$G_N\left(\frac{q}{1+qx}\right) = \sum_{k=0}^{\infty} G_N((-1)^k q^{k+1} x^k)$$

в силу линейности квадратуры. Но квадратуры Гаусса точны на полиномах степени $M \leq 2N - 1$, поэтому

$$G_N\left(\frac{q}{1+qx}\right) = \sum_{k=0}^{2N-1} \frac{(-1)^k q^{k+1}}{k+1} + O(q^{2N+1}) = \ln(1 + q) + O(q^{2N+1}).$$

Слева здесь стоит рациональная функция от q нужного вида, т.е. отношение полиномов степени N . Что и требовалось доказать.

Например, подставив $q = 1$ в (12), получаем последовательность рациональных приближений для $\ln 2$:

$$\left\{ \frac{2}{3}, \frac{9}{13}, \frac{131}{189}, \frac{445}{642}, \frac{34997}{50490}, \frac{62307}{62307}, \frac{2359979}{3404730}, \frac{25786503}{37202060}, \dots \right\}, \quad (13)$$

которая совпадает с последовательностью, полученной в [11] с помощью ϵ -алгоритма Винна (Wynn). Пользуясь случаем, исправляем там опечатку: $(8/13 \rightarrow 9/13)$.

Эта последовательность соответствовала в [11] поддиагональной паде-аппроксимации, так как там рассматривалась функция $\ln(1 + x)/x$. Отметим также, что в [11] было выбрано неудачное преобразование эквивалентности, в результате чего полученная цепная дробь выглядела сложнее, чем она есть (см. разд. 4).

Совпадение паде-аппроксимаций функции, представленной рядом Стильеса, т.е. где коэффициенты разложения – это моменты интеграла Стильеса, с квадратурами Гаусса не является новым результатом (см. [12]) (авторы там перепутали $[n - 1, n]$ - и $[n, n - 1]$ -аппроксимации).

Таким образом, существенно нелинейная операция — вычисление диагональной паде-аппроксимации $\ln(1+x)$ с помощью ε -алгоритма Винна сводится нашим методом к линейным операциям с матрицами.

Также теперь ясно, почему не удалось получить точную паде-аппроксимацию для арктангенса. Дело в том, что наш алгоритм линеен, поэтому мы получили сумму двух различных паде-аппроксимаций. А именно,

$$\frac{1}{1+x^2} = \frac{1}{2(1+ix)} + \frac{1}{2(1-ix)}, \quad \int_0^x \frac{dt}{2(1+it)} = \frac{1}{2} \arctan x - \frac{i}{4} \ln(1+x^2) = v(x),$$

т.е. последовательность $\{S_i[N]\}$ соответствует сумме двух паде-аппроксимаций,

$$S_i[N] = P[N, N](v(x))\big|_{x=1} + \overline{P[N, N](v(x))\big|_{x=1}}, \quad N = 1, 2, \dots$$

Хотя любая рациональная функция представима в виде суммы простых дробей, следствие, которое мы вывели из теоремы, не следует из совпадения квадратур Гаусса с паде-аппроксимациями простых дробей по очевидным причинам.

Как показано в [11], из последовательности (13) следует иррациональность числа $\ln 2$ по критерию Дирихле, т.е.

$$|p - q \ln 2| \rightarrow 0,$$

где p/q — это рациональные приближения в (13). Этим свойством обладают некоторые (но далеко не все) рациональные последовательности, полученные с помощью паде-аппроксимаций, и мы займемся этим в следующем разделе.

Таким образом, наш метод дает весьма точные приближения коэффициентов разложения интегралов некоторых рациональных функций по полиномам Лежандра (см. фиг. 1).

Оказалось, что если эти коэффициенты исходно рациональны, то наш метод в ряде случаев вычисляет их точно в рациональной арифметике. Рассмотрим уравнение

$$xy'(x) = 1,$$

решением которого является функция $y(x) = \ln x$ при выборе краевого условия $y(1) = 0$. Это также первый пример, где мы рассматриваем более сложный оператор, чем просто дифференцирование. Однако способ решения этого уравнения точно такой же.

Берем матрицу $A = X \cdot D$ и заменяем последнюю строку на вектор b_0 . Последний элемент вектора $r = e$ здесь ноль. Вычисляем вектор $y = A^{-1} \cdot r$, который дает приближенное разложение $\ln x$. Однако в данном случае все коэффициенты разложения, кроме последнего, $y[N]$, оказываются точными. Таким образом, получаем

$$\ln x = -1 + \sum_{n=1}^{\infty} (-1)^{n-1} \frac{2n+1}{n(n+1)} P(n, x), \quad x \in (0, 1]. \quad (14)$$

Если бы мы решали уравнение $y'(x) = 1/x$ этим способом, что возможно, так как матрица X обратима, то мы получили бы почти такой же результат, т.е. все коэффициенты находятся правильно, кроме (по неизвестным пока причинам) предпоследнего.

Приведем еще один пример. Рассмотрим уравнение

$$xy'(x) - y(x) = x,$$

общим решением которого является функция $y(x) = x(C + \ln x)$. Поэтому возьмем краевое условие $y(1) = 0$ и решим эту задачу как предыдущую. Тогда при каждом N мы получим все коэффициенты разложения $y(x) = x \ln x$, кроме последнего, $y[N]$. Таким образом, получаем

$$x \ln x = -\frac{1}{4} + \frac{1}{12} P(1, x) + \sum_{n=2}^{\infty} (-1)^n \frac{2n+1}{n(n^2-1)(n+2)} P(n, x), \quad x \in [0, 1].$$

Полученные разложения могут быть использованы для решения неалгебраических ОДУ нашим методом (см. разд. 5). Хотя равномерная аппроксимация решений окажется при этом соответствующей ожидаемой, т.е. (вообще говоря) не может быть лучше, чем аппроксимация $\ln x$ по-

линомами Лежандра, но аппроксимация решений в концах интервала может быть очень хорошей, что мы уже наблюдали (см. фиг. 1).

4. ЦЕПНЫЕ ДРОБИ И ИРРАЦИОНАЛЬНОСТЬ

Как известно, быстро сходящиеся последовательности рациональных чисел играют ключевую роль в доказательствах иррациональности их предела (см. [11] и ссылки там).

При этом, как оказалось во всех без исключения случаях (а их можно пересчитать по пальцам одной руки), если последовательность $\{p_n/q_n\}$ дает иррациональность своего предела x (например, константы Апери) по критерию Дирихле, то она сходится к x быстрее (в обычном, численном смысле), чем обыкновенная цепная дробь числа x . А значит (см. [11]), эта последняя может быть однозначно восстановлена по последовательности $\{p_n/q_n\}$ с помощью функций $K(x, n)$, которые мы определили в [11].

Иными словами, в этом случае последовательность $K(p_n/q_n, n)$ совпадает с последовательностью подходящих дробей числа x , т.е. $K(p_n/q_n, n) = K(x, n)$, $n > \text{const}$.

Но это последнее свойство может также выполняться, если критерий Дирихле для последовательности $\{p_n/q_n\}$ не выполняется (см. [11]). Например, для последовательности $\{S_1[N] = r_1\}$ рациональных приближений числа $\pi/4$, которую мы получили в задаче (4), $K(S_1[N], N) = K(\pi/4, N)$, $N = 1, 2, \dots$, т.е. мы получаем все подходящие дроби числа $\pi/4$. При этом, например,

$$\frac{\pi}{4} - S_1[30] \approx 1.36 \times 10^{-40}, \quad \frac{\pi}{4} - K(\pi/4, 30) \approx 9.51 \times 10^{-31}.$$

Однако наибольший интерес все же представляют последовательности, в структуре которых удается выявить некоторые закономерности. Этим мы и займемся в этом разделе.

Последовательности диагональных, а также над- или поддиагональных паде-аппроксимаций и их комбинации (qd -алгоритм) некоторых функций могут давать очень хорошие рациональные приближения иррациональных чисел. Например, последовательность (13) дает иррациональность числа $\ln 2$ по критерию Дирихле (см. [11]).

Но у нас также имеется более общая последовательность (12). Как оказалось, эта последняя значительно более удобна для выявления закономерностей в ее структуре, чем ее частный вариант при $q = 1$. Это связано с тем, что параметр q не дает CAS приводить дроби к наименьшему знаменателю, что в данном случае контрпродуктивно.

Закономерность в последовательности (12) становится видна, если преобразовать ее в цепную дробь Эйлера (по формулам Д. Бернулли). Обозначим (12) (заменив там $q \rightarrow t$) как последовательность $s(n) = p_n/q_n$ подходящих дробей, где p_n и q_n — это, соответственно, числитель и знаменатель дробей (12). Вычислим

$$\begin{aligned} a_1 &= p_1, & a_n &= \frac{p_{n-1}q_n - p_nq_{n-1}}{p_{n-1}q_{n-2} - p_{n-2}q_{n-1}}, & n > 1, \\ b_1 &= q_1, & b_n &= \frac{p_nq_{n-2} - p_{n-2}q_n}{p_{n-1}q_{n-2} - p_{n-2}q_{n-1}}, & n > 1, \end{aligned} \quad (15)$$

где $p_0 = 0$ и $q_0 = 1$. Тогда

$$s(n) = \mathbf{K}_{k=1}^n \frac{a_k}{b_k},$$

или

$$\ln(1+t) = \frac{2t}{2+t - \frac{t^2}{6+3t - \frac{4t^2}{10+5t - \frac{9t^2}{14+7t - \frac{16t^2}{18+9t - \dots}}}}}, \quad (16)$$

где закономерность в частных дробях очевидна.

Хотя нам не удалось получить Паде-аппроксимации арктангенса с помощью нашего алгоритма напрямую, но применение формулы

$$\arctan x = \frac{i}{2} \ln \left(\frac{1-ix}{1+ix} \right)$$

в (16) дает такое разложение. После некоторых преобразований получаем известное разложение

$$\arctan x = \frac{x}{1 + \frac{x^2}{3 + \frac{4x^2}{5 + \frac{9x^2}{7 + \dots}}}}.$$

Применим наш подход к некоторым ОДУ, зависящим от параметра. Рассмотрим задачу Коши

$$y'(x) - ty(x) = 0, \quad y(0) = 1, \quad y(x) = \exp(tx),$$

где t — это независимый параметр. Это также пример, как формально однородное уравнение решается как неоднородное за счет начального условия, т.е. вектор в правой части $r = 0$, но затем $r[N] = 1$.

Применив наш алгоритм и преобразование (15), получим

$$e^t = \frac{t}{1 - \frac{2t}{2 + t + \frac{t^2}{6 + \frac{t^2}{10 + \frac{t^2}{14 + \frac{t^2}{18 + \frac{t^2}{22 + \dots}}}}}}}, \quad (17)$$

т.е. после начального куска цепная дробь становится “квазипериодической”.

Однако в данном случае мы получили феномен, который ранее (насколько нам известно) не наблюдался. То есть мы получили разложение, которое (при $t = 1$) содержит обыкновенную цепную дробь некоторой константы. Более того, дробь (17) (при $t = 1$) сходится на несколько порядков быстрее, чем обыкновенная цепная дробь константы e .

Данный феномен объясняется довольно просто. Избавившись от начального куска дроби (17) и построив ее сверху согласно очевидным закономерностям, получим известное разложение

$$\frac{e^t - 1}{e^t + 1} = \frac{t}{2 + \frac{t^2}{6 + \frac{t^2}{10 + \frac{t^2}{14 + \frac{t^2}{18 + \frac{t^2}{22 + \dots}}}}}}. \quad (18)$$

Таким образом, преобразование Мёбиуса с рациональными коэффициентами иррационального числа может радикальным образом изменить скорость сходимости обыкновенной цепной дроби новой константы по сравнению с исходной. Например,

$$e - K(e, 10) \approx 1.102 \times 10^{-7}, \quad \frac{e-1}{e+1} - K\left(\frac{e-1}{e+1}, 10\right) \approx 4.216 \times 10^{-26}.$$

Это свойство могло бы найти применение в доказательствах иррациональности, но (насколько нам известно) никогда не применялось.

Применим теперь формулу

$$i \tan x = \frac{\exp(2ix) - 1}{\exp(2ix) + 1}$$

в (18). Тогда (после преобразования эквивалентности) получим классическое разложение для тангенса:

$$\tan x = \frac{x}{1 - \frac{x^2}{3 - \frac{x^2}{5 - \frac{x^2}{7 - \frac{x^2}{9 - \dots}}}}}. \quad (19)$$

Это разложение Ламберт и Лежандр использовали в доказательстве иррациональности числа π (см. [13], р. 289–296). А именно, согласно (очень простому) критерию иррациональности Ламберта–Лежандра формула (19) может давать только иррациональные числа для рациональных x . Но $\tan(\pi/4) = 1$, поэтому π не может быть рациональным.

По тем же причинам из дроби (18) следует иррациональность $\exp(r)$ для $0 \neq r \in \mathbb{Q}$. Поэтому и $\ln(r)$ иррационально для $1 \neq r \in \mathbb{Q}$.

Заметим, что основную часть доказательства Лежандра в [13] занимал как раз вывод цепной дроби (19). Лежандр использовал для этого функции Бесселя, причем, по-видимому, до самого Бесселя. А мог бы, как теперь ясно, применить свои полиномы.

Приведем еще один пример. Рассмотрим задачу Коши

$$(1 + tx)y'(x) - sty(x) = st, \quad y(0) = 0, \quad y(x) = (1 + tx)^s - 1,$$

где s и t — это независимые параметры. Применив наш алгоритм и преобразование (15), получим

$$(1 + t)^s = 1 + \frac{st}{1 + \frac{(1-s)t}{2 - \frac{2(s+1)t}{st - 2t - 6 - \frac{(s^2-4)t^2}{10 + 5t + \frac{(s^2-9)t^2}{14 + 7t + \frac{(s^2-16)t^2}{18 + 9t \dots}}}}}}, \quad (20)$$

т.е., как и раньше, после начального куска цепная дробь становится “квазипериодической”. Избавившись от начального куска дроби (20) и построив ее сверху согласно очевидным закономерностям, получим разложение

$$\frac{(1+t)^s - 1}{(1+t)^s + 1} = \frac{st}{2 + t + \frac{(s^2-1)t^2}{6 + 3t + \frac{(s^2-4)t^2}{10 + 5t + \frac{(s^2-9)t^2}{14 + 7t + \frac{(s^2-16)t^2}{18 + 9t + \dots}}}}}, \quad (21)$$

которое весьма напоминает (18).

Применим теперь тождество $i^{-2i} = e^\pi$, т.е. сделаем подстановку $s = -2i$, $t = i - 1$ в (21). Тогда получим (после преобразования эквивалентности)

$$\frac{e^\pi - 1}{e^\pi + 1} = \frac{1}{2} \prod_{n=1}^{\infty} \frac{n^2 - 2n + 5}{2n - 1} = \frac{2}{1 + \frac{5}{3 + \frac{8}{5 + \frac{13}{7 + \frac{20}{9 + \dots}}}}}. \quad (22)$$

Цепная дробь для e^π была получена недавно в [14]. Причем, как полагает автор в [14], это было сделано впервые. Однако наша дробь устроена проще.

Заметим также, что цепная дробь (22) сходится к своему пределу медленнее, чем его подходящие дроби, хотя число $e^\pi = (-1)^{-i}$ трансцендентно (как алгебраическое число в степени мнимая квадратичная иррациональность (см. [15] и ссылки там)).

5. КОНСТАНТЫ ЭЙЛЕРА И ЭЙЛЕРА–ГОМПЕРТЦА

Как было отмечено в [16], константа Эйлера γ и константа Эйлера–Гомпертца δ имеют весьма похожие интегральные представления. В этом разделе мы продолжим эту аналогию.

Рассмотрим функцию

$$y(x) = \int_0^{\infty} \ln(x+t) \exp(-t) dt = \ln(x) + \exp(x) \operatorname{Ei}(1, x),$$

где $\operatorname{Ei}()$ — это интегральная экспонента. Функция $y(x)$ удовлетворяет ОДУ:

$$y'(x) = y(x) - \ln(x), \quad (23)$$

общим решением которого является функция

$$y(x) = \ln(x) + \exp(x) \operatorname{Ei}(1, x) + C \exp(x).$$

Легко проверить, что $y(0) = -\gamma + C$ и $y(1) = y'(1) = \delta + C \exp(1)$. Поэтому, к сожалению, из уравнения (23) нельзя получить сразу обе константы, γ и δ , а только их линейную комбинацию с константой e . Иными словами, задача Коши для уравнения (23), $y(0) = 0$, дает рациональные приближения для константы

$$\eta = \delta + e\gamma = \sum_{n=1}^{\infty} \frac{H(n)}{n!} \approx 2.1653822153269. \quad (24)$$

Эта формула следует из представления (см. [17])

$$\delta = e(F([1,1],[2,2],-1) - \gamma),$$

где $F()$ — это гипергеометрическая функция, т.е., как легко проверить,

$$\exp(x) F([1,1],[2,2],-x) = \sum_{n=1}^{\infty} \frac{H(n)x^{n-1}}{n!}.$$

Задача Коши для уравнения (23) решается точно так же, как мы уже делали неоднократно. Отличие состоит в том, что эта задача сингулярна ($y'(0) = +\infty$), а также что ОДУ не является алгебраическим.

Таким образом, решается линейная алгебраическая задача $Ay = r$, где $A = D - E$ с последующей корректировкой последней строки, а вектор r получается с помощью формулы (14), и $r[N] = 0$, согласно начальному условию.

Мы проделали эти вычисления в рациональной арифметике до размерности аппроксимации $N = 100$. Полученная в результате последовательность рациональных приближений $\{s(n), n = 1, \dots, 100\}$ константы η не дает иррациональности этой константы по критерию Дирихле, но приближает ее лучше в численном смысле, чем последовательность подходящих дробей

константы η , т.е., согласно [11], $K(\eta, n) = K(s(n), n)$, $n > 11$. Иными словами, получена последовательность подходящих дробей константы η начиная с 12-й. Например,

$$\eta - s(50) \approx 4.3461 \times 10^{-95}, \quad \eta - K(\eta, 50) \approx 1.8460 \times 10^{-59}. \quad (25)$$

Здесь мы воспользовались модификацией алгоритма, т.е. замедлили скорость сходимости полученной последовательности рациональных приближений с помощью функций $K(x, n)$, введенных в [11].

Данный модифицированный алгоритм (т.е. вычисление n -й подходящей дроби от n -го рационального приближения) обладает сертификатом качества (см. [11]). Он гарантированно дает подходящие дроби нужной константы либо не дает, о чем сигнализирует. При этом не надо знать численного значения данной константы, а все вычисления проводятся в рациональной арифметике.

В случае если алгоритм “дает сбой”, т.е. очередное рациональное приближение оказывается хуже, чем приближение подходящей дробью этой константы (что может быть, например, если в цепной дроби встретился нетипично большой знаменатель), то это никак не влияет на последующие подходящие дроби, если данный член просто опустить. Такой случай встречался в разложении $\zeta(6)$ (см. [11]). Полученная по формулам (15) цепная дробь просто окажется некоторым сжатием исходной.

Таким образом, свойство последовательности рациональных чисел $\{s(n)\}$ сходиться к своему пределу x быстрее (в численном смысле), чем последовательность подходящих дробей числа x , $\{K(x, n)\}$, является внутренним свойством данной последовательности и проверяется в рациональной арифметике независимо от знания ее предела x .

Воспользуемся этими соображениями для проверки скорости сходимости ряда (24). Этот ряд весьма напоминает ряд Энгеля (см. [18]), а такие ряды обычно сходятся к своему пределу быстрее, чем подходящие дроби этого предела (см. [11]).

Гармоническое число в числителе (24) не должно сильно повлиять на скорость сходимости, так как оно растет как $\ln n$, что пренебрежимо мало по сравнению с факториалом.

Как оказалось, последовательность частичных сумм ряда (24), $\{h(m)\}$, и в самом деле стремится к η быстрее, чем подходящие дроби этой константы, т.е. $K(\eta, m) = K(h(m), m)$, $m > 40$. Например,

$$\eta - h(100) \approx 5.5684 \times 10^{-160}, \quad \eta - K(\eta, 100) \approx 9.0078 \times 10^{-112}.$$

В завершение этого примера решим задачу Коши $y(0) = -\gamma$ для уравнения (23), но на сей раз в символьном виде. То есть γ и δ теперь просто символы. Найдем последовательность приближений для $y(1) = \delta$, $N = 1, 2, \dots$. Эта последовательность имеет вид $\tilde{\eta}(N) - \tilde{\epsilon}(N)\gamma$, где тильда означает, что это рациональные приближения соответствующих констант. Они оказались в точности равны полученным ранее, т.е. в формулах (25) и (17).

Таким образом, если иметь последовательность хороших рациональных приближений для γ (или δ), то по данному алгоритму можно получить (почти) столь же хорошую последовательность рациональных приближений для δ (или γ).

6. СУММИРОВАНИЕ ГОЛОНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Как мы видели, наш метод дает эффективные рациональные приближения коэффициентов разложения Фурье–Лежандра для функций, которые являются решениями линейных однородных ОДУ с полиномиальными коэффициентами. Такие функции называются *голономными* (см., например, [19]). Те неоднородные ОДУ, которые мы рассматривали, очевидно приводятся к однородным.

Оригинальный τ -метод Ланцоша также предназначался для этого класса уравнений (см. [4]).

Последовательность чисел $\{a_n, n = 0, 1, \dots\}$ называется *голономной*, если она удовлетворяет линейному разностному уравнению с полиномиальными (от n) коэффициентами.

Если в качестве функции понимать формальный степенной ряд, то каждой голономной функции соответствует голономная последовательность чисел — коэффициентов ее формального разложения, а также наоборот: каждой голономной последовательности соответствует голо-

номная функция — обыкновенная производящая функция этой последовательности (см. [19, р. 140]).

Хотя многие математики, продолжая традиции 19-го столетия, полагают, что “не каждый формальный ряд соответствует аналитической функции” (см. [19, р. 29]), однако это не так. Каждый формальный ряд соответствует аналитической функции, правда, возможно, не одной (теорема Бореля—Ритта, см. [20, р. 43]). Все зависит от интерпретации слова “соответствует”.

Возьмем ряд из [19, р. 29] в качестве примера. Мы использовали этот ряд в [21] как пример “голономного” суммирования последовательности. Рассмотрим формальный ряд

$$y(x) = \sum_{n=0}^{\infty} n! x^n \quad (26)$$

и найдем для него аналитическую функцию $y(x)$, такую, что $y(1) = 1$.

Формальный голономный ряд (26) удовлетворяет дифференциальному уравнению

$$x^2 y' - (1 - x)y + 1 = 0, \quad (27)$$

которое неголономно, но приводится к нужному виду. Функция

$$y(x) = \frac{1}{x} \exp\left(-\frac{1}{x}\right) \left(C - \operatorname{Ei}\left(1, -\frac{1}{x}\right)\right),$$

где C — это константа интегрирования, является общим аналитическим решением этого уравнения.

Для любой константы C эта аналитическая функция “соответствует” своему формальному разложению (26), которое является асимптотическим в полуплоскости $\operatorname{Re}(x) > 0$, с трансфинитной экспоненциально малой добавкой.

Это означает, что $y(1)$, например, может принимать любое значение. Если взять

$$C = \operatorname{Ei}(1, -1) + \exp(1) \approx 0.8231640121032 - i\pi,$$

то получим $y(1) = 1$.

Чтобы получить разложение этой функции по полиномам Лежандра нашим методом, нужно взять матрицу $X2.D - E + X$ и заменить ее последнюю строку на вектор b_0 (см. разд. 2). Вектор правой части — это $r = -e$ с подстановкой $r[N] = 1$, согласно выбранному краевому значению. Заметим, что $X2.D = X.X.D$. Для $N = 10$ получим

$$y(0) = 2693025253/2682336916, \quad y(0) - 1 \approx 3.98471 \times 10^{-3},$$

что весьма близко к верхней равномерной оценке погрешности для этой функции на интервале $[0, 1]$.

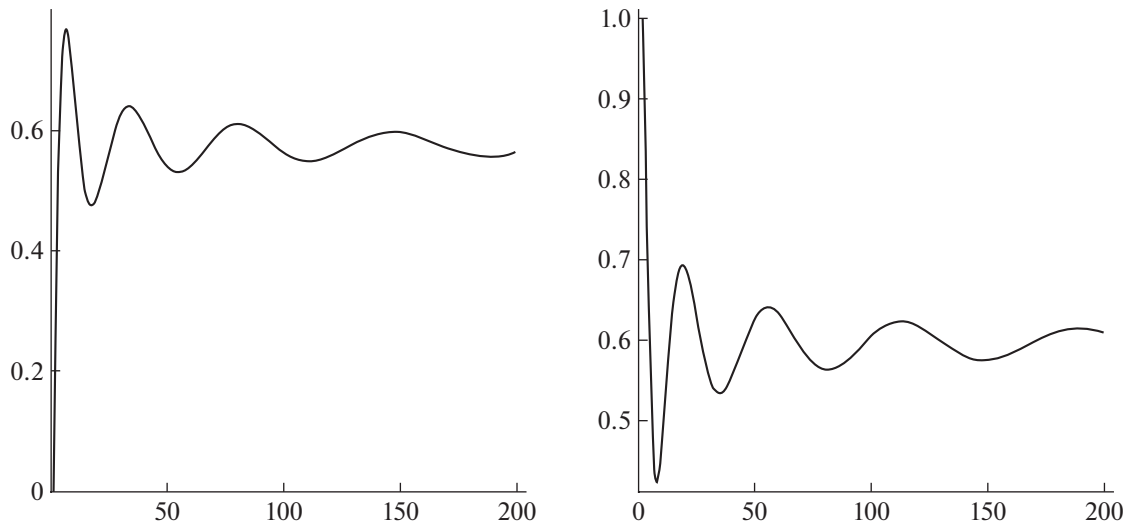
Отметим, что неединственность функции, которая соответствует своему формальному степенному разложению (а такие разложения всегда асимптотические), не имеет никакого отношения к сходимости или расходимости этого разложения. Это же относится к наличию либо отсутствию экспоненциально малой (трансфинитной) добавки. Предыдущий пример переносится без существенных изменений на уравнение

$$x^2 y' - y = -\frac{1+x+x^2}{(1+x)^2}, \quad y(x) = \frac{1}{1+x} + C \exp\left(-\frac{1}{x}\right).$$

Таким образом, мы имеем способ суммирования голономных последовательностей и рядов, причем необязательно сходящихся в обычном смысле, который сводится к решению соответствующих голономных ОДУ и подстановке $x = 1$ в полученное решение.

Здесь наш метод приближенного решения голономных ОДУ может быть весьма полезен, так как эти уравнения, вообще говоря, неинтегрируемы. Но даже в случае их интегрируемости в квадратурах (с применением CAS) эти квадратуры могут быть очень сложно устроены и (почти) совершенно бесполезны.

Вычислим для примера константу Эйлера—Гомпертца путем суммирования расходящегося ряда (26) при $x = -1$. Это классическая задача, которая решалась многими способами. Ланцош ее также использовал для демонстрации эффективности τ -метода (см. [4]).



Фиг. 2. Сходимость частичных сумм $u(n)$ (слева) и $v(n)$ (справа).

$u(n)$ и $v(n)$ соответственно. Пользуясь известным рекуррентным соотношением для полиномов Лагерра, получаем $u(1) = 0$, $u(2) = 1/4$, $u(3) = 17/36$ и

$$n^2 u(n) = (3n^2 - 4n + 2)u(n-1) - (n-1)(3n-4)u(n-2) + (n-1)(n-2)u(n-3),$$

а также $v(1) = 1$, $v(2) = 1$, $v(3) = 5/6$ и

$$(n-1)nv(n) = (n-1)(3n-4)v(n-1) - (3n-4)(n-2)v(n-2) + (n-2)^2 v(n-3).$$

На фиг. 2 показаны графики частичных сумм $u(n)$ и $v(n)$ до $n = 200$. Из фиг. 2 видно, что эти суммы осциллируют с медленно уменьшающейся амплитудой и увеличивающимся периодом. Такие ряды не могут суммироваться обычными средствами.

В данном случае мы уже просуммировали ряды (29), пользуясь явным видом полученных решений. Однако, как было отмечено, интегрируемость ОДУ для голономных функций является, вообще говоря, редким явлением. Кроме того, явные формулы не дают рациональных приближений. Поэтому применим наш метод аппроксимации решений.

Мы уже неоднократно описывали алгоритм решения таких уравнений. Уравнения (31) решаются точно так же. В результате для размерностей аппроксимации $N = 10$ и $N = 20$ получим для γ :

$$\gamma - u(1) \approx 1.1167 \times 10^{-5}, \quad \gamma - u(1) \approx -1.2006 \times 10^{-7},$$

а также для δ :

$$\delta - v(1) \approx -2.8034 \times 10^{-8}, \quad \delta - v(1) \approx 2.9407 \times 10^{-13}.$$

Таким образом, хотя голономные последовательности для γ и δ выглядят почти идентично (см. (29)), сам вид уравнений (31), а также результаты аппроксимации их решений (32) свидетельствуют, что константа γ имеет более сложную природу в некотором “голономном” смысле. На это же указывает тот факт, что для рациональных приближений константы δ существуют линейные формы второго порядка, а для γ пока что только третьего (см. [16]).

В заключение заметим, что наши методы обобщаются также на другие системы ортогональных полиномов, например, на полиномы Чебышёва. Последние гораздо удобнее для численного решения голономных ОДУ, для итерационного решения нелинейных уравнений, а также для решения разностных уравнений, которые возникают при обобщенном суммировании рядов (см. [21]). Однако это является темой отдельной работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Pashkovskii S.* Computational Application of Chebyshev Polynomials and Series. Moscow: Nauka, 1983.
2. *Бабенко К.И.* Основы численного анализа. Ижевск: РИХД, 2002.
3. *Gottlieb D., Orszag S.A.* Numerical analysis of spectral methods: theory and applications. CBMS Regional Conference Series in Applied Mathematics. 6th printing, 1996.
4. *Lanczos C.* Applied Analysis. New-York: Dover Publications, 1956.
5. *Варин В.П., Петров А.Г.* Гидродинамическая модель ушной улитки // Ж. вычисл. матем. и матем. физ. 2009. Т.49. № 9. С. 1708–1723.
6. *Wilf H.S.* Mathematics for the physical sciences. NewYork: Wiley, 1962.
7. *Johnson D.* Chebyshev Polynomials in the Spectral Tau Method and Applications to Eigenvalue Problems // NASA Contractor Report 198451. 1996.
8. *Ortiz E.L., Samara H.* An Operational Approach to the Tau Method for the Numerical Solution of Non-Linear Differential Equations // Computing. 1981. V. 27. P. 15–25.
9. *Krylov V.I.* Approximate calculation of integrals. New-York, London: Macmillan, 1962.
10. *Gantmacher F.R.* Application of the Theory of Matrices. New-York: Chelsea Press, 1960.
11. *Варин В.П.* Преобразование последовательностей в доказательствах иррациональности некоторых фундаментальных констант // Ж. вычисл. матем. и матем. физ. 2022. Т. 62. № 10. С. 1587–1614.
12. *Allen G.D. et al.* Padé approximation and Gaussian quadrature // Bull. Austral. Math. Soc. 1974. V. 11. P. 63–69.
13. *Legendre A.M.* Éléments de Géométrie. (2em. ed). Chez Fermin Didot Père et Fils. Paris, 1823.
14. *Rivoal T.* Polynomial continued fractions for $\exp(\pi)$ // hal-03269677v3. 2022. <https://hal.archives-ouvertes.fr/hal-03269677v3>.
15. *Кузьмин Р.О.* Об одном новом классе трансцендентных чисел // Известия Акад. наук СССР. VII сер. Отд. физ.-мат. наук. 1930. Вып. 6. С. 585–597.
16. *Aptekarev A.I.* On linear forms containing the Euler constant // [arXiv:0902.1768v2]. 2009. <http://arxiv.org/abs/0902.1768v2>.
17. *Варин В.П.* Факториальное преобразование некоторых классических комбинаторных последовательностей // Ж. вычисл. матем. и матем. физ. 2018. Т. 59. № 6. С. 1747–1770.
18. *Perron O.* Irrationalzahlen. Berlin, Leipzig: Göschen's Lehrbücherei, 1921.
19. *Kauers M., Paule P.* The Concrete Tetrahedron. Symbolic Sums, Recurrence Equations, Generating Functions, Asymptotic Estimates. Wien: Springer, 2011.
20. *Wasow W.* Asymptotic expansions for ordinary differential equations. New-York: Dover Publications, 1987.
21. *Варин В.П.* Функциональное суммирование рядов // Ж. вычисл. матем. и матем. физ. 2023. Т. 63. № 1. С. 3–17.
22. *Wimp J.* Sequence transformations and their applications. New-York, etc.: Academic Press, 1981.