ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 91 за 2020 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Н.Н. Калиткин, С.А. Колганов

Построение аппроксимаций, удовлетворяющих чебышевскому альтернансу

Рекомендуемая форма библиографической ссылки: Калиткин Н.Н., Колганов С.А. Построение аппроксимаций, удовлетворяющих чебышевскому альтернансу // Препринты ИПМ им. М.В.Келдыша. 2020. № 91. 33 с. https://doi.org/10.20948/prepr-2020-91
https://library.keldysh.ru/preprint.asp?id=2020-91

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук

Н.Н. Калиткин, С.А. Колганов

Построение аппроксимаций, удовлетворяющих чебышевскому альтернансу

Москва — 2020

Калиткин Н.Н., Колганов С.А.

Построение аппроксимаций, удовлетворяющих чебышевскому альтернансу

Предложен эффективный алгоритм построения аппроксимирующих формул для достаточно плавно меняющихся функций. Аппроксимирующие формулы могут иметь вид многочлена, обобщенного многочлена, отношения многочленов или обобщенных многочленов, а также некоторой функции от перечисленных выражений. Метод позволяет находить коэффициенты аппроксимирующих формул, обеспечивающие достижение чебышевского альтернанса либо для абсолютной погрешности, либо для относительной. Алгоритм основан на итерационном процессе нахождения узлов интерполяции. На каждой итерации узлы интерполяции сдвигаются так, чтобы сошедшийся процесс обеспечивал чебышевский альтернанс. Метод проиллюстрирован на задаче аппроксимации функций Ферми-Дирака, играющих важную роль в задачах квантовой механики.

Ключевые слова: аппроксимация, чебышевский альтернанс, функции Ферми-Дирака

Nikolai Nikolaevich Kalitkin, Semen Andreevich Kolganov

The construction of approximations satisfying the Chebyshev alternance

An efficient algorithm for constructing approximating formulas for sufficiently smoothly changing functions is proposed. Approximating formulas can take the form of a polynomial, a generalized polynomial, a relation of polynomials or generalized polynomials, as well as some function of the listed expressions. The method allows you to find the coefficients of approximating formulas that ensure the achievement of the Chebyshev alternance either for the absolute error or for the relative one. The algorithm is based on an iterative process of finding interpolation nodes. At each iteration, the interpolation nodes are shifted so that the converged process provides the Chebyshev alternance. The method is illustrated on the problem of approximation of the Fermi-Dirac functions, which play an important role in problems of quantum mechanics.

Key words: approximation, Chebyshev alternance, Fermi-Dirac functions

Работа поддержана грантом РФФИ №18-01-00175.

1. Задачи аппроксимации

Очень многие специальные функции математической физики определяются как решения некоторых дифференциальных уравнений. Для их практического использования составляют таблицы, каждая точка которых находится в результате трудоемких вычислений. В докомпьютерную эпоху пользовались интерполяцией по таким таблицам. В настоящее время аппроксимируют эти таблицы некоторыми разложениями по базису элементарных функций, легко вычисляемых на компьютере. Обычно такими функциями служат многочлены, экспоненты, тригонометрические и другие функции, для которых имеются экономичные компьютерные подпрограммы. Само разложение строят так, чтобы оно было наилучшим в некоторой норме, причем выбор этой нормы зависит от требуемых практических приложений. Обсудим последний вопрос подробнее.

В простейшем случае функция $u(x)$ задана на конечном, причем не слишком большом, отрезке значений x. В этом случае для аппроксимации пригодны самые различные системы функций, начиная с полиномов. Если $u(x)$ знакопеременна, то обычно ищут наилучшее приближение в нормах C или L_2. Для нормы L_2 наилучшее приближение дают обобщенные ряды Фурье, использующие тригонометрические функции для периодических $u(x)$ и различные ортогональные системы многочленов для непериодических $u(x)$. Способы построения рядов Фурье общеизвестны и несложны. Эти разложения дают наилучшую аппроксимацию в норме L_2.

Приближения, наилучшие в норме C, должны удовлетворять условию ***чебышевского альтернанса***: знаки последовательных экстремумов погрешности чередуются, а все модули экстремумов одинаковы. Такие приближения отрезком ряда нельзя найти за конечное число действий. Их получают итерационными алгоритмами. Построению таких алгоритмов посвящен ряд работ (см. [1- 5] и цитированную там литературу).

Если $u(x)$ знакопостоянна, причем достаточно сильно меняется на заданном отрезке, то целесообразнее строить приближения, наилучшие в смысле относительной погрешности аппроксимации $\delta u(x)/u(x)$. Способы построения таких наилучших приближений намного хуже исследованы.

Особенно труден случай, когда требуется аппроксимировать $u(x)$ на прямой или полупрямой. Во-первых, величина $u(x)$ может меняться на много порядков. Поэтому надо минимизировать относительную погрешность. Во-вторых, редко удается подобрать систему функций, пригодную для аппроксимации $u(x)$ во всей требуемой области. Приходится разбивать эту область на отдельные подобласти, в каждой из которых используется свой

способ аппроксимации. При этом возникает естественное дополнительное требование: аппроксимации должны образовывать единую непрерывную функцию, т.е. совпадать на границах соседних подобластей. Проще всего добиться непрерывности, требуя обращения $\delta u(x)$ в нуль на границах каждой подобласти. В-третьих, в этом случае одна или обе крайние подобласти полубесконечны. Введение нормы L_2 на полубесконечных областях проблематично. Приходится использовать норму C, что требует нахождения чебышевского альтернанса.

В данной работе предложен эффективный алгоритм решения задач последнего типа. Он применен к решению в реальных прикладных задачах.

2. Чебышевская аппроксимация многочленом

Сначала рассмотрим вспомогательную задачу – нахождение многочлена $P(x)$, наименее уклоняющегося от нуля (т.е. мы полагаем $u(x) \equiv 0$). Областью задания функции для определенности выберем отрезок $x \in [-1;1]$. Наша задача будет отличаться от классической постановки Чебышева: мы требуем, чтобы на концах отрезка $x = \pm 1$ многочлен обращался в нуль. Обозначим нули соответствующего многочлена через $x_n, 0 \le n \le N$; для получения чебышевского альтернанса все нули многочлена x_n должны быть вещественными, причем $x_n \in [-1;1]$; согласно нашим дополнительным условиям, $x_0 = -1, x_N = 1$.

Искомый многочлен имеет $N+1$ нуль, поэтому он является многочленом степени $N+1$; поскольку он определен с точностью до множителя, для однозначного определения будем считать, что его старший член есть x^{N+1} с коэффициентом $a_{N+1} = 1$. Между каждой парой узлов x_{n-1}, x_n лежит один экстремум многочлена; обозначим его положение через $x_{n-1/2}$ и введем величину $p_{n-1/2} = P(x_{n-1/2}), 1 \le n \le N$. В этих обозначениях задача ставится так:

требуется найти такие положения внутренних узлов $x_n, 1 \le n \le N-1$, чтобы знаки $p_{n-1/2}$ чередовались, а модули $p_{n-1/2}$ были одинаковы.

3. Кубическая кривая

Представляется разумной следующая идея алгоритма. Рассмотрим 3 соседних узла x_{n-1}, x_n, x_{n+1}. Сравним значения двух экстремумов $p_{n-1/2}$ и $p_{n+1/2}$. Интуитивно очевидно, что если сдвинуть узел x_n в сторону большего по

модулю экстремума, то модуль этого экстремума уменьшится, а модуль другого возрастет. Вопрос в том, на какую величину нужно сдвигать узел x_n.

Для оценки такого сдвига рассмотрим простейший случай: кубическая интерполяционная кривая проходит через эти 3 узла. Запишем эту кривую так, чтобы она автоматически удовлетворяла условию интерполяции в узле x_n:

$$P(x) = \xi + a\xi^2 - b\xi^3, \xi = x - x_n. \tag{1}$$

Из условий интерполяции в узлах x_{n-1}, x_{n+1} получаем коэффициенты

$$b = -\frac{1}{\xi_{n-1}\xi_{n+1}}, a = -\frac{\xi_{n-1} + \xi_{n+1}}{\xi_{n-1}\xi_{n+1}}, \xi_{n\pm1} = x_{n\pm1} - x_n. \tag{2}$$

Поскольку ξ_{n-1} и ξ_{n+1} имеют разные знаки, то $b > 0$.

Определим положения правого и левого экстремумов из условия $P'(\xi_{n\pm1/2}) = 0$. Это дает квадратное уравнение, решения которого равны

$$\xi_{n\pm1/2} = \frac{1}{3b}\left(a \pm \sqrt{a^2 + 3b}\right). \tag{3}$$

Напомним, что поскольку $b > 0$, то при любом знаке a выполняется $0 < \xi_{n+1/2} < \xi_{n+1}$ и $\xi_{n-1} < \xi_{n-1/2} < 0$. Экстремальные значения кубического многочлена равны

$$p_{n\pm1/2} = \frac{1}{27b^2}\left[a\left(2a^2 + 9b\right) \pm 2\left(a^2 + 3b\right)^{3/2}\right]. \tag{4}$$

Запишем отношение экстремумов с обратным знаком и подставим в него значения a, b:

$$\begin{gathered} -\frac{p_{n+1/2}}{p_{n-1/2}} = \frac{2\left(a^2 + 3b\right)^{3/2} + a\left(2a^2 + 9b\right)}{2\left(a^2 + 3b\right)^{3/2} - a\left(2a^2 + 9b\right)} = \\ = \frac{2\left(\xi_{n-1}^2 - \xi_{n-1}\xi_{n+1} + \xi_{n+1}^2\right)^{3/2} + \left(\xi_{n-1} + \xi_{n+1}\right)\left(2\xi_{n-1}^2 - 5\xi_{n-1}\xi_{n+1} + 2\xi_{n+1}^2\right)}{2\left(\xi_{n-1}^2 - \xi_{n-1}\xi_{n+1} + \xi_{n+1}^2\right)^{3/2} - \left(\xi_{n-1} + \xi_{n+1}\right)\left(2\xi_{n-1}^2 - 5\xi_{n-1}\xi_{n+1} + 2\xi_{n+1}^2\right)}; \end{gathered} \tag{5}$$

напомним, что по смыслу вывода дробь в правой части должна быть положительной. Поскольку $\xi_{n-1}\xi_{n+1} < 0$, то в последнем выражении все скобки с

квадратичными членами положительны, а скобка с линейными членами может иметь любой знак.

Преобразуем формулу (4) к следующему виду:

$$-\frac{p_{n+1/2}}{p_{n-1/2}}=\frac{1+\alpha}{1-\alpha},\alpha=\frac{\left(\xi_{n-1}+\xi_{n+1}\right)\left(2\xi_{n-1}^2-5\xi_{n-1}\xi_{n+1}+2\xi_{n+1}^2\right)}{2\left(\xi_{n-1}^2-\xi_{n-1}\xi_{n+1}+\xi_{n+1}^2\right)^{3/2}}. \tag{6}$$

Из формулы (6) получается $\alpha=\left(p_{n+1/2}+p_{n-1/2}\right)\Big/\left(p_{n+1/2}-p_{n-1/2}\right)$. Подставляя сюда выражение (6) для α, получаем

$$\xi_{n-1}+\xi_{n+1}=\frac{p_{n+1/2}+p_{n-1/2}}{p_{n+1/2}-p_{n-1/2}}\frac{2\left(\xi_{n-1}^2-\xi_{n-1}\xi_{n+1}+\xi_{n+1}^2\right)^{3/2}}{\left(2\xi_{n-1}^2-5\xi_{n-1}\xi_{n+1}+2\xi_{n+1}^2\right)}. \tag{7}$$

Учтем, что $\xi_{n-1}+\xi_{n+1}=x_{n-1}+x_{n+1}-2x_n$, а $\xi_{n+1}-\xi_{n-1}=x_{n+1}-x_{n-1}$. Тогда формулу (7) можно переписать так:

$$\begin{gathered}\frac{1}{2}\left(x_{n-1}+x_{n+1}\right)-x_n=\left(x_{n+1}-x_{n-1}\right)\frac{p_{n+1/2}+p_{n-1/2}}{p_{n+1/2}-p_{n-1/2}}\cdot w_n,\\ w_n=\frac{\left(\xi_{n-1}^2-\xi_{n-1}\xi_{n+1}+\xi_{n+1}^2\right)^{3/2}}{\left(\xi_{n+1}-\xi_{n-1}\right)\left(2\xi_{n-1}^2-5\xi_{n-1}\xi_{n+1}+2\xi_{n+1}^2\right)}.\end{gathered} \tag{8}$$

Итерации. Мы хотим аппроксимировать $u(x)$, которая является достаточно гладкой функцией (т.е. имеющей столько непрерывных производных, сколько потребуется по ходу изложения). Рассмотрим задачу её наилучшей в норме C аппроксимации в некоторой конечной подобласти значений x. Для простоты преобразуем координаты так, чтобы в новых координатах это был отрезок $-1\le x\le 1$. Для аппроксимации выберем многочлен степени $N+1$.

Точное решение задачи наилучшей чебышевской аппроксимации не найдено. Поэтому будем строить итерационный процесс. Выберем начальную сетку $x_n, 0\le n\le N$, у которой крайние узлы лежат на границах отрезка. Построим на этой сетке интерполяционный многочлен $P_{N+1}(x)$ и вычислим все интервальные экстремумы погрешности $p_{n-1/2}, 1\le n\le N$. Для практического вычисления этих экстремумов достаточно ввести на каждом интервале $\left(x_{n-1},x_n\right)$

равномерную сетку из ~ 20 вспомогательных точек, непосредственно вычислить погрешность $P_{N+1}(x)-u(x)$ во всех вспомогательных точках и прямым перебором определить экстремум; такая процедура обеспечивает достаточную точность определения интервальных экстремумов погрешности.

Для следующей итерации надо сместить узлы интерполяции. Как это сделать? Для задачи с кубической кривой ответ очевиден: новое значение $\hat{x}_n = (x_{n-1} + x_{n+1})/2$. Попробуем принять подобное движение узла для общей задачи интерполяции:

$$\hat{x}_n = x_n + (x_{n+1} - x_{n-1}) \frac{p_{n+1/2} + p_{n-1/2}}{p_{n+1/2} - p_{n-1/2}} \cdot w_n . \qquad (9)$$

Величина w_n сама зависит не только от узлов x_{n-1}, x_{n+1}, но и от положения неизвестного узла x_n. Однако легко проверить, что зависимость w_n от x_n очень слабая. Зависимость косвенная через ξ_{n-1}, ξ_{n+1}. В соответствующей дроби (8) числитель и знаменатель одного порядка $\sim \xi^3$. Поэтому дробь (8) зависит только от отношения $q \equiv -\xi_{n+1}/\xi_{n-1}$:

$$w_n(q) = \frac{(1+q+q^2)^{3/2}}{(1+q)(2+5q+2q^2)}. \qquad (10)$$

Заметим, что $0 < q < +\infty$ и $w_n(q) = w_n\left(\frac{1}{q}\right)$. Тогда экстремумы таковы:

$$\min w_n = w_n(1) = \frac{1}{2\sqrt{3}}, \max w_n = w_n(0) = w_n(+\infty) = \frac{1}{2}. \qquad (11)$$

Формулу (11) можно рассматривать как итерационный процесс расчета $\hat{x}_n$: вычислим все w_n с учетом их зависимости от x_n и найдем $\hat{x}_n$. Сдвинем так все точки, построим новую интерполяцию, найдем новые $p_{n\pm 1/2}$. Затем по новым данным пересчитаем w_n и найдем новые $\hat{x}_n$. Итерации будем вести до сходимости.

Однако основное количество итераций будет происходить при значениях $q \approx 1$. Причина в том, что описанный процесс является простой итерацией и его сходимость линейная; она медленная, и основное число итераций происходит вблизи корня. Поэтому вместо общей формулы можно взять $w = 1/2\sqrt{3}$, и сходимость итераций практически не ухудшится.

Детали алгоритма. Описанный выше алгоритм рассчитан на малую окрестность корня. Вдали от корня возможна ситуация, когда $\hat{x}_n > \hat{x}_{n+1}$ или $\hat{x}_n < \hat{x}_{n-1}$, т.е. происходит "рокировка" узлов. Это математически бессмысленное состояние часто называют "перехлест".

Предложим страхующую поправку – введение шага τ. Тогда предварительная формула сдвига узлов будет иметь следующий вид:

$$\hat{x}_n = x_n + v_n\tau, v_n = \left(x_{n+1} - x_{n-1}\right) w_n \frac{p_{n+1/2} + p_{n-1/2}}{p_{n+1/2} - p_{n-1/2}}, 1 \le n \le N; w_n = \frac{1}{2\sqrt{3}}; \qquad (12)$$

условные скорости в граничных узлах $v_0 = v_N = 0$. Вариант алгоритма с более общей формулой для w_n далее рассматривать не будем.

Предыдущие теоретические соображения дают значение $\tau = 1$. Но при этом возможен "перехлест". Поэтому сначала сделаем виртуальную итерацию. Возьмем соседнюю пару узлов $\left(x_{n-1}, x_n\right)$ и определим, при каком значении $\tau_{n-1/2}$ возможен "перехлест":

$$x_{n-1} + v_{n-1}\tau_{n-1/2} = x_n + v_n\tau_{n-1/2}. \qquad (13)$$

Отсюда получаем

$$\tau_{n-1/2} = \begin{cases} \left(x_n - x_{n-1}\right) / \left(v_{n-1} - v_n\right), v_{n-1} > v_n; \\ +\infty, v_{n-1} \le v_n. \end{cases} \qquad (14)$$

Для получения разумного результата шаг должен быть в несколько раз меньше $\tau_{n-1/2}$. Поэтому будем умножать $\tau_{n-1/2}$ на настроечный параметр $a < 1$; разумные пределы для этого параметра составляют $0.1 - 0.5$.

Кроме того, вблизи корня должно выполняться $\tau \approx 1$. Последнюю величину также целесообразно умножить на некоторый настроечный параметр b; разумными представляются пределы $0.8 < b < 1.2$. Таким образом, для всех узлов будем выбирать единый окончательный шаг итерации по формуле

$$\overline{\tau} = \min(b, a\tau_{n-1/2}). \qquad (15)$$

Далее на пробных расчетах будет показано, что целесообразно выбирать $b = 1, a \sim 0.2$.

Критерий сходимости. Целью итерационного процесса является получение чебышевского альтернанса, т.е. чередования знаков экстремумов

погрешности $p_{n-1/2}$ с одновременным выравниванием их модулей. Поэтому скорость сходимости целесообразно определять по убыванию величины

$$L = \frac{\max\limits_{1\le n\le N-1}\left|p_{n-1/2}\right|}{\min\limits_{1\le n\le N-1}\left|p_{n-1/2}\right|} \tag{16}$$

от итерации к итерации. По определению $L \ge 1$. Если на новой итерации значение $\hat{L}$ не превышает значения L на предыдущей итерации, то сходимость итераций монотонная. При изучении сходимости необходимо добиваться возможно большей близости L к 1. Но для практических расчетов вполне достаточно получить $L < 1.01$, и даже $L < 1.1$.

Доказать сходимость предложенного процесса не удалось. Поэтому сходимость была изучена и проиллюстрирована на представительных примерах, причем не только для аппроксимации многочленами, но и отношением многочленов.

4. Нули чебышевского альтернанса; пример 1

Формулировка задачи. Начнем исследование с поиска нулей интерполяционного многочлена чебышевского типа. Это $N+1$ нуль $x_n, 0 \le n \le N$, так что это нули многочлена степени $N+2$:

$$P_{N+2}(x) = \sum_{k=0}^{N+2} a_k x^k, -1 \le x \le 1. \tag{17}$$

Однако он не является классическим многочленом Чебышева 1-го или 2-го рода, т.к. мы дополнительно требуем, чтобы крайние нули были концами отрезка: $x_0 = -1, x_N = 1$.

Рассмотрим, как целесообразно вычислять внутренние нули такого многочлена. Для этого положим $u(x) \equiv 0$. Тогда погрешность интерполяции будет просто равна самому многочлену $P_{N+2}(x)$. При этом умножение $P_{N+2}(x)$ на произвольный множитель не меняет узлов интерполяции $u(x) \equiv 0$. Следовательно, мы можем для простоты положить $a_{N+2} = 1$. Тогда при заданном расположении узлов интерполяции задача нахождения других коэффициентов многочлена принимает следующий вид:

$$\sum_{k=0}^{N+1} a_k x_n^k = -x_n^{N+2}, 0 \le n \le N. \tag{18}$$

Для нахождения коэффициентов a_k получается линейная система с определителем Вандермонда. Этот определитель не равен нулю, поэтому система всегда имеет решение.

Однако реально обусловленность подобных систем быстро ухудшается с увеличением N. При 64-битовых числах обычно $N<10$ дает удовлетворительные результаты, а $N>20$ – совершенно неприемлемые. То, что мы выбрали симметричный отрезок $-1\le x\le 1$ и положили $u(x)\equiv 0$, позволяет ощутимо уменьшить ошибки округления. Для нахождения требуемых нулей мы вычисляем интервальные экстремумы многочлена $P_{N+2}(x)$ и применяем к ним процедуру выравнивания экстремумов, описанную выше. Опишем результаты этих численных экспериментов.

Результаты расчетов. Были проведены расчеты для числа узлов от $N=4$ до тех пор, пока ошибки округления не становились преобладающими (это наступает при $N>15$). Во всех случаях начальное распределение узлов бралось равномерным, и всегда итерационный процесс сходился. При этом на начальной равномерной сетке отношение L при больших N достигало 3000! Чтобы выразительно изобразить профиль погрешности в этих случаях, пришлось выработать специальные масштабы построения графиков, описанные ниже.

Сетки и профили. На рис. 1 показано начальное и конечное расположение узлов для $N+1=16$. Поскольку задача симметрична относительно $x=0$, на рисунке приведена половина распределения $x\ge 0$. Видно, что окончательное распределение узлов сильно отличается от начального равномерного распределения. Окончательная сетка разрежена в центре и сильно сгущена около края.

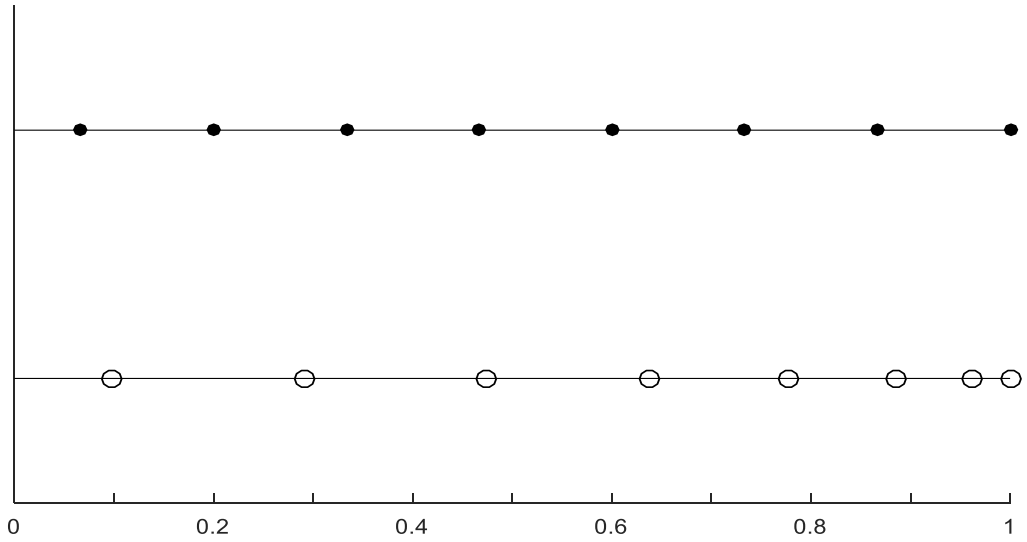

Рис. 1. Распределение узлов для $N+1=16$: черные кружки – начальное приближение, светлые кружки – окончательное распределение.

На рис. 2 показаны профили погрешности для $N+1=16$; в силу симметрии показана только часть профилей для $x \ge 0$. По оси абсцисс выбран равномерный масштаб. Но амплитуды осцилляций меняются в очень широких пределах, причем сам многочлен является знакопеременной функцией. Поэтому масштаб по оси ординат выбран следующим образом. Выбран наименьший (по модулю) экстремум многочлена на равномерной сетке $\tilde{p} = \min_n \left| p_{n-1/2} \right|$. Тогда по оси ординат отложим величину

$$\delta(u) = arcsh\left(\frac{u(x)}{\tilde{p}}\right) \equiv sign(u) \cdot \ln\left(\frac{|u|}{\tilde{p}} + \sqrt{1+\left(\frac{u}{\tilde{p}}\right)^2}\right). \qquad (19)$$

Когда $|u(x)| \ll \tilde{p}$, то $\delta(u) \approx u$ и масштаб по ординате почти линеен и хорошо передает участки изменения знака $u(x)$. Но когда $|u(x)| \gg \tilde{p}$, масштаб по ординате становится логарифмическим и на графике хорошо размещаются наибольшие экстремумы $u(x)$.

На рис. 2 показаны профили $u(x)$ на начальной равномерной сетке (тонкая линия) и на окончательной сетке (жирная линия). Видно, что на начальной сетке экстремумы отличаются на много порядков, а на окончательной сетке величины экстремумов одинаковы, т.е. достигнуто условие чебышевского альтернанса.

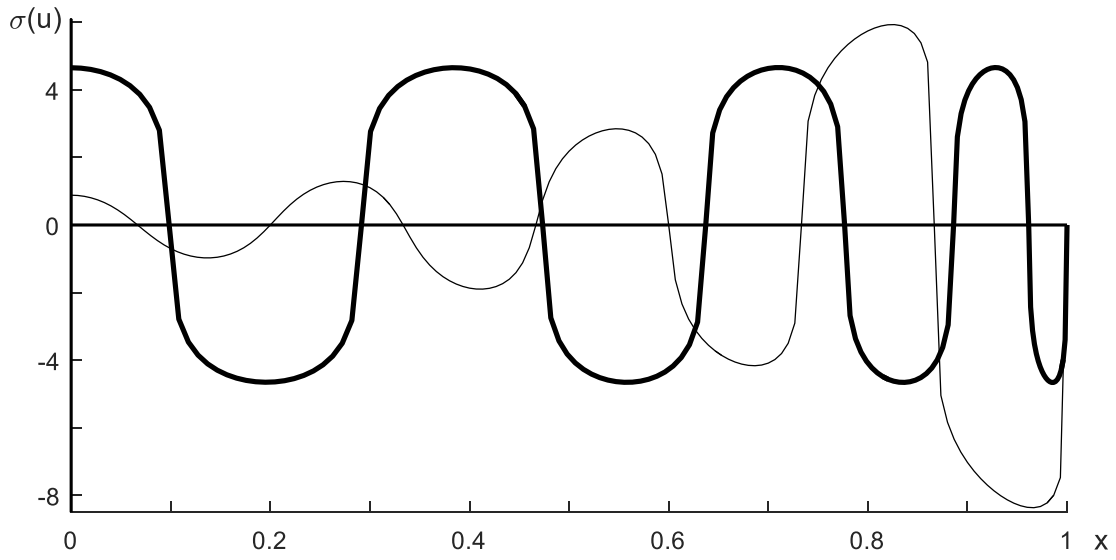

Рис. 2. Профили погрешности $u(x)$ для $N+1=16$;
тонкая линия – нулевое приближение, жирная – окончательный результат.

Улучшение точности. На рис. 3 показаны зависимости максимального и минимального экстремумов $u(x)$ от числа улов $N+1$. Масштаб по абсциссе $N+1$ выбран равномерный, а по ординате $|p_{n-1/2}|$ – логарифмический. На рис. 3 показаны 3 линии. Верхняя дает $\max|p_{n-1/2}|$ на начальной равномерной сетке (черные кружки), нижняя показывает $\min|p_{n-1/2}|$ на начальной равномерной сетке (также черные кружки), а средняя показывает $|p_{n-1/2}|$ на конечной сетке (светлые кружки). Эти линии оказались практически прямыми, т.е. зависимость указанных экстремумов от $N+1$ является экспоненциальной. Для верхней, средней и нижней линий соответствующие аппроксимации будут

$$\lg|p_{n-1/2}|\approx\begin{cases}-0,705-0,181\cdot(N-3),\\-0,764-0,310\cdot(N-3),\\-0,954-0,438\cdot(N-3).\end{cases} \qquad (20)$$

Линии исходят примерно из одной точки, т.е. при $N+1=4$ разница между равномерной и конечными сетками невелика. Однако при $N+1=16$ расстояние между верхней и нижними кривыми составляет ~ 2000 раз; это показывает невыгодность равномерной сетки для целей аппроксимации. Расстояние между верхней и средней кривыми является выигрышем в точности при переходе от равномерной сетки к оптимальной; для $N+1=16$ этот выигрыш составляет ~ 40 раз. Такой выигрыш в точности хорошо иллюстрирует преимущество оптимальной сетки.

Заметим, что средняя линия на рис. 3 лежит почти посередине между верхней и нижней. С учетом логарифмического масштаба по ординате это означает, что выигрыш в точности при переходе к оптимальной сетке близок к $\sqrt{\max|p_{n-1/2}|/\min|p_{n-1/2}|}$ на равномерной сетке.

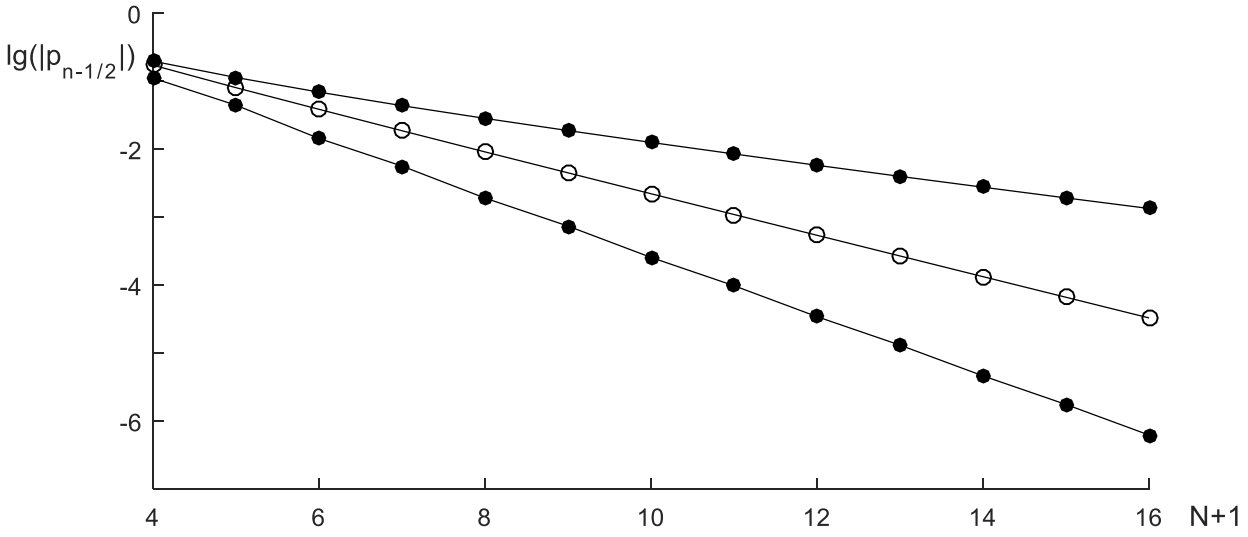

Рис. 3. Зависимости экстремумов $u(x)$ от числа улов $N+1$; темные кружки – равномерная сетка, светлые кружки – оптимальная сетка.

Сходимость. В формуле выбора шага (15) в районе корня теоретическое значение для параметра $b=1$. Расчеты с варьированием b показали, что небольшие варьирования $b=1\pm 0.1$ почти не влияют на скорость сходимости. При этом сходимость является монотонной. Однако дальнейшее уменьшение b начинает заметно замедлять сходимость, не нарушая ее монотонности. Опаснее увеличение b: сходимость не только замедляется, но может стать немонотонной; при $b>1{,}5$ наблюдалась даже расходимость итераций. Поэтому в дальнейших расчетах принималось $b=1$.

Влияние параметра a проиллюстрируем на рис. 4, построенном для $N=15$. На нем изображена зависимость шага τ от номера итерации. Видно, что на первых итерациях шаг τ существенно меньше 1. Он монотонно возрастает с номером итерации и довольно быстро стремится к значению 1. При параметре $a=0.1$ (черные кружки) выход на 1 происходит за 11 итераций, при $a=0.2$ (светлые кружки) выход происходит быстрее (за ~ 5 итераций). Дальнейшее увеличение a не имеет смысла: выход на $\tau=1$ и так достаточно быстрый, но возникает риск “перехлестов”. Поэтому мы рекомендуем для расчетов значение $a=0.2$.

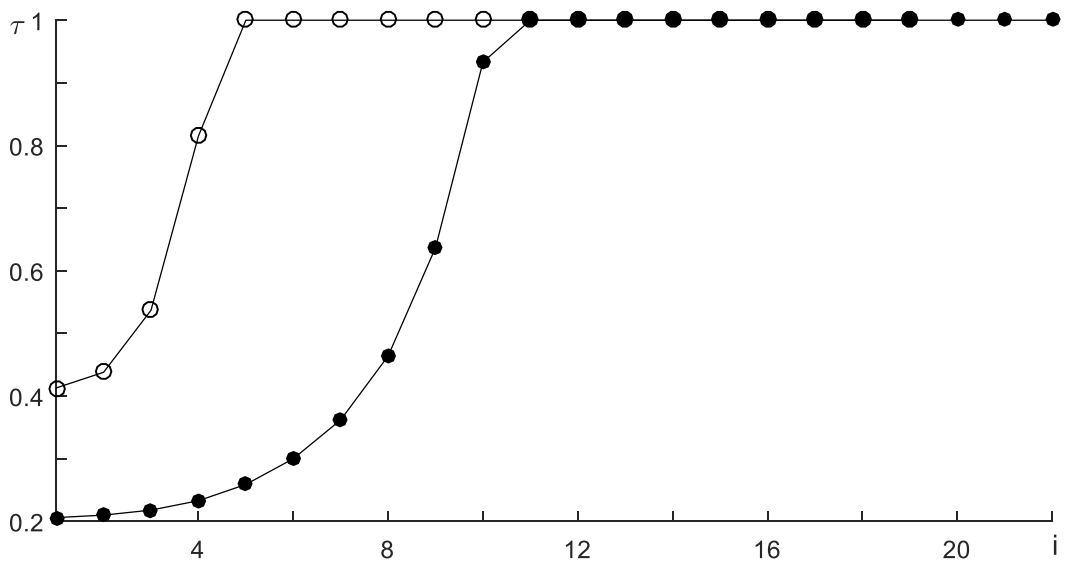

Рис. 4. Зависимость τ от номера итерации; черные кружки – $a = 0.1$, светлые кружки – $a = 0.2$.

Выравнивание экстремумов характеризуется величиной L, определяемой формулой (16). Зависимость этой величины от номера итерации представлена на рис.5 в полулогарифмическом масштабе также для $N = 15$. Линии для $a = 0.2$ (светлые кружки) и $a = 0.1$ (черные кружки) начинаются из одной точки; затем первая кривая идет ниже. Обе кривые асимптотически выходят на параллельные прямые. Их искривленные участки соответствуют значениям $\tau < 1$ на рис. 4, а параллельные прямые соответствуют значениям $\tau = 1$. Асимптотический выход на прямые означает наступление экспоненциальной сходимости, что характерно для простых итераций.

Наклоны прямых на рис. 5 составляют примерно $1.56 - 1.57$. Эти наклоны определяют скорость сходимости. Разумеется, значения этих наклонов зависят от величины N; этот вопрос будет исследован позднее.

Заметим, что в теоретических исследованиях мы брали $L < 1.01$ (различие экстремумов не превышает 1%). При построении практических аппроксимаций допустимо брать менее жесткий критерий $L < 1.1$ (различие экстремумов не превышает 10%).

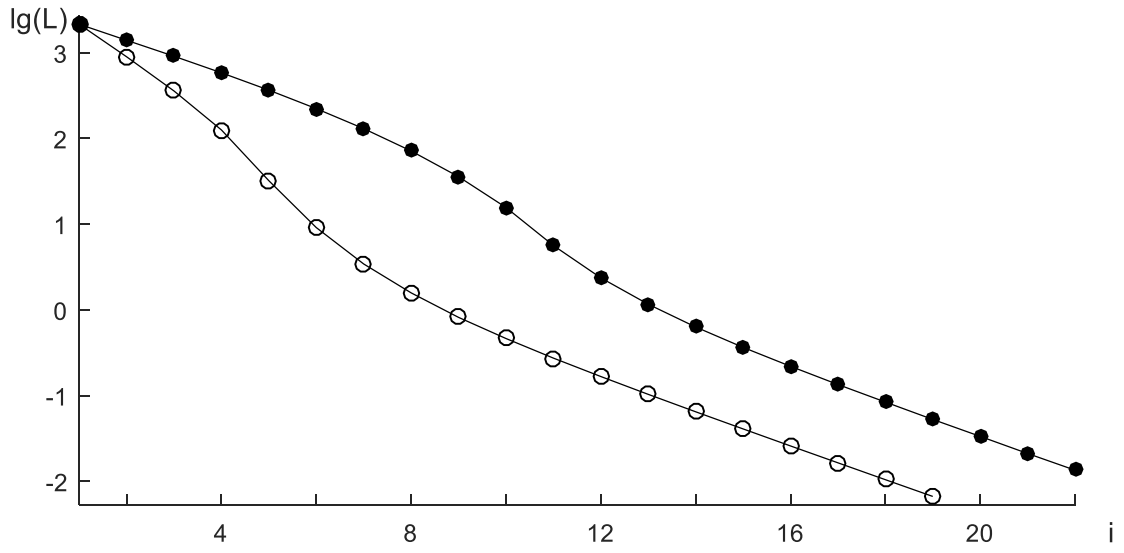

Рис. 5. Зависимость $\lg L$ от номера итерации; черные кружки – $a = 0.1$, светлые кружки – $a = 0.2$.

5. Аппроксимация многочленом; пример 2

Формулировка задачи. Полиномы, наименее уклоняющиеся от нуля, используют для построения наилучших аппроксимаций функции полиномами. Процедура построения подобных аппроксимаций непосредственно не связана с нулями полиномов. Она также основана на нахождении чебышевского альтернанса: для заданной $u(x)$ найти такие узлы интерполяции полиномами $P(x)$, чтобы знаки погрешностей чередовались, а модули погрешностей были бы одинаковыми.

Напомним, что такие аппроксимации хороши лишь для приближенного вычисления самой функции $u(x)$. Использовать дифференцирование таких аппроксимаций для вычисления производных $u(x)$ не рекомендуется: погрешность производной при этом может быть маленькой в середине отрезка, но становится очень большой вблизи границ отрезка.

В задаче аппроксимации возможны различные варианты:

1) требуется аппроксимировать знакопостоянную или знакопеременную функцию $u(x)$;

2) нужно обеспечить минимизацию абсолютной или относительной погрешности;

3) следует ли требовать обращения погрешности в нуль на границах отрезка.

Здесь мы подробно рассмотрим тот вариант, который требуется для наших приложений: $u(x)$ знакопостоянна, требуется минимизировать относительную погрешность $\delta(x)=\dfrac{P(x)}{u(x)}-1$, причем $\delta(x)$ должно обращаться в нуль на границах отрезка.

Для определенности выберем отрезок $x\in[-1;1]$. Разместим на нем узлы интерполяции $x_n, 0\le n\le N$, так, чтобы они монотонно возрастали, причем $x_0=-1$ и $x_N=1$. Возьмем многочлен

$$P_N(x)=\sum_{k=0}^{N}a_k x^k \tag{21}$$

степени N. В точках x_n многочлен должен совпадать с заданной функцией:

$$\sum_{k=0}^{N}a_k x_n^k=u(x_n), 0\le n\le N. \tag{22}$$

Требуется найти такое расположение внутренних узлов $x_n, 1\le n\le N-1$, чтобы относительная погрешность

$$\delta(x)=\frac{P_N(x)}{u(x)}-1 \tag{23}$$

удовлетворяла условию чебышевского альтернанса.

Алгоритм. Строим итерационный алгоритм, описанный в разделе 4. Пусть на очередной итерации известно некоторое расположение внутренних узлов. Из системы уравнений (22) находим коэффициенты a_k. Как и ранее, это линейная система относительно коэффициентов a_k, а ее определитель есть определитель Вандермонда. Он отличен от нуля, так что система имеет решение, притом единственное. В отличие от предыдущей задачи, правые части системы ненулевые, поэтому ошибки округления будут несколько больше, чем при нахождении нулей чебышевского альтернанса.

Заметим, что диапазон изменения аргумента x на практике может быть любым. Этот диапазон влияет на ошибки округления. Ошибки округления будут наименьшими, если $x_0=-x_N$. Поэтому если первоначальный отрезок x не удовлетворяет этому условию, то целесообразно предварительно провести соответствующее линейное преобразование аргумента.

Когда коэффициенты a_k найдены, вводим ~ 20 вспомогательных точек x_j между каждой парой узлов x_{n-1}, x_n; их расположение может быть равномерным на этом интервале. Во всех вспомогательных точках находим профиль погрешности $\delta(x_j)$ по формуле (23). Затем на каждом интервале (x_{n-1}, x_n)

простым перебором находим экстремум относительной погрешности $p_{n-1/2}$. Напомним, что это не точные значения экстремумов, а приближенные. Но при ~ 20 вспомогательных узлах относительная погрешность экстремума составляет $\sim 0,25\%$. Такой точности вполне достаточно для уравнивания экстремумов с точностью 1%.

Далее находим внутреннее расположение узлов $\hat{x}_n, 1 \le n \le N$, используя алгоритм, построенный в разделе 3. Эти итерации повторяем, пока интервальные экстремумы $p_{n-1/2}$ не станут удовлетворять с требуемой точностью условию чебышевского альтернанса.

На предельно больших N ошибки округления могут привести к расходимости алгоритма. Поэтому помимо обычного критерия остановки итераций по выполнению чебышевского альтернанса был введен “аварийный” останов: если на очередной итерации максимальный экстремум становился больше, чем на предыдущей итерации, то расчет прекращался. Последняя итерация отбрасывалась, и в качестве окончательного результата бралась предыдущая итерация.

Пример расчета. Выберем функцию, величина которой меняется на заданном отрезке $[-1,+1]$ почти в 8 раз: $u(x) = e^x$. Начальное расположение узлов бралось равномерным. Степень многочлена увеличивалась от $N = 3$ до тех пор, пока процесс не срывался из-за ошибок округления при решении линейной системы (это происходило при $N > 12$, т.е. на 3 меньше, чем в задаче об определении нулей многочлена). Результаты расчетов представлены на рис. 6–7.

На рис. 6 показана зависимость логарифма максимальной относительной погрешности на начальной сетке (черные кружки) и на окончательной сетке (светлые кружки) в зависимости от N. В обеих линиях начальные участки искривлены, но далее линии выходят на прямые. Это свидетельствует об экспоненциальной зависимости погрешности от степени многочлена. Наклон второй кривой больше. Это означает, что оптимальная сетка по отношению к равномерной дает тем больший выигрыш в точности, чем больше степень многочлена. При $N = 12$ на оптимальной сетке достигается отличная относительная погрешность $4,4 \cdot 10^{-14}$. Дальнейшему увеличению точности препятствуют ошибки округления при решении линейной системы (на 64-разрядных числах).

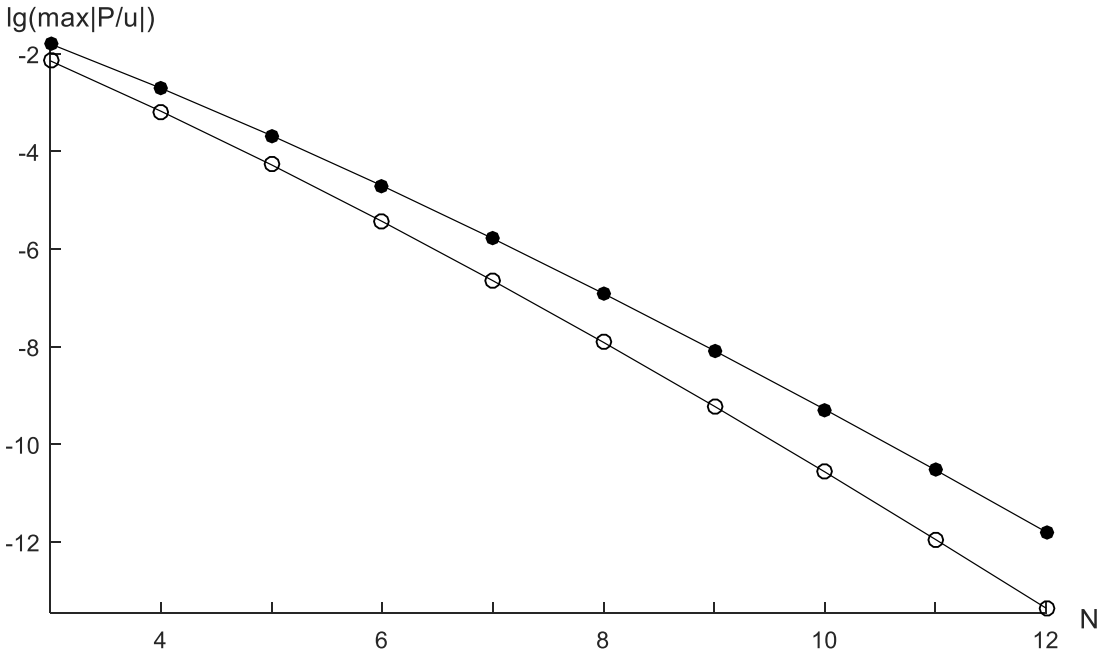

Рис. 6. Максимум относительной погрешности: черные кружки – равномерное расположение узлов, светлые кружки – оптимальное.

На рис. 7 показана зависимость числа итераций от числа узлов. График практически линеен, причем число итераций остается небольшим при всех степенях многочлена. Здесь мы требовали выравнивания экстремумов с точностью 1%; это условие выполнялось при $N \le 11$. Но при $N = 12$ итерации окончились по “аварийному” варианту, однако точность все равно была достигнута высокая.

Поэтому алгоритм на этом тесте показал себя вполне надежным.

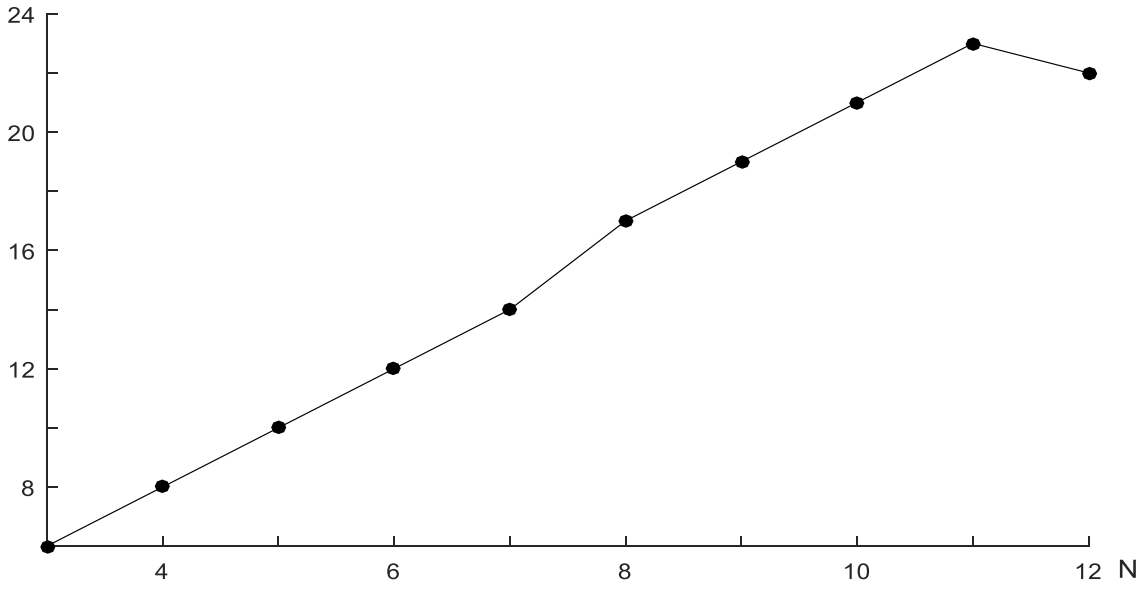

Рис. 7. Зависимость числа итераций от числа узлов.

6. Аппроксимация отношением многочленов; пример 3

Многочлен высокой степени очень быстро возрастает при увеличении $|x|$. Поэтому он зачастую неудобен для аппроксимации, если сама $u(x)$ не очень сильно возрастает (тем более если $|x|$ убывает при увеличении аргумента). В этих случаях для аппроксимации часто используют функцию $Q(x)$, являющуюся отношением двух многочленов:

$$Q(x)=\frac{\sum_{k=0}^{N} a_k x^k}{\sum_{m=0}^{M} b_m x^m}, b_0=1. \qquad (24)$$

Число свободных коэффициентов при этом равно $N+M+1$. Разность степеней N и M обычно выбирают так, чтобы качественно передать поведение $u(x)$ при больших аргументах, а сумму степеней подбирают достаточно большой для обеспечения хорошей точности аппроксимации.

Для определенности потребуем, чтобы $Q(x)$ совпадала с $u(x)$ в узлах x_n. Число таких узлов должно быть равно числу свободных коэффициентов, что дает пределы индекса $0 \le n \le N+M$. Сами условия интерполяции имеют следующий вид:

$$Q(x_n)\equiv\frac{\sum_{k=0}^{N} a_k x_n^k}{\sum_{m=0}^{M} b_m x_n^m}=u_n, u_n=u(x_n), 0\le n\le N+M\ . \qquad (25)$$

Эти условия с учетом $b_0=1$ преобразуются к форме

$$\sum_{k=0}^{N} a_k x_n^k - u_n\sum_{m=1}^{M} b_m x_n^m = u_n, 0\le n\le N+M\ . \qquad (26)$$

Получена система линейных уравнений для определения неизвестных коэффициентов $a_k, 0\le k\le N$ и $b_m, 1\le m\le M$. Ее обусловленность обычно несколько хуже, чем при аппроксимации многочленом.

Алгоритм нахождения чебышевского альтернанса аналогичен тому, который был построен для аппроксимации многочленами. Единственное отличие состоит в другой линейной системе для нахождения свободных коэффициентов.

Замечание 1. Увеличение числа свободных коэффициентов можно производить по-разному, в зависимости от конкретных особенностей $u(x)$.

Если вид функции требует сохранения величины $N-M$, то в новом расчете надо одновременно увеличивать N и M на 1; число свободных коэффициентов при этом увеличивается на 2. Если же сохранение разности $N-M$ необязательно, то можно поочередно прибавлять по одному коэффициенту то в числителе, то в знаменателе. Число свободных коэффициентов при этом будет каждый раз увеличиваться на 1.

Замечание 2. Для аппроксимации (24) в принципе возможна одна неприятность: в расчете могут получиться такие коэффициенты b_m, что многочлен в знаменателе будет обращаться в нуль в какой-то внутренней точке отрезка. Такая ситуация означает срыв расчета. Ее можно опознать по поведению погрешности $\delta(x)$ в дополнительных точках x_j. Соответствующую диагностику необходимо включать в программу.

В случае подобного срыва расчета нужно выбрать другое нулевое приближение. Однако в случае аппроксимации многочленом описанный в разделе 3 алгоритм сходился при любом нулевом приближении. А при аппроксимации отношением многочленов не просматривается способа найти нулевое приближение, обеспечивающее сходимость. Не удается также видоизменить алгоритм так, чтобы он обеспечивал сходимость от любого нулевого приближения.

Расчеты были проведены для $u(x)=e^x$ на отрезке $-1\le x\le 1$, как и в примере аппроксимации многочленом. Для начального расчета полагалось $N=0, M=1$; при этом в числителе и знаменателе было по одному свободному коэффициенту. Далее поочередно прибавлялось по одному коэффициенту то в числителе, то в знаменателе. Расчеты происходили без срыва до $N=4, M=4$ включительно (9 свободных коэффициентов). При добавлении следующего коэффициента расчет сорвался. В то же время при аппроксимации многочленом алгоритм сохранял устойчивость при 12 свободных коэффициентах (т.е. на 3 коэффициента больше).

На рис. 8 показана зависимость логарифма максимальной относительной погрешности от числа свободных коэффициентов на начальной равномерной сетке (черные кружки) и погрешность после выравнивания экстремумов (светлые кружки). На линии, соответствующей начальной равномерной сетке, можно заметить слабый зигзагообразный ход; он обусловлен поочередным прибавлением коэффициента то в числитель, то в знаменатель. На линии, соответствующей сошедшимся итерациям, зигзагообразность практически незаметна. Видно, что при 9 свободных коэффициентах достигается высокая относительная точность 10^{-13}.

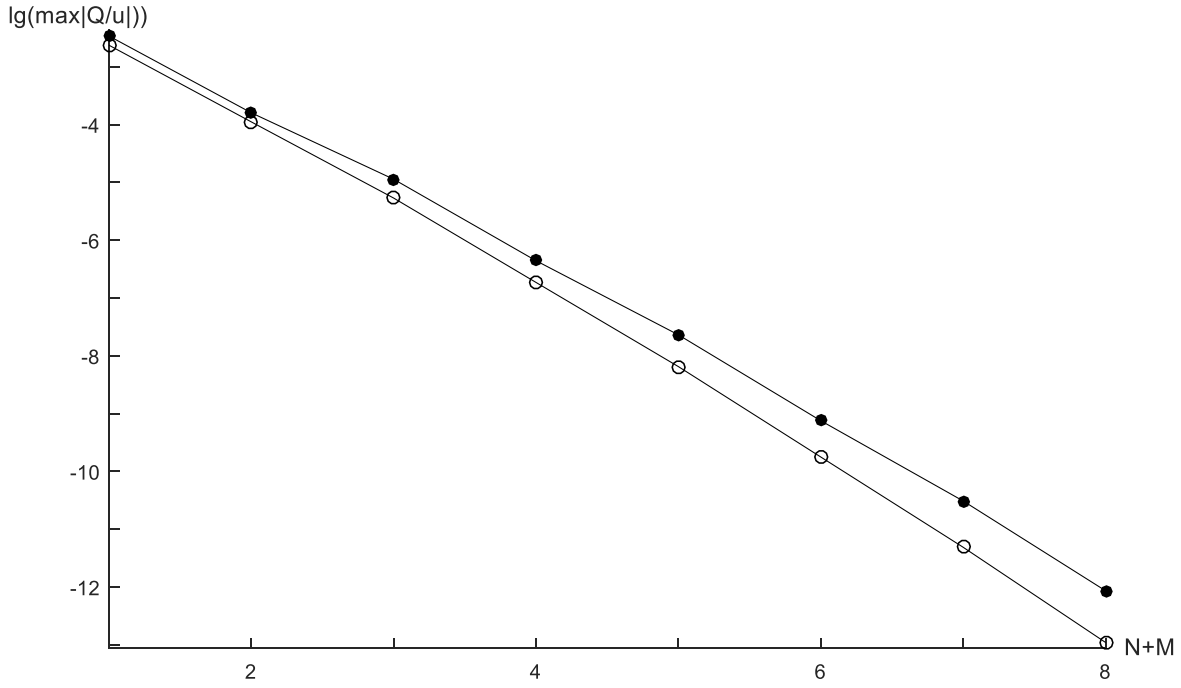

Рис. 8. Зависимость логарифма максимальной относительной погрешности от числа свободных коэффициентов; черные кружки – равномерное расположение узлов, светлые кружки – оптимальное.

На рис. 9 показана зависимость числа итераций от числа свободных коэффициентов. Здесь отчетливо видна зигзагообразная зависимость. Число итераций остается небольшим. Оно в среднем в $\sim 1{,}5$ раза меньше, чем для аппроксимации многочленом с таким же числом коэффициентов (сравните с рис. 7).

Наиболее интересно сравнение погрешностей, которые дают аппроксимации многочленом и отношением многочленов при одинаковом числе свободных коэффициентов. На рис. 10 показана зависимость логарифма этих погрешностей от числа свободных коэффициентов. Видно, что отношение тем выгодней, чем больше число свободных коэффициентов. При 9 коэффициентах выигрыш в точности составляет ~ 250 раз. Поэтому аппроксимация отношением многочленов предпочтительней, несмотря на худшую обусловленность и опасность обращения знаменателя в нуль.

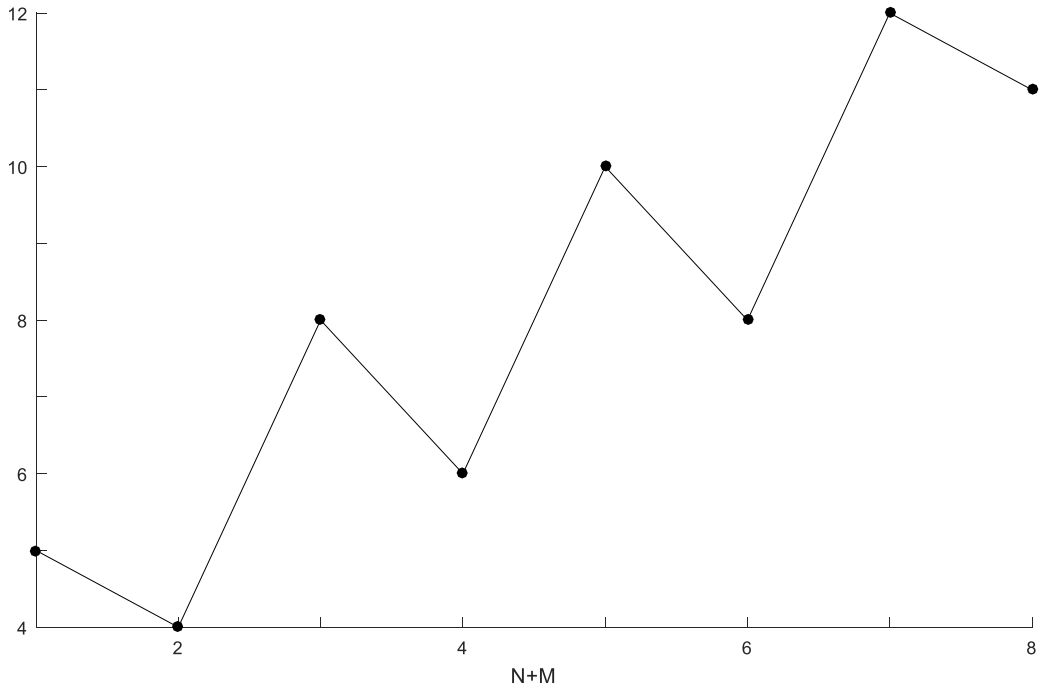

Рис. 9. Зависимость числа итераций от числа свободных коэффициентов.

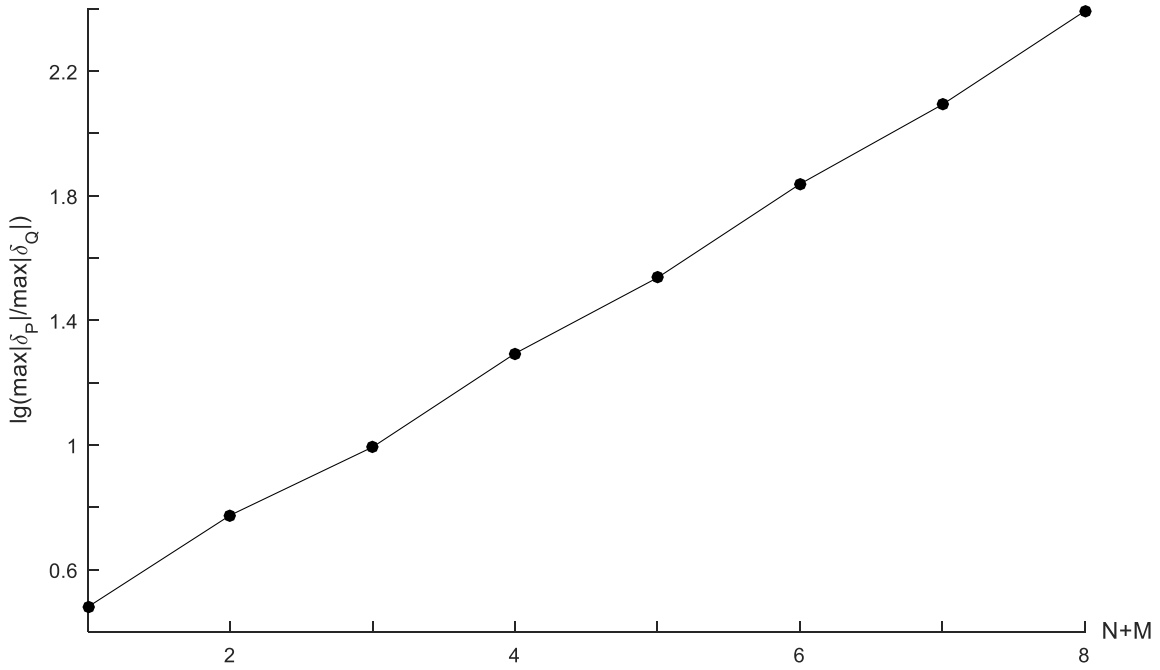

Рис. 10. Зависимость $\lg\left(\max|\delta_P|/\max|\delta_Q|\right)$ от числа свободных коэффициентов.

7. Аппроксимации функций Ферми-Дирака

Проблема. В качестве практического приложения данного метода рассмотрим построение аппроксимаций функций Ферми-Дирака (далее ФД) полуцелого индекса. Эти функции были введены в науку в работах Паули и Зоммерфельда [6-7]. Они играют важную роль в задачах квантовой механики, так как к ним сводится вычисление различных моментов импульса для фермиевского распределения частиц (фермиевскому распределению подчиняются частицы с полуцелым спином – электроны, протоны, нейтроны и многие другие).

Функции ФД индекса k определяются через интеграл, зависящий от параметра:

$$I_k(x) = \int_0^\infty \frac{t^k dt}{1+e^{t-x}}, -\infty < x < +\infty. \qquad (27)$$

В физических приложениях индекс k принимает либо целые значения (для четных моментов импульса), либо полуцелые (для нечетных моментов). В современной математической теории рассматриваются произвольные значения k, а также комплексные значения аргумента x; однако никаких естественно-научных приложений для такого обобщения не найдено.

Первые таблицы функций ФД для важнейших полуцелых индексов $k = -1/2, 1/2, 3/2$ были представлены в [8]. Там же были предложены первые несложные аппроксимации этих функций. Эти аппроксимации имели невысокую по современным понятиям относительную точность $\sim 10^{-4}$. При этом область значений аргумента разбивалась на несколько подобластей, в каждой из которых выбирался свой вид аппроксимирующей формулы. В дальнейшем строились более точные аппроксимации. Самая точная из современных аппроксимаций [9] содержит разбиение на 6 подобластей. При $-\infty < x \le -1$ используется сходящийся ряд, при $30 < x < +\infty$ берется асимптотическое разложение; средняя область разбивается на 4 отрезка точками $x^* = 1, x^* = 4, x^* = 10,$ и в каждом отрезке строится чебышевская аппроксимация многочленом от x. Суммарное число коэффициентов аппроксимации превышает 70. Это не очень благоприятно для вычислений.

Составление аппроксимаций всего из 2 подобластей было проведено в [10-11]. При этом использовался специфический прием: все функции полуцелого индекса выражались через промежуточный аргумент, в качестве которого выбиралась функция $I_{1/2}(x)$. При выборе вида аппроксимирующих формул учитывался характер разложений при $x \to +\infty$ или при $x \to -\infty$ соответственно. Благодаря этому удалось добиться относительной погрешности $\sim 10^{-6} - 10^{-7}$ при суммарном числе коэффициентов 10-15.

В данной работе преследовались три цели. Первая – увеличить число полуцелых индексов, для которых построена хорошая аппроксимация, включающая все полуцелые индексы от $k=-\frac{3}{2}$ до $k=\frac{7}{2}$. Вторая – составить аппроксимацию, состоящую только из двух подобластей. Третья – сделать формулы аппроксимации более естественными и удобными, чем в [10-11].

Замечание. Проблема целых индексов k была решена ранее в [12]. Был построен ряд, который сходится при любых x и k. Сходимость этого ряда неравномерная и быстро падает при $x>0$. Однако при $x\le 0$ его сходимость достаточно быстрая, и поэтому он пригоден для прямых вычислений с любой требуемой точностью. В [13-14] показано, что для целого k функция $I_k(x)$ при $x>0$ легко выражается через значение при $x<0$. Это полностью решает проблему вычисления функций ФД целого индекса с любой требуемой точностью.

Выбор вида аппроксимации. Напомним, что качественное поведение $I_k(x)$ на левой и правой полупрямых принципиально отличается [8]: $I_k(x)\to\Gamma(k+1)e^x$ при $x\to-\infty$, а $I_k(x)\to x^{k+1}\big/(k+1)$ при $x\to+\infty$. Поэтому использовать аргумент x при построении аппроксимирующих формул нецелесообразно. Адекватным является вспомогательный аргумент

$$y(x)=I_0(x)\equiv\ln(1+e^x). \qquad (28)$$

В качестве вспомогательного аргумента взята единственная из функций ФД целого индекса $k=0$, которая точно выражается через элементарные функции от x. Асимптотики этой формы очевидны: $y(x)\to e^x$ при $x\to-\infty$, $y(x)\to x$ при $x\to+\infty$. Эти асимптотики качественно таковы же, как для полуцелых индексов k. Это существенно облегчает выбор аппроксимирующих формул. Разделим прямую на 2 области некоторой точкой x^* и опишем выбор аппроксимирующей формулы в левой и правой областях.

Левая формула. Сравним главные члены асимптотик $I_k(x)$ и $y(x)$ при $x\to-\infty$. Нетрудно видеть, что при этом $I_k(x)\to\Gamma(k+1)y(x)$; далее аргумент x в промежуточном аргументе y будем опускать. Поэтому аппроксимирующую формулу целесообразно строить, умножая $\Gamma(k+1)y$ на корректирующий множитель, являющийся функцией y.

Простейший множитель – это многочлен достаточно высокой степени N; его младший коэффициент равен 1, а следующие коэффициенты являются свободными параметрами, подбираемыми для минимизации относительной погрешности (надо минимизировать не абсолютную, а относительную погрешность, т.к. в этих пределах функция ФД может меняться на много

порядков величины). Однако такой вид аппроксимации для обеспечения хорошей точности потребует слишком много коэффициентов. Вдобавок, асимптотика такой формулы при $x \to +\infty$ будет $I_k(x) \sim y^{N+1} \sim x^{N+1}$ с $N \gg 1$, что принципиально неверно. Поэтому такая формула может обеспечить хорошую аппроксимацию лишь при небольших x^*, чего нам недостаточно.

Можно взять корректирующий множитель в виде отношения многочленов от y степеней N и M соответственно. Здесь так же, как и в случае с многочленом, младшие коэффициенты равны 1, а следующие коэффициенты являются свободными параметрами. Тогда при $x \to +\infty$ асимптотика будет $I_k(x) \sim y^{N-M+1} \sim x^{N-M+1}$. Это целая степень x, поэтому ее нельзя сделать равной полуцелой степени $k+1$. Однако соответствующим выбором разности $N-M$ можно довести отличие этой степени от нужной до всего $1/2$. Такая формула имеет большие шансы на успех при существенно большем значении границы x^*.

Однако можно построить такой корректирующий множитель, который точно передает степень асимптотики при $x \to +\infty$. Для этого достаточно положить

$$I_k(x) \approx \Gamma(k+1) y \left(\frac{\sum_{n=0}^{N} a_n y^n}{\sum_{m=0}^{M} b_m y^m} \right)^k, N = M+1, a_0 = b_0 = 1, -\infty < x \le x^*. \qquad (29)$$

Эта аппроксимация правильно передает степень y при $x \to +\infty$, что должно позволить ещё увеличить значение границы x^*. Если в качестве новой функции взять величину

$$z = \left(\frac{I_k(x)}{\Gamma(k+1) y} \right)^{\frac{1}{k}}, \qquad (30)$$

то z является отношением двух многочленов от y степеней N и M. Таким образом, задача аппроксимации сводится к случаю, рассмотренному в разделе 6.

В задаче аппроксимации целесообразно ввести ещё одно дополнительное условие. Мы предполагаем составить аппроксимацию из двух формул, «склеенных» в точке x^*. Каждая из формул есть непрерывная функция, поэтому склейку также следует сделать непрерывной. Но включение просто требования непрерывности в точке x^* заставило бы нас совместно находить коэффициенты правой и левой формул. Такой алгоритм оказывается достаточно сложным, а его сходимость проблематична. Поэтому проще всего

взять точку x^* как неподвижный узел интерполяции, вычислить в нем значения $y(x^*)$ и $z(y^*)$ и включить этот узел в алгоритм нахождения «левых» коэффициентов a_n, b_m.

Правая формула. При построении формулы при $x > x^*$ мы руководствовались аналогичными соображениями. Отличия заключались в следующем.

Во-первых, при $x \to +\infty$ недостаточно передать только главный член асимптотики $x^{k+1}/k+1 \approx y^{k+1}/k+1$. Для практически приложений надо ещё точно передать следующий член асимптотического разложения (например, теплоемкость электронного газа при низких температурах определяется именно следующим членом разложения функции $I_{3/2}(x)$).

Во-вторых, асимптотическое разложение при $x \to +\infty$ производится по величине $x^{-2} \approx y^{-2}$. Поэтому все многочлены в числителе и знаменателе аппроксимации должны строиться по степеням величины y^{-2}.

В-третьих, для возможно лучшего продолжения формулы в сторону малых x нужно правильно передавать асимптотику $I_k(x) \sim y$ при $y \to 0$, т.е. $x \to -\infty$.

Этим требованиям удовлетворяет следующая аппроксимация:

$$I_k(x) \approx \frac{y}{k+1}\left(y^2 + \frac{\pi^2(k+1)}{3}\frac{\sum_{n=0}^{N}\bar{a}_n y^{-2n}}{\sum_{m=0}^{N}\bar{b}_m y^{-2m}} \right)^{k/2}, \bar{a}_0 = \bar{b}_0 = 1, x^* \le x < +\infty. \qquad (31)$$

Вводя вспомогательную функцию

$$\bar{z} = \frac{3}{\pi^2(k+1)}\left[\left(\frac{k+1}{y} I_k(x) \right)^{\frac{2}{k}} - y^2 \right], \qquad (32)$$

получаем задачу представления функции $\bar{z}$ как отношение двух многочленов по степеням величины y^{-2}. Эта задача, так же как и для левой формулы, описана в разделе 6. Как и для левой формулы, надо брать точку x^* за узел интерполяции, чтобы обеспечить непрерывную склейку формул.

Замечание. Для обеих формул (29) и (31) нас интересует относительная погрешность не вспомогательных функций z или $\bar{z}$, а непосредственно самой функции $I_k(x)$. Описанные выше алгоритмы движения узлов интерполяции допускают использование такой погрешности.

Численные расчеты. ***Ошибки округления.*** В данных численных расчетах влияние ошибок округления оказалось гораздо большим, чем в методических расчетах. Это было связано с тем, что при увеличении степеней многочленов N, M обусловленность линейной системы для нахождения коэффициентов стремительно ухудшалась. Стандартные программы решения линейных систем обычно дают диагностику обусловленности, основанную на спектре матрицы линейной системы: под числом обусловленности понимается отношение наибольшего и наименьшего из модулей собственных значений. Но этот принцип был получен для вычислений с фиксированной точкой, а современные вычисления ведутся с плавающей точкой. Поэтому использовать штатный критерий обусловленности не вполне корректно.

Для данных расчетов мы выработали следующее практическое правило. В ходе вычислений находилась погрешность не только во вспомогательных узлах сетки, а также в узлах интерполяции. При полном отсутствии ошибок округления погрешность во всех узлах интерполяции должна равняться нулю. Если погрешность в узлах интерполяции оказывалась на несколько порядков меньше, чем экстремумы погрешности между ними, то ошибками округления можно было пренебрегать. Если погрешности в узлах интерполяции достигали $\sim 1\%$ от величин экстремумов, то результатам расчетов ещё можно было доверять. Если ошибки в узлах превышали 10% от величин экстремумов, то расчеты считались недостоверными. Уточним, что доверять или не доверять при этом можно самим графикам погрешности. Найденные при этом коэффициенты действительно обеспечивают указанную погрешность. Однако сами величины коэффициентов могут при этом более заметно отличаться от действительно оптимальных коэффициентов, т.е. найденных при расчетах с очень большой разрядностью чисел.

Мы проводили данные расчеты с 64-разрядными числами. При этом нам удалось добиться относительной погрешности $\sim 10^{-8}$. Этого достаточно для прикладных расчетов с точностью single precision. Но чтобы добиться существенно более высокой точности, нужно провести расчеты с числами гораздо большей разрядности.

Начальное расположение узлов. Даже при аппроксимации многочленом мы не полностью застрахованы от срыва алгоритма. Если функция $u(x)$ имеет быстро осциллирующий характер, то возможны ситуации, когда между парой соседних узлов может оказаться не один, а два или более экстремумов погрешности. Тогда алгоритм даст сбой. Алгоритм устойчиво работает лишь тогда, когда хорошо выбрано начальное расположение узлов интерполяции.

Для аппроксимации отношением многочленов влияние начального положения узлов сказывается ещё сильнее. При неудачном расположении узлов возможно возникновении нулей знаменателя, т.е. полюсов аппроксимации. Это также приводит к сбоям алгоритма.

Для левой формулы (29) мы выбирали начальное расположение узлов, используя линейно-тригонометрическое распределение [15] по y. Для правой формулы (31) использовалось аналогичное распределение узлов, но по величине y^{-2}. Однако в отдельных расчетах возникали срывы алгоритма и требовалась ручная корректировка начального приближения.

Таблицы коэффициентов. Для построения аппроксимаций требовалось проводить вычисления функции ФД в узлах интерполяции и во вспомогательных узлах. При $x \le 0$ эти вычисления производились с помощью всюду сходящегося ряда из [12]. При $x \ge 50$ использовалось асимптотическое разложение [8]. При промежуточных значениях аргумента формулы (27) использовалось непосредственное вычисление квадратур сеточными методами. При этом использовалась замена переменной, приводящая интеграл к форме с экспоненциальной по числу узлов сетки сходимостью [16]. Эти способы слишком трудоемки для повседневного вычисления функций, зато они обеспечивают малую относительную погрешность не хуже 10^{-16}.

Целесообразно выбирать x^* так, чтобы число свободных коэффициентов в правой и левой формулах было примерно одинаковым для обеспечения одинаковой точности. Пробные расчеты показали, что значение $x^* = 4$ для всех требуемых индексов k почти удовлетворяет указанным условиям.

Заметим также, что в формулах (29, 31) число членов в числителе и знаменателе жестко связано: при добавлении одного члена в числитель надо добавлять один член в знаменатель. Поэтому суммарные числа коэффициентов могут меняться только на 2. Практика показала, что прибавление каждой пары коэффициентов уменьшает относительную погрешность в ~ 100 раз.

Наилучших результатов в расчетах с 64-разрядными числами нам удалось добиться, когда в левой формуле бралось 7 свободных коэффициентов, а в правой 8. Эти коэффициенты для функций полуцелых индексов от $k = -3/2$ до $k = 7/2$ приведены в Табл. 1-6. Коэффициенты приведены с 10 значащими цифрами; несколько избыточное число знаков взято из-за возможных ошибок округления, возникающих при вычислениях со знакопеременными коэффициентами.

Относительные погрешности для этих наборов коэффициентов приведены в отдельной Табл. 7. Видно, что при $x < x^*$ аппроксимация несколько точнее, чем при $x > x^*$. Это означает, что выбор x^* не оптимален. Немного увеличив x^*, можно было бы уравнять погрешности аппроксимации слева и справа. Однако вряд ли удалось бы получить итоговую точность лучше, чем $2 \cdot 10^{-8}$. Видно также, что наилучшая точность получается для индекса $k = 1/2$. При увеличении индекса погрешность возрастает довольно медленно, а при уменьшении индекса – намного быстрее.

Таблица 1

Коэффициенты формул (29) и (31) для $k = -3/2$

n,m	a_n	b_m	$\bar{a}_n$	$\bar{b}_m$
1	-	-	+4.859830216e+1	3.906201556e+1
2			+5.442288503e+3	5.373283816e+3
3			+2.918221502e+5	1.096397824e+5
4			–7.641044945e+5	1.755046237e+6

Таблица 2

Коэффициенты формул (29) и (31) для $k = -1/2$

n,m	a_n	b_m	$\bar{a}_n$	$\bar{b}_m$
1	4.987087019e-1	8.449799936e-2	2.950244466e+1	2.404833299e+1
2	1.131683972e-1	2.329679817e-2	4.185482375e+3	4.041199712e+3
3	1.466616901e-2	1.159558692e-3	9.668327721e+4	5.961830573e+4
4	1.032378227e-3		1.831673616e+6	1.066504147e+6

Таблица 3

Коэффициенты формул (29) и (31) для $k = 1/2$

n,m	a_n	b_m	$\bar{a}_n$	$\bar{b}_m$
1	3.820834771e-1	8.919175996e-2	6.482980116e+1	6.209431334e+1
2	8.000730853e-2	2.130398793e-2	4.616550102e+3	4.451666901e+3
3	1.004533482e-2	1.076853962e-3	9.809221572e+4	8.094151205e+4
4	6.727185749e-4		3.893359136e+5	4.228938642e+5

Таблица 4

Коэффициенты формул (29) и (31) для $k = 3/2$

n,m	a_n	b_m	$\bar{a}_n$	$\bar{b}_m$
1	3.061333833e-1	9.065231353e-2	7.996251865e+1	7.921410408e+1
2	6.501495275e-2	2.105772040e-2	5.068072837e+3	5.019081662e+3
3	7.906246720e-3	1.029109407e-3	1.158448755e+5	1.106556941e+5
4	5.098901498e-4		3.720135346e+5	4.729702604e+5

Таблица 5

Коэффициенты формул (29) и (31) для $k = 5/2$

n,m	a_n	b_m	$\bar{a}_n$	$\bar{b}_m$
1	2.557236808e-1	9.108008340e-2	8.673851621e+1	8.730933671e+1
2	5.622249451e-2	2.168841174e-2	5.207177175e+3	5.254837248e+3
3	6.785462523e-3	1.077376848e-3	1.328507740e+5	1.343392101e+5
4	4.447657909e-4		4.396602435e+5	5.784887241e+5

Таблица 6

Коэффициенты формул (29) и (31) для $k = 7/2$

n,m	a_n	b_m	$\bar{a}_n$	$\bar{b}_m$
1	2.055638979e-1	7.533494870e-2	1.096825119e+2	1.109124215e+2
2	4.677106802e-2	2.112606808e-2	5.993702472e+3	6.114038798e+3
3	5.438477164e-3	1.001816462e-3	2.016335465e+5	2.065735954e+5
4	3.663039961e-4		7.295591244e+5	9.520099659e+5

Таблица 7

Относительные погрешности формул (29) и (31)

k	$-3/2$	$-1/2$	$1/2$	$3/2$	$5/2$	$7/2$
$\delta_{лев} \cdot 10^8$	–	0,3	0,2	0,5	0,7	1,3
$\delta_{прав} \cdot 10^8$	39	3,0	1,2	2,1	3,0	4,1

8. Заключение

В данной работе предложен метод аппроксимации функции многочленом, реализующий чебышевский альтернанс. Хотя сходимость алгоритма не доказана, но численные примеры убедительно показывают наличие сходимости для сравнительно плавно меняющихся функций $u(x)$. Метод интересен тем, что в отличие от классического чебышевского альтернанса он позволяет минимизировать не только абсолютную, но и относительную погрешность. Он также позволяет вводить в задачу дополнительные ограничения: например, обращение погрешности в нуль на одной или обеих границах интервала.

Метод естественно обобщается на более широкий круг задач. Предложенный алгоритм нахождения чебышевского альтернанса фактически не изменяется, если вместо обыкновенных многочленов пользоваться обобщенными многочленами:

$$P_N(x)=\sum_{k=0}^{N}a_k\varphi_k(x),Q(x)=\sum_{k=0}^{N}a_k\varphi_k(x)/\sum_{m=0}^{M}b_m\psi_m(x); \qquad (33)$$

здесь $\varphi_k(x)$ и $\psi_m(x)$ – некоторые системы линейно-независимых функций. В частности, возможно $\varphi_k(x)\equiv\psi_m(x)$, т.е. использование одной системы функций.

Другое возможное обобщение – это аппроксимация $u(x)$ некоторой монотонной функцией от обобщенного многочлена или отношения обобщенных многочленов:

$$u(x)\approx F\left(P_N(x)\right),u(x)\approx F\left(Q(x)\right). \qquad (34)$$

Разумеется, для практической реализации этого метода не только функция F должна выражаться через элементарные функции своего аргумента, но и обратная к ней функция также должна обладать этим свойством. Фактический пример применения подобной аппроксимации был приведен в разделе 7. При этом также возможна минимизация абсолютной или относительной погрешности.

Все это делает предложенный алгоритм в достаточной степени универсальным и ценным для практических приложений. Разумеется, данный алгоритм рассчитан на достаточно плавно меняющиеся функции. Применять его к сильно осциллирующим функциям вроде $u(x)=\sin\left(1/x\right)$ не следует.

Работа поддержана грантом РФФИ №18-01-00175.

Библиографический список

1. Гончаров В.Л. Теория интерполирования и приближения функций, Гостехиздат, 1954.
2. Бахвалов Н.С. Численные методы, «Наука», т. I, 1975.
3. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы // 3-е издание, Москва, БИНОМ. Лаборатория знаний, 2004.
4. Калиткин Н.Н Численные методы, «Наука», М., 1978, 504 с.
5. Калиткин Н.Н., Альшина Е.А. Численные методы, книга 1, Численный анализ // Москва, «Академия», 2013, 304 с.

6. Pauli W. Uber Gasentartung und Paramagnetismus // Zeutschrift für Physik, 1927, v.41, p.81-102. URL:https://link.springer.com/article/10.1007%2FBF01391920
7. Sommerfeld A. Zur Electronentheore der Metalle aQuf Grundder Fermishen Statistik // Zeutschrift für Physik, 1928, v.47, p.1-3. URL: https://link.springer.com/article/10.1007%2FBF01391052
8. Stoner E. C., McDougall J. The computation of fermi-dirac functions // Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1938, v237(773), p.67-104. URL: https://scinapse.io/papers/2024538905
9. Theiler J. et al. Galassi M., Davies J. GNU scientific library, Network Theory Ltd., 2002.
10. Калиткин Н.Н., Кузьмина Л.В. Интерполяционные формулы для функций Ферми–Дирака, Москва, ИПМ АН СССР, 1972, № 62.
11. Кузьмина Л.В. Численный расчет термодинамических функций веществ в статистической модели атома с квантово-обменными поправками, канд. дисс., Москва, ИПМ АН СССР, 1978.
12. Калиткин Н.Н. О вычислении функций Ферми–Дирака, Ж. вычисл. матем. и матем. физ., 1968, том 8, № 1, с.173–175. URL: http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=zvmmf&paperid=7223&option_lang=rus
13. Калиткин Н.Н., Колганов С.А. Функции Ферми-Дирака. I. Свойства функций., 2018, v.41, p.81-102. URL:https://link.springer.com/article/10.1007%2FBF01391920
14. Калиткин Н.Н., Колганов С.А. Функции Ферми-Дирака. II. Прямое вычисление функций., 2018, v.41, p.81-102. URL:https://link.springer.com/article/10.1007%2FBF01391920
15. Калиткин Н.Н., Колганов С.А. Прецизионные аппроксимации функций Ферми-Дирака целого индекса. Матем. моделирование, 28:3 (2016), 23-32; URL: http://mi.mathnet.ru/mm3708
    Kalitkin N.N, Kolganov S.A., Precision approximations of Fermi-Dirac functions of an integer index, Math. Models and Comput. Simul, 2016, 8:6, 607-614; URL: https://doi.org/10.1134/S2070048216060090
16. Калиткин Н.Н., Колганов С.А. Вычисление функций Ферми–Дирака экспоненциально сходящимися квадратурами // Матем. моделирование, 29:12 (2017),134–146; URL: http://mi.mathnet.ru/rus/mm/v29/i12/p134
    Kalitkin N.N, Kolganov S.A., Computing the Fermi−Dirac Functions by Exponentially Convergent Quadratures, Math. Models Comput. Simul.,10:4 (2018),472–482;

Оглавление