Н. Н. Калиткин, Л. В. Кузьмина, Об аппроксимации неортогональными системами, *Матем. моделирование*, 2004, том 16, номер 3, 95–108

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ 2004 год, том 16, номер 3, стр. 95-108

ОБ АППРОКСИМАЦИИ НЕОРТОГОНАЛЬНЫМИ СИСТЕМАМИ

© *Н.Н.Калиткин, Л.В.Кузьмина*

Институт математического моделирования РАН, Москва

Работа поддержана грантами НШ –1918.2003.1, РФФИ 02-01-00066 и 03-01-00439

Рассмотрена задача среднеквадратичной аппроксимации непериодических функций некоторыми неортогональными базисами – степенным и так называемым двойным периодом. Основной трудностью здесь является решение плохо обусловленных линейных систем для коэффициентов разложения. Исследованы погрешности округления, возникающие при решении этих систем прямыми методами Гаусса и квадратного корня. Определены оптимальные порядки систем и даны рекомендации для практических расчетов. Обнаружено, что метод двойного периода является очень перспективным и может быть альтернативой вейвлет-методу.

ON APPROXIMATION WITH NONORTHOGONAL SYSTEMS

N.N.Kalitkin, L.V.Kuzmina

Institute for Mathematical Modelling of Rus. Acad. Sci., Moscow

A problem of the least square approximation for non-periodic function was discussed when some non-orthogonal systems were used (powers and so called double period). The main difficulty was how to solve an ill-possed linear equations system for coefficients of approximation. Round-off errors were investigated for explicite methods of Gauss and square root. Optimal orders of systems were found, and some recomendations were proposed for practical calculations. The double period method occurs very perspective and may be an alternative to the wavelet-analyses.

1. Проблема. Пусть на $[a, b]$ задана достаточно гладкая (то есть имеющая достаточно много непрерывных производных) непериодическая функция $u(x)$. Рассмотрим задачу наилучшей среднеквадратичной аппроксимации $u(x)$ с помощью некоторого обобщенного многочлена

$$\Phi_M(x) = \sum_{m=0}^{M} c_m \varphi_m(x), \tag{1}$$

где $\{\varphi_m(x),\ m=0,1,2,\ldots\}$ есть некоторая система функций на $[a, b]$. Для этого на $[a, b]$ вводят скалярное произведение в смысле суммы по точкам (при сеточном задании функции) или интеграла

$$(u, v) = \int_a^b u(x)v(x)\rho(x)dx, \qquad \rho(x) > 0; \tag{2}$$

далее будем для определенности брать вес $\rho(x)\equiv 1$. Тогда задача

$$\|u - \Phi_M\|_{L_2} = \min \tag{3}$$

сводится, как известно, к решению системы линейных уравнений для коэффициентов c_m:

$$\sum_{m=0}^{M}(\varphi_k,\varphi_m)c_m=(\varphi_k,u),\qquad 0\le k\le M. \tag{4}$$

Её матрица есть матрица Грама. Если функции $\varphi_m(x)$ линейно независимы, то эта матрица симметрична и положительно определена, так что она невырожденная, и система (4) имеет решение, притом единственное.

Желательно использовать ортогональные системы $\{\varphi_m\}$: тогда матрица Грама диагональна, и система (4) явно решается:

$$c_m=(\varphi_m,u)/(\varphi_m,\varphi_m),\qquad 0\le m\le M. \tag{5}$$

Задача сводится к разложению в обобщенный ряд Фурье. Но для практической ценности метода функции $\varphi_m(x)$ должны легко, точно и устойчиво вычисляться на компьютерах даже при очень больших $m\sim100$ и более. Этим требованиям реально удовлетворяют только две системы. Одна – тригонометрический ряд Фурье для периодических $u(x)$. Для непериодических $u(x)$ остаются только многочлены Чебышева $T_m(x)=\cos(m\arccos x)$, вычисляемые именно по этой формуле (соответствующие тригонометрические подпрограммы есть на всех компьютерах). Это ограничение не всегда удобно, тем более, что разложение по $T_m(x)$ плохо дифференцируется вблизи границ отрезка.

Поэтому на практике нередко приходится пользоваться неортогональными системами $\{\varphi_m\}$. Тогда матрица Грама недиагональна, и зачастую плохо обусловлена при больших и даже небольших M. Это приводит к существенным ошибкам округления при решении линейной системы (4), даже если используется многоразрядный компьютер. При увеличении M точность аппроксимации сначала улучшается; но затем ошибки округления возрастают настолько, что при дальнейшем увеличении M точность ухудшается (нередко быстро). Визуально это почти невозможно определить. Практикам-вычислителям важно знать, сколько членов разложения они могут взять, не рискуя. Но четкого ответа в литературе нет. Поэтому здесь проведено исследование двух практически наиболее важных систем – степеней x^m на отрезках [0,1] и [–1, +1], и функций двойного периода [1]. Их решение проводилось прямыми методами Гаусса и квадратного корня. Разработана система тестов и исследовано возрастание ошибок округления при увеличении порядка системы. Найдены оптимальные значения M и даны рекомендации практикам.

Все расчеты проводились на компьютерах с 64-разрядными числами (это широко распространенные PC при режиме double precision).

2. Система степеней. При обработке экспериментального материала практики широко используют метод наименьших квадратов с системой степеней. Рассмотрим простейший случай, когда $a=0$, $b\ne0$, и за счет изменения масштаба отрезок приводится к единичному:

$$\varphi_m(x)=x^m,\quad m=0,1,\ldots,\quad x\in[0,1]. \tag{6}$$

Их скалярные произведения в смысле интеграла равны

$$(\varphi_k,\varphi_m)=\int_0^1 x^{k+m}dx=\frac{1}{k+m+1},\qquad k,m=0,1,2,\ldots\ . \tag{7}$$

Матрицей Грама (7) оказалась матрица Гильберта. Она очень плохо обусловлена. В докомпьютерное время, когда вычисления велись с 3–5 десятичными знаками, практики сталкивались с большими ошибками округления уже при $M\approx3\div4$, так что обычно ограничивались степенями $M\approx2\div3$.

Если $a\ne0$, и $b/a\approx+1$, обусловленность становится настолько плохой, что практические расчеты вообще невозможны (кроме неинтересного случая $M=0$).

Ситуация существенно улучшается, если за начало координат выбрать середину отрезка $[a, b]$. Это эквивалентно использованию симметризованной системы степеней

$$\varphi_m(x) = x^m, \quad m = 0,1,2,..., \quad x \in [-1,1]. \tag{8}$$

Тогда

$$\frac{1}{2}(\varphi_k, \varphi_m) = \begin{cases} 1/(k+m+1) & \text{при четном} \quad k+m, \\ 0 & \text{при нечетном} \quad k+m. \end{cases} \tag{9}$$

Эта матрица Грама получается из матрицы Гильберта заменой каждой второй косой линии на нули; её обусловленность существенно лучше, что важно при обработке экспериментов.

Приведем вид матрицы Гильберта (7) и матрицы (9):

$$\begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 & \cdots \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & \cdots \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & \cdots \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & \cdots \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \quad \text{и} \quad \begin{pmatrix} 1 & 0 & 1/3 & 0 & 1/5 & \cdots \\ 0 & 1/3 & 0 & 1/5 & 0 & \cdots \\ 1/3 & 0 & 1/5 & 0 & 1/7 & \cdots \\ 0 & 1/5 & 0 & 1/7 & 0 & \cdots \\ 1/5 & 0 & 1/7 & 0 & 1/9 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{10}$$

3. Метод двойного периода [1]. В нем отрезок $[a,b]$ линейно преобразуется в симметричный отрезок $[-\pi/2, \pi/2]$. Рассматривается непериодическая достаточно гладкая $u(x)$ на этом отрезке. Вводится тригонометрическая система, состоящая из двух подсистем. Первая подсистема

$$\{\varphi_n(x), \quad 0 \le n \le 2N\} = \{1, \sin 2x, \cos 2x, \sin 4x, \cos 4x, ..., \sin 2Nx, \cos 2Nx\}; \tag{11}$$

это обычная система Фурье с периодом π, равным длине отрезка задания $u(x)$. Вторая – это функции с удвоенным периодом 2π, равным удвоенной длине отрезка:

$$\{\psi_m(x), \quad 1 \le m \le M\} = \{\sin x, \cos x, \sin 3x, \cos 3x, \sin 5x, ...\}; \tag{12}$$

их число M может иметь любую четность.

Подсистема (11) при $N \to \infty$ полна на $[-\pi/2, \pi/2]$. Поэтому можно разложить $u(x)$ по этой подсистеме в ряд Фурье, вообще не используя подсистему (12). Однако этот ряд будет сходиться к периодическому продолжению $u(x)$, а оно разрывно на границах в точках $x=\pm\pi/2$. Поэтому сходимости в норме C не будет, а в норме L_2 скорость сходимости будет очень медленной: $O(N^{-1/2})$. Кроме того, частичные суммы ряда Фурье при этом сильно осциллируют вблизи разрывов (эффект Гиббса). Это неприемлемо.

Рассмотрим роль подсистемы (12), считая $u(x)$ достаточно гладкой на $[-\pi/2, \pi/2]$. Включение в расчет $\psi_1(x)$ можно интерпретировать как вычитание из $u(x)$ некой величины с таким коэффициентом, чтобы полученная разность $v(x)$ принимала одинаковые значения на границах отрезка $x=\pm\pi/2$. Тогда периодическое продолжение $v(x)$ будет непрерывно, но с разрывом $v'(x)$. Разложение $v(x)$ по первой подсистеме сходится в норме L_2 как $O(N^{-3/2})$ и в норме C как $O(N^{-1})$. Это уже заметно лучше.

Аналогично, включение в расчет $\psi_2(x)$ интерпретируется, как компенсация разрыва первой производной, $\psi_3(x)$ – как второй и т.д. Скорость сходимости разложения по N при этом каждый раз увеличивается на один порядок (пока это позволяет гладкость самой $u(x)$).

Однако при $N=\infty$ добавление даже $\psi_1(x)$ делает задачу разложения переопределенной. Поэтому следует ожидать, что при больших N линейная система (3) будет плохо обусловленной, тем хуже, чем больше M. Исследуем этот вопрос.

Матрица Грама для системы (11)-(12) легко вычисляется; она изображена на рис.1. Матрица состоит из четырех клеток: двух квадратных порядков $2N+1$ и M соответственно, и двух прямоугольных $M\times(2N+1)$ и $(2N+1)\times M$. Большинство матричных элементов оказывается

равным нулю. Каждая из подсистем (11) и (12) ортогональна на [−π/2, π/2], так что в обеих квадратных клетках отличны от нуля только диагональные элементы:

$$a_{00} = \pi; \quad a_{nn} = \pi/2 \quad \text{при} \quad 1 \le n \le 2N + M. \tag{13}$$

Подсистемы (11) и (12) взаимно не ортогональны. Но половина матричных элементов соответствующих скалярным произведениям синусов на косинусы, обращается в нуль. Ненулевыми остаются только скалярные произведения

$$\begin{aligned} &(\cos kx, \cos lx) = (-1)^{(k-l-1)/2}\left(\frac{1}{k-l} + \frac{(-1)^l}{k+l}\right), \\ &\qquad\qquad\qquad\qquad k+l \text{ нечетное}; \\ &(\sin kx, \sin lx) = (-1)^{(k-l-1)/2}\left(\frac{1}{k-l} - \frac{(-1)^l}{k+l}\right), \end{aligned} \tag{14}$$

они расположены в шахматном порядке (см. рис.1).

Видно, что в каждой четной строке матрицы отличны от нуля только четные элементы, а в нечетной – только нечетные. Значит, линейная система (4) расщепляется на две независимые системы, составленные из четных (считая с нулевой) или нечетных строк соответственно. Они отвечают за аппроксимацию четной составляющей $u(x)$ косинусами обеих подсистем (11) и (12), а нечетной – синусами. Матрицы расщепленных систем также имеют клеточную структуру, причем прямоугольные клетки у них полностью заполнены. Такая структура упрощает решение линейной системы и уменьшает ошибки округления.

●	o	o	o	o	o	o	o	o	o	●	o
o	●	o	o	o	o	o	o	o	●	o	●
o	o	●	o	o	o	o	o	o	o	●	o
o	o	o	●	o	o	o	o	o	●	o	●
o	o	o	o	●	o	o	o	o	o	●	o
o	o	o	o	o	●	o	o	o	●	o	●
o	o	o	o	o	o	●	o	o	o	●	o
o	o	o	o	o	o	o	●	o	●	o	●
o	o	o	o	o	o	o	o	●	o	●	o
o	●	o	●	o	●	o	●	o	●	o	o
●	o	●	o	●	o	●	o	●	o	●	o
o	●	o	●	o	●	o	●	o	o	o	●

Рис.1. Матрица Грама двойного периода для N=4 и M=3; кружки – нулевые элементы, точки – ненулевые.

Обусловленность. Её можно теоретически оценить для случаев M=1 и 2 (при M=0 обусловленность идеальная, но аппроксимация очень плохая). При M=1 расщепленная система для косинусов вообще не содержит прямоугольных клеток и диагональна. У системы для синусов эти клетки превращаются в строку и столбец; тогда строка легко исключается методом Гаусса, а диагональный матричный элемент в ней a=π/2 преобразуется в

$$a' = \frac{\pi}{2} - \frac{2}{\pi}\sum_{n=1}^{N}\left(\frac{1}{2n-1} + \frac{1}{2n+1}\right)^2. \tag{15}$$

Последнюю сумму с помощью довольно громоздких выкладок удается разложить в ряд по нечетным степеням $1/(2N+1)$, что дает

$$a' = \frac{4/\pi}{2N+1}\left[1+\frac{1/3}{(2N+1)^2}+\ldots\right]. \tag{16}$$

Отношение исходного и преобразованного элементов можно считать мерой обусловленности

$$æ \equiv a/a' = \frac{\pi^2}{8}(2N+1)\left[1-\frac{1/3}{(2N+1)^2}+\ldots\right], \tag{17}$$

поскольку именно при вычислении a' происходят вычитания, приводящие к потере точности. Видно, что при $N\to\infty$ матрица стремится к вырожденной, но обусловленность ухудшается весьма медленно, так что можно пользоваться очень большими значениями N.

При $M=2$ матрица для системы синусов остается той же, что при $M=1$, но появляются ненулевые строка и столбец в матрице косинусов. Аналогичное исключение строки приводит к следующему ряду для преобразованного по Гауссу диагонального матричного элемента:

$$a' = \frac{4}{3\pi(2N+1)^3}\left[1-\frac{4/3}{(2N+1)^2}+\ldots\right]. \tag{18}$$

Ему соответствует мера обусловленности

$$æ \equiv a/a' = \frac{3\pi^2}{8}(2N+1)^3\left[1+\frac{4/3}{(2N+1)^2}+\ldots\right]. \tag{19}$$

Эта величина много больше чем (17), так что здесь обусловленность ухудшается с увеличением N много быстрее, чем при $M=1$. Поэтому здесь допустимы существенно меньшие значения N.

Для $M>2$ аналогичные оценки не удалось получить. Судя по данным результатам, следует ожидать такой закономерности. Наибольшая потеря точности должна происходить при вычислении последнего диагонального преобразования по Гауссу матричного элемента, а число обусловленности будет $a/a'=\text{const}\cdot(2N+1)^{2M-1}[1+O(N^{-2})]$. Эта гипотеза неплохо подтверждается численными расчетами. Таким образом, увеличение M приводит к очень быстрому ухудшению обусловленности и ужесточает ограничение по N.

Тактика расчета. Подсистема двойного периода (12) служит для ликвидации граничных разрывов самой периодически продолженной $u(x)$ и возможно большего числа ее производных. Подсистема основного "периода" (11) используется для разложения этой улучшенной функции в обычный ряд Фурье. Для хорошей аппроксимации деталей поведения функции – экстремумов (особенно нешироких и многочисленных) необходимо брать достаточно большое N. Кроме того, функции первой подсистемы следует включать в расчет парами синус косинус: это гармоники одной частоты, и коэффициенты при них примерно одинаковы.

Наоборот, функции второй подсистемы целесообразно включать в расчет по одной, постепенно увеличивая M до тех пор, пока при достаточно больших N сохраняется хорошая обусловленность. Попробуем сформулировать критерии обрезания N при выбранном M.

Такие числа обусловленности, как (17) и (19), характеризуют относительную ошибку преобразованного элемента a'. Относительная ошибка последнего коэффициента Фурье по первой подсистеме будет не меньше этой величины (ибо при её расчете используется еще преобразованный член в правой части (4), вносящий дополнительную ошибку). Пусть компьютер имеет относительную погрешность представления чисел ε (для 64-разрядных чисел $\varepsilon\approx 10^{-16}$). Тогда погрешность вычисления последнего коэффициента Фурье будет $\sim\varepsilon æ$.

Если сама $u(x)$ достаточно гладкая, то улучшенная с помощью M членов второй подсистемы функция имеет $M-1$ непрерывную производную, а M-я производная кусочно-непрерывна. Тогда коэффициенты Фурье убывают как $\text{const}\cdot N^{-M-1}$, где константа сравнима с

нормой M-й производной. Члены ряда Фурье разумно суммировать до тех пор, пока коэффициенты ряда превышают ошибки. Отсюда следует критерий обрезания ряда по N:

$$\varepsilon\text{æ} \le \text{const} \cdot N^{-M-1}. \tag{20}$$

Константа здесь зависит от $u(x)$, точнее, M-й производной улучшенной функции; но в дальнейших оценках будем заменять ее единицей.

Подставляя в (20) обусловленности (17) и (19), получим априорную оценку обрезания ряда по N:

$$N \le N_{\max} = \begin{cases} (2.5\varepsilon)^{-1/3} \to 10^5 & \text{при} \quad M = 1, \\ (30\varepsilon)^{-1/6} \to 260 & \text{при} \quad M = 2, \\ (\gamma_M \cdot \varepsilon)^{-1/3M} & \text{для сделанной выше гипотезы;} \end{cases} \tag{21}$$

стрелкой показаны значения для 64-разрядных вычислений. Действительно видно быстрое убывание $N_{\max}$ с ростом M.

При необходимости можно учесть истинное значение опущенной константы. Мы ставим функции $\psi_m(x)$ после функций $\varphi_n(x)$, так что коэффициенты перед ними в обобщенном многочлене будут c_{2N+m}. Тогда улучшенная функция равна

$$v(x) = u(x) - \sum_{m=1}^{M} c_{2N+m} \psi_m(x). \tag{22}$$

Особенно просты случаи $M \le 2$. Тогда легко выписать явные выражения коэффициентов, ликвидирующие разрывы $u(x)$ и $u'(x)$:

$$c_{2N+1} = \left[u(\pi/2) - u(-\pi/2)\right]/2, \quad c_{2N+2} = \left[u'(-\pi/2) - u'(\pi/2)\right]/2 \tag{23}$$

строго говоря, решение системы (4) дает несколько иные, но близкие значения. Для $M>2$ нахождение коэффициентов сложнее, причем первые два коэффициента уже сильно отличаются от (23).

Возьмем M функций $\psi_m(x)$, и исключим разрывы у $u^{(q)}(x)$, $0 \le q \le M-1$. Тогда в ряд Фурье по гармоникам одинарного периода разлагается $v(x)$, имеющая $(M-1)$ непрерывную производную и кусочно-непрерывную M-ю производную. Выпишем её коэффициенты Фурье.

Тогда многократное интегрирование по частям (см. Приложение) дает

$$c_n = \pm \frac{1}{\pi \cdot (2n)^{M+1}} \left[v^{(M)}\left(\frac{\pi}{2}\right) - v^{(M)}\left(-\frac{\pi}{2}\right) \right] \tag{24}$$

и квадратная скобка выражается через граничные значения $u^{(M)}(\pm\pi/2)$. Чем они больше, тем больше константа и $N_{\max}$ при выбранном M; это облегчает задачу аппроксимации функций со многими экстремумами.

4. Решение линейной системы. Для произвольных плотно заполненных матриц существует много методов [2–4]. Мы ограничились двумя.

Первый – это прямой метод Гаусса с выбором главного элемента. Он прост, считается достаточно устойчивым по отношению к ошибкам округления, а также удобно применим к матрицам с плотными массивами нулевых элементов (почти треугольным, квазитреугольным, ящичным, ленточным и т.п.): он позволяет обходить нули и существенно уменьшать трудоемкость вычислений. В литературе нет указаний, чтобы какой-либо общий прямой метод превосходил метод Гаусса по устойчивости к ошибкам округления.

Для этого метода использовалась программа GAUSS, составленная Д.С.Гужевым и адаптированная Д.И.Асоцким к последним версиям математического обеспечения.

Заметим, что в задаче (4) матрица не произвольная $A=A^T>0$. Для таких матриц в методе Гаусса формально не требуется выбора главного элемента: он всегда оказывается

положительным и лежащим на главной диагонали. Но при очень плохой обусловленности ошибки округления могут привести после очередного цикла прямого хода к отрицательному диагональному элементу. Это означает, что результаты становятся недостоверными.

Вторым был взят метод квадратного корня, применимый только к симметричным (точнее, эрмитовым) матрицам. Он основан на приведении матрицы к произведению $A=SDS^T$, где S – треугольная матрица, а D – диагональная с элементами $d_{nn}=\pm 1$; для $A>0$ эти элементы $d_{nn}=+1$, так что в нашем случае матрицу D можно не вводить; однако при плохой обусловленности в ходе процесса на диагонали могут возникать отрицательные элементы, так что для надежной работы программы матрицу D целесообразно сохранять.

Этот метод был опробован в надежде на то, что он окажется более устойчивым к ошибкам округления, чем метод Гаусса (в литературе нет ответа на этот вопрос). Программу SQUART для этого метода составила Л.В.Кузьмина. Использовался вариант без выбора главного элемента, поскольку при $A>0$ здесь также главный элемент автоматически должен оказываться на главной диагонали.

В литературе имелись указания, что итерационные методы типа сопряженных градиентов позволяют получить до 4–5 верных знаков при решении очень плохо обусловленных систем с матрицей Гильберта порядка $N>1000$ (при использовании 64-разрядных вычислений [5]). Проверкой этого подхода занимается сейчас другой коллектив.

Регуляризация нередко рекомендуется для решения плохо обусловленных систем [6]. Особенно просто она выглядит для линейных систем при $A=A^T>0$. Тогда достаточно заменить $A\to A+\alpha E$, где $\alpha>0$ – некоторая достаточно малая величина, называемая параметром регуляризации. Однако оптимальный выбор параметра регуляризации всегда считается непростой задачей. Изложим некоторые соображения.

Пусть относительная погрешность компьютерных чисел есть ε. Тогда абсолютная погрешность матричного элемента есть εa_{nm}, а за величину погрешности всей матрицы можно принять $\|\delta A\|_E=\varepsilon\|A\|_E$; эвклидова норма удобна тем, что она легко вычисляется и одновременно не очень превышает минимальную из норм матрицы – спектральную. Если матрица плотно заполнена, то число операций в методе Гаусса равно $\sim 2N^3/3$, где N – порядок матрицы. По статистике, при таком количестве операций единичная ошибка возрастает в $N^{3/2}$ раз. Поэтому примем регуляризацию

$$A\to A+\alpha N^{3/2}\varepsilon\,\|A\|_E,\quad \text{где}\quad \|A\|_E=\left(\sum_{n,m=1}^{N} a_{nm}^2\right)^{1/2},\quad 0<\alpha\ll 1. \tag{25}$$

При этом можно надеяться подобрать константу (параметр) α, одинаковый для всех плотно заполненных матриц (разумеется, симметричных и положительных). Если матрица неплотно заполнена, и число операций меньше, то вместо $N^{3/2}$ надо ставить корень из числа операций.

5. Методика тестирования. Линейную систему (4) символически запишем как $A\mathbf{c}=\mathbf{b}$. Выберем в качестве точного решения произвольный вектор $\mathbf{c}$. Вычислим соответствующую ему правую часть $\mathbf{b}$ прямым перемножением $\mathbf{b}=A\mathbf{c}$; даже для плохо обусловленной матрицы такое вычисление устойчиво, и отличие вычисленного $\mathbf{b}$ от точного результата составляет $\sim \varepsilon N^{1/2}\,\|\mathbf{b}\|_{l_2}$, им можно пренебречь.

Теперь подставим найденное $\mathbf{b}$ в систему уравнений (4) и решим ее тестируемым методом. Найденное приближенное решение обозначим через $\tilde{\mathbf{c}}$. За меру погрешности примем величину его отличия от исходного вектора $\mathbf{c}$:

$$\delta=\|\tilde{\mathbf{c}}-\mathbf{c}\|_{l_2}\,/\,\|\mathbf{c}\|_{l_2}; \tag{26}$$

нормировка на длину вектора естественна – она делает погрешность инвариантной относительно умножения вектора или матрицы на константу.

При включении регуляризации будем по-прежнему находить $\mathbf{b}=A\mathbf{c}$, но приближенное решение будем восстанавливать по регуляризованной матрице: $\left(A+\alpha\varepsilon\,\|A\|_E\,N^{3/2}\right)\tilde{\mathbf{c}}=\mathbf{b}$.

Остается выбрать такие наборы столбцов **c**, чтобы тестирование было достаточно представительным. Были рассмотрены разные наборы.

1°. Для решений произвольного вида все коэффициенты c_n имеют, вообще говоря, одинаковый порядок; однако их конкретные величины могут существенно различаться. Поэтому использовались тесты следующего типа:

$$c_n = 2\gamma_n - 1, \quad 1 \le n \le N; \tag{27}$$

здесь γ_n – случайные числа, равномерно распределенные на [0, 1]. Таким образом, значения c_n суть случайные числа на отрезке [−1, +1].

Разумеется, расчеты с разными наборами случайных чисел дают разные погрешности (26). Поэтому бралось J наборов случайных чисел, для каждого j-го набора определялась погрешность δ_j, и находились средняя погрешность и оценка ее точности:

$$\delta_c = \frac{1}{J}\sum_{j=1}^{J}\delta_j, \quad \sigma = \left[\frac{1}{J}\sum_{j=1}^{J}(\delta_j-\delta_c)^2\right]^{1/2}. \tag{28}$$

В расчетах достаточным оказалось J=10, при этом обычно получалось $\sigma\approx(1-1.5)\delta$, то есть отличия отношений $\delta_j/\delta_{ср}$ лежали в пределах 2÷3. Это вполне удовлетворительный результат. Заметим, что для графического изображения удобно использовать значение $\lg\delta_c$ как ординату точки и $\lg(1+\delta/\delta_c)$ как величину погрешности этой ординаты.

В качестве γ_n брались псевдослучайные числа [6]. Первая 1000 чисел этой последовательности отбрасывалась, поскольку там заметно влияние простого набора начальных параметров метода. Дальнейшие J отрезков по N+1 чисел (для метода двойного периода – $2N+M+1$ числа) брались как наборы для (27). Отметим только одно смущающее обстоятельство: среди чисел $2\gamma_n-1$ часто попадались подозрительно длинные куски одного знака – по 4–5 чисел подряд. Однако в литературе последовательность [6] считается надежно проверенной.

2°. Практический интерес представляют также коэффициенты c_n, моделирующие разложение $u(x)$ в ряд, приводящее к данной матрице. Например, для метода двойного периода коэффициенты c_n $(0\le n\le 2N)$ моделируют разложение функции в тригонометрический ряд Фурье. Улучшенная функция имеет $M-1$ непрерывную производную и разрывную M-ю; поэтому коэффициенты убывают как $c_n=O(n^{-M-1})$. Остальные коэффициенты определяются граничными скачками функции и ее производных, так что имеют величину $O(1)$. Соответственно выбирался тест

$$c_n = (2\gamma_n-1)/n^{M+1} \quad \text{при} \quad 0\le n\le 2N, \qquad c_n = 2\gamma_n - 1 \quad \text{при} \quad 2N+1\le n\le 2N+M. \tag{29}$$

Многовариантная обработка проводилась согласно (28).

3°. Если разлагать $u(x)$ по многочленам $T_n(x)$, то коэффициенты убывают как $O(n^{-p-1})$, где p – порядок старшей кусочно-непрерывной производной функции (см. Приложение). Для разложения по степеням (6) или (8) лишь старший коэффициент c_N совпадает с коэффициентом разложения по $T_n(x)$. Однако порядки всё же близки; поэтому для матриц (10) мы приняли тест

$$c_n = (2\gamma_n-1)/n^{p+1}, \quad 0\le n\le N; \tag{30}$$

здесь рассматривались разные значения p=0,1,2,…, чтобы промоделировать разложение функций $u(x)$ разной гладкости. Заметим, что p=0 соответствует случаю (27) и означает аппроксимацию кусочно-непрерывной функции.

6. Результаты расчетов. Для разных матриц и методов расчеты проводились в диапазоне $1 \le N \le 60$. Строились графики зависимости $\lg\delta_0$ от N. Поскольку ошибки округления имеют хаотический характер, кривые не вполне плавные; они немного пилообразные. На самом деле каждая точка кривой имеет погрешность σ, которая в переменных графика соответствует примерно $\pm(0.3 \div 0.5)$. Эта погрешность примерно такова, как амплитуды зубцов. Поэтому следовало бы заменить каждую ломанную гладкой полосой ширины $\sim(0.6 \div 1.0)$.

Матрица Гильберта. Результаты расчетов представлены на рис.2. При решении линейной системы методом Гаусса в диапазоне $2 \le N \le 12$ график является почти прямой линией для всех вариантов выбора **c**, и хорошо аппроксимируется формулой

$$\lg\delta_c \approx 1.5N - 18.0 \quad \text{для} \quad \varepsilon = 10^{-16}; \tag{31}$$

при возрастании порядка системы N на единицу погрешность возрастает в ~30 раз (на полтора порядка). При N=12 у решения не остается ни одного верного знака. Затем ошибка резко, с изломом кривой, выходит на уровень $\lg\delta_c \approx 0$ и далее практически перестает расти (при возрастании N до 30 наблюдается еле заметный рост).

Для решений (30), моделирующих коэффициенты Фурье, уровень "полки" слегка зависит от p. Но эта зависимость статистически мало достоверна, поскольку каждая линия является на самом деле полосой.

Был проведен также ряд расчетов с регуляризацией (25). Лучшие результаты дал параметр регуляризации α=0.002. При нем качественный характер кривой не менялся, участок прямолинейного роста сохранялся, а "полка" немного снижалась до $\lg\delta_c \approx -1$ для всех гладкостей $0 \le p \le 6$. Если значение α увеличить в 10–30 раз, то "полка" не улучшается, а точность на участке роста ухудшается. При уменьшении α в 10–30 раз повышается уровень "полки".

Поэтому далее все расчеты методом Гаусса проводились с регуляризацией (25) и параметром α=0.002.

Линейная система с матрицей Гильберта решалась также методом квадратного корня (рис.2): здесь мы ограничились только гладкостью p=0, поскольку влияние p для этой матрицы невелико. Регуляризация не включалась (α=0). Участок линейного роста практически совпал с расчетом методом Гаусса; но при N>12 вместо "полки" наблюдался менее быстрый, но достаточно значительный рост погрешности (при N=30 погрешность достигала $\lg\delta_c \approx +6$, то есть ошибка была в миллион раз больше самого решения!). В рабочей области N<12 метод квадратного корня не уступает методу Гаусса по точности и вдвое экономичнее; но если расчет выходит за рабочую область, он оказывается менее надежным.

Причину этого понять не удалось. Возможно, регуляризация улучшила бы ситуацию. Не исключено также влияние чисто программистких аспектов. Но выяснить это было трудно. Поскольку метод квадратного корня не оправдал надежд, для дальнейшего тестирования он не использовался

Для матрицы Гильберта можно сделать оценку $N_{\max}$ для функций $u(x)$ различной гладкости p аналогично (21). Пусть точность представления чисел есть ε. Тогда сравним $c_N = O(N^{-p-1})$ с погрешностью (31) и получим

$$(p+1)\lg N_{\max} \approx 2 - \lg\varepsilon - 1.5N_{\max}. \tag{32}$$

В табл.1 приведены значения $N_{\max}$ при разных гладкостях p для 64-разрядных расчетов, а также достигаемая при этом погрешность коэффициентов. Видно, что чем выше гладкость, тем меньше $N_{\max}$ и одновременно сильно возрастает точность вычисления коэффициентов (но погрешность аппроксимации $u(x)$ этим отрезком ряда может оказаться много больше, а увеличивать N уже нельзя!).

Таблица 1.

Вычисления с матрицей Гильберта на 64-разрядном компьютере.

p	1	2	4	6	8	11	14
$N_{\max}$	11	10	9	8	7	6	5
δ_c	−1.5	−3.0	−4.5	−6.0	−7.5	−9.0	−10.5

Видно, что при 64-разрядных вычислениях можно получить хорошую точность аппроксимации несимметричным степенным рядом (6), взяв 6–8 членов ряда. Однако при 32-разрядных вычислениях допустимы лишь $N \approx 3 \div 5$, и точность аппроксимации резко падает; этот случай пригоден лишь при обработке экспериментальных измерений, собственные ошибки которых составляют не менее ~1%.

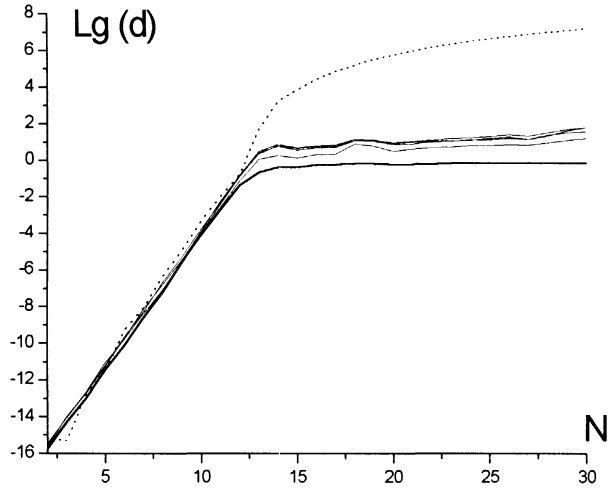

Рис.2. Погрешность решения линейных систем с матрицей Гильберта порядка N. Пунктирная линия – расчет методом квадратного корня без регуляризации (p=0). Сплошные тонкие линии – расчет методом Гаусса без регуляризации (p=0,1,3,6), сплошная толстая линия – с оптимальной регуляризацией (p=0÷6).

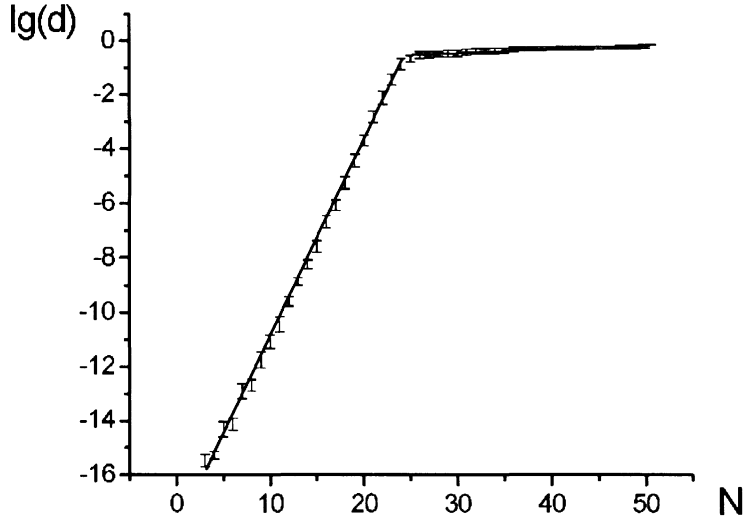

Рис.3. Погрешность решения линейной системы с матрицей (9) методом Гаусса с выбором главного элемента и оптимальной регуляризацией. При каждом N показаны среднее значение погрешности и разброса, а прямые линии изображают сглаженную зависимость от N.

Симметризованная система степеней (8) дает существенно лучшие результаты. Расчеты для неё показаны на рис.3. Здесь график также линейно растет на начальном участке и описывается формулой

$$\lg\delta_c = \begin{cases} -18+0.7N & N \le 25, \\ -0.5 & N > 25, \end{cases} \tag{33}$$

$\varepsilon = 10^{-16}$. Видно, что рабочий участок кривой увеличился примерно вдвое, его наклон уменьшился тоже вдвое, и все значения $N_{\max}$ возрасли вдвое по сравнению с табл.1; "полка" осталась примерно на прежнем уровне.

Поэтому симметричная система степеней (8) позволяет получить хорошие аппроксимации на 64-разрядном компьютере и удовлетворительные на 32-разрядном.

Двойной период. Расчеты погрешности для его матрицы (13)–(14) методом Гаусса с оптимальной регуляризацией для различных M и N показаны на рис.4. В регуляризации (25) вместо $N^{3/2}$ бралось $2N^{1/2}M$, поскольку для этой матрицы число действий в методе Гаусса не $\sim N^3$, а $\sim 4NM^2$. Диапазон изменения N составлял $M/2 \le N \le 60$. Обсудим результаты.

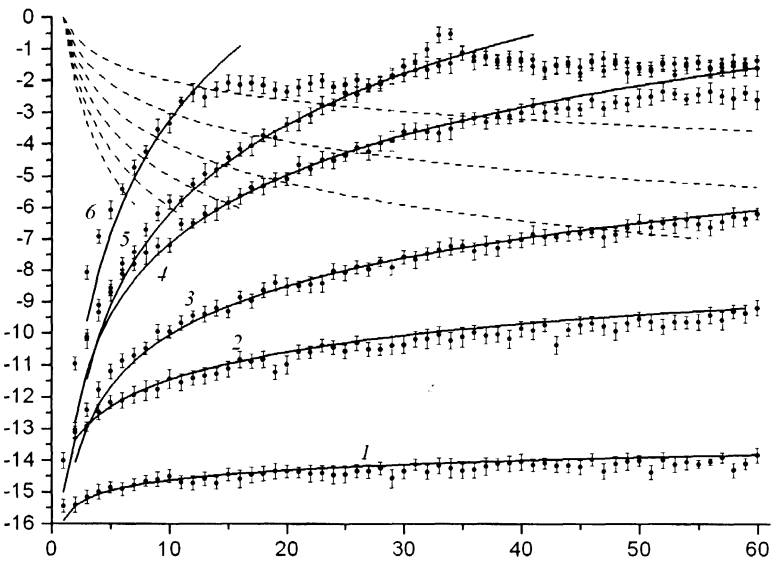

Рис.4. Погрешность решения системы двойного периода: сплошные линии – усредненные кривые для разных M, точки – расчеты для отдельных N с указанием разброса, штриховые линии – коэффициенты Фурье для разных M_x ($u^{(M)} \sim 1$). Цифры около кривых – значения M.

Зависимость погрешности $\lg\delta_c$ от N при M=1 и 2 хорошо соответствует теоретическим оценкам (17) и (19). Для M>2 эти зависимости также неплохо описываются формулой

$$æ \approx a_M \cdot (2N+1)^{2M-1}; \tag{34}$$

константы этой зависимости приведены в табл.2, причём для $M \le 2$ они также подобраны и слегка отличаются от теоретических. Эти закономерности действуют до тех пор, пока $\lg\delta_c \le -0.5$. Далее кривые переходят в "полку" на уровне $\lg\delta_c \approx -0.5$. Видно, что расчеты с M>6 на 64-разрядном компьютере нецелесообразны, ибо ошибки округления слишком велики.

Пунктиром показаны кривые примерно описывающие убывание коэффициентов ряда Фурье $c_N = O(N^{-M-1})$. Их пересечения с соответствующими сплошными линиями дают значения $N_{\max}$ для данного M. Они также хорошо соответствуют теоретическим оценкам (21). В табл.3

приведены зависимости $N_{max}(M)$ для компьютеров с различной разрядностью чисел. Видно, что увеличение разрядности сильно увеличивает возможности метода двойного периода; наоборот, на 32-разрядных компьютерах его затруднительно применять.

Таблица 2.
Коэффициенты формулы (34).

M	1	2	3	4	5	6
$\lg a_M$	−0.2	−0.4	−0.6	−0.2	−1.9	−1.6

Таблица 3.

Значения N_{max} для различных M и разрядности чисел.

число \ M	1	2	3	4	5	6	7	8
32 разряда	210	15	6	4	3			
64 разряда	10^5	260	45	15	10	7	5	4
80 разрядов	$2\cdot10^6$	1500	130	38	18	11	8	6
96 разрядов	$2\cdot10^7$	6800	360	83	34	19	12	9
132 разряда	$2\cdot10^{10}$	$1.5\cdot10^5$	2800	380	120	53	30	19

На основе табл.3 дадим практические рекомендации для вычислений с 64 разрядами (это double precision на широко распространенных PC). Значение $M\geq5$ почти невозможно использовать – слишком мало гармоник основного периода можно включить в аппроксимацию. Значение $M=4$ позволяет аппроксимировать функции $u(x)$ высокой гладкости и простого поведения – с одним–двумя экстремумами. Значение $M=3$ позволяет использовать заметно больше гармоник и аппроксимировать функции с 2–4 экстремумами. При $M=2$ можно описывать $u(x)$ с большим числом экстремумов, но ценой сильного увеличения трудоемкости за счет большого числа гармоник. Наконец, $M=1$ и огромное число гармоник удобно для описания особо сложного поведения функций.

Надо помнить однако, что при $M=1$ и 2 следует ожидать меньшей точности аппроксимации (трудно получить много верных знаков). Кроме того, это резко ухудшает экстраполяционные свойства аппроксимации. При $M=1$ экстраполяция будет иметь на границе отрезка излом (разрыв первой производной), что совершенно не позволяет экстраполировать. При $M=2$ излома не будет, но останется разрыв $u''(x)$, что позволяет экстраполировать лишь на очень небольшое расстояние. Лишь при $M=3\div4$ можно рассчитывать на неплохую экстраполяцию.

Следует помнить, что оценки N_{max} в табл.3 сделаны в предположении $\|u^{(M)}\|\approx1$. Для функций с большим числом экстремумов $\|u^{(M)}\|\gg1$; при этом коэффициенты Фурье намного больше, и N_{max} соответственно увеличиваются. Однако при этом увеличиваются и погрешности коэффициентов Фурье. Все это позволяет при $M=1$ или 2 передать более сложное поведение функций, хотя с меньшей точностью.

Таким образом, тщательное тестирование ошибок округления заметно трансформировало метод двойного периода по сравнению с [1]. Поскольку при малых M можно использовать много гармоник, то появляется возможность аппроксимировать непериодические функции с большим (беря $M=2$) и даже огромным (беря $M=1$) числом экстремумов. Это означает, что метод двойного периода может составить серьезную конкуренцию вейвлет-анализу. Это применение представляется нам очень перспективным.

Заметим, что хотя разложение непериодических функций по многочленам Чебышева $T_n(x)$ позволяет использовать также очень много членов ряда, оно не может конкурировать с вейвлет-анализом. Причина в том, что "всплески" многочленов $T_n(x)$ очень узки вблизи границ отрезков, но широки в центре. Экстремумы же каждой гармоники имеют одинаковую ширину на всем отрезке. Это важное преимущество метода двойного периода.

7. Приложение. Рассмотрим скорость убывания коэффициентов обобщенных рядов Фурье, предполагая непрерывность $u^{(q)}(x)$, $0 \le q \le p$ на $[a, b]$ и существование кусочно-непрерывной $u^{(p+1)}(x)$. Ограничимся двумя примерами.

1°. Пусть $u(x)$ периодическая на $[-\pi, \pi]$ и разлагается в тригонометрический ряд Фурье. Рассмотрим коэффициент Фурье для определенности при косинусе и вычислим его рекуррентным интегрированием по частям. При этом в силу периодичности $u^{(q)}(-\pi)=u^{(q)}(\pi)$, $0 \le q \le p$. Граничные члены сокращаются, так что получим

$$\pi b_n = \int_{-\pi}^{\pi} u(x)\cos ux = \frac{1}{n}u'(x)\sin nx\Big|_{-\pi}^{\pi} - \frac{1}{n}\int_{-\pi}^{\pi} u'(x)\sin uxdx = \ldots = \\ = \pm\frac{1}{np}\int_{-\pi}^{\pi} u^{(p)}(x)\{\sin nx \quad \text{или} \quad \cos nx\}dx; \tag{35}$$

здесь знак и выбор функции в фигурных скобках зависят от p. Последний интеграл в (35) берется аналогично по частям, но выражается через сумму скачков $u^{(p+1)}(x)$, что дает

$$\pi b_n = \pm\frac{1}{n^{p+1}}\sum_j \left[u^{(p+1)}(x_j+0) - u^{(p+1)}(x_j-0)\right]\{\cos nx_j \quad \text{или} \quad \sin nx_j\} = O(n^{-p-1}), \tag{36}$$

где x_j – точки разрыва $u^{(p+1)}(x)$.

Аналогично вычисляется коэффициент для синуса. Напомним, что этот результат общеизвестен.

2°. Пусть задана непериодическая $u(x)$ на $[-1, 1]$. Разложим ее по многочленам Чебышева 1-го рода $T_n(x)=\cos(n \arccos x)$. Эти многочлены ортогональны с весом $\rho(x)=(1-x^2)^{-1/2}$. Напишем коэффициент разложения и перейдем к угловой переменной $\theta=\arccos x$; получим

$$\frac{\pi}{2}c_n = \int_{-1}^{1} u(x)T_n(x)\frac{dx}{\sqrt{1-x^2}} = \int_0^{\pi} v(\theta)\cos(n\theta)d\theta, \quad v(\theta) = u(\cos\theta). \tag{37}$$

Функция $v(\theta)$ непериодична на $[0, \pi]$. Однако последний интеграл (37) можно взять по частям аналогично (35). При q-м интегрировании возникает разность граничных значений для выражения

$$\pm\frac{1}{n^q}\{\sin(n\theta) \quad \text{или} \quad \cos(n\theta)\}\frac{d^q}{d\theta^q}v(\theta), \quad 0 \le q \le p; \tag{38}$$

в фигурных скобках берется синус при четном q, и косинус – при нечетном. Явно выполняя дифференцирование по θ, последовательно получаем из (38) с точностью до знака:

$$q = 0, \quad \frac{1}{n}\sin(n\theta) \quad \text{и} \quad (\cos\theta);$$

$$q = 1, \quad \frac{1}{n^2}\cos(n\theta)\cdot u_x(\cos\theta)\sin\theta;$$

$$q = 2, \quad \frac{1}{n^3}\sin(n\theta)\cdot\left[u_{xx}(\cos\theta)\sin^2\theta - u_x(\cos\theta)\cos\theta\right]; \tag{39}$$

$$q = 3, \quad \frac{1}{n^4}\cos(n\theta)\cdot\left[u_{xxx}(\cos\theta)\sin^3\theta - 3u_{xx}(\cos\theta)\sin\theta\cos\theta - u_x(\cos\theta)\sin\theta\right];$$

и т.д. При четных q в (39) выходит общим множителем $\sin(n\theta)$, а при нечетных – $\sin\theta$; они обращаются в нуль на границе отрезка, аннулируя соответствующие члены. Поэтому после p интегрирований по частям остается

$$\frac{\pi}{2}c_n = \pm\frac{1}{n^p}\int_0^\pi \{\sin(n\theta) \quad \text{или} \quad \cos(n\theta)\}\, v^{(p)}(\theta)d\theta. \tag{40}$$

Последний интеграл также берется по частям и выражается через сумму скачков $v^{(p+1)}(\theta)$ на разрывах. Поэтому $c_n=O(n^{-p-1})$, как и для тригонометрических рядов Фурье.

СПИСОК ЛИТЕРАТУРЫ

1. *Н.Н. Калиткин, Л.В. Кузьмина.* Аппроксимация и экстраполяция табулированных функций. ДАН, 2000, т.374, №4, с.464-468.
2. *Дж. Х. Уилкинсон* Алгебраическая проблема собственных значений. – М.: Наука, 1970.
3. *Д.К. Фадеев, В.Н. Фадеева* Вычислительные методы линейной алгебры. – М.: Физматлит, 1963, 734с.
4. *В.В.Воеводин В.В.* Линейная алгебра. – М.: Наука, 1979, 400с.
5. *А.А. Абрамов, В.И. Ульянова, Л.Ф. Юхно.* О применении метода Крейга к решению линейных уравнений с неточно заданными исходящими данными. // .Ж. выч. матем. и матем. физ. 2002, т.42, №12, с.1763-1770.
6. *И.М. Соболь, Ю.Л. Левитан.* О датчике псевдослучайных чисел для персональных компьютеров. Мат.моделирование, 1990, т.2, №8, с.119-126.

Поступила в редакцию 10.11.2003.