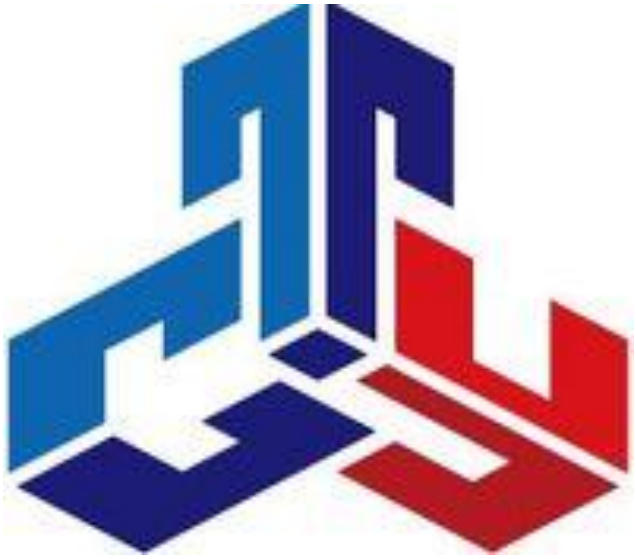
СОВРЕМЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ



Численные методы моделирования

Учебно-методическое пособие для студентов технического вуза

Рязань
2021

УДК 519,6

ББК 22.1

Ч39

Издается по решению ученого совета СТИ

Численные методы моделирования.Учебно-методическое пособие для студентов технического вуза. / Составители Никитина С. Ю., Фроловский М.Ю.–

Совр. техн. универ-т. – Рязань, 2021. – 46 с. – Электронное издание.

Рецензент: к.п.н., доцент кафедры Математики, информатики и лингвистики РФ МИЭМП А.Н. Конюхов

В учебно-методическом пособии изложены теоретические сведения по основным разделам численного моделирования: численному дифференцированию и интегрированию, теории интерполяции функций, преобразованию Фурье, теории матриц и теории решения систем линейных уравнений.

Пособие предназначено студентам технических вузов.

УДК 519,6

ББК 22.1

Ч39

§ 1. Методы численного дифференцирования функций

1.1. Дискретная функция. Методы односторонней разности

Существуют такие функции $f(x)$, аналитическое вычисление производных которых представляет собой сложную задачу и более выгодным является численное дифференцирование.

Производная функции определяется выражением:

$$f'(x_0) = \frac{df}{dx} = \lim_{dx \to 0} \frac{f(x_0 + dx) - f(x_0)}{dx}$$

Заменяя приращение *dx* на конечную величину Δ*x*, называемую шагом дифференцирования, получаем выражение:

$$f'(x_0) = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

Если дифференцируемая функция задана уравнением (рис.1.1), то для вычисления значения дифференциала необходимо получить значение функции $f(x)$ в точке x_0 и в точке $x_0 + \Delta x$. После чего можно вычислить значение производной функции $f'(x_0)$.

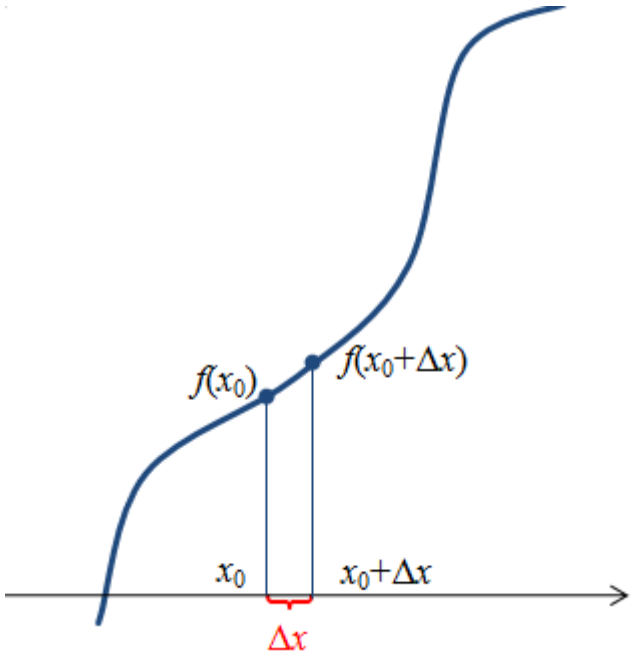


Рис.1.1.Непрерывная функция

Если функция задана выборкой, т.е. набором значений функции в точках (рис.1.2), то выражение для численного дифференцирования (при условии, что x образуют

возрастающую последовательность) можно переписать в виде:

$$f_i' = \frac{f_{i+1} - f_i}{(x_{i+1} - x_i)}$$

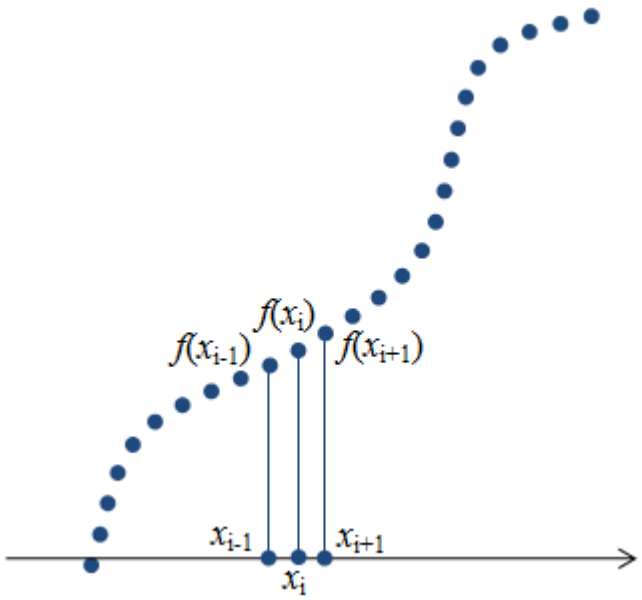


Рис.1.2. Дискретная функция

f_1	x_1
f_2	x_2
…	…
f_i	x_i
…	…
f_n	x_n

Особенно такой подход актуален в том случае, когда набор данных имеет случайное распределение, для которого неизвестна подходящая функциональная зависимость.

Как видно из этих выражений, значение производной в точке x_i, оценивается по значению функции в этой точке и в следующей точке x_{i+1}. Такой способ можно условно назвать правосторонней разностью. Нетрудно записать выражение для левосторонней разности:

$$f'(x_0) = \frac{f(x_0) - f(x_0 - \Delta x)}{\Delta x} \quad \text{или} \quad f_i' = \frac{f_i - f_{i-1}}{(x_i - x_{i-1})}$$

1.2. Метод двусторонней разности

С точки зрения точности оба эти подхода равнозначны. Более точное значение дает метод двусторонней разности (что особенно справедливо для гладких функций). Теорема Лагранжа говорит о том, что уравнение

$$f'(x_0) = \frac{f(b) - f(a)}{b - a}$$

(при условии, что (a,b) – замкнутый промежуток, на котором функция $f(x)$ дифференцируема) имеет по меньшей мере один корень $x = \zeta$. Положение этого корня, вообще говоря, зависит от вида функции $f(x)$. Если она квадратичная, то уравнение первой степени и его корень лежит в точности на середине отрезка (a,b), то есть

$$\zeta = \frac{b+a}{2}$$

Для остальных функций это свойство осуществляется приблизительно. Именно, если *a* имеет постоянное значение, а *b* стремится к *a*, то один из корней, как правило (исключение составляют лишь те случаи, когда вторая производная $f''(a)$ равна нулю или не существует), стремиться к середине отрезка, то есть $\lim_{b \to a} \frac{\zeta - a}{b - a} = \frac{1}{2}$.

Поэтому более точное приближение к искомому значению производной функции в точке x_0 можно получить, воспользовавшись следующим выражением:

$$f'(x_0) = \frac{f(x_0 + \Delta x) - f(x_0 - \Delta x)}{2 \cdot \Delta x}$$

или, для функций заданных в виде выборки:

$$f'_i = \frac{f_{i+1} - f_{i-1}}{(x_{i+1} - x_{i-1})}$$

Эти выражения и носят название формул двусторонней разности.

1.3. Частное дифференцирование функции от многих переменных

Отдельно следует отметить случай численного определения частных дифференциалов функций от многих переменных. В этом случае, все аргументы функции становятся константами, кроме аргумента, по которому проводится дифференцирование, а требуемый порядок производной получается путем последовательного вычисления производных, вплоть до требуемого порядка.

$$\frac{df}{dx_i} = \frac{f(\ldots, x_i + \Delta x_i, \ldots) - f(\ldots, x_i, \ldots)}{\Delta x_i}$$

1.4. Производная высоких порядков

На практике существуют также задачи вычисления производных высоких порядков. При этом один из вариантов вычисления заключается в том, что производная n-го порядка считается первой производной от n-1-го порядка. Так вторая производная функции $f(x)$

является первой производной от первой производной $f(x)$:

$$f''(x) = (f'(x))' \text{ или } \frac{d^2 f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right)$$

При этом формула для вычисления производной может быть представлена следующим образом:

$$\frac{d^2 f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right) = \frac{f'_1 - f'_{-1}}{2h} = \frac{\frac{f'_2 - f'_0}{2h} - \frac{f'_0 - f'_{-2}}{2h}}{2h} = \frac{f'_2 - 2f'_0 + f'_{-2}}{(2h)^2}$$

§ 2. Задача численного интегрирования

В ряде задач возникает необходимость вычисления определенного интеграла от некоторой функции:

$$I = \int_a^b f(x) \cdot dx \qquad (2.1)$$

Где $f(x)$ - подынтегральная функция, непрерывная на отрезке $[a,b]$.

Геометрический смысл интеграла заключается в том, что если $f(x) \geq 0$ на отрезке $[a,b]$, то интеграл $\int_a^b f(x) \cdot dx$ численно равен площади фигуры, ограниченной графиком функции $y = f(x)$, отрезком оси абсцисс, прямой $x = a$ и прямой $x = b$ (рис.2.1). Таким образом, вычисление интеграла равносильно вычислению площади криволинейной трапеции.

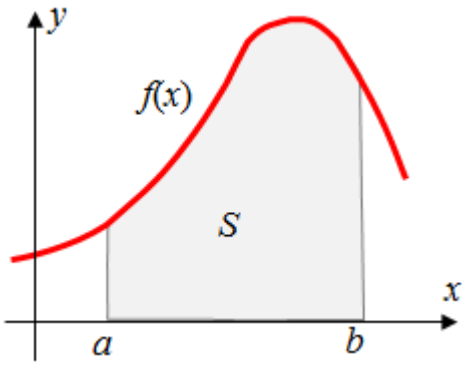


Рис.2.1. Геометрический смысл интеграла

Задача численного интегрирования (историческое название - квадратура) состоит в замене исходной подинтегральной функции некоторой аппроксимирующей функцией (обычно полиномом).

Численное интегрирование применяется, когда:

- сама подынтегральная функция не задана аналитически, а например, представлена в виде таблицы значений;
- аналитическое представление подынтегральной функции известно, но её первообразная не выражается через аналитические функции.

Способы численного вычисления определенных интегралов основаны на замене интеграла конечной суммой:

$$\int_a^b f(x)\cdot dx \approx \sum_{j=1}^{N} c_j \cdot f(x_j) \qquad (2.2)$$

Где c_j – числовые коэффициенты, выбор которых зависит от выбранного метода численного интегрирования, x_j – узлы интегрирования ($x_j \in [a,b], j = 1,\ldots,N$). Выражение (2.2) называют квадратурной формулой.

Разделим отрезок $[a,b]$ на *N* равных частей, т.е. на *N* элементарных отрезков. Длина каждого элементарного отрезка:

$$h = \frac{b-a}{N} \qquad (2.3)$$

Тогда значение интеграла можно представить в виде:

$$\int_a^b f(x)\cdot dx \approx \sum_{j=1}^{N} \int_{x_{j-1}}^{x_j} f(x)\cdot dx \qquad (2.4)$$

Из этого выражения видно, что для численного интегрирования на отрезке $[a,b]$, достаточно построить квадратурную формулу на каждом частичном отрезке $[x_{j-1}, x_j]$.

Погрешность квадратурной формулы определяется выражением:

$$\Psi_N = \int_a^b f(x)\cdot dx - \sum_{j=1}^{N} c_j \cdot f(x_j) \qquad (2.5)$$

и зависит от выбора коэффициентов c_j и от расположения узлов x_j.

Погрешность численного интегрирования определяется шагом разбиения. Уменьшая этот шаг, можно добиться большей точности. Однако, увеличивать число точек не всегда возможно. Если функция задана в табличном виде, приходится ограничиваться заданным множеством точек. Повышение точности может быть в этом случае достигнуто за счет повышения степени используемых интерполяционных многочленов.

Формулы Ньютона-Котеса получаются путем замены подынтегральной функции интерполяционным многочленом Лагранжа с разбиением каждого частичного отрезка интегрирования на *n* равных частей. Получившиеся формулы используют значения подынтегральной функции в узлах интерполяции и являются точными для всех многочленов степени *x* зависящей от числа узлов. Точность формул растет с увеличением степени интерполяционного многочлена.

Метод Гаусса не предполагает разбиения отрезка интегрирования на равные промежутки. Формулы численного интегрирования интерполяционного типа ищутся таким образом, чтобы они обладали наивысшим порядком точности при заданном числе узлов. Узлы и коэффициенты формул численного интегрирования находятся из условий обращения в нуль их остаточных членов для всех многочленов максимально высокой степени.

2.1 Методы Ньютона-Котеса

2.1.1. Метод прямоугольников

Одним из простейших методов численного интегрирования является метод прямоугольников. На частичном отрезке $[x_{j-1}, x_j]$ заменяют подынтегральную функцию полиномом Лагранжа нулевого порядка, построенным в одной точке. Естественно в качестве этой точки выбрать среднюю: $x_{j-0.5} = x_j - 0.5h$. Тогда значение интеграла на частичном отрезке:

$$\int_{x_{j-1}}^{x_j} f(x) \cdot dx \approx f(x_{j-0.5}) \cdot h \qquad (2.6)$$

Подставив это выражение в (2.5), получим составную формулу средних прямоугольников:

$$\int_a^b f(x) \cdot dx \approx \sum_{j=1}^{N} f(x_{j-0.5}) \cdot h \qquad (2.7)$$

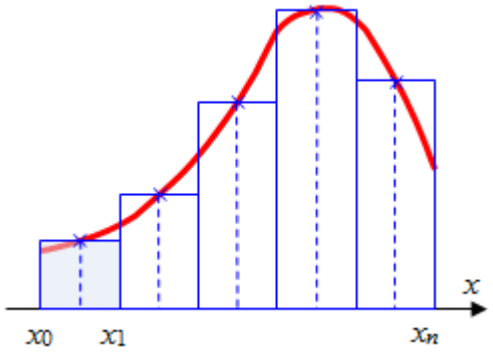
Графическая иллюстрация метода средних прямоугольников представлена на рис.2.2(а). Из рисунка видно, что площадь криволинейной трапеции приближенно заменяется площадью многоугольника, составленного из *N* прямоугольников. Таким образом, вычисление определенного интеграла сводится к нахождению суммы *N* элементарных прямоугольников.

Формулу (2.7) можно представить в ином виде:

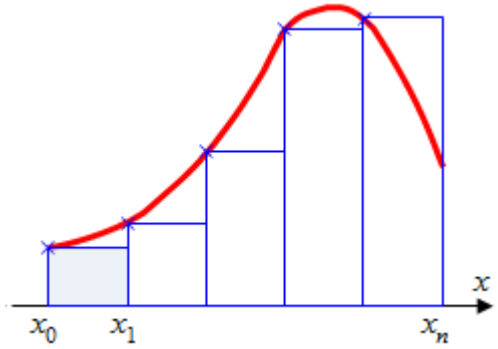
$$\int_a^b f(x) \cdot dx \approx \sum_{j=1}^{N} h \cdot f(x_{j-1}) \quad \text{или} \quad \int_a^b f(x) \cdot dx \approx \sum_{j=1}^{N} h \cdot f(x_j) \qquad (2.8)$$

Эти формулы называются формулой левых и правых прямоугольников соответственно. Графически метод левых и правых прямоугольников представлен на рис.2.2(б, в). Однако

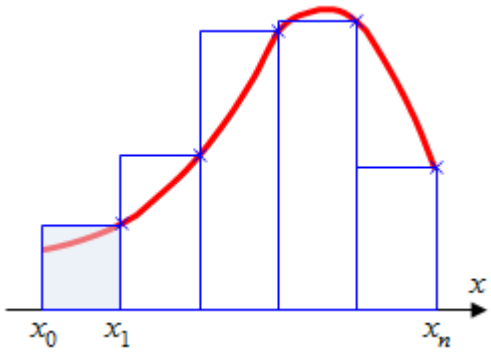
из-за нарушения симметрии в формулах правых и левых прямоугольников, их погрешность значительно больше, чем в методе средних прямоугольников



а) средние прямоугольники



б) левые прямоугольники



в) правые прямоугольники

Рис.2.2. Интегрирование методом прямоугольников

2.1.2. Метод трапеций

Если на частичном отрезке $[x_{j-1}, x_j]$ подынтегральную функцию заменить полиномом Лагранжа первой степени, то есть:

$$f(x) = L_{1,j}(x) = \frac{1}{h}\left[(x - x_{j-1})f(x_j) - (x - x_j)f(x_{j-1})\right] \quad (2.9)$$

то искомый интеграл на частичном отрезке запишется следующим образом:

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx \frac{1}{h}\left[f(x_j)\int_{x_{j-1}}^{x_j}(x - x_{j-1})dx - f(x_{j-1})\int_{x_{j-1}}^{x_j}(x - x_j)dx \right] = \frac{f(x_{j-1}) + f(x_j)}{2}h \quad (2.10)$$

И, составная формула трапеций на всем отрезке интегрирования $[a,b]$ примет вид:

$$\int_a^b f(x)dx \approx \sum_{j=1}^{N}\frac{f(x_j) + f(x_{j-1})}{2}h = h\left[\frac{1}{2}(f_1 + f_N) + f_2 + \ldots + f_{N-1}\right] \quad (2.11)$$

Графически метод трапеций представлен на рис.2.3. Площадь криволинейной трапеции заменяется площадью многоугольника, составленного из *N*трапеций, при этом кривая заменяется вписанной в нее ломаной. На каждом из частичных отрезков функция аппроксимируется прямой, проходящей через конечные значения, при этом площадь трапеции на каждом отрезке определяется по формуле 2.10.

Погрешность метода трапеций выше, чем у метода средних прямоугольников. Однако на практике найти среднее значение на элементарном интервале можно только у функций, заданных аналитически (а не таблично), поэтому использовать метод средних прямоугольников удается далеко не всегда.

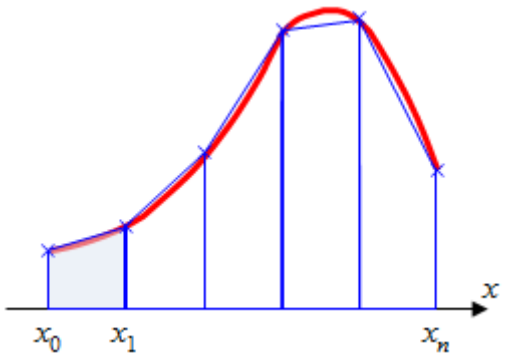


Рис.2.3. Интегрирование методом методом трапеций

2.1.3. Метод Симпсона

В этом методе подынтегральная функция на частичном отрезке $[x_{j-1}, x_j]$ аппроксимируется параболой, проходящей через три точки x_{j-1}, $x_{j-0.5}$, x_j, то есть интерполяционным многочленом Лагранжа второй степени:

$$f(x) = L_{2,j}(x) = \frac{2}{h^2}\left[(x - x_{j-0.5})(x - x_j) f(x_{j-1}) - \right.$$
$$\left. - 2 \cdot (x - x_{j-1})(x - x_j) f(x_{j-0.5}) + (x - x_{j-1})(x - x_{j-0.5}) f(x_j)\right] \qquad (2.12)$$

Проведя интегрирование, получим:

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx \frac{h}{6}\left(f_{j-1} + 4f_{j-0.5} + f_j\right) \qquad (2.13)$$

Это и есть формула Симпсона или формула парабол. На отрезке $[a,b]$ формула Симпсона примет вид:

$$\int_a^b f(x)dx \approx \frac{h}{6}\left[f_0 + f_N + 2\left(f_1 + f_2 + \ldots + f_{N-1}\right) + 4\left(f_{0.5} + f_{1.5} + f_{2.5} + \ldots + f_{N-0.5}\right)\right] =$$
$$= \frac{h}{6}\left[f_0 + f_N + 2 \cdot \sum_{j=1}^{N-1} f_j + 4 \cdot \sum_{j=0.5}^{N-0.5} f_j \right]$$
(2.14)

Можно разбить отрезок интегрирования $[a,b]$ на четное количество 2*N* равных частей с шагом $h = \frac{b-a}{2N}$. Тогда можно построить параболу на каждом сдвоенном частичном отрезке $[x_{j-1}, x_j]$ и переписать выражения (2.12-2.14) без дробных индексов. Тогда формула Симпсона примет вид:

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f_0 + f_{2N} + 2\left(f_2 + f_4 + \ldots + f_{2N-2}\right) + 4\left(f_1 + f_3 + f_5 + \ldots + f_{2N-1}\right)\right] =$$
$$= \frac{h}{3}\left[f_0 + f_{2N} + 2 \cdot \sum_{j=2,2}^{2N-2} f_j + 4 \cdot \sum_{j=1,2}^{2N-1} f_j \right]$$
(2.15)

Графическое представление метода Симпсона показано на рис.2.4. На каждом из сдвоенных частичных отрезков заменяем дугу данной кривой параболой.

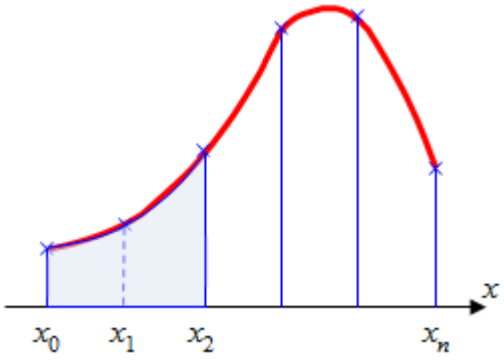


Рис.2.4. Метод Симпсона

2.1.4. Семейство методов Ньютона-Котеса

Выше мы рассмотрели три схожих метода интегрирования функций – метод прямоугольников, метод трапеций, метод Симпсона. Их объединяет общая идея – интегрируемая функция интерполируется на отрезке интегрирования по равноотстоящим узлам многочленом Лагранжа, для которого аналитически вычисляется значение интеграла. Семейство методов, основанных на таком подходе, называется методами Ньютона-Котеса.

$$\int_a^b f(x)\cdot dx \approx \sum_{j=1}^{N} c_j \cdot f(x_j)$$

В выражении мы называли c_j коэффициентами, исходя из их смысла. Однако, правильнее эти величины называть весовыми коэффициентами.

$$\Psi_N = \int_a^b f(x)dx - \sum_{j=0}^{N} c_j f(x_j)$$

Величину , определяющую погрешность численного интегрирования, также называют остатком.

Для семейства методов Ньютона-Котеса можно записать общее выражение:

$$\int_a^b f(x)dx \approx \frac{n\cdot h}{C_n}\sum_{j=1}^{N}\sum_{i=0}^{n} c_{in} f(x_i) \qquad (2.16)$$

где n – порядок метода Ньютона-Котеса, N – количество частичных отрезков, $h = \frac{x_j - x_{j-1}}{n}$, $C_n = \sum_{i=0}^{n} c_{in}$, $x_i = x_j + i\cdot h$.

Коэффициенты c_{in} могут быть заданы в табличной форме:

n	C_n	c_{0n}	c_{1n}	c_{2n}	c_{3n}	c_{4n}	c_{5n}
0	**1**	1					
1	**2**	1	1				
2	**6**	1	4	1			

3	**8**	1	3	3	1		
4	**90**	7	32	12	32	7	
5	**288**	19	75	50	50	75	19

Рис.2.4. Весовые коэффициенты метода Ньютона-Котеса

Из выражения (2.16) легко можно получить формулу прямоугольников для $n=0$, формулу трапеций для $n=1$, и формулу Симпсона для $n=2$.

2.2 Метод Гаусса

В формулах численного интегрирования Ньютона-Котеса используются равноотстоящие узлы. В случае квадратурных формул Гаусса узлы интегрирования x_i на отрезке $[x_{j-1}, x_j]$ располагаются не равномерно, а выбираются таким образом, чтобы при наименьшем возможном числе узлов точно интегрировать многочлены наивысшей возможной степени.

$$\int_a^b f(x)\cdot dx \approx \frac{a-b}{2N}\sum_{j=1}^{N}\sum_{i=0}^{n} c_{in}\cdot f(x_i) \qquad (2.17)$$

Узлы x_i являются корнями полинома Лежандра степени *n*, а веса вычисляются интегрированием полиномов Лежандра по формуле $a_i = \frac{2}{(1-x_i^2)[P_n'(x_i)]^2}$, где P_n' - первая производная полинома Лежандра.

Приведенные в таблице 2.5 данные рассчитаны для отрезка $[-1,1]$, для интегрирования на произвольном частичном отрезке необходимо пересчитать значения узлов для данного отрезка $[x_{j-1}, x_j]$:

$$x_i = x_{j-1} + \frac{(x_{i[-1;1]}+1)(x_{j-1}-x_j)}{2} \qquad (2.18)$$

n	*i*	$x_{i[-1;1]}$	c_i
1	1	0	2
2	1	-0.5773503	1
	2	0.5773503	1
3	1	-0.7745967	0.5555556
	2	0	0.8888889
	3	0.7745967	0.5555556

4	1	-0.8611363	0.3478548
	2	-0.3399810	0.6521451
	3	0.3399810	0.6521451
	4	0.8611363	0.3478548
5	1	-0.9061798	0.4786287
	2	-0.5384693	0.2369269
	3	0	0.5688888
	4	0.5384693	0.2369269
	5	0.9061798	0.4786287
6	1	-0.9324700	0.1713245
	2	-0.6612094	0.3607616
	3	-0.2386142	0.4679140
	4	0.2386142	0.4679140
	5	0.6612094	0.3607616
	6	0.9324700	0.1713245

Рис.2.5. Весовые коэффициенты метода Гаусса

Правила Гаусса относятся к правилам открытого типа. Это означает, что ни один и узлов не совпадает ни с одним из концов отрезка интегрирования *a* или *b*. "Открытость" полезна тем, что трудности, которые могут возникнуть при вычислении значений подынтегральной функции, связаны обычно именно с концевыми точками. Некоторые функции не определены при $x = 0$, но практически всегда существует $\lim_{x \to 0} f(x)$, поэтому функция может быть доопределена, такая особенность называется устранимой.

Веса квадратур Гаусса всегда положительны и при увеличении числа узлов, точность приближения почти всегда возрастает.

2.3 Методы Монте–Карло

Рассмотренные методы называются *детерминированными*, то есть лишенными элемента случайности.

Методы Монте–Карло – это численные методы решения математических задач с помощью моделирования случайных величин. Методы Монте-Карло позволяют успешно решать математические задачи, обусловленные вероятностными процессами. Более того, при решении задач, не связанных с какими-либо вероятностями, можно искусственно придумать вероятностную модель (и даже не одну), позволяющую решать эти задачи.

При вычислении интеграла по формуле прямоугольников интервал $[a, b]$ разбивается на *N* одинаковых интервалов, в серединах которых вычисляются значения подынтегральной функции. Вычисляя значения функции в случайных узлах, можно получить более точный результат:

$$\int_a^b f(x)dx \approx \frac{b-a}{N}\sum_{j=1}^{N} f(x_j)$$, где $x_j = a + \gamma_j(b-a)$. (2.20)

Где γ_j – случайное число, равномерно распределенное на интервале $[0,1]$.

Погрешность вычисления интеграла методом Монте-Карло значительно больше, чем у ранее рассмотренных детерминированных методов. Однако, при вычислении кратных интегралов детерминированными методами оценка погрешности перерастает в задачу порой более сложную, чем вычисление интеграла. В то же время погрешность вычисления кратных интегралов методом Монте-Карло слабо зависит от кратности и легко вычисляется в каждом конкретном случае практически без дополнительных затрат.

§ 3. Интерполяция

3.1. Задача интерполяции

Пусть функция $f(x)$ задана таблицей своих значений x_i, y_i на интервале $[a,b]$:

$$y_i = f(x_i),\ i = 0,1,\ldots,n,\ a \le x_i \le b \quad (3.1)$$

Задача интерполяции - найти функцию $F(x)$, принимающую в точках x_i те же значения y_i.

Условие интерполяции:

$$F(x_i) = y_i \quad (3.2)$$

При этом предполагается, что среди значений x_i нет одинаковых. Точки x_i называют узлами интерполяции.

Если $F(x)$ ищется только на отрезке $[a,b]$ - то это задача интерполяции, а если за пределами первоначального отрезка, то это задача экстраполяции.

- Интерполяция – определение промежуточных значений функции по известному дискретному набору значений функции.
- Экстраполяция – определение значений функции за пределами первоначально известного интервала.
- Аппроксимация – определение в явном виде параметров функции, описывающей распределение точек.

Задача нахождения интерполяционной функции $F(x)$ имеет много решений, так как через заданные точки x_i, y_i можно провести бесконечно много кривых, каждая из которых будет графиком функции, для которой выполнены все условия интерполяции. Для практики важен случай аппроксимации функции многочленами:

$$F(x) = P_m(x_i) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_m \cdot x^m, \; i = 0,1,\ldots,m \qquad (3.3)$$

При этом искомый полином называется интерполяционным полиномом.

При построении одного многочлена для всего рассматриваемого интервала $[a,b]$, для нахождения коэффициентов многочлена необходимо использовать все уравнения системы (3.3). Данная система содержит $n+1$ уравнение, следовательно, с ее помощью можно определить $n+1$ коэффициент. Поэтому максимальная степень интерполяционного многочлена $m = n$, и многочлен принимает вид:

$$P_n(x_i) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_n \cdot x^n, \; i = 0,1,\ldots,n \qquad (3.4)$$

3.2. Локальная и глобальная интерполяция

Если задан $n+1$ узел интерполяции, то на этих узлах можно построить один интерполяционный многочлен *n*-й степени, $n-1$ многочленов первой степени и большой набор многочленов степени меньше *n*, опирающиеся на некоторые из этих узлов.

Теоретически максимальную точность обеспечивает многочлен более высокой степени. Однако на практике наиболее часто используют многочлены невысоких степеней, во избежание погрешностей при расчетах коэффициентов при больших степенях многочлена.

Если функция $f(x)$ интерполируется на отрезке $[a,b]$ с помощью единого многочлена $P_m(x)$ для всего отрезка, то такую интерполяцию называют глобальной. В случае локальной интерполяции на каждом интервале $[x_i, x_{i+1}]$ строится интерполяционный отдельный интерполяционный полином невысокой степени.

3.3. Кусочно-линейная интерполяция

Простейшим и часто используемым видом локальной интерполяции является линейная (или кусочно-линейная) интерполяция. Она заключается в том, что узловые точки соединяются отрезками прямых (Рис.3.1), то есть через каждые две точки (x_i, y_i) и (x_{i+1}, y_{i+1}) проводится полином первой степени:

$$F(x) = a_0 + a_1 \cdot x, \text{ при } x_{i-1} \le x \le x_i \qquad (3.5)$$

Коэффициенты a_0 и a_1 разные на каждом интервале $[x_i, x_{i+1}]$, и находятся из выполнения условий интерполяции на концах отрезка:

$$\begin{cases} f_{i-1} = a_0 + a_1 \cdot x_{i-1} \\ f_i = a_0 + a_1 \cdot x_i \end{cases} \qquad (3.6)$$

Из системы уравнений (3.6) можно найти коэффициенты:

$$a_0 = f(x_{i-1}) - a_1 \cdot x_{i-1}, \quad a_1 = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \qquad (3.7)$$

При использовании кусочно-линейной интерполяции сначала нужно определить интервал, в который попадает значение *x*, а затем подставить его в выражение (3.5), используя коэффициенты для данного интервала.

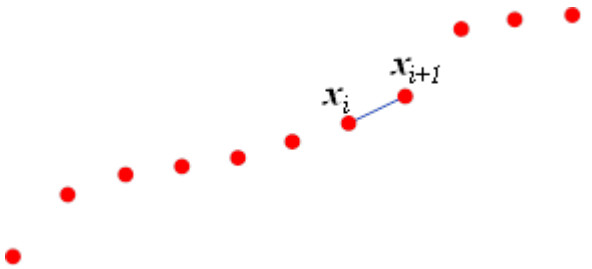


Рис. 3.1. Кусочно-линейная интерполяция

3.4. Кусочно-квадратичная интерполяция

В случае квадратичной интерполяции, для каждых трех узловых точек (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{i+1}, y_{i+1}), строится уравнение параболы:

$$F(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2, \text{ при } x_{i-1} \le x \le x_{i+1} \qquad (3.8)$$

Здесь коэффициенты a_0, a_1 и a_2 разные на каждом интервале $[x_{i-1}, x_{i+1}]$ и определяются решением системы уравнений для условия прохождения параболы через три точки:

$$\begin{cases} f_{i-1} = a_0 + a_1 \cdot x_{i-1} + a_2 \cdot x_{x-i}^2 \\ f_i = a_0 + a_1 \cdot x_i + a_2 \cdot x_x^2 \\ f_{i+1} = a_0 + a_1 \cdot x_{i+1} + a_2 \cdot x_{x+i}^2 \end{cases} \qquad (3.9)$$

Из системы уравнений (3.9) можно найти коэффициенты:

$$\begin{aligned} a_0 &= f(x_{i-1}) - a_1 \cdot x_{i-1} - a_2 \cdot x_{i-1}^2 \\ a_1 &= \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} - a_2 \cdot (x_i + x_{i-1}) \\ a_2 &= \frac{f(x_{i+1}) - f(x_{i-1})}{(x_{i+1} - x_{i-1}) \cdot (x_{i+1} - x_i)} - \frac{f(x_i) - f(x_{i-1})}{(x_i - x_{i-1}) \cdot (x_{i+1} - x_i)} \end{aligned} \qquad (3.10)$$

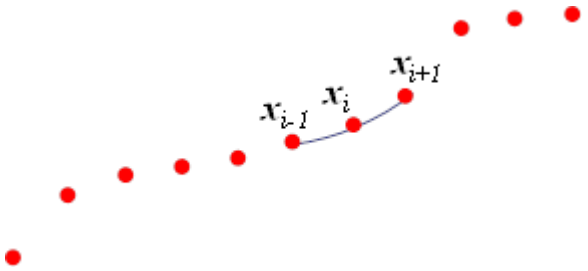


Рис.3.2. Кусочно-квадратичная интерполяция

3.5. Многочлен Лагранжа

При глобальной интерполяции на всем интервале $[a,b]$ строится единый многочлен. Одной из форм записи интерполяционного многочлена для глобальной интерполяции является многочлен Лагранжа:

$$L_n(x)=\sum_{i=0}^{n} y_i \cdot l_i(x) \qquad (3.11)$$

где $l_i(x)$ – базисные многочлены степени *n*:

$$l_i(x)=\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}=\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \qquad (3.12)$$

То есть многочлен Лагранжа:

$$L_n(x)=\sum_{i=0}^{n} y_i \cdot \prod_{\substack{j=1\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j} \qquad (3.13)$$

Многочлен $l_i(x)$ удовлетворяет условию $l_i(x_j)=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$. Это условие означает, что многочлен равен нулю при каждом x_j кроме x_i, то есть $x_0,x_1,\ldots x_{i-1},x_{i+1},\ldots x_n$ – корни этого многочлена. Таким образом, степень многочлена $L_n(x)$ равна *n* и при $x\neq x_i$ в сумме обращаются в нуль все слагаемые, кроме слагаемого с номером $i=j$, равного y_i.

Выражение (3.11) применимо как для равноотстоящих, так и для не равноотстоящих узлов. Погрешность интерполяции методом Лагранжа зависит от свойств функции $f(x)$, от расположения узлов интерполяции и точки *x*. Полином Лагранжа имеет малую погрешность при небольших значениях *n* (*n*<20). При больших *n* погрешность начинает расти, что свидетельствует о том, что метод Лагранжа не сходится (т.е. его погрешность не убывает с ростом *n*).

Многочлен Лагранжа в явном виде содержит значения функций в узлах интерполяции, поэтому он удобен, когда значения функций меняются, а узлы интерполяции неизменны. Число арифметических операции, необходимых для построения многочлена Лагранжа, пропорционально n^2 и является наименьшим для всех форм записи. К недостаткам этой формы записи можно отнести то, что с изменением числа узлов приходится все вычисление проводить заново.

Кусочно-линейная и кусочно-квадратичная локальные интерполяции являются частными случаями интерполяции многочленом Лагранжа.

3.6. Многочлен Ньютона

Другая форма записи интерполяционного многочлена – интерполяционный многочлен Ньютона с разделенными разностями. Пусть функция $f(x)$ задана с произвольным шагом и точки таблицы значений занумерованы в произвольном порядке.

Разделенные разности нулевого порядка совпадают со значениями функции в узлах. Разделенные разности первого порядка определяются через разделенные разности нулевого порядка:

$$f(x_i, x_{i+1}) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \qquad (3.14)$$

Разделенные разности второго порядка определяются через разделенные разности первого порядка:

$$f(x_i, x_{i+1}, x_{i+2}) = \frac{f(x_{i+1}, x_{i+2}) - f(x_i, x_{i+1})}{x_{i+2} - x_i} \qquad (3.15)$$

Разделенные разности *k*-го порядка определяются через разделенную разность порядка $k-1$:

$$f(x_i, x_{i+1}, \ldots, x_{i+k}) = \frac{f(x_{i+1}, \ldots, x_{i+k}) - f(x_i, \ldots, x_{i+k-1})}{x_{i+k} - x_i} \qquad (3.16)$$

Используя понятие разделенной разности интерполяционный многочлен Ньютона можно записать в следующем виде:

$$P_n(x) = f(x_0) + f(x_0, x_1)\cdot(x - x_0) + f(x_0, x_1, x_2)\cdot(x - x_0)\cdot(x - x_1) + \ldots + \\ + f(x_0, x_1, \ldots, x_n)\cdot(x - x_0)\cdot(x - x_1)\ldots\cdot(x - x_{n-1}) \qquad (3.17)$$

За точностью расчета можно следить по убыванию членов суммы (3.17). Если функция достаточно гладкая, то справедливо приближенное равенство $f(x) - P_n(x) \approx P_{n+1}(x) - P_n(x)$. Это приближенное равенство можно

использовать для практической оценки погрешности интерполяции: $\varepsilon_n = |P_{n+1}(x) - P_n(x)|$.

Для повышения точности интерполяции в сумму могут быть добавлены новые члены, что требует подключения дополнительных узлов. При этом для формулы Ньютона безразлично, в каком порядке подключаются новые узлы, в то время как для формулы Лагранжа при добавлении новых узлов все расчеты надо производить заново.

Предположим, что необходимо увеличить степень многочлена на единицу, добавив в таблицу еще один узел x_{n+1}. Для вычисления $P_{n+1}(x)$ достаточно добавить к $P_n(x)$ лишь одно слагаемое $f(x_0,\dots,x_n,x_n)\cdot(x-x_0)\cdot(x-x_1)\dots\cdot(x-x_n)$.

Для повышения точности интерполяции в сумму могут быть добавлены новые члены, что требует подключения дополнительных узлов. При этом для формулы Ньютона безразлично, в каком порядке подключаются новые узлы, в то время как для формулы Лагранжа при добавлении новых узлов все расчеты надо производить заново.

Предположим, что необходимо увеличить степень многочлена на единицу, добавив в таблицу еще один узел x_{n+1}. Для вычисления $P_{n+1}(x)$ достаточно добавить к $P_n(x)$ лишь одно слагаемое $f(x_0,\dots,x_n,x_n)\cdot(x-x_0)\cdot(x-x_1)\dots\cdot(x-x_n)$.

§ 4. Преобразование Фурье и его свойства

4.1. Непрерывное преобразование Фурье и его свойства

4.1.1. Непрерывное преобразование Фурье

Пусть x и y – это пространственные декартовы координаты, а $f(x,y)$ – произвольная двумерная функция. Преобразование Фурье произвольной двумерной функции также является двумерной функцией и определяется следующим интегралом:

$$\tilde{f}(\nu_x,\nu_y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\cdot e^{-2\pi i(x\nu_x+y\nu_y)}dxdy = F[f(x,y)], \quad (4.1)$$

где ν_x и ν_y – частотные декартовы координаты, $\tilde{f}(\nu_x,\nu_y)$ – фурье-образ функции $f(x,y)$, F – оператор преобразования Фурье.

роизвольная двумерная функция и её фурье-образ связаны обратным преобразованием Фурье:

$$f(x,y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\tilde{f}(\nu_x,\nu_y)\cdot e^{2\pi i(\nu_x x+\nu_y y)}\,d\nu_x d\nu_y = F^{-1}\left[\tilde{f}(\nu_x,\nu_y)\right]. \quad (4.2)$$

4.1.2. Основные свойства фурье-образов произвольной функции

В таблице приведены основные свойства фурье-образа произвольной функции. a, b и c – произвольные константы, $g(x,y)$ – произвольная функция, а $\tilde{g}(\nu_x,\nu_y)$ – её фурье-образ.

Функция	Фурье-образ
$\sum_n c_n \cdot f_n(x,y)$	$\sum_n c_n \cdot \tilde{f}_n(\nu_x,\nu_y)$
$f(ax,by)$	$\frac{1}{\lvert ab\rvert}\cdot\tilde{f}\left(\frac{\nu_x}{a},\frac{\nu_y}{b}\right)$
$f(x-a,y-b)$	$\tilde{f}(\nu_x,\nu_y)\cdot e^{-2\pi i\cdot(a\nu_x+b\nu_y)}$
$f(x,y)\otimes g(x,y)$	$\tilde{f}(\nu_x,\nu_y)\cdot\tilde{g}(\nu_x,\nu_y)$

Значение фурье-образа в точке с координатой $\nu=0$ можно представить как сумму всех значений функции, а значение функции в точке с координатой $x=0$ можно представить как сумму всех значений фурье-образа (*теорема о центральном значении*):

$$\tilde{f}(0)=\int_{-\infty}^{+\infty} f(x)dx \quad , \quad f(0)=\int_{-\infty}^{+\infty}\tilde{f}(\nu)d\nu \qquad (4.3)$$

Модуль фурье-спектра убывает пропорционально $\sim\frac{1}{\nu^{n+1}}$, где *n* – порядок дифференцируемости исходной функции. То есть, чем более гладкая функция, тем быстрее убывает ее фурье-спектр (*теорема о производной*):

$$\frac{\partial^n f(x)}{\partial x^n}\overset{F}{\longleftrightarrow}\tilde{f}(\nu)\cdot(2\pi i\nu)^n \qquad (4.5)$$

Количество энергии (сумма всех значений функции), содержащееся в функции после преобразования Фурье (сумма всех значений фурье-спектра) остается неизменной (теорема Парсеваля или закон сохранения энергии):

$$\int_{-\infty}^{+\infty}\lvert f(x)\rvert^2 dx=\int_{-\infty}^{+\infty}\lvert\tilde{f}(\nu)\rvert^2 d\nu$$

4.1.3. Свойства симметрии преобразования Фурье

Функция	Фурье-образ
вещественная и четная	вещественный и четный
вещественная и нечетная	мнимый и нечетный
вещественная и не симметричная	комплексный: вещественная часть четная, мнимая часть нечетная

Фурье-образ функций, обладающих осевой симметрией, может быть найден с использованием преобразования Ганкеля [6, 12]:

$$\tilde{f}(\nu_r) = 2\pi \int_0^\infty f(r) \cdot r \cdot J_0(2\pi r \nu_r) dr$$

, (4.6)

где $r = \sqrt{x^2 + y^2}$ – радиус в полярной пространственной системе координат; $\nu_r = \sqrt{\nu_x^2 + \nu_y^2}$ – радиус в полярной частотной системе координат, а J_0 – функция Бесселя. Справедливо и обратное преобразование:

$$f(r) = 2\pi \int_0^\infty \tilde{f}(\nu_r) \cdot \nu_r \cdot J_0(2\pi \nu_r r) d\nu_r$$

. (4.7)

4.1.4. Фурье-образ двумерной функций с разделяющимися переменными

Фурье-образ двумерной функций с разделяющимися переменными можно определить как произведение фурье-образов составляющих её множителей [6]:

$$\tilde{f}(\nu_x, \nu_y) = F[f(x,y)] = F[f_x(x)] \cdot F[f_y(y)] = \tilde{f}_x(\nu_x) \cdot \tilde{f}_y(\nu_y)$$

. (4.8)

В разделе 4.2.2. приведены аналитические выражения и графики некоторых одномерных функций и их фурье-образов, которые используются для представления функций с разделяющимися переменными.

4.2. Дискретное преобразование Фурье

4.2.1. Спектр периодической функции

Рассмотрим преобразование Фурье от периодической функции $f(x) = g(x - nT)$, где $g(x)$ – функция одного периода, повторяющаяся с периодом T. Такую периодическую функцию можно описать как свертку функции $comb(x/T)$ с функцией одного периода $g(x)$:

$$f(x) = g(x) \otimes comb(x/T)$$

(4.9)

Тогда, согласно свойствам преобразованию Фурье (раздел 4.1.2), спектр периодической функции $f(x)$ можно вычислить как произведение фурье-образа функции $comb(x/T)$ и фурье-образа функции одного периода $g(x)$:

$$\tilde{f}(\nu) = \tilde{g}(\nu) \cdot comb(T\nu)$$

(4.10)

Теорема о спектре периодической функции: спектр периодической функции с периодом T существует только в отдельных точках, то есть является дискретным с шагом 1/T (рис.4.1), причем огибающая дискретного спектра – фурье-образ одного периода функц

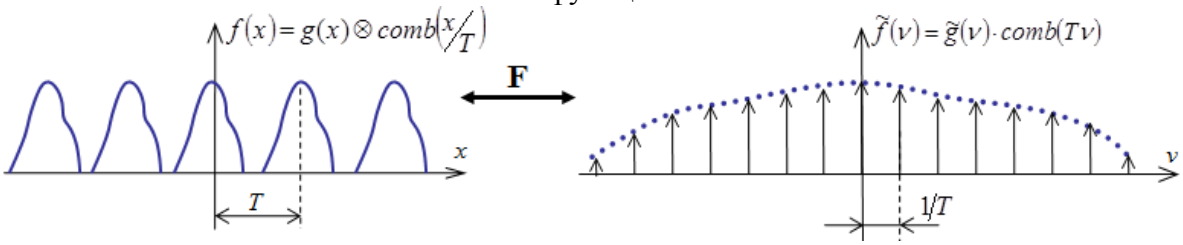


Рис.4.1. Периодическая функция и ее спектр

4.2.2. Спектр дискретной функции

Чтобы описать дискретную функцию, можно представить исходную функцию в виде произведения огибающей $g(x)$ и функции отсчетов $comb\left(x/\Delta x\right)$:

$$f(x)=g(x)\cdot comb\left(x/\Delta x\right) \qquad (4.11)$$

Теорема о спектре дискретной функции: спектр дискретной функции с шагом дискретизации Δx будет периодической функцией с периодом $T=1/\Delta x$, а в пределах одного периода – спектр огибающей выборки (рис.4.2).

При этом частота Найквиста $\nu=1/2\Delta x$ - предельная частота, на которой еще имеет смысл говорить о спектре выборки, дальше будет просто его повторение.

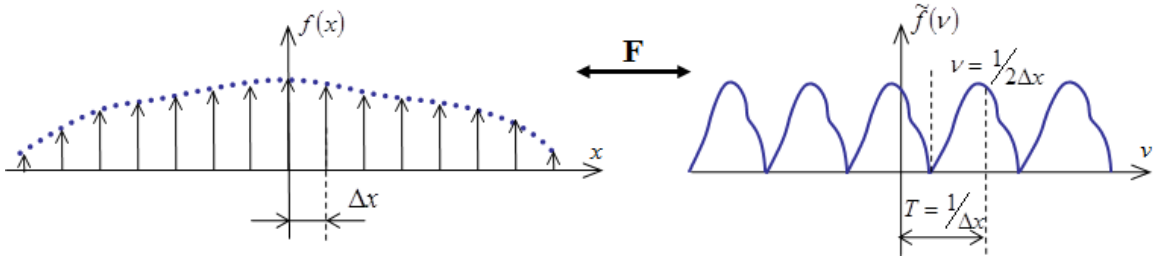


Рис.4.2. Дискретная функция и ее спектр

4.2.3. Теорема о выборке

Теорема о выборке определяет условия, при которых возможно по выборке $f(x)=g(x)\cdot comb\left(x/\Delta x\right)$ восстановить непрерывную функцию $g(x)$. В общем случае восстановить по выборке непрерывную функцию невозможно. Однако если исходная функция имеет финитный спектр Фурье (конечный по протяженности), то при соблюдении определенных условий для шага выборки Δx функцию можно восстановить однозначно.

Теорема о выборке (известна так же как теорема Уиттекера – Шеннона или теорема Котельникова): любая двумерная функция с финитным фурье-образом однозначно определяется выборкой с шагами Δx и Δy, величина которых удовлетворяет неравенствам:

$$\Delta x \le \frac{1}{2\nu_x}; \quad \Delta y \le \frac{1}{2\nu_y}, \tag{4.12}$$

где ν_x и ν_y - предельные частоты в фурье-образе этой функции.

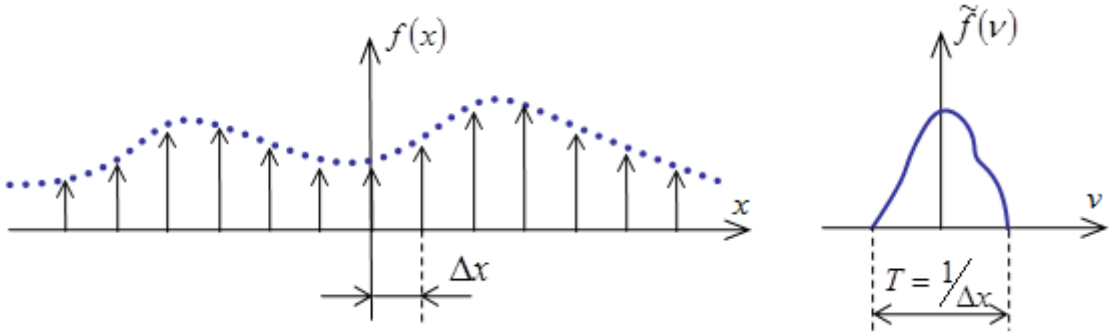


Рис.4.3. Теорема о выборке

4.2.4. Дискретное преобразование Фурье (ДПФ)

При численной реализации преобразования Фурье непрерывные функции заменяются дискретными, а их интегральные преобразования – соответствующими суммами. Двумерное дискретное преобразование Фурье (ДПФ) выборки некоторой функции размером $N \times N$ описывается следующим выражением:

$$\tilde{f}_{mn} = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} f_{kl} \cdot e^{-2\pi i \frac{(km+ln)}{N}}, \tag{4.13}$$

где m – номер элемента в выборке функции, k – номер элемента в выборке фурье-спектра, N – размерность выборок.

Обратное ДПФ определяется выражением:

$$f_{kl} = \frac{1}{N}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} \tilde{f}_{mn} \cdot e^{2\pi i \frac{(mk+nl)}{N}}. \tag{4.14}$$

Вычисление преобразования Фурье путём выполнения непосредственно суммирования является неэффективным, поэтому для вычисления дискретного фурье-образа функций, как правило, используется один из так алгоритмов быстрого преобразования Фурье (БПФ). Разработано большое количество быстрых алгоритмов для вычисления преобразования Фурье (алгоритм Кули-Тьюки, алгоритм Гуда-Томаса, алгоритм Винограда и другие), многие из которых реализованы в виде библиотек на различных языках программирования.

Быстрые алгоритмы работают наиболее эффективно с выборками, размерность которых является 2n, т.е. 2, 4, 16, 32, 64, 128, 256, 512, 1024, 2048 и т.д.

Одной из наиболее эффективных и удобных в использовании библиотек является FFTW. Существенное увеличение производительности вычислений в этой библиотеке достигается за счёт того, что во время выполнения выбирается наиболее подходящий алгоритм БПФ для данной аппаратной и программной среды. Это позволяет оптимизировать выполнение программ на различных платформах и использовать вычислительные ресурсы с максимальной эффективностью.

4.2.5. Сдвиговое дискретное преобразование Фурье (СДПФ)

Свойства ДПФ таковы, что после выполнения вычислений нулевые отсчёты фурье-образа попадают в нулевой элемент выборки (рис.4.4) [8, 22]. Первые $N/2-1$ элементов выборки воспроизводят положительную частотную область фурье-образа, а следующие $N/2$ элементов соответствуют отрицательным частотам. Между тем, при работе с дискретными функциями и их фурье-образами удобнее получать выборку с обычным расположением элементов. Для этого необходимо выполнить циклическое смещение элементов на $N/2$.

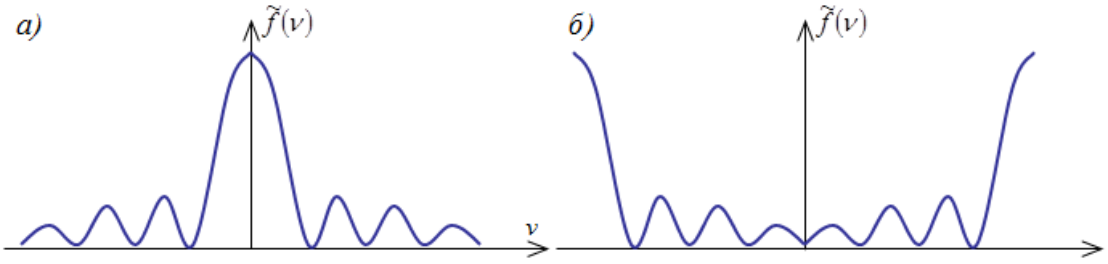


Рис.4.4. Вид спектра функции:
а) после непрерывного преобразования Фурье
б) после дискретного преобразования Фурье

Кроме того, вообще желательно иметь возможность осуществлять произвольное (нецелочисленное) смещение положения элементов относительно начала координат. В частности, при моделировании формирования частично когерентного изображения это особенно важно, так как в вычислениях необходимо использовать фурье-образ функции комплексного пропускания предмета смещённый относительно зрачковой функции.

Преодолеть неудобства использования ДПФ позволяет его модификация, которая осуществляется на основе теоремы о сдвиге преобразования Фурье. Пусть сдвиг нулевого отсчёта функции относительно начала координат составляет $(k_s \cdot \Delta x, l_s \cdot \Delta y)$, а сдвиг нулевого отсчёта фурье-образа относительно начала его координат составляет $(m_s \cdot \Delta \nu_x, n_s \cdot \Delta \nu_y)$. Тогда дискретное представление прямого преобразования Фурье примет вид:

$$\tilde{f}_{mn} = \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} f_{kl} \cdot e^{-2\pi i \frac{(k+k_s)(m+m_s)}{N}} \cdot e^{-2\pi i \frac{(l+l_s)(n+n_s)}{N}}. \qquad (4.15)$$

где m_s, n_s – величина сдвига функции; k_s, l_s – величина сдвига спектра.

Преобразование такого вида называется сдвиговым дискретным преобразованием Фурье (СДПФ). Введение параметров сдвига придаёт СДПФ свойства, которые сближают его с непрерывным преобразованием Фурье. Для получения фурье-образов с привычным расположением начала координат в центре выборки в соответствии с выражением (4.4) необходимо выполнить СДПФ с параметрами сдвига $k_s = N/2$, $l_s = N/2$ и $m_s = N/2$, $n_s = N/2$. СДПФ с параметрами сдвига $k_s = 0$, $l_s = 0$ и $m_s = 0$, $n_s = 0$, совпадает с ДПФ и обладает всеми его свойствами. С помощью пары преобразований: СДПФ и обратного СДПФ, – с надлежащим образом подобранными параметрами сдвига k_s, l_s и m_s, n_s можно получать интерполированные произвольно расположенные промежуточные отсчёты функций.

СДПФ легко выражается через ДПФ тривиальным раскрытием скобок в выражении (4.15):

$$\tilde{f}_{mn} = \left\{ \left[\sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \left(f_{kl} \cdot e^{-2\pi i \frac{km_s}{N}} \cdot e^{-2\pi i \frac{ln_s}{N}} \right) \cdot e^{-2\pi i \frac{km}{N}} \cdot e^{-2\pi i \frac{ln}{N}} \right] \cdot e^{-2\pi i \frac{mk_s}{N}} \cdot e^{-2\pi i \frac{nl_s}{N}} \right\} \cdot e^{-2\pi i \frac{m_s k_s}{N}} \cdot e^{-2\pi i \frac{n_s l_s}{N}}. \quad (4.16)$$

Как видно из выражения (4.16), вычисление СДПФ можно выполнить в 3 этапа:

1. Домножение выборки функции на сдвиговые экспоненты $e^{-2\cdot\pi\cdot i\frac{l_s n}{N}}$, $e^{-2\cdot\pi\cdot i\frac{km_s}{N}}$, обеспечивающие смещение спектра.
2. Выполнение ДПФ, с использованием любого алгоритма БПФ.
3. Домножение выборки спектра на сдвиговые экспоненты $e^{-2\cdot\pi\cdot i\frac{k_s m}{N}}$, $e^{-2\cdot\pi\cdot i\frac{ln_s}{N}}$ компенсирующие смещение выборки.

При этом два последних множителя в выражении (4.16) являются постоянными и не учитываются. Аналогичное выражение и процедура вычисления используется для обратного СДПФ.

§ 5 Основы матричного анализа

Матричные вычисления основываются на иерархии операций линейной алгебры. Скалярные произведения состоят из скалярных операций сложения и умножения; умножение матрицы на вектор составлено из скалярных произведений; перемножение матриц сводится к набору умножений матрицы на вектор.

Основные обозначения

Введем обозначение: $\mathbf{R}$ – множество вещественных чисел, Обозначим через $\mathbf{R}^{M\times N}$ векторное пространство всех вещественных $M \times N$ матриц:

$$\mathbf{A}\in\mathbf{R}^{M\times N} \Leftrightarrow \mathbf{A}=\left(a_{ij}\right)=\begin{bmatrix} a_{00} & \cdots & a_{0n} \\ \vdots & & \vdots \\ a_{m0} & \cdots & a_{mn} \end{bmatrix}, a_{ij}\in\mathbf{R},$$

где M – количество строк, N – количество столбцов, m=M-1 – максимальный индекс строки, N=n-1 – максимальный индекс столбца. Для обозначения матриц обычно используют заглавные буквы $\mathbf{A},\mathbf{B},\mathbf{\Delta}$ (в литературе обычно выделяются жирным шрифтом), соответствующая строчная буква с индексом ij относится к (i,j)-му элементу этой матрицы (например, $a_{ij},b_{ij},\delta_{ij}$), причем первый индекс i означает номер строки, второй индекс j означает номер столбца. Кроме того, для (i,j)-го элемента матрицы **A** может использоваться обозначение $[\mathbf{A}]_{ij}$. При необходимости детальной записи алгоритма для обозначения ij элемента матрицы используют обозначение A(i,j).

Частным случаем матрицы является вектор. Пусть $\mathbf{R}^N$ – векторное пространство вещественных векторов:

$$\mathbf{x}\in\mathbf{R}^N \Leftrightarrow \mathbf{x}=\begin{bmatrix} x_0 \\ \vdots \\ x_n \end{bmatrix}, x_i\in\mathbf{R}$$

При этом, в отличии от матриц, вектора обозначаются, как правило, строчными буквами латинского алфавита (в литературе обычно выделяются жирным шрифтом). Обозначение x_i означает i-ую компоненту вектора **x**, так же используют обозначение x(i).

Следует отметить, что $\mathbf{R}^N$ отождествляется с $\mathbf{R}^{N\times 1}$, так что элементы $\mathbf{R}^N$ – это *векторы-столбцы*. С другой стороны, $\mathbf{R}^{1\times N}$ состоит из*векторов-строк*:

$$\mathbf{x}\in\mathbf{R}^{1\times N} \Leftrightarrow \mathbf{x}=(x_0,\ldots,x_n) .$$

Если **x** – вектор-столбец, то $\mathbf{y}=\mathbf{x}^T$ – вектор-строка. Символ ^{T} означает операцию транспонирования, о которой будет сказано позже.

Разбиение матриц на строки и столбцы есть ни что иное, как частный случай разбиения матриц на блоки. В общем случае, разбиение матриц на блоки выглядит следующим образом:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{00} & \cdots & \mathbf{A}_{0q} \\ \vdots & & \vdots \\ \mathbf{A}_{p0} & \cdots & \mathbf{A}_{pq} \end{bmatrix} \begin{matrix} M_0 \\ \\ M_p \end{matrix}$$
$$\begin{matrix} N_0 & & N_q \end{matrix}$$

Здесь $M_0 + \ldots + M_p = M$, $N_0 + \ldots + N_p = N$, $\mathbf{A}_{ij}$ означает (i,j)-блок, или подматрицу. Блок $\mathbf{A}_{ij}$ имеет размерность M_i на N_j , и будем говорить, что $\mathbf{A} = (\mathbf{A}_{ij})$ есть $P \times Q$ блочная матрица (здесь и далее для обозначения индекса и размерности используются строчные и заглавные буквы соответственно, так, что $m = M - 1$, $n = N - 1$, $p = P - 1$, $q = Q - 1$).

Основные операции над матрицами

Основные манипуляции с матрицами включают в себя следующее:

Транспонирование ($\mathbf{R}^{M \times N} \to \mathbf{R}^{N \times M}$): $\mathbf{C} = \mathbf{A}^T \Rightarrow c_{ij} = a_{ji}$.

$$\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \Rightarrow \begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix}$$

Сложение ($\mathbf{R}^{M \times N} \times \mathbf{R}^{M \times N} \to \mathbf{R}^{M \times N}$): $\mathbf{C} = \mathbf{A} + \mathbf{B} \Rightarrow c_{ij} = a_{ij} + b_{ij}$

Умножение матрицы на число (скаляр) ($\mathbf{R} \times \mathbf{R}^{M \times N} \to \mathbf{R}^{M \times N}$): $\mathbf{C} = \alpha\mathbf{A} \Rightarrow c_{ij} = \alpha a_{ij}$

Умножение матрицы на *матрицу* ($\mathbf{R}^{M \times R} \times \mathbf{R}^{R \times N} \to \mathbf{R}^{M \times N}$): $\mathbf{C} = \mathbf{AB} \Rightarrow c_{ij} = \sum_{k=0}^{R-1} a_{ik} b_{kj}$

Типовые матрицы

Нулевая матрица. Обозначение – [0]. Нулевая матрица размера $M \times N$ есть матрица этого размера, все элементы которой равны 0.

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Для нулевой матрицы справедливы следующие соотношения:

- $\mathbf{A} + [0] = \mathbf{A}$, где **A** – произвольная матрица размера $M \times N$.
- $[0]\mathbf{B} = [0]$, где **B** – произвольная матрица, имеющая N строк.
- $\mathbf{C}[0] = [0]$, где **C** – произвольная матрица, имеющая M столбцов.

Матрица A размера $N \times N$ называется квадратной матрицей порядка N .

Матрица A называется диагональной, если из $i \neq k$ следует $a_{ik} = 0$.

$$\begin{bmatrix} \times & 0 & 0 \\ 0 & \times & 0 \\ 0 & 0 & \times \end{bmatrix}$$

Если **D** – диагональная, а **A** – произвольная матрица, то произведение **DA** масштабирует **A** по строкам, а **AD** – по столбцам.

Матрица **A** называется *верхнетреугольной* (*наддиагональной*), если из $i > k$ следует $a_{ik} = 0$.

$$\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}$$

Матрица **A** называется *строго верхнетреугольной*, если из $i \geq k$ следует $a_{ik} = 0$.

$$\begin{bmatrix} 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{bmatrix}$$

Матрица **A** называется *нижнетреугольной* (*поддиагональной*), если из $i < k$ следует $a_{ik} = 0$.

$$\begin{bmatrix} \times & 0 & 0 \\ \times & \times & 0 \\ \times & \times & \times \end{bmatrix}$$

Матрица **A** называется *строго нижнетреугольной*, если из $i \le k$ следует $a_{ik} = 0$.

$$\begin{bmatrix} 0 & 0 & 0 \\ \times & 0 & 0 \\ \times & \times & 0 \end{bmatrix}$$

Единичная матрица. Обозначается **I**. Единичная матрица порядка N есть диагональная матрица размера $N \times N$ с единичными диагональными элементами:

$\mathbf{I} \equiv [\delta_{ik}]$, где $\delta_{ik} = \begin{cases} 0, & i \neq k \\ 1, & i = k \end{cases}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Для единичной матрицы справедливы следующие соотношения:

- $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$, где **A** – произвольная квадратная матрица порядка N .
- $\mathbf{IB} = \mathbf{B}$, где **B** – произвольная матрица, имеющая N строк.
- $\mathbf{C}[0] = [0]$, где **C** – произвольная матрица, имеющая N столбцов.

Квадратная матрица **A** порядка $N \times N$ называется *симметрической* (*симметричной*), если $\mathbf{A}^T = \mathbf{A}$, то есть если $a_{ik} = a_{ki}$.

$$\begin{bmatrix} \times & a & b \\ a & \times & c \\ b & c & \times \end{bmatrix}$$

Квадратная матрица **A** порядка $N \times N$ называется *кососимметрической* (*антисимметрической*), если $\mathbf{A}^T = -\mathbf{A}$, то есть если $a_{ik} = -a_{ki}$.

$$\begin{bmatrix} \times & a & b \\ -a & \times & c \\ -b & -c & \times \end{bmatrix}$$

Ленточная матрица. Будем говорить, что матрица **A** размера $M \times N$ имеет *нижнюю ширину ленты* p, если $a_{ij} = 0$ для $i > j + p$, и *верхнюю ширину ленты* q, если из $j > i + q$ следует $a_{ij} = 0$. Пример матрицы 8×5 с нижней шириной ленты 1 и верхней шириной ленты 2:

$$\begin{bmatrix} \times & \times & \times & 0 & 0 \\ \times & \times & \times & \times & 0 \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Определения линейной алгебры

Набор векторов $\{\mathbf{a}_0, \ldots, \mathbf{a}_n\}$ из пространства $\mathbf{R}^M$ называется *линейно независимым*, если из выражения $\sum_{j=0}^{n} \alpha_j \mathbf{a}_j = 0$ следует, что $\alpha(0, n) = 0$. В противном случае существует равная 0 нетривиальная линейная комбинация векторов a_i, и говорят, что набор $\{\mathbf{a}_0, \ldots, \mathbf{a}_n\}$ *линейно зависим*.

Подпространство $\mathbf{R}^M$ – это такое подмножество $\mathbf{R}^M$, которое также является линейным пространством. Если заданы векторы $\mathbf{a}_0, \ldots, \mathbf{a}_n \in \mathbf{R}^M$, то множество всевозможных линейных комбинаций этих векторов является линейным пространством, называемым *линейной оболочкой* $\{\mathbf{a}_0, \ldots, \mathbf{a}_n\}$:

$$span\{\mathbf{a}_0, \ldots, \mathbf{a}_n\} = \left\{ \sum_{j=0}^{n} \beta_j \mathbf{a}_j : \beta_j \in \mathbf{R} \right\}.$$

Если $\{\mathbf{a}_0, \ldots, \mathbf{a}_n\}$ – линейно независимый набор векторов и $\mathbf{b} \in span\{\mathbf{a}_0, \ldots, \mathbf{a}_n\}$, то **b** единственным образом представляется в виде линейной комбинации векторов $\mathbf{a}_j$.

Подмножество $\{\mathbf{a}_{i_0},\ldots,\mathbf{a}_{i_k}\}$ называется *максимальным линейно независимым подмножеством* $\{\mathbf{a}_0,\ldots,\mathbf{a}_n\}$, если оно линейно независимо и не является собственным подмножеством никакого другого линейно независимого подмножества $\{\mathbf{a}_0,\ldots,\mathbf{a}_n\}$. Если $\{\mathbf{a}_{i_0},\ldots,\mathbf{a}_{i_k}\}$ – максимальное линейно независимое подмножество, то $span\{\mathbf{a}_0,\ldots,\mathbf{a}_n\} = span\{\mathbf{a}_{i_0},\ldots,\mathbf{a}_{i_k}\}$ и $\{\mathbf{a}_{i_0},\ldots,\mathbf{a}_{i_k}\}$ является *базисом* для $span\{\mathbf{a}_0,\ldots,\mathbf{a}_n\}$. Если $S \in \mathbf{R}^M$ – подпространство, то можно найти базисные вектора $\mathbf{a}_0,\ldots,\mathbf{a}_k \in S$ так, что $S = span\{\mathbf{a}_0,\ldots,\mathbf{a}_k\}$. Все базисы подпространства S имеют одинаковое количество элементов. Это число называется *размерностью* S и обозначаются $\dim(S)$.

С каждой матрицей **A** размера $M \times N$ связаны два важных подпространства: *область значений* и *ядро (нуль-пространство)*.

Область значений матрицы **A** определяется так:

$$range(\mathbf{A}) = \{y \in \mathbf{R}^M : \mathbf{y} = \mathbf{Ax}, \quad \mathbf{x} \in \mathbf{R}^N\}.$$

Если $\mathbf{A} = [\mathbf{a}_0,\ldots,\mathbf{a}_n]$ есть разбиение по столбцам, то

$$range(\mathbf{A}) = span\{\mathbf{a}_0,\ldots,\mathbf{a}_n\}.$$

Ядро (нуль-пространство) матрицы **A** задается следующим образом:

$$null(\mathbf{A}) = \{\mathbf{x} \in \mathbf{R}^N : \mathbf{Ax} = \mathbf{0}\}.$$

Ранг матрицы **A** определяется следующим образом:

$$rank(\mathbf{A}) = \dim(range(\mathbf{A})).$$

Обратная матрица для матрицы $\mathbf{A}$ обозначается $\mathbf{A}^{-1}$ определяется условием:

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Если $\mathbf{A}^{-1}$ существует, то матрица $\mathbf{A}$ называется *невырожденной*, в противном случае $\mathbf{A}$ называется *вырожденной*.

Обратная к произведению матрица является произведением обратных к сомножителям, взятым в обратном порядке:

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Транспонирование обратной матрицы – это то же самое, что обращение транспонированной:

$$\left(\mathbf{A}^{-1}\right)^{T} = \left(\mathbf{A}^{T}\right)^{-1} \equiv \mathbf{A}^{-T}$$

Тождество

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

показывает, как изменяется обратная матрица при изменении самой матрицы.

Если $\mathbf{A} = (a) \in \mathbf{R}^{1\times 1}$, то ее *детерминант* дается равенством $\det(\mathbf{A}) = a$. Детерминант $N \times N$ матрицы определяется через детерминанты $(N-1)\times(N-1)$-матриц. Для $\mathbf{A} \in \mathbf{R}^{N\times N}$ можно записать:

$$\det(\mathbf{A}) = \sum_{j=0}^{n} (-1)^{j} a_{0j} \det(\mathbf{A}_{1j}),$$

где $\mathbf{A}_{1j}$ – это $(N-1)\times(N-1)$-матрица, получаемая из $\mathbf{A}$ вычеркиванием первой строки и j-го столбца. Детерминант обладает некоторыми полезными свойствами:

$$\begin{aligned}
&\det(\mathbf{AB}) &&= \det(\mathbf{A})\det(\mathbf{B}) && \mathbf{A},\mathbf{B} \in \mathbf{R}^{N\times N},\\
&\det(\mathbf{A}^{T}) &&= \det(\mathbf{A}) && \mathbf{A} \in \mathbf{R}^{N\times N},\\
&\det(c\mathbf{A}) &&= c^{N}\det(\mathbf{A}) && \mathbf{A} \in \mathbf{R}^{N\times N}, c \in \mathbf{R},\\
&\det(\mathbf{A}) \neq 0 &&\Leftrightarrow \mathbf{A}\ \text{невырожденная} && \mathbf{A} \in \mathbf{R}^{N\times N}.
\end{aligned}$$

Нормы

Нормой вектора $\mathbf{x} \in \mathbf{R}^{N}$ называют такое действительное число, обозначаемое $\|\mathbf{x}\|$, что:

1. $\|\mathbf{x}\| \geq 0$, причем $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = 0$;
2. $\|\lambda\mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|$ для $\forall \lambda \in \mathbf{R}$;
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ для $\forall \mathbf{y} \in \mathbf{R}^{N}$.

Линейное пространство $\mathbf{R}^N$ с введенной в нем нормой называют *нормированным пространством* (точнее, вещественным нормированным пространством).

Обычное понятие длины (модуля) вектора удовлетворяет всем аксиомам, определяющим норму, т.е. длина есть норма. Существует большое количество конструкций $\|\mathbf{x}\|$, удовлетворяющих всем трем аксиомам нормы вектора. Полезный класс векторных норм – это так называемая *p-норма*(*норма Гёльдера*), определяемые как:

$$\|\mathbf{x}\|_p = \left(\sum_{i=0}^{n} |x_i|^p \right)^{\frac{1}{p}}, \quad p \geq 1$$

Наиболее важными из p-норм являются 1, 2 и ∞ -нормы:

$$\|\mathbf{x}\|_1 = \sum_{i=0}^{n} |x_i| = |x_0| + \ldots |x_n|$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=0}^{n} |x_i|^2} = \sqrt{|x_0|^2 + \ldots |x_n|^2} = \sqrt{\mathbf{x}^T \mathbf{x}}$$

$$\|\mathbf{x}\|_\infty = \max_{0 \leq i \leq n} |x_i|$$

Норму при $p = 1$ называют *норма-сумма*. Получаемая при $p = 2$ норма называется *евклидовой нормой*, множество $\mathbf{R}^N$ с введенной в нем евклидовой нормой называют *евклидовым пространством* и часто обозначают через E^N. Евклидово пространство характерно тем, что в нем определено скалярное произведение векторов $\mathbf{x} = (x_0, \ldots, x_n)^T$ и $\mathbf{y} = (y_0, \ldots, y_n)^T$ равенством $(\mathbf{x}, \mathbf{y}) := \sum_{i=0}^{n} x_i y_i$ и при этом имеет место связь $\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x}, \mathbf{x})}$. Получаемая при переходе к пределу $p \to \infty$ норма называется *норма-максимум*.

Отметим некоторые полезные свойства векторных норм. Классический результат о *p-нормах – неравенство Гёльдера*:

$$\left| \mathbf{x}^T \mathbf{y} \right| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Очень важным частным случаем этого неравенства является *неравенство Коши-Шварца*:

$$\left|\mathbf{x}^T\mathbf{y}\right| \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

Все нормы в $\mathbf{R}^N$ *эквивалентны*, т.е. для двух норм $\|\cdot\|_\alpha$ и $\|\cdot\|_\beta$ в $\mathbf{R}^N$ существуют положительные константы c_1 и c_2, такие, что:

$$c_1\|\mathbf{x}\|_\alpha \le \|\mathbf{x}\|_\beta \le c_2\|\mathbf{x}\|_\alpha$$

для всех $\mathbf{x} \in \mathbf{R}^N$. Например, при $\mathbf{x} \in \mathbf{R}^N$

$$\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{N}\|\mathbf{x}\|_2$$

$$\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \sqrt{N}\|\mathbf{x}\|_\infty$$

$$\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_1 \le N\|\mathbf{x}\|_\infty$$

Нормой матрицы называется действительное число $\|\mathbf{A}\|$, удовлетворяющее условиям ($\mathbf{A} \in \mathbf{R}^{M\times N}$):

1. $\|\mathbf{A}\| \ge 0$, причем $\|\mathbf{A}\| = 0 \Leftrightarrow \mathbf{A} = \mathbf{0}$;
2. $\|\mathbf{A}+\mathbf{B}\| \le \|\mathbf{A}\| + \|\mathbf{B}\|$, где $\mathbf{B} \in \mathbf{R}^{M\times N}$;
3. $\|\lambda\mathbf{A}\| = |\lambda|\cdot\|\mathbf{A}\|$ для $\forall \lambda \in \mathbf{R}$.

Чаще всего в вычислительной алгебре используется норма Фробениуса

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=0}^{m}\sum_{j=0}^{n}\left|a_{ij}\right|^2}$$

и p-нормы

$$\|\mathbf{A}\|_p = \sup_{x\neq 0}\frac{\|\mathbf{Ax}\|_p}{\|\mathbf{x}\|_p},$$

где sup означает *точную верхнюю границу* (*наименьшую верхнюю границу*) множества. Так например $\sup|\cos(x)| = 1$.

Матричные p-нормы определяются через векторные p-нормы, рассмотренные выше. Видно, что $\|\mathbf{A}\|_p$ есть p-норма наибольшего вектора, полученного действием $\mathbf{A}$ на векторы единичной (в p-норме) длины:

$$\|\mathbf{A}\|_p = \sup_{x\neq 0}\left\|\mathbf{A}\frac{\mathbf{x}}{\|\mathbf{x}\|_p}\right\|_p = \max_{\|\mathbf{x}\|_p=1}\|\mathbf{Ax}\|_p$$

Важно отметить, что норма Фробениуса и p-нормы определяют целые семейства норм: 2-норма в $\mathbf{R}^{3\times 2}$ – это другая функция, чем 2-норма в $\mathbf{R}^{5\times 6}$. Таким образом, легко проверяемое неравенство

$$\|\mathbf{AB}\|_p \le \|\mathbf{A}\|_p\|\mathbf{B}\|_p, \quad \mathbf{A}\in\mathbf{R}^{M\times N}, \quad \mathbf{B}\in\mathbf{R}^{N\times Q}$$

в действительности описывает соотношение между тремя различными нормами. Формально будем говорить, что нормы f_1, f_2 и f_3 в пространствах $\mathbf{R}^{M\times Q}$, $\mathbf{R}^{M\times N}$ и $\mathbf{R}^{N\times Q}$ соответственно, *взаимно согласованы*, если для всех $\mathbf{A}\in\mathbf{R}^{M\times N}$, $\mathbf{B}\in\mathbf{R}^{M\times N}$ выполняется условие $f_1(\mathbf{AB})\le f_2(\mathbf{A})f_3(\mathbf{B})$.

Нужно сказать, что не все матричные нормы удовлетворяют мультипликативному свойству

$$\|\mathbf{AB}\| \le \|\mathbf{A}\|\|\mathbf{B}\|$$

Например, если $\|\mathbf{A}\|_\Delta = \max|a_{ij}|$ и

$$\mathbf{A}=\mathbf{B}=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix},$$

то

$$\|\mathbf{AB}\|_\Delta > \|\mathbf{A}\|_\Delta\|\mathbf{B}\|_\Delta .$$

p-Нормы обладают тем важным свойством, что для каждой матрицы $\mathbf{A}\in\mathbf{R}^{M\times N}$ и вектора $\mathbf{x}\in\mathbf{R}^N$ выполняется $\|\mathbf{Ax}\|_p \le \|\mathbf{A}\|_p\|\mathbf{x}\|_p$. Более общо, для любых векторных норм $\|\cdot\|_\alpha$ в $\mathbf{R}^N$ и $\|\cdot\|_\beta$ в $\mathbf{R}^M$ мы имеем $\|\mathbf{Ax}\|_\beta \le \|\mathbf{A}\|_{\alpha,\beta}\|\mathbf{x}\|_\alpha$, где матричная норма $\|\mathbf{A}\|_{\alpha,\beta}$ определяется как

$$\|\mathbf{A}\|_{\alpha,\beta} = \sup_{x \neq 0} \frac{\|\mathbf{Ax}\|_\beta}{\|\mathbf{x}\|_\alpha}$$.

В этом случае $\|\cdot\|_{\alpha,\beta}$ *подчинена* векторным нормам $\|\cdot\|_\alpha$ и $\|\cdot\|_\beta$.

Ортогональность

Набор векторов $\{\mathbf{x}_0, \ldots, \mathbf{x}_n\}$ в $\mathbf{R}^M$ называется *ортогональным*, если $\mathbf{x}_i^T \mathbf{x}_j = 0$ при $i \neq j$, и *ортонормированным*, если $\mathbf{x}_i^T \mathbf{x}_j = \delta_{ij}$. Ортогональные векторы максимально независимы, поскольку направлены в разные стороны.

Подпространства $S_0, \ldots, S_p$ в $\mathbf{R}^M$ *попарно ортогональны*, если $\mathbf{x}^T \mathbf{y} = 0$ для всех $\mathbf{x} \in S_i$ и $\mathbf{y} \in S_j$, таких что $i \neq j$. Ортогональное дополнение подпространства $S \subseteq \mathbf{R}^M$ определяется как

$$S^\perp = \left\{ \mathbf{y} \in \mathbf{R}^M : \mathbf{y}^T \mathbf{x} = 0, \quad \mathbf{x} \in S \right\}$$.

Нетрудно видеть, что $range(\mathbf{A})^\perp = null(\mathbf{A}^T)$.

Матрица $\mathbf{Q} \in \mathbf{R}^{M \times M}$ *ортогональна*, если $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$. Если $\mathbf{Q} = [\mathbf{q}_0, \ldots, \mathbf{q}_m]$ ортогональна, то векторы $\mathbf{q}_i$ образуют ортонормированный базис в $\mathbf{R}^M$.

Следует отметить, что 2-норма инвариантна относительно ортогональных преобразований, поскольку если $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$, то $\|\mathbf{Qx}\|_2^2 = \mathbf{x}^T \mathbf{Q}^T \mathbf{Qx} = \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|_2^2$. Этим же свойством обладают матричная 2-норма и норма Фробениуса. Можно показать, что для всех ортогональных матриц **Q** и **Z** подходящих размеров:

$$\|\mathbf{QAZ}\|_F = \|\mathbf{A}\|_F$$

и

$$\|\mathbf{QAZ}\|_2 = \|\mathbf{A}\|_2$$.

Особенности программной реализации

При программной реализации класса матрицы и его методов желательно учитывать тот факт, что матрица, являющаяся по сути двумерным массивом, будет храниться в памяти ЭВМ в виде набора чисел. Большинство языков программирования позволяет создавать многомерные, или уж во всяком случае, двумерные массивы, однако расположение элементов в памяти при этом скрыто от программиста. Поэтому часто класс матрицы реализует хранение всех элементов в виде одномерного массива, при этом подбирают способ более выгодный с вычислительной (или интерфейсной) точки зрения – либо матрица вытягивается в одномерный массив построчно, либо постолбцово. При этом в функциях, реализующих интерфейс матрицы производится пересчет двумерных индексов к одномерному:

- при построчном хранении матрицы индексы строки i и столбца j пересчитываются к одномерному индексу k следующим образом: $k = i \cdot N + j$, где N – количество столбцов в матрицы;
- при постолбцовом хранении матрицы $k = i + j \cdot M$, где M – количество строк матрицы.

Здесь следует отметить, что такой пересчет индексов предполагает, что элементы матрицы нумеруются с 0 (что обычно в практике программирования, в отличие от математических традиций начинать нумерацию с 1).

Также хотелось бы отметить, что в функциях, принимающих матрицы в списке аргументов, желательно использовать либо ссылки, либо указатели, так как копирование матрицы может быть весьма ресурсоемкой процедурой в зависимости от ее размерности, т.е. если объявлен некий класс матрицы Matrix, то функция function1(Matrix matrix1, Matrix matrix2) нежелательна, предпочтительнее function2(Matrix& matrix1, Matrix* matrix2).

Следует уделять особое внимание операциям, в которых производится перебор всех элементов матрицы – здесь желательно учитывать способ хранения (построчно или постолбцово) матрицы. Так при работе с построчно хранимыми матрицами предпочтительно организовывать во внешнем цикле перебор строк, а во внутреннем – перебор столбцов; для постолбцово хранящихся матриц – наоборот.

При реализации некоторых матричных операций, например транспонирования, имеет смысл ввести проверку «квадратности» матрицы, так как в случае квадратной матрицы некоторые операции значительно упрощаются, что позволяет снижать вычислительные затраты.

§ 6. Системы линейных уравнений (СЛАУ)

Сейчас, для простоты, будем говорить о случае, когда количество неизвестных и количество уравнений совпадают. Стоит отметить, что решение избыточных или недостаточных СЛАУ сводится к решению достаточных СЛАУ.

Матричная запись систем уравнений

Рассмотрим систему линейных уравнений с набором неизвестных:

$$\begin{cases} a_{00}x_0 + a_{01}x_1 + \ldots + a_{0n}x_n = b_0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{i0}x_0 + a_{i1}x_1 + \ldots + a_{in}x_n = b_i \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n0}x_0 + a_{n1}x_1 + \ldots + a_{nn}x_n = b_n \end{cases}$$

Оперировать с такой системой, очевидно, не очень удобно. Эту же систему уравнений можно представить в более компактном матричном виде:

$$\mathbf{Ax} = \mathbf{b}$$

где $\mathbf{b} = (b_0, b_1, \ldots, b_n)^T$ – вектор свободных членов и $\mathbf{x} = (x_0, x_1, \ldots, x_n)^T$ – вектор неизвестных (вектор-решение) с вещественными координатами, а $\mathbf{A} = (a_{ij})_{i,j=0}^{n}$ – вещественная $N \times N$ -матрица коэффициентов данной системы**.**

Методы решения СЛАУ

Все методы решения линейных алгебраических задач (наряду с задачей решения СЛАУ, это и вычисление определителей, и обращение матриц, и задачи на собственные значения) можно разбить на два класса: прямые и итерационные. Прямые методы – это методы, которые приводят к решению за конечное число арифметических операций. Если операции реализуются точно, то и решение будет точным (поэтому к классу прямых методов применяют еще название точные методы). Итерационные методы – это методы, точное решение в которых может быть получено лишь в результате бесконечного повторения единообразных (как правило, простых) действий.

Рассмотрим некоторые прямые методы. Самым простым является метод, в основе которого лежат формулы Крамера, когда решение вычисляется по формулам:

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}} \quad (i = 0,1,\ldots,N),$$

где $\mathbf{A}_i$ – матрица, полученная из $\mathbf{A}$ заменой i-го столбца коэффициентов при вычисляемом неизвестном столбцом свободных членов (вектором b). Однако такой подход характеризуется катастрофически высоким ростом вычислительных затрат при росте размерности матрицы $\mathbf{A}$ – на вычисление определителя N-го порядка будет затрачиваться N! операций умножения. Например при N=100 эта величина составит 100!=10158. Поэтому такой подход приемлем на практике только при небольших N.

Наиболее известным и популярным способом решения линейных систем является метод Гаусса, суть которого заключается в последовательном исключении неизвестных –

поэтапном приведении системы эквивалентными преобразованиями к треугольному виду. На каждом этапе обнуляют коэффициенты при неизвестных: на первом этапе – при x_0 во втором, третьем, ... , n-ом уравнениях; на втором этапе – при x_1 в третьем, четвертом, ... , n-ом уравнениях преобразованной системы и т.д. Для этого, на первом этапе, заменяют второе, третье, ... , n-ое уравнения на уравнения, получающиеся сложением этих уравнений с первым, домноженным соответственно на $-\frac{a_{10}}{a_{00}}$, $-\frac{a_{20}}{a_{00}}$, ... , $-\frac{a_{n0}}{a_{00}}$. На втором этапе работают с системой, полученной на первом этапе, проделывая аналогичные преобразования с подсистемой, исключающей первой уравнение. После окончания этой процедуры получают систему верхнетреугольного вида, решение которой получается обратной

производится деление в методе Гаусса называют ведущими или главными элементами. Такая модификация метода носит название метода Гаусса с постолбцовым выбором главного элемента (или с частичным упорядочиванием подстановкой.

Чтобы избежать ошибок округления, неизбежных в машинной арифметике, и исключить деление на ноль, на каждом этапе уравнения рассматриваемой подсистемы обычно переставляют так, чтобы деление производилось на наибольший по модулю в данном столбце (обрабатываемом подстолбце) элемент. Числа, на которые по столбцам).

Немного иной подход основан на LU разложении матрицы A. В этом верхнетреугольной случае матрица A представляется в виде произведения $\mathbf{A} = \mathbf{LU}$ нижнетреугольной матрицы L и матрицы U. При этом, для однозначности, диагональные элементы матрицы L (либо U) устанавливают равными 1, после этого несложно найти выражение для элементов матриц L и U через элементы матрицы A. Когда разложение найдено, система преобразуется к виду:

$$\mathbf{LUx} = \mathbf{b}$$

и, через вспомогательный вектор $y = (y_0, \ldots, y_n)^T$, решают обратной подстановкой последовательно две системы: $\mathbf{Ly} = \mathbf{b}$, а затем $\mathbf{Ux} = \mathbf{y}$.

Суть итерационных методов заключается в решении эквивалентной системы вида:

$$\mathbf{x} = \mathbf{Bx} + \mathbf{c},$$

где x – тот же вектор неизвестных, а B и c – некоторые новые матрица и вектор соответственно (способ такого преобразования существует бесконечное множество). Такую систему можно трактовать как задачу о неподвижной точке линейного отображения B в пространстве $\mathbf{R}^N$ и определить последовательность приближений $\mathbf{x}^{(k)}$ к неподвижной точке $\mathbf{x}^*$ рекуррентным равенством:

$$\mathbf{x}^{(k+1)} = \mathbf{Bx}^{(k)} + \mathbf{c}, \quad k = 0,1,2,\ldots .$$

Такие методы также носят название методов простых итераций (МПИ). В качестве простого примера можно привести метод Якоби, который заключается в том, что матрица A представляется в виде:

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{R} \ ,$$

где D – диагональная, L и R – соответственно левая и правая строго треугольные матрицы. Тогда в рекуррентном равенстве:

$$\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R}), \quad \mathbf{c} = \mathbf{D}^{-1}\mathbf{b} \ .$$

Типовые обозначения матриц

В литературе часто можно встретить следующие обозначения матриц специального вида:

L – нижняя [lower] (или левая [left]) треугольная матрица.

U (или R) – верхняя [upper] (или правая [right]) треугольная матрица.

Q – ортогональная матрица.

Матрицы перестановки:

1. Если матрица B получена из единичной заменой числа 1 в i-ой строке на некоторое число a, то матрица AB (матрица A – соответствующей для B размерности) получается из матрицы A умножением всех элементов i-го столбца на a. Матрица BA получается из матрицы Aумножением всех элементов i-ой строки на a.
2. Если матрица B получена из единичной заменой недиагонального элемента $\delta_{ik} = 0$ на 1, то матрица AB получается из матрицы A заменой k-го столбца на сумму k-го и i-го столбцов. Матрица BA получается из матрицы A заменой i-ой строки на сумму i-ой и k-ой строк.
3. Если матрица B – матрица перестановки, полученная из единичной матрицы перестановкой каких либо ее двух столбцов (или, что тоже самое, двух ее строк с теми же номерами), то матрица AB получается из матрицы A перестановкой соответствующих столбцов, а матрица BA– перестановкой соответствующих строк.

Устойчивость задач

Учитывая распространенность СЛАУ, имеет смысл попытаться количественно охарактеризовать степень неопределенности этих задач. Знание таких характеристик позволяет обоснованно судить о корректности моделей, грамотно подбирать методы и строить алгоритмы, правильно трактовать полученные результаты.

Рассмотрим линейную алгебраическую систему:

$$\mathbf{Ax} = \mathbf{b} \ ,$$

где A – невырожденная $N \times N$ матрица коэффициентов данной системы; b – ненулевой N-мерный вектор свободных членов; x – N-мерный вектор неизвестных.

Пусть правая часть системы получила приращение (возмущение) $\Delta\mathbf{b}$, тогда реакцией решения будет вектор поправок $\Delta\mathbf{x}$:

$$\mathbf{A}(\mathbf{x}+\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b} \ .$$

Получим оценку сверху для относительной погрешности вектора-решения через погрешность вектора свободных членов:

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \ .$$

Положительное число $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ – коэффициент связи – называют числом (мерой) обусловленности матрицы A и обозначают $cond(\mathbf{A})$(conditioned), так же встречаются обозначения $\nu(\mathbf{A})$ и $\chi(\mathbf{A})$. Можно показать, что $cond(\mathbf{A})$ служит также коэффициентом роста при неточном задании элементов матрицы A (возмущение $\Delta\mathbf{A}$):

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le cond(\mathbf{A}) \cdot \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}+\Delta\mathbf{A}\|} \text{ и } \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}+\Delta\mathbf{x}\|} \le cond(\mathbf{A}) \cdot \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \ .$$

Из приведенных неравенств видно, что чем больше число обусловленности, тем сильнее сказывается на решении линейной системы ошибка в исходных данных. Грубо говоря, если $cond(\mathbf{A}) \approx \mathbf{10}^{p}$ и исходные данные имеют погрешность в l-ом знаке после запятой, то независимо от способа решения системы в результате можно гарантировать не более $l - p$ знаков после запятой.

Если число $cond(\mathbf{A})$ велико, то система считается плохо обусловленной. Оценка снизу для числа обусловленности дает 1, т.е. $cond(\mathbf{A})$ не может быть меньше 1. Для конкретной ЭВМ можно указать также верхнюю границу, превышение которой может привести к заведомо ложным решениям: решение считается ненадежным, если $cond(\mathbf{A}) \ge (u)^{-1}$ или даже $cond(\mathbf{A}) \ge (u)^{-0.5}$, где u – единичная ошибка округления (машинная точность). При этом важно отметить, что масштабирование матрицы A путем умножения на скаляр a не меняет ее число обусловленности.

Здесь стоит отметить, что малость невязки $\boldsymbol{\rho} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ плохо обусловленной системы еще не говорит о близости приближенного решения $\tilde{\mathbf{x}}$ к точному $\mathbf{x}$.

Двумерный случай допускает простую геометрическую трактовку понятия обусловленности. Плохая обусловленность системы двух уравнений с двумя неизвестными означает, что прямые, являющиеся геометрическими образами уравнений,

пересекаются на координатной плоскости под очень острым углом. В этом случае небольшое искажение в данных, интерпретируемое как параллельный перенос (при возмущении свободного члена) или поворот прямых (при возмущении матрицы коэффициентов) приводит к значительному перемещению их точки пересечения, т.е. геометрического образа решения.

Для системы

$$\begin{cases} x + 10y & = & 11 \\ 100x + 1001y & = & 1101 \end{cases}$$

или, в матричных обозначениях $\mathbf{Az} = \mathbf{b}$, где

$$\mathbf{A} = \begin{bmatrix} 1 & 10 \\ 100 & 1001 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 11 \\ 1101 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} x \\ y \end{bmatrix}$$

получаем число обусловленности в матричной норме, индуцированной векторной нормой-макисмум $cond_{\infty}(\mathbf{A}) = 1113111 > 10^6$. Тогда, взяв некоторое малое возмущение данных $\triangle\mathbf{b} = [0.01, \quad 0]^T$ получаем оценку относительной погрешности решения ≤ 10.11 , и абсолютную ≤ 10.11 , что и подтверждается при решении: $\mathbf{x} = [1, \quad 1]^T$, $\mathbf{x} + \triangle\mathbf{x} = [11.01, \quad 0]^T$.

ЛИТЕРАТУРА

Основная литература

1	Сост. Никитина С.Ю.	Вычислительная математика. Методические рекомендации для выполнения контрольной работы.	Рязань: СТИ, 2013.
2	Формалев В.Ф., Ревизников	*ЭБС Книгафонд: Численные методы: учебное пособие*	М: Физматлит, 2014.- 399 с
3	Турчак Л.И., Плотников	*ЭБС Книгафонд: Основы численных методов: учебное пособие*	М: Физматлит, 2014.- 304 с
4	Рябенький В.С.	*ЭБС Книгафонд: Введение в вычислительную математику: уч. пособие*	М: Физматлит, 2008.- 285 с
5	Ракитин В.И.	*ЭБС Книгафонд: Руководство по методам вычислений и приложения MATHCAD: уч.пос.*	М: Физматлит, 2014.- 264 с

Дополнительная литература

1	Воеводин В.В.	*ЭБС Книгафонд: Вычислительная математика и структура алгоритмов: Учебник*	М: МГУ, 2010.- 166 с.
2	Лебедев В.И.	*ЭБС Книгафонд: Функциональный анализ и вычислительная математика: Учебное пособие*	М: Физматлит, 2014.- 294 с
3	Шипачёв В.С.	Курс Высшей математики	М.:ТК Велби, 2004
4	Письменный Д.Т.	Конспект лекций по высшей математике. Ч 1-2.	М.: Айрис-пресс, 2007

Содержание

Подписано в печать 21.03.21.
Электронное издание.

Издательство Современного технического университета

390048, г. Рязань, ул. Новоселов, 35А.

(4912) 300630, 30 08 30