Ляхов Александр Федорович

УДК 519

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ ЭНТРОПИИ В МАТЕМАТИЧЕСКИХ ВЫРАЖЕНИЯХ

Аннотация

В работе показано, что различные формы записи алгебраических выражений несут разное количество информации о распределении вычислительных погрешностей. В случае развернутых формул вычислений интервал оценивания расширяется, но при этом распределение погрешности становится неравномерным. Для оценки степени неравномерности распределения погрешности используется понятие энтропии.

Ключевые слова: погрешность, информация, энтропия, распределение.

ВВЕДЕНИЕ

При выполнении вычислений, как правило, приходится сталкиваться с возникновением и ростом погрешности вычислений. Если вычисления не очень громоздкие и производятся вручную, то в процессе счета человек может контролировать рост погрешности вычислений и предпринимать меры для её уменьшения. Заметим, что ручные вычисления обычно производятся способом, который может быть назван «вычисления с переменной длиной числа, с квазификсированной-квазиплавающей запятой», то есть длина используемых чисел регулируется.

Объёмы современных машинных вычислений при решении сложных задач могут содержать 10^{10} и более элементарных арифметических операций. Машинные вычисления обычно осуществляются с плавающей запятой и с фиксированной длиной числа. В этом случае оценки погрешностей могут быть получены только в результате сложных исследований и вычислений.

Известно, что, с одной стороны, все оценки погрешности измерений величин, с которыми производятся вычисления, носят вероятностный характер, то есть наряду с интервалом погрешности указывается соответствующая доверительная вероятность, с другой стороны, реальные погрешности вычислений всегда много меньше теоретических оценок. Всё это подталкивает исследователей к построению статистических подходов оценки погрешности. Эти подходы базируется на идее, что процесс округления есть случайный процесс, и, следовательно, можно построить его модель, основываясь на теории вероятностей. Для того чтобы формально применить методы математической статистики, необходимо создать вероятностное пространство результатов вычислений. Это может быть осуществлено путем создания некоторой умозрительной модели или многократного повторения процесса вычислений с использованием современных технологий распараллеливания и многопроцессорных вычислительных машин[1] .

[1] Погрешности вычислений в одной и той же задаче при её повторных выполнениях на многопроцессорных компьютерах различны. Это является следствием того, что распараллеливание задачи определяется общим состоянием занятости процессоров, участвующих в работе.

Построим вероятностное пространство, полагая числа, над которыми производятся арифметические операции, варьирующимися. Представим, что вычисление может быть сделано с бесконечной точностью, но на каждом арифметическом шаге осуществляется округление, и присоединяется некоторая погрешность [2]. Будем полагать, что эта погрешность – случайная величина, равномерно распределенная на интервале длиной, равной единице последнего значащего разряда.

Заметим, что в этой модели допускается непрерывное распределение и игнорируется тот факт, что действительное распределение машинного округления дискретно, так как вычислительные машины оперируют с числами конечной длины.

Одним из главных свойств случайных величин является отсутствие уверенности в их значении. Эта неопределенность изменяется при выполнении связанных с этими величинами операций.

В теории информации за меру неопределенности случайной величины X с плотностью распределения $f(x)$ принимается величина, называемая энтропией и равная

$$H(x) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx\,.$$

Заметим, что максимальную энтропию имеет равномерно распределенная случайная величина [1, 3]. Следовательно, чем меньше можно сказать о значении, которое примет случайная величина, то есть чем меньше информации о ней мы имеем, тем энтропия больше.

Выполняя арифметические действия над числами, погрешности которых имеют некоторые распределения, получим новые погрешности с распределениями, отличными от исходных. В зависимости от порядка выполнения действий, как величина погрешности, так и мера неопределенности результата, то есть энтропия, будут различными.

ПОСТАНОВКА ЗАДАЧИ

Рассмотрим различную запись одного и того же выражения:

$$I_1 = (a+b)^2, \qquad I_2 = a^2 + 2ab + b^2.$$

Оценим энтропию этих выражений с точки зрения предложенной модели округления чисел.

Пусть a – случайная величина, равномерно распределенная на интервале $(a-\Delta a; a+\Delta a)$ и имеющая математическое ожидание, равно a. Плотность распределения

$$f_a(x) = \begin{cases} \dfrac{1}{2\Delta a}, x \in (a-\Delta a; a+\Delta a), \\ 0, \quad x \notin (a-\Delta a; a+\Delta a), \end{cases}$$

Пусть b – случайная величина, равномерно распределенная на интервале $(b-\Delta b; b+\Delta b)$ с математическим ожиданием, равным b:

Плотность распределения $f_b(x) = \begin{cases} \dfrac{1}{2\Delta b}, x \in (b-\Delta b; b+\Delta b), \\ 0, \quad x \notin (b-\Delta b; b+\Delta b). \end{cases}$

Для того чтобы ответить на вопрос, какая запись – I_1 или I_2 содержит в себе больше информации, требуется найти плотности распределения $f_{(a+b)^2}(x)$ и $f_{a^2+2ab+b^2}(x)$, а затем вычислить энтропию этих распределений.

Из теории вероятностей известно [2], что, если имеется непрерывная случайная величина X с плотностью $f(x)$, то случайная величина $Y = \varphi(X)$ имеет плотность распределения

$$g(y) = \sum_{i=1}^{k} f(\psi_i(y))\,|\psi_i'(y)|, \qquad (1)$$

где k – число значений функции, обратной к $\varphi(x)$, соответствующее данному y, $\psi_1(y), \psi_2(y), ..., \psi_k(y)$ – значения обратной функции, соответствующие данному y [1].

Найдём законы распределения случайных величин входящих в искомые выражения. Определим плотность распределения квадрата случайной величины $\varphi(x) = x^2$. Функция $y = x^2$ не монотонна; $\psi_1(y) = -\sqrt{y}$, $\psi_2(y) = \sqrt{y}$. Из (1) получим

$$g(y) = \frac{1}{\sqrt{2y}}(f(-\sqrt{y}) + f(\sqrt{y})), y > 0. \quad (2)$$

Рассмотрим функцию двух случайных аргументов

$$Y = \varphi(X_a, X_b),$$

функция распределения случайной величины Y равна

$$G(y) = \iint\limits_{\varphi(x_a, x_b) < y} f(x_a, x_b) dx_a dx_b, \quad (3)$$

где область интегрирования на плоскости $x_a O x_b$ определяется из условия $\varphi(x_a, x_b) < y$. Дифференцируя (3) по величине y, найдем плотность распределения случайной величины Y: $g(y) = \frac{dG(y)}{dy}$.

Полагаем, что случайные величины X_a и X_b независимы, то есть $f(x_a, x_b) = f_a(x_a) f_b(x_b)$, и положительны.

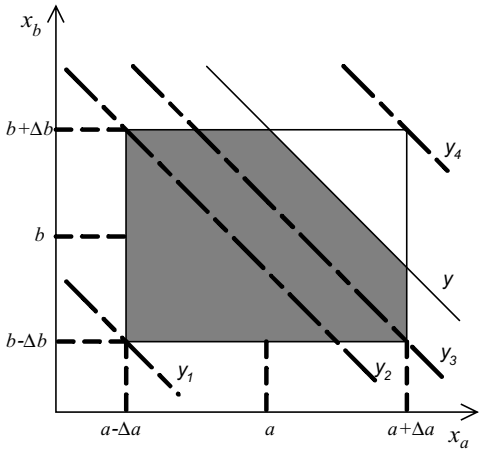
Найдем функцию распределения случайной величины

$$Y = \varphi(X_a, X_b) = X_a + X_b, \quad G(y) = \iint\limits_{x_a + x_b < y} f_a(x_a) f_b(x_b) dx_a dx_b.$$

В этом случае область интегрирования – часть прямоугольника, отсеченная прямой $x_a + x_b = y$ (рис. 1).

Предположим, что $\Delta a > \Delta b$. Покажем, как определяется функция $G(y)$ на интервале $y_1 \le y \le y_2$

$$G(y) = \frac{1}{4\Delta a \Delta b} \int\limits_{-a-\Delta a}^{-y-(b-\Delta b)} dx_a \int\limits_{b-\Delta b}^{y-x_a} dx_b = \frac{1}{4\Delta a \Delta b}\left[\frac{1}{2}(y - b + \Delta b)^2 - (y - b + \Delta b)(a - \Delta a) + \frac{1}{2}(a - \Delta a)^2\right].$$



$y_1 = a - \Delta a + b - \Delta b, y_2 = a - \Delta a + b + \Delta b,$
$y_3 = a + \Delta a + b - \Delta b, y_4 = a + \Delta a + b + \Delta b.$

Рис. 1

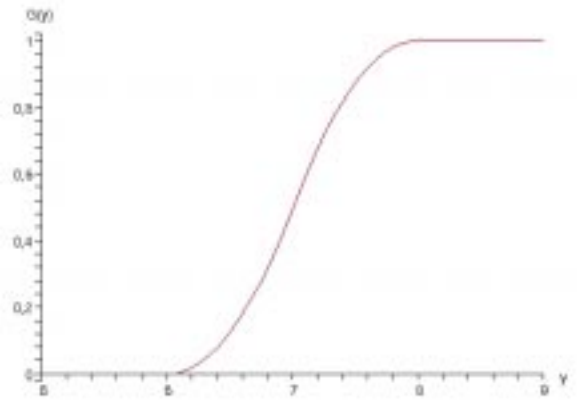


Рис. 2

На остальных интервалах $G(y)$ определяется аналогичным образом.

$$G(y)=\frac{1}{4\Delta a\Delta b}\begin{cases}0, & y<y_1,\\ \frac{1}{2}(y-b-\Delta b)^2-(y-b+\Delta b)(a-\Delta a)+\frac{1}{2}(a-\Delta a)^2, & y_1\le y<y_2,\\ 2\Delta b^2+2\Delta b(y-a+\Delta b-b-\Delta b), & y_2\le y<y_3,\\ 4\Delta a\Delta b-\frac{1}{2}(y-b-\Delta b)^2+(y-b-\Delta b)(a+\Delta a)-\frac{1}{2}(a+\Delta a)^2, & y_3\le y<y_4,\\ 4\Delta a\Delta b, & y_4\le y.\end{cases}$$

График функции $G(y)$ изображен на рис. 2.

Дифференцируя по y, находим плотность распределения суммы двух случайных величин:

$$g(y)=\frac{1}{4\Delta a\Delta b}\begin{cases}0, y<y_1,\\ (y-b+\Delta b), & y_1\le y<y_2,\\ 2\Delta b, & y_2\le y\le y_3,\\ -(y-b-\Delta b)+(a+\Delta a), & y_3\le y<y_4,\\ 0, & y_4\le y.\end{cases}$$

График плотности распределения $g(y)$ изображен на рис. 3.

Функция распределения произведения случайных величин $Y=\varphi(X_a,X_b)=X_a\cdot X_b$

$$G(y)=\iint\limits_{x_a\cdot x_b<y} f_a(x_a)\,f_b(x_b)\,dx_a\,dx_b\ .$$

Область интегрирования – часть прямоугольника, отсеченная кривой $x_a\cdot x_b=y$ (рис. 4).

Предположим, что $a\Delta b<b\Delta a$, тогда $y_1\le y_2\le y_3\le y_4$.

В этом случае получим:

$$G(y)=\frac{1}{4\Delta a\Delta b}\begin{cases}0, y<y_1,\\ y\left(\ln\left|\frac{y}{(a-\Delta a)(b-\Delta b)}\right|-1\right)+(a-\Delta a)(b-\Delta b), & y_1\le y<y_2,\\ y\ln\left|\frac{b+\Delta b}{b-\Delta b}\right|-2\Delta b(a-\Delta a), & y_2\le y<y_3,\\ 4\Delta a\Delta b-y\left(\ln\left|\frac{y}{(a+\Delta a)(b+\Delta b)}\right|-1\right)-(a+\Delta a)(b+\Delta b), & y_3\le y<y_4,\\ 4\Delta a\Delta b, & y_4\le y.\end{cases}$$

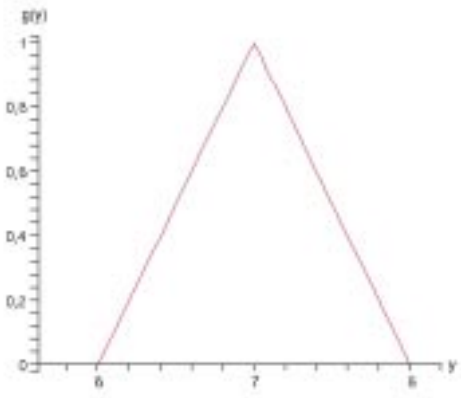


Рис. 3

График функции $G(y)$ изображен на рис. 5.

Дифференцируя его по y, находим плотность распределения произведения двух случайных величин:

$$g(y)=\frac{1}{4\Delta a\Delta b}\begin{cases}0, y<y_1,\\ \ln\left|\frac{y}{(a-\Delta a)(b-\Delta b)}\right|, & y_1\le y<y_2,\\ \ln\left|\frac{b+\Delta b}{b-\Delta b}\right|, & y_2\le y<y_3,\\ -\ln\left|\frac{y}{(a+\Delta a)(b+\Delta b)}\right|, & y_3\le y<y_4,\\ 0, & y_4\le y.\end{cases}$$

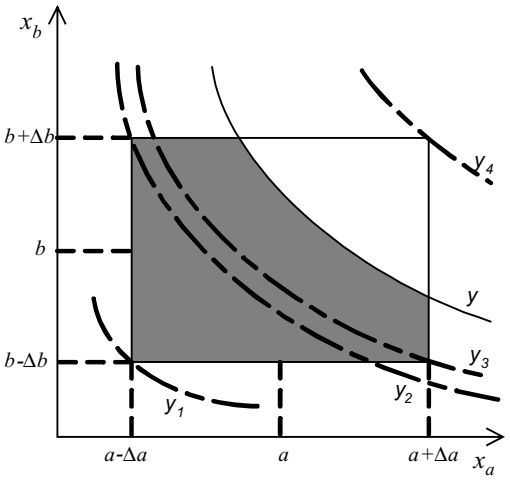
График плотности распределения $g(y)$ показан на рис. 6.

ЧИСЛЕННЫЙ ПРИМЕР НАХОЖДЕНИЯ ЭНТРОПИИ ВЫРАЖЕНИЯ

Поскольку определить плотность распределения $f_{(a+b)^2}(x)$ и $f_{a^2+2ab+b^2}(x)$ в общем виде невозможно, приведём пример решения поставленной задачи для конкретных значений $a, b, \Delta a, \Delta b$.

Расчеты были проведены в системе Maple. На каждом этапе вычислений производилась проверка выполнения основных свойств плотности распределения для полученных функций.

Примем $a = 2,\ b = 5,\ \Delta a = \Delta b = 0{,}5$, $I_1 = (2+5)^2$.



$y_1 = (a-\Delta a)(b-\Delta b),\quad y_2 = (a-\Delta a)(b+\Delta b),$

$y_3 = (a+\Delta a)(b-\Delta b),\quad y_4 = (a+\Delta a)(b+\Delta b)$

Рис. 4

Согласно ранее выведенным формулам плотность распределения случайной величины (2 + 5) имеет следующий график (рис. 7).

Из графика видно, что результатом сложения может быть любое число из интервала (6; 8), возводя его в квадрат, получим плотность выражения I_1 $f_{(2+5)^2(x)}$ (рис. 8).

Энтропия выражения I_1 численно равна $H(I_1) \approx 3{,}1373$.

Вычислим энтропию выражения $I_2 = 2^2 + 2 \cdot 2 \cdot 5 + 5^2$.

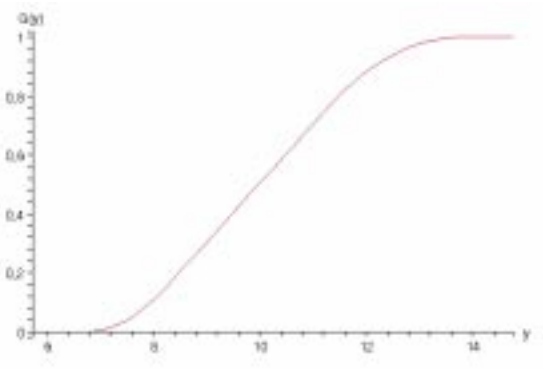


Рис. 5

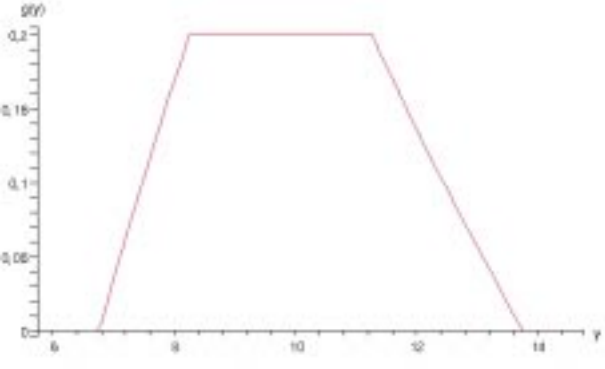


Рис. 6

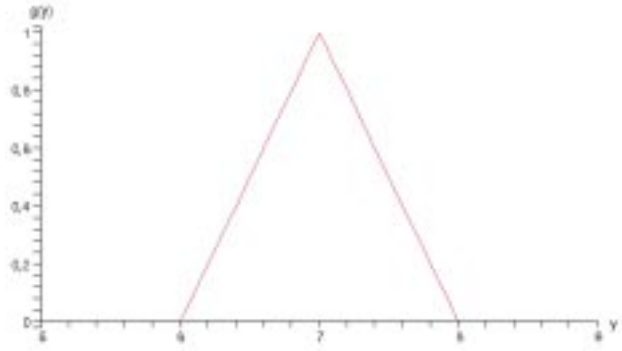


Рис. 7

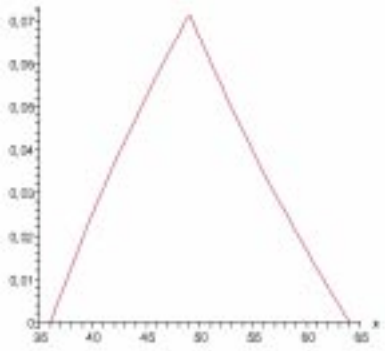


Рис. 8

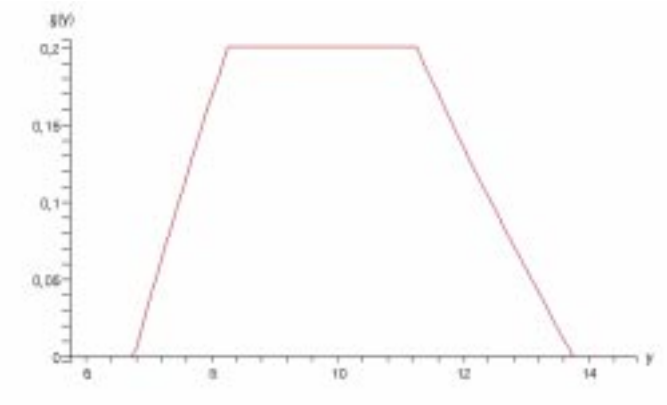


Рис. 9

График плотности распределения случайной величины $2 \cdot 5$ имеет следующий вид (рис. 9)

Произведение может принимать все значения из интервала (6,75; 13,75). Максимальная плотность вероятностей имеет место для интервала (8,25; 11,25).

Плотности распределения величин 2^2 и 2^5 показаны на рис. 10.

Плотность распределения суммы квадратов $2^2 + 2^5$ имеет следующий вид (рис. 11).

Плотность распределения выражения I_2 показана на рис. 12.

Энтропия такого распределения численно равна $H(I_2) \approx 2{,}8923$.

Можно видеть, что закон распределения выражения I_2 близок к нормальному распределению. Это обусловлено тем, что выражение I_2 содержит 3 операции сложения. Из центральной предельной теоремы теории вероятностей известно, что при сложении большого количества независимых случайных величин закон распределения их суммы приближается к нормальному [1].

Математическое ожидание и среднеквадратичное отклонение полученного распределения $m_I \approx 49{,}167, \sigma_I \approx 4{,}4014$.

Значение энтропии выражения I_2 можно вычислить приближённо как энтропию соответствующего нормального распределения. Энтропия нормального закона имеет простую аналитическую запись [1]: $H(N(m,\sigma)) = \log(\sigma\sqrt{2\pi \cdot e})$.

Вычислим энтропию нормального распределения с параметрами m_I, σ_I :
$H\left(N\left(m_I, \sigma_I\right)\right) \approx 2{,}9009$.

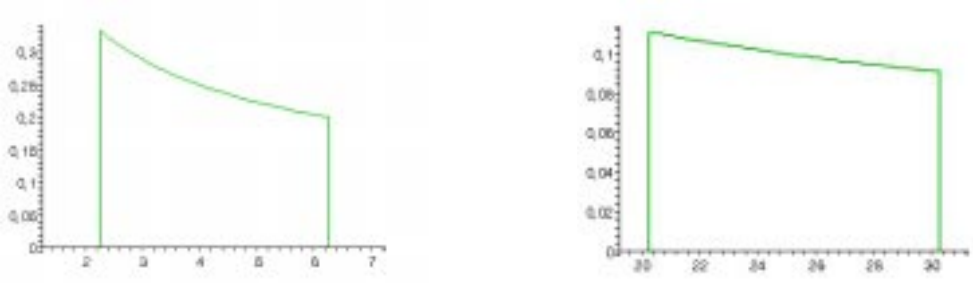


Рис. 10

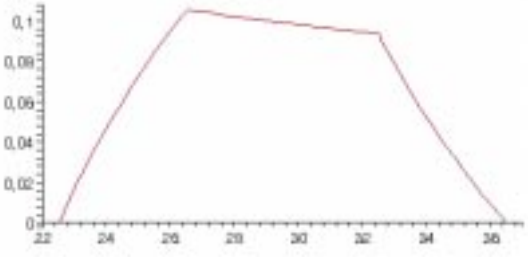


Рис. 11

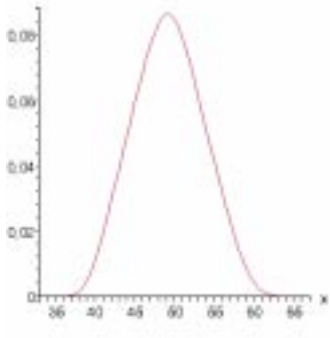


Рис.12

Можно видеть, что энтропии, вычисленные по определению и с помощью предельной теоремы, отличаются не более чем на 0,01.

Сравнивая энтропии выражений $I_1 = (2+5)^2$ и $I_2 = 2^2 + 2 \cdot 2 \cdot 5 + 5^2$, можно видеть, что $H(I_1) \approx 3{,}1373 > H(I_2) \approx 2{,}8923$.

Приведём оценки погрешности выражений по правилам элементарной теории погрешности. Погрешностями округления после выполнения арифметических операций будем пренебрегать.

Для первого выражения можно записать

$$\delta_{(a+b)^2} = 2\delta_{(a+b)} = 2\left(\frac{\Delta a + \Delta b}{a+b}\right),$$

для второго выражения формула погрешности будет сложнее

$$\delta_{a^2+b^2+2ab} = 2\left(\frac{a\Delta a + b\Delta b + 2(b\Delta a + a\Delta b)}{a^2 + b^2 + 2ab}\right).$$

Заметим, что первая оценка несколько занижена, так как при её получении пренебрегли членом $\frac{(\Delta a + \Delta b)^2}{(a+b)^2}$.

Численные значения погрешности $\delta_{(a+b)^2} = 0{,}28$, $\delta_{a^2+b^2+2ab} = 0{,}42$.

Количественный анализ показывает, что, с точки зрения предложенной модели округления чисел, запись I_2 содержит в себе больше информации, чем запись I_1, так как запись I_2 является развернутой формой I_1. Следовательно, результат во втором случае более определён.

Аналогичные вычисления для ряда других чисел показал, что во всех случаях развернутые формы выражений имеют меньшую энтропию, чем формы свёрнутые.

Проводился анализ следующего тождества

$$Sin(x+y) = Sin(x) * Cos(y) + Cos(x) * Sin(y).$$

В этом случае $I_1 = Sin(x+y)$ и $I_2 = Sin(x) * Cos(y) + Cos(x) * Sin(y)$, X, Y случайные величины, равномерно распределённые на интервале $[0, \pi/4]$, то есть $x = \pi/8, y = \pi/8$, а $\Delta x = \pi/8, \Delta y = \pi/8$. Интервалы изменения случайных величин $I_1 \in [0;1]$, $I_2 \in [0;\sqrt{2}]$.

Основные вероятностные характеристики для правой части тождества: математическое ожидание $M[I_1] = 0{,}6715$, дисперсия $D[I_1] = 0{,}4909$, энтропия $H[I_1] = 0{,}2216$. Вероятностные характеристики левой части тождества: математическое ожидание $M[I_2] = 0{,}6703$, дисперсия $D[I_2] = 0{,}0718$, энтропия $H[I_2] = 0{,}1835$.

Можно видеть, что и в этом случае интервал погрешности расширяется для развёрнутого выражения, а энтропия уменьшается.

ЗАКЛЮЧЕНИЕ

Проведенное исследование показывает, что в случае развернутых формул вычислений интервал оценивания расширяется, но при этом распределение погрешности становится неравномерным. Следовательно, применяя методы статистической оценки погрешности с учётом её доверительного интервала вероятности, можно уменьшить интервал учитываемой погрешности.

Заметим, что если операции сложения преобладают в исследуемом выражении, то, энтропия такого выражения может быть найдена как энтропия нормального закона с соответствующими параметрами распределения.

Литература

1. *Вентцель Е.С., Овчаров Л.А.* Теория вероятностей и её инженерные приложения. М.: Наука, 1976. С. 480.
2. *Хемминг Р.В.* Численные методы для научных работников и инженеров. М.: Наука, 1968. С. 400.
3. *Яглом А.М., Яглом И.М.* Вероятность и информация. М.: Наука, 1973. С. 511.

Abstract

In the paper it is shown that different notations of algebraic expressions contain different amounts of information about distribution of calculating errors. For unfold calculating formulae estimating interval broadens but the distribution of error becomes nonuniform. The concept of entropy is used to estimate degree of nonuniformness of distribution.

Keywords: error, information, entropy, distribution.

Ляхов Александр Федорович,
кандидат физико-математических наук, доцент кафедры теоретической механики механико-математического факультета НГУ им. Н.И. Лобачевского,
Lyakhov@mm.unn.ru