V. V. Voevodin, The asymptotic distribution of rounding off errors in linear transformations, *Zh. Vychisl. Mat. Mat. Fiz.*, 1967, Volume 7, Number 5, 965–976

ЖУРНАЛ
ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Том 7 Сентябрь 1967 Октябрь № 5

УДК 518:512.25

# ОБ АСИМПТОТИЧЕСКОМ РАСПРЕДЕЛЕНИИ ОШИБОК ОКРУГЛЕНИЯ ПРИ ЛИНЕЙНЫХ ПРЕОБРАЗОВАНИЯХ

*В. В. ВОЕВОДИН*

*(Москва)*

1. Пусть в $n$-мерном вещественном пространстве $R$ задана односвязная выпуклая замкнутая область $G$. Будем считать, что векторы $z$ из $G$ — случайные величины, плотность распределения которых есть непрерывная функция $P(z)$, причем $P(z) \geqslant c > 0$.

Предположим, что над векторами $z$ совершается последовательность $U_1, U_2, \ldots, U_k$, $k < n$, преобразований, матрицы которых имеют вид

$$U_r = E - a_r b'_r, \qquad r = 1, 2, \ldots, k. \tag{1}$$

Здесь $a_r$ и $b_r$ — прямоугольные матрицы размером $n \times 1$ (векторы-столбцы). Отметим сразу, что на подобных преобразованиях основано большинство прямых и итерационных методов решения задач линейной алгебры.

Обозначим через $z_r$ вектор, полученный из вектора $z \subset G$ после первых $r$ преобразований; тогда

$$z_r = z_{r-1} - (b_r, z_{r-1}) a_r. \tag{2}$$

В действительности мы не можем выполнить эти преобразования, так как все вычисления приходится осуществлять с конечным числом знаков. Допустим, что вычисления ведутся с $t$ двоичными знаками после запятой, а промежуточные результаты, имеющие больше $t$ знаков, округляются по обычному правилу (см. ниже). Сами преобразования будут определяться матрицами

$$U_r^{(t)} = E - a_r^{(t)} b_r^{(t)'},$$

где векторы $a_r^{(t)}$ и $b_r^{(t)}$ получены из векторов $a_r$ и $b_r$ округлением их координат до $t$ знаков.

Во многих задачах линейной алгебры (например, при решении систем линейных алгебраических уравнений) безразлично, выполнять ли преобразования (2) с векторами $a_r$, $b_r$ или $a_r^{(t)}$, $b_r^{(t)}$, важно лишь, чтобы само преобразование было выполнено достаточно точно. Поэтому в дальнейшем мы будем считать, что векторы $z_r$ вычисляются по формуле

$$z_r = z_{r-1} - (b_r^{(t)}, z_{r-1}) a_r^{(t)}. \tag{3}$$

Однако и в этом случае реально будет вычислен лишь вектор $z_r^{(t)}$, где

$$z_r^{(t)} = z_{r-1}^{(t)} - (b_r^{(t)}, z_{r-1}^{(t)}) a_r^{(t)} + \varepsilon_r^{(t)}, \tag{4}$$

а $\varepsilon_r^{(t)}$ — ошибка, внесенная на $r$-м шаге за счет неточной реализации формулы (3). Имеем

$$z_r^{(t)} = U_r^{(t)} z_{r-1}^{(t)} + \varepsilon_r^{(t)} = \ldots$$

$$\ldots = U_r^{(t)} U_{r-1}^{(t)} \ldots U_1^{(t)} (z + \varepsilon_0^{(t)} + U_1^{(t)^{-1}} \varepsilon_1^{(t)} + U_1^{(t)^{-1}} U_2^{(t)^{-1}} \varepsilon_2^{(t)} + \ldots$$

$$\ldots + U_1^{(t)^{-1}} U_2^{(t)^{-1}} \ldots U_r^{(t)^{-1}} \varepsilon_r^{(t)}).$$

Здесь $\varepsilon_0^{(t)}$ — ошибка, полученная от округления компонент вектора $z$ до $t$ знаков после запятой.

Из последней формулы следует, что вектор $z_r^{(t)}$ можно рассматривать как результат точного преобразования вектора, стоящего в круглых скобках. Этот вектор отличается от заданного вектора $z$ на величину

$$\eta_r^{(t)} = \varepsilon_0^{(t)} + U_1^{(t)^{-1}} \varepsilon_1^{(t)} + \ldots + U_1^{(t)^{-1}} U_2^{(t)^{-1}} \ldots U_r^{(t)^{-1}} \varepsilon_r^{(t)}, \tag{5}$$

которую согласно [1] будем называть эквивалентным возмущением. В [1] исследуются мажорантные оценки эквивалентного возмущения. В настоящей работе $\eta_r^{(t)}$ исследуется как функция случайного аргумента $z$. Кроме этого, в отличие от [1], здесь исследуется лишь влияние ошибок на эквивалентность преобразования, что позволяет не учитывать ошибки вычисления векторов $a_r^{(t)}$, $b_r^{(t)}$ и получить за счет этого несколько лучшие оценки. Ниже доказывается, что при $t \to \infty$ главный член $\eta_r^{(t)}$ есть сумма независимых равномерно распределенных величин. Приводятся оценки для главного члена и членов более высокого порядка малости.

2. Предположим, что при реализации формулы (3) скалярное произведение вычисляется с удвоенным числом знаков; тогда

$$\varepsilon_r^{(t)} = \delta_r^{(t)} a_r^{(t)} + \sigma_r^{(t)}, \tag{6}$$

где $\delta_r^{(t)}$ — ошибка от округления скалярного произведения $(b_{r-1}^{(t)}, z_{r-1}^{(t)})$, а $\sigma_r^{(t)}$ — ошибка от округления произведения округленного скалярного произведения на вектор $a_r^{(t)}$.

Очевидно, что

$$|\delta_r^{(t)}| \leqslant 1/2\, 2^{-t}, \qquad |\sigma_r^{(t)}| \leqslant 1/2\, 2^{-t}.$$

Второе неравенство означает, что модули всех координат вектора $\sigma_r^{(t)}$ не превосходят $1/2 2^{-t}$.

Из (5), (6) следует

$$\eta_r^{(t)} = \varkappa_r^{(t)} + \nu_r^{(t)},$$

где

$$\varkappa_r^{(t)} = \varepsilon_0^{(t)} + U_1^{(t)^{-1}} \delta_1^{(t)} a_1^{(t)} + \ldots + U_1^{(t)^{-1}} U_2^{(t)^{-1}} \ldots U_r^{(t)^{-1}} \delta_r^{(t)} a_r^{(t)},$$

$$\nu_r^{(t)} = U_1^{(t)^{-1}} \sigma_1^{(t)} + \ldots + U_1^{(t)^{-1}} U_2^{(t)^{-1}} \ldots U_r^{(t)^{-1}} \sigma_r^{(t)}$$

Как будет показано ниже, распределение ошибок для основных численных методов определяется функцией $\nu_r^{(t)}$. Поэтому $\nu_r^{(t)}$ мы будем исследовать как функцию случайного аргумента $z$, а $\varkappa_r^{(t)}$ оценим мажорантно.

Заметим, что при фиксированном векторе $a_r^{(t)}$ ошибка $\sigma_r^{(t)}$ зависит только от величины округленного скалярного произведения $(b_r^{(t)}, z_{r-1}^{(t)})$. Так как векторов, подвергающихся преобразованию, гораздо больше, чем возможных округленных значений величины $(b_r^{(t)}, z_{r-1}^{(t)})$, то обязательно некоторое количество векторов $z_{r-1}^{(t)}$ будет иметь при $r$-м преобразовании одну и ту же ошибку $\sigma_r^{(t)}$.

Обозначим через $G^{(t)}$ множество тех векторов, в которые переходят векторы из $G$ после округления их координат до $t$ знаков. Очевидно, что $G^{(t)} \subset G$, за исключением точек, близких к границе. Разобьем множество $G^{(t)}$ на непересекающиеся подмножества $G^{(t)}_{k_1k_2\ldots k_r}$, относя к каждому из них только те векторы из $G^{(t)}$, которые при всех преобразованиях от 1-го до $r$-го имеют одинаковые значения округленных скалярных произведений, равных, соответственно, $k_1, k_2, \ldots, k_r$.

Подмножества $G^{(t)}_{k_1k_2\ldots k_r}$ будем называть пучками $r$-го порядка. В силу сказанного выше, по крайней мере некоторые из пучков состоят более чем из одного вектора. Как следует из определения, все векторы, принадлежащие одному пучку $r$-го порядка, будут давать одно и то же значение ошибки $\nu_r^{(t)}$. Кроме того, ясно, что

$$\bigcup_{k_{r+1}} G^{(t)}_{k_1k_2\ldots k_rk_{r+1}} = G^{(t)}_{k_1k_2\ldots k_r}, \qquad \bigcup_{k_1} G^{(t)}_{k_1} = G^{(t)}.$$

Наши ближайшие задачи: исследование ошибки, которую несет пучок, исследование структуры пучка и вычисление вероятности попадания в тот или иной пучок.

Определим отклонение числа $x$ от ближайшего целого следующим соотношением:

$$\lfloor x \rceil = \begin{cases} \{x\}, & \text{если } \{x\} < 1/2, \\ \{x\} - 1, & \text{если } \{x\} \geqslant 1/2, \end{cases}$$

где $\{x\}$ означает дробную долю числа $x$ [2]. Через $\langle x \rangle$ будем обозначать округленное до $t$ знаков число $x$. Если под знаками $\{\ \}$, $\lfloor\ \rceil$, $\langle\ \rangle$ будет стоять вектор, то результатом такой операции будет также вектор, каждая координата которого получена применением соответствующей операции к координате исходного вектора.

Легко подсчитать, что

$$\langle a \rangle = a - 2^{-t}\lfloor 2^t a \rceil,$$

поэтому

$$\sigma_r^{(t)} = -2^{-t}\lfloor 2^t \langle (b_r^{(t)}, z_{r-1}^{(t)}) \rangle a_r^{(t)} \rceil.$$

Отклонение от ближайшего целого есть кусочно-линейная функция дробной доли, число $2^t\langle (b_r^{(t)} z_{r-1}^{(t)}) \rangle$ — целое. Следовательно, изучение ошибки, которую несет пучок, сводится к изучению распределения дробных

долей векторов $ka_r^{(t)}$, где $k$ пробегает некоторую последовательность целых чисел.

Пусть для вектора $z_{r-1}^{(t)}$ округленное значение скалярного произведения равно $l$. Ясно, что одну и ту же ошибку $\sigma_r^{(t)}$ будут иметь все векторы $z_{r-1}^{(t)}$, для которых

$$l - {}^1/_2 2^{-t} \leqslant (b_r^{(t)}, z_{r-1}^{(t)}) < l + {}^1/_2 2^{-t},$$

т. е. все векторы, принадлежащие слою с одной исключенной границей. Причем толщина этого слоя равна $2^{-t}\|b_r^{(t)}\|_E^{-1}$ и его плоские грани перпендикулярны вектору $b_r^{(t)}$.

Из формул (4), (6) получаем

$$z_r^{(t)} = z_{r-1}^{(t)} - \langle (b_r^{(t)}, z_{r-1}^{(t)}) \rangle a_r^{(t)} + \sigma_r^{(t)},$$

откуда следует, что преобразование векторов, принадлежащих одному пучку, заключается только в операции переноса, так как для всех векторов пучка вектор $-\langle (b_r^{(t)}, z_{r-1}^{(t)}) \rangle a_r^{(t)} + \sigma_r^{(t)}$ будет постоянным. Отсюда вытекает, что все векторы пучка $G_{k_1 k_2 \ldots k_r}^{(t)}$ принадлежат некоторой области $G_{k_1 k_2 \ldots k_r}$, являющейся пересечением области $G$ и $r$ слоев, перпендикулярных, соответственно, векторам $b_1^{(t)}$, $b_2^{(t)}, \ldots, b_r^{(t)}$ и имеющих толщину $2^{-t}\|b_1^{(t)}\|_E^{-1}$, $2^{-t}\|b_2^{(t)}\|_E^{-1}, \ldots, 2^{-t}\|b_r^{(t)}\|_E^{-1}$. Ни одна из областей $G_{k_1 k_2 \ldots k_r}$ не содержит векторов из двух различных пучков, все области непересекающиеся и

$$\bigcup_{k_{r+1}} G_{k_1 k_2 \ldots k_r, r+1} = G_{k_1 k_2 \ldots k_r}, \qquad \bigcup_{k_1} G_{k_1} = G.$$

Обозначим через $E_z$ множество векторов, принадлежащих кубу с центром в точке $z$ со сторонами, параллельными осям координат и равными $2^{-t}$. Если $z \subset G^{(t)}$, то все векторы из области $G \cap E_z$, за исключением некоторых граничных, после округления их координат до $t$ знаков после запятой переходит в вектор $z$. Вероятность $p(z \subset G_{k_1 k_2 \ldots k_r}^{(t)})$ попадания в пучок $G_{k_1 k_2 \ldots k_r}^{(t)}$ определяется формулой

$$p(z \subset G_{k_1 k_2 \ldots k_r}^{(t)}) = \int_{G_{k_1 k_2 \ldots k_r}^{(E)}} P(z)\, dz,$$

где

$$G_{k_1 k_2 \ldots k_r}^{(E)} = \bigcup_{z \subset G_{k_1 k_2 \ldots k_r}^{(t)}} (G \cap E_z).$$

3. Для дальнейшего исследования нам потребуются некоторые сведения и результаты из теории чисел. Предположим [2], что дано конечное число векторов $z^{(q)}$, $1 \leqslant q \leqslant Q$, с координатами $z_j^{(q)}$, распределенных каким-то образом в единичном кубе, т. е.

$$0 \leqslant z_j^{(q)} < 1, \qquad 1 \leqslant j \leqslant n.$$

Пусть $\alpha$ и $\beta$ суть $n$-мерные векторы с координатами $\alpha_j$ и $\beta_j$, $0 \leqslant \alpha_j < \beta_j < 1$. Обозначим через $F(\alpha, \beta)$ число векторов $z^{(q)}$, лежащих в параллелепипеде

$$\alpha_j \leqslant z_j^{(q)} < \beta_j$$

объема $\Pi(\beta_j - \alpha_j)$. Тогда

$$D = \sup_{\alpha,\beta} |Q^{-1}F(\alpha, \beta) - \Pi(\beta_j - \alpha_j)|$$

называется отклонением векторов $z^{(q)}$.

Если задана бесконечная последовательность векторов, то через $D_Q$ обозначим отклонение первых $Q$ из них. В случае, когда

$$\lim_{Q\to\infty} D_Q = 0,$$

будем говорить, что последовательность векторов $z^{(q)}$ равномерно распределена в единичном кубе.

Известно [2], что последовательность векторов $\{k\theta\}$, $k = 1, 2, \ldots$, равномерно распределена, если только координаты вектора $\theta$ рационально независимы вместе с числом 1. Возьмем некоторую последовательность векторов $\theta_p$, $p = 1, 2, \ldots$, сходящуюся к $\theta$, и для каждого $p$ построим систему векторов $\{k_p\theta_p + \gamma\}$, $\{(k_p + 1)\theta_p + \gamma\}, \ldots, \{(k_p + l_p)\theta_p + \gamma\}$. Обозначим через $D_p$ отклонение этой системы. Справедлива

Теорема 1. *Если вектор* $\theta$ *имеет рационально независимые вместе с числом* 1 *координаты и*

$$\lim_{p\to\infty} \theta_p = \theta, \qquad \lim_{p\to\infty} l_p = \infty,$$

*то*

$$\lim_{p\to\infty} D_p = 0$$

*равномерно по* $k_p$ *и* $\gamma$.

Мы не будем останавливаться на доказательстве этой теоремы, так как оно весьма громоздко и не представляет особого интереса для дальнейшего.

Следующая лемма устанавливает соотношение между объемом области $G_{k_1k_2\ldots k_r}$ ($\operatorname{mes} G_{k_1k_2\ldots k_r}$) и объемом области $G^{(E)}_{k_1k_2\ldots k_r}$ ($\operatorname{mes} G^{(E)}_{k_1k_2\ldots k_r}$).

Лемма 1. *Почти для всех векторов* $b_1, b_2, \ldots, b_r$ *имеет место предельное соотношение*

$$\lim_{t\to\infty} \frac{\operatorname{mes} G_{k_1k_2\ldots k_r}}{\operatorname{mes} G^{(E)}_{k_1k_2\ldots k_r}} = 1, \tag{7}$$

*причем равномерно почти по всем* $k_1, k_2, \ldots, k_r$.

Основная идея доказательства ясна из рассмотрения случая $n = 2$, $r = 1$. Для простоты мы ограничимся лишь этим случаем.

Не уменьшая общности, можно считать, что элементы $G^{(t)}$ совпадают с узлами сетки, построенной с шагом, равным 1. Тогда для доказательства леммы достаточно показать, что количество элементов из $G^{(t)}$, попавшее в любой слой, асимптотически равно объему слоя.

Будем называть лучом все точки $G^{(t)}$, лежащие на одной прямой. Количество точек, попавших в слой, совпадает с количеством их проекций (с учетом кратности) на вектор $b_1^{(t)}$, попавших в проекцию самого слоя на тот же вектор. Проекция слоя представляет собой полуинтервал длины $\|b_1^{(t)}\|_E^{-1}$.

Спроектируем точки одного луча $G^{(t)}$ на вектор $b_1^{(t)}$. Эти проекции образуют на $b_1^{(t)}$ некоторую равномерную сетку с шагом $\cos\varphi^{(t)}$, где $\varphi^{(t)}$ — угол между лучом и вектором $b_1^{(t)}$. Проектирование точек соседнего луча равносильно сдвигу первой сетки на $\sin\varphi^{(t)}$ вправо или влево. Если отношение $\sin\varphi^{(t)} / \cos\varphi^{(t)}$ сходится к иррациональному числу (а это определяется лишь вектором $b_1^{(t)}$), то, согласно теореме 1, количество точек, принадлежащих любым соседним $p$ лучам и проекции которых на $b_1^{(t)}$ попадают в фиксированный отрезок, асимптотически (при $p \to \infty$) пропорционально $p$, причем независимо от положения самого отрезка.

Независимо от положения слоя его объем асимптотически пропорционален количеству пересекающих его лучей. Поэтому утверждение леммы следует из того, что

$$\sum_{k_1} \operatorname{mes} G_{k_1}^{(E)} = \sum_{k_1} \operatorname{mes} G_{k_1} = \operatorname{mes} G.$$

Сходимость (7) не будет равномерной только для тех слоев, которые примыкают к точкам пересечения векторов $b_r^{(t)}$ с границей области $G$.

Теорема 2. *В условиях выполнения леммы* 1 *имеет место предельное соотношение*

$$\lim_{t\to\infty} \frac{P(z \subset G_{k_1 k_2 \ldots k_r})}{P(z \subset G^{(E)}_{k_1 k_2 \ldots k_r})} = 1, \tag{8}$$

*причем равномерно почти по всем* $k_1, k_2, \ldots, k_r$.

Функция $P(z)$ равномерно непрерывна в $G$, поэтому по заданному $\varepsilon$ можно найти такое разбиение области $G$ на непересекающиеся области $g_1(\varepsilon), g_2(\varepsilon), \ldots, g_N(\varepsilon)$, что отклонение $P(z)$ от среднего значения на каждой $g_i(\varepsilon)$ не будет превосходить $\varepsilon$. Очевидно, что области $g_i(\varepsilon)$ можно рассматривать в качестве области $G$ леммы 1. Следовательно,

$$\lim_{t\to\infty} \frac{\operatorname{mes} g_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_r}}{\operatorname{mes} g_i(\varepsilon) \cap G^{(E)}_{k_1 k_2 \ldots k_r}} = 1. \tag{9}$$

Пусть среднее значение $P(z)$ в области $g_i(\varepsilon)$ равно $p_i \geqslant c > 0$. Используя теорему о среднем значении интеграла и соотношение (9), имеем

$$\lim_{t\to\infty} \frac{P(z \subset g_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_r})}{P(z \subset g_i(\varepsilon) \cap G^{(E)}_{k_1 k_2 \ldots k_r})} =$$

$$= \lim_{t\to\infty} \frac{(p_i + O(\varepsilon)) \operatorname{mes} g_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_r}}{(p_i + O(\varepsilon)) \operatorname{mes} g_i(\varepsilon) \cap G^{(E)}_{k_1 k_2 \ldots k_r}} = 1 + O(\varepsilon).$$

Отсюда получаем, что по заданному $\varepsilon$ можно найти такое $t_0(\varepsilon)$, при

котором для всех $t \geqslant t_0(\varepsilon)$ будут выполняться неравенства

$$(1 - O(\varepsilon))P(z \subset g_i(\varepsilon) \cap G^{(E)}_{k_1 k_2 \ldots k_r}) \leqslant P(z \subset g_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_r}) \leqslant$$

$$\leqslant (1 + O(\varepsilon))P(z \subset g_i(\varepsilon) \cap G^{(E)}_{k_1 k_2 \ldots k_r}), \qquad O(\varepsilon) \geqslant 0.$$

Далее,

$$\frac{P(z \subset G_{k_1 k_2 \ldots k_r})}{P(z \subset G^{(E)}_{k_1 k_2 \ldots k_r})} = \left[\sum_{i=1}^{N} P(z \subset g_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_r})\right] \times$$

$$\times \left[\sum_{i=1}^{N} P(z \subset g_i(\varepsilon) \cap G^{(E)}_{k_1 k_2 \ldots k_r})\right]^{-1},$$

поэтому для всех $t \geqslant t_0(\varepsilon)$

$$1 - O(\varepsilon) \leqslant \frac{P(z \subset G_{k_1 k_2 \ldots k_r})}{P(z \subset G^{(E)}_{k_1 k_2 \ldots k_r})} \leqslant 1 + O(\varepsilon).$$

В силу произвольности выбора $\varepsilon$ отсюда вытекает утверждение теоремы.

Зафиксируем теперь произвольно малое число $\mu > 0$ и рассмотрим область $G^{(t)}_{\mu}$, получающуюся из области $G$ с помощью вырезания тех граничных областей $G_{k_1 k_2 \ldots k_r}$, для которых нарушается равномерное выполнение предельного равенства (8). При этом будем предполагать, что

$$P(z \subset G^{(t)}_{\mu}) \geqslant 1 - \mu.$$

Не ограничивая общности, можно считать, что ошибка $\sigma^{(t)}_r$ распределена в единичном кубе, размерность которого совпадает с числом ненулевых коэффициентов вектора $a_r$. Пусть $S_r$ есть некоторый параллелепипед в этом кубе. Справедлива

Лемма 2. *Имеет место предельное соотношение*

$$\lim_{t \to \infty} \frac{P(\sigma_r^{(t)} \subset S_r,\ z \subset G^{(t)}_{k_1 k_2 \ldots k_{r-1}})}{\operatorname{mes} S_r P(z \subset G^{(t)}_{k_1 k_2 \ldots k_{r-1}})} = 1,$$

*причем равномерно для всех пучков из* $G^{(t)}_{\mu}$.

Имеем [3]

$$P(\sigma_r^{(t)} \subset S_r, \quad z \subset G^{(t)}_{k_1 k_2 \ldots k_{r-1}}) =$$

$$= P(z \subset G^{(t)}_{k_1 k_2 \ldots k_{r-1}}) P(\sigma_r \subset S_r / z \subset G^{(t)}_{k_1 k_2 \ldots k_{r-1}}),$$

поэтому для доказательства леммы достаточно показать, что

$$\lim_{t \to \infty} P(\sigma_r^{(t)} \subset S_r / z \subset G^{(t)}_{k_1 k_2 \ldots k_{r-1}}) = \operatorname{mes} S_r \tag{10}$$

равномерно для всех пучков из $G^{(t)}_{\mu}$.

Заметим, что в области $G^{(t)}_{\mu}$ каждый пучок $(r-1)$-го порядка при $t \to \infty$ порождает бесконечно много пучков $r$-го порядка, если только векторы $b_1, b_2, \ldots, b_r$ линейно-независимы. Если бы все пучки $r$-го порядка были равновероятны, то соотношение (10) тривиально следовало бы из теоремы 1. В общем случае доказательство этого соотношения во многом сходно с доказательством теоремы 2.

По заданному $\varepsilon > 0$ находим разбиение области $G_\mu^{(t)}$ на непересекающиеся области $q_i(\varepsilon)$, ограниченные плоскостями, перпендикулярными векторам $b_1^{(t)}, b_2^{(t)}, \ldots, b_r^{(t)}$ и границей области $G_\mu^{(t)}$. При этом будем предполагать, что в каждой из $q_i(\varepsilon)$ все пучки $r$-го порядка равновероятны с точностью до множителя $(1 + O(\varepsilon))$ для всех $t \geqslant t_1(\varepsilon)$. Получаем [3]

$$P(\sigma_r^{(t)} \subset S_r / z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)}) = \frac{P(z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)}, \sigma_r^{(t)} \subset S_r)}{P(z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)})} =$$

$$= \frac{\sum_i P(z \subset q_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_{r-1}}^{(t)}, \sigma_r^{(t)} \subset S_r)}{P(z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)})} =$$

$$= \sum_i \frac{P(z \subset q_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_{r-1}}^{(t)})}{P(z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)})} \; P(\sigma_r^{(t)} \subset S_r / z \subset q_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_{r-1}}^{(t)}).$$

Согласно разбиению $G_\mu^{(t)}$ на области $q_i(\varepsilon)$, будем иметь

$$\lim_{t \to \infty} P(\sigma_r^{(t)} \subset S_r / z \subset q_i(\varepsilon) \cap G_{k_1 k_2 \ldots k_{r-1}}^{(t)}) = \operatorname{mes} S_r (1 + O(\varepsilon))$$

равномерно для всех пучков и $q_i(\varepsilon)$. Отсюда (10) следует в силу произвольности числа $\varepsilon$.

Теорема 3. *Ошибки* $\sigma_r^{(t)}$, $r = 1, 2, \ldots, n-1$, *асимптотически являются независимыми равномерно распределенными величинами почти для всех векторов* $a_r$ *и* $b_r$.

Используя лемму 2, находим

$$\lim_{t \to \infty} P(\sigma_r^{(t)} \subset S_r) = \lim_{t \to \infty} [P(\sigma_r^{(t)} \subset S_r, z \subset G_\mu^{(t)}) + P(\sigma_r^{(t)} \subset S_r, z \overline{\subset} G_\mu^{(t)})] =$$

$$= \lim_{t \to \infty} \Big[ \sum_{G_{k_1 k_2 \ldots k_{r-1}}^{(t)}} P(\sigma_r^{(t)} \subset S_r, \; z \subset G_\mu^{(t)} \cap G_{k_1 k_2 \ldots k_{r-1}}^{(t)}) +$$

$$+ P(\sigma_r^{(t)} \subset S_r, \; z \overline{\subset} G_\mu^{(t)}) \Big] = \operatorname{mes} S_r + O(\mu)$$

для любого $\mu$. Следовательно,

$$\lim_{t \to \infty} P(\sigma_r^{(t)} \subset S_r) = \operatorname{mes} S_r.$$

Далее,

$$\lim_{t \to \infty} P(\sigma_r^{(t)} \subset S_r, \; \sigma_{r-1}^{(t)} \subset S_{r-1}, \ldots, \sigma_1^{(t)} \subset S_1) = \lim_{t \to \infty} [P(\sigma_{r-1}^{(t)} \subset S_{r-1}, \; \sigma_{r-2}^{(t)} \subset$$

$$\subset S_{r-2}, \ldots, \sigma_1^{(t)} \subset S_1) P(\sigma_r^{(t)} \subset S_r / \sigma_{r-1}^{(t)} \subset S_{r-1}, \ldots, \sigma_1^{(t)} \subset S_1)] =$$

$$= \lim_{t \to \infty} P(\sigma_{r-1}^{(t)} \subset S_{r-1}, \ldots$$

$$\ldots, \sigma_1^{(t)} \subset S_1) \lim_{t \to \infty} \frac{\sum P(z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)}) P(\sigma_r^{(t)} \subset S_r / z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)})}{\sum P(z \subset G_{k_1 k_2 \ldots k_{r-1}}^{(t)})} =$$

$$= \operatorname{mes} S_r \lim_{t \to \infty} P(\sigma_{r-1}^{(t)} \subset S_{r-1}, \ldots, \sigma_1^{(t)} \subset S_1) = \operatorname{mes} S_r \operatorname{mes} S_{r-1} \ldots \operatorname{mes} S_1.$$

В последних суммах суммирование ведется лишь по тем пучкам, которые дают ошибки из $S_1, S_2, \ldots, S_{r-1}$.

4. Пусть $\nu = \sigma_1 + \sigma_2 + \ldots + \sigma_r$ есть случайная величина, равная сумме случайных независимых величин $\sigma_i$, каждая из которых равномерно распределена в параллелепипеде с центром в нуле и со сторонами длиной $\sigma_{i1}, \sigma_{i2}, \ldots, \sigma_{ir}$. Тогда математическое ожидание величины $\|\nu\|_E^2$ определяется формулой

$$M\|\nu\|_E^2 = {}^1/_{12} \sum_{i,j} \sigma_{ij}^2.$$

Этот факт легко установить непосредственной проверкой. Рассмотрим теперь несколько конкретных преобразований, укладывающихся в схему (1) и встречающихся в численных методах решения задач линейной алгебры [4].

Метод Гаусса (без деления строки на ведущий элемент). Не ограничивая общности, можно считать, что матрицы $U_r$ в этом методе имеют вид

$$U_r = \begin{Vmatrix} 1 & & & & \\ & \ddots & & 0 & \\ & & 1 & & \\ & & \alpha_{r+1,\,r} & 1 & \\ 0 & & \vdots & & \ddots \\ & & \alpha_{n,\,r} & & 1 \end{Vmatrix}.$$

Следовательно,

$$a_r' = (\overbrace{0, \ldots, 0}^{r-1}, 0, \alpha_{r+1,\,r}, \ldots, \alpha_{nr}),$$

$$b_r' = (0, \ldots, 0, 1, 0, \ldots, 0).$$

Для единичных векторов $b_r$ справедливы все высказанные выше утверждения и, кроме этого, $\delta_r^{(t)} = 0$, $r = 1, 2, \ldots, n-1$. Далее, в матрице $U_1^{(t)-1} U_2^{(t)-1} \ldots U_r^{(t)-1}$ отличны от единичных лишь первые $r$ столбцов, а первые $r$ координаты вектора $\sigma_r^{(t)}$ равны нулю. Поэтому

$$U_1^{(t)-1} U_2^{(t)-1} \ldots U_r^{(t)-1} \sigma_r^{(t)} = \sigma_r^{(t)}.$$

Так как для реализации метода Гаусса нужно выполнить $n-1$ преобразований, то

$$\frac{\|x_{n-1}^{(t)}\|_E^2}{2^{2t}} \leqslant \frac{n}{4}, \qquad \lim_{t\to\infty} \frac{M\|\nu_{n-1}^{(t)}\|_E^2}{2^{2t}} \leqslant \frac{n^2}{24}. \tag{11}$$

Метод отражений. В этом методе матрицы $U_r$ ортогональные и имеют вид

$$U_r = E - 2w_r w_r', \quad \|w_r\|_E = 1.$$

Несложные вычисления показывают, что

$$\lim_{t\to\infty}\frac{\|\varkappa_{n-1}^{(t)}\|_E^2}{2^{2t}}\leqslant\frac{(2n+1/n)^2}{4},\qquad \lim_{t\to\infty}\frac{M\|\nu_{n-1}^{(t)}\|_E^2}{2^{2t}}\leqslant\frac{n^2}{24}.$$

Метод ортогонализации (без нормировки строк). В этом методе

$$U_r=\begin{Vmatrix} 1 & & & & 0 & \\ & \ddots & & & & \\ & & 1 & & & \\ \alpha_{r+1,1}&\dots&\alpha_{r+1,r}&1&& \\ & & & & \ddots & \\ 0 & & & & & 1 \end{Vmatrix}.$$

Следовательно,

$$a_r'=(\overbrace{0,\dots,0}^{r},\ 1,\ 0,\dots,0),$$
$$b_r'=(\alpha_{r+1,1},\dots,\alpha_{r+1,r},\ 0,\ 0,\dots,0).$$

Отсюда вытекает, что $\sigma_r^{(t)}=0$.

Далее, в матрице $U_1^{(t)-1}U_2^{(t)-1}\dots U_r^{(t)-1}$ лишь первые $r$ столбцов отличны от столбцов единичной матрицы, а первые $r$ координат вектора $a_r^{(t)}$ равны нулю. Поэтому

$$U_1^{(t)-1}U_2^{(t)-1}\dots U_r^{(t)-1}\delta_r^{(t)}a_r^{(t)}=\delta_r^{(t)}a_r^{(t)}.$$

Окончательно находим

$$\frac{\|\varkappa_{n-1}^{(t)}\|_E^2}{2^{2t}}\leqslant n,\qquad \|\nu_{n-1}^{(t)}\|_E^2=0.$$

Интересно отметить, что если бы мы пытались оценить $M\|\varkappa_{n-1}^{(t)}\|_E^2$, то получили бы улучшение оценки не более чем в 6 раз.

5. На примере преобразований Гаусса покажем теперь, как оценить наиболее вероятное значение главного члена нормы эквивалентного возмущения. Имеем

$$\max_{z\subset G}\|\nu_{n-1}^{(t)}\|_E\leqslant 2^{-t}\sqrt{\frac{n^3}{12}}. \tag{12}$$

Пусть $\sigma_{ir}^{(t)}$ означает $i$-ю координату вектора $\sigma_r^{(t)}$. Заметим, что на каждом пучке $(r-1)$-го порядка ошибка $\sigma_{ir}^{(t)}$ будет не чем иным, как ошибкой от умножения двух чисел, из которых одно есть $i$-я координата вектора $a_r^{(t)}$, а второе — округленное значение скалярного произведения $(b_r^{(t)}, z_{r-1}^{(t)})$. Поэтому, используя результаты [5], находим

$$|M(\sigma_{ir}^{(t)}/z\subset G_{k_1k_2\dots k_{r-1}})|=O((x_{k_1\dots k_{r-1}}2^t)^{-1/2})2^{-t},$$
$$M(\sigma_{ir}^{(t)2}/z\subset G_{k_1k_2\dots k_{r-1}}^{(t)})=(1+O((x_{k_1\dots k_{r-1}}2^t)^{-1/2}))\frac{2^{-2t}}{12},$$

где $x_{k_1\dots k_{r-1}}$ — «длина» пучка $G_{k_1\dots k_{r-1}}^{(t)}$ в направлении вектора $b_r^{(t)}$.

Далее, на каждом пучке $(r-1)$-го порядка функция $\nu_{r-1}^{(t)}$ постоянна, следовательно,

$$M\|\sigma_r^{(t)}\|_E^2 = \sum_{i=r+1}^{n} M\sigma_{ir}^{(t)2} = \sum_{i=r+1}^{n} \sum_{G_{k_1\dots k_{r-1}}^{(t)}} P(z \subset G_{k_1\dots k_{r-1}}^{(t)}) M(\sigma_{ir}^{(t)2}/z \subset$$

$$\subset G_{k_1\dots k_{r-1}}^{(t)}) = \frac{(n-r)2^{-2t}}{12}(1+O(2^{-t/2})),$$

$$|M(\sigma_r^{(t)}, \nu_{r-1}^{(t)})| \leqslant \sum_{i=r+1}^{n} |M\sigma_{ir}^{(t)}\nu_{i,r-1}^{(t)}| =$$

$$= \sum_{i=r+1}^{n} \Big| \sum_{G_{k_1\dots k_{r-1}}^{(t)}} P(z \subset G_{k_1\dots k_{r-1}}^{(t)}) M[\sigma_{ir}^{(t)}\nu_{i,r-1}^{(t)}/z \subset G_{k_1\dots k_{r-1}}^{(t)}] =$$

$$= \sum_{i=r+1}^{n} \Big| \sum_{G_{k_1\dots k_{r-1}}^{(t)}} P(z \subset G_{k_1\dots k_{r-1}}^{(t)}) \nu_{i,r-1}(z \subset G_{k_1\dots k_{r-1}}^{(t)}) \times$$

$$\times M(\sigma_{ir}^{(t)}/z \subset G_{k_1\dots k_{r-1}}^{(t)})| \leqslant \sum_{i=r+1}^{n} \max|\nu_{i,r-1}^{(t)}| \sum_{G_{k_1\dots k_{r-1}}^{(t)}} P(z \subset G_{k_1\dots k_{r-1}}^{(t)}) \times$$

$$\times |M(\sigma_{ir}^{(t)}/z \subset G_{k_1\dots k_{r-1}}^{(t)})| \leqslant 2^{-2t}n^2 O(2^{-t/2}).$$

Здесь константы величин $O(2^{-t/2})$ зависят только от области $G$ и плотности распределения $P(z)$ и не зависят от $n$.

Мы не будем подробно останавливаться на закономерности последних выводов в этих соотношениях; отметим лишь, что она вытекает из существования математического ожидания величин $O((x2^t)^{-1/2})$, где $x$ — «длина» пучков. Полученные соотношения позволяют оценить $M\|\nu_{n-1}^{(t)}\|_E^2$:

$$M\|\nu_{n-1}^{(t)}\|_E^2 = M\|\nu_{n-1}^{(t)}\|_E^2 + M\|\sigma_{n-1}^{(t)}\|_E^2 + 2M(\sigma_{n-1}^{(t)}, \nu_{n-2}^{(t)}) \leqslant$$

$$\leqslant M\|\nu_{n-2}^{(t)}\|_E^2 + \frac{2^{-2t}}{12}(1+O(2^{-t/2})) + n^2 2^{-2t}O(2^{-t/2}) \leqslant \dots$$

$$\dots \leqslant \frac{2^{-2t}n^2}{24}(1+nO(2^{-t/2})).$$

Отсюда получаем оценки для математического ожидания и дисперсии величины $\|\nu_{n-1}^{(t)}\|_E$. Именно [3]:

$$M\|\nu_{n-1}^{(t)}\|_E \leqslant \sqrt{(M\|\nu_{n-1}^{(t)}\|_E^2)} \leqslant \frac{2^{-t}n}{\sqrt{24}}(1+nO(2^{-t/2})),$$

$$\sqrt{(D\|\nu_{n-1}^{(t)}\|_E)} \leqslant \sqrt{(M\|\nu_{n-1}^{(t)}\|_E^2)} \leqslant \frac{2^{-t}n}{\sqrt{24}}(1+nO(2^{-t/2})).$$

Окончательное решение поставленной задачи дает неравенство Чебышева [3]

$$P(\|v_{n-1}^{(t)}\|_E \leqslant y\frac{2^{-t}n}{\sqrt{24}}(1+nO(2^{-t/2}))) = 1 - P(\|v_{n-1}\|_E > $$

$$> y\frac{2^{-t}n}{\sqrt{24}}(1+nO(2^{-t/2}))) \geqslant 1 - P(|\,\|v_{n-1}^{(t)}\|_E - M\|v_{n-1}^{(t)}\|_E| > $$

$$> (y-1)\frac{2^{-t}n}{\sqrt{24}}(1+nO(2^{-t/2}))) \geqslant 1 - \frac{1}{(y-1)^2}.$$

В частности,

$$P(\|v_{n-1}^{(t)}\|_E \leqslant n2^{-t}) \geqslant 0.93.$$

Полученные результаты хорошо согласуются как с (11), так и с (12).

*Поступила в редакцию*
*9.12.1966*

**Цитированная литература**

1. J. H. Wilkinson. The algebraic eigenvalue problem. Oxford, Clarendon Press, 1965.
2. Дж. В. С. Касселс. Введение в теорию диофантовых приближений. М., Изд-во ин. лит., 1961.
3. Б. В. Гнеденко. Курс теории вероятностей. М., Гостехиздат, 1954.
4. В. В. Воеводин. Численные методы алгебры. М.— Л., «Наука», 1966.
5. Г. Ким, Д. М. Чибисов. Распределение ошибок округления при умножении двух чисел на вычислительной машине с фиксированной запятой. Матем. заметки, 1967, 1, № 2, 225—234.