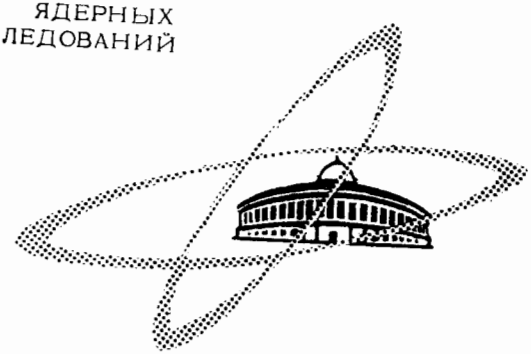
ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
Дубна



Р-1873

Я. Каутски, И. Фриш

ОКРУГЛЕНИЯ И НЕКОТОРЫЕ ПСЕВДООПЕРАЦИИ НА ВЫЧИСЛИТЕЛЬНЫХ МАШИНАХ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

1964

Р-1873

ОКРУГЛЕНИЯ И НЕКОТОРЫЕ ПСЕВДООПЕРАЦИИ НА ВЫЧИСЛИТЕЛЬНЫХ МАШИНАХ

В В Е Д Е Н И Е

С увеличением и усложнением вычислений на электронных вычислительных машинах происходит потеря точности окончательных результатов вследствие накопления ошибок округления, возникающих при замене точных чисел так называемыми машинными, то есть числами, которые в действительности вводятся в машину. До появления быстродействующих машин эта особенность больших цифровых процессов явно не обнаруживалась, поскольку вычисления были очень короткими.

Для обычных вычислительных машин такие ошибки вполне естественны, поскольку в этих машинах как данные, так и промежуточные числа выражаются лишь с определенной точностью. При этом, вообще говоря, используются вычислительные методы, обоснованные лишь в случае точных числовых значений.

При исследовании этих вопросов вводится понятие так называемой цифровой устойчивости заданного вычислительного метода (процесса или алгоритма); найдены, например, некоторые критерии для определенного типа цифровой устойчивости в случае разностных формул с постоянными коэффициентами для приближенного решения обыкновенных дифференциальных уравнений. Исследуются и другие более сложные проблемы.

Оказывается, что на цифровую устойчивость влияет не только вычислительный метод сам по себе, но и, например, начальные условия, а также способ, по которому выполняются отдельные операции (сложение, умножение, деление),и округление их результатов. Приведем короткий пример.

Пусть последовательность $\{y_n\}$ вычисляется по формуле

$$y_n = (y_{n-1} : n) \times n$$

при начальном условии y_0 (точный результат $y_n = y_0$) на вычислительной машине с плавающей запятой, причем мантисса содержит M двоичных разрядов. Если машина будет округлять при помощи так называемого "отрезания", т.е. мантиссой результата операции (умножения или деления) считается просто M старших цифр точного результата, то понятно, что y_n будет невозрастающей последовательностью, члены которой будут медленно уменьшаться. Однако если машина будет действительно округлять (точное определение приводится позже - суть дела в отрезании после прибавления половины младшего разряда), то можно ожидать, что под влиянием случайности ошибки округления должны исчезнуть (точнее, последо-

вательность y_n должна незначительно возрастать, потому что при указанном округлении более вероятна ошибка вверх). В действительности это не так. При $y_0 = 1$ последовательность y_n, медленно колеблясь, по существу окажется убывающей (для $M = 36$ имеем $y_{10^6} = 0,999\,999\,994$), но при $y_0^* = 1 + 2^{1-M}$ (это число, мантисса которого больше мантиссы числа 1 и ближе всех к ней) получается $y_n = y_0^*$ для всех n. Это явление можно объяснить только после подробного разбора машинных операций умножения и деления и их округления.

Анализом аналогичных операций для машины с фиксированной запятой занимались Нейман и Голдштайн [1] при исследовании цифровой устойчивости решения системы линейных алгебраических уравнений и, в частности, показали, что для нашей последовательности всегда $y_n = y_0$ на машине с фиксированной запятой.

В этой работе подобная задача рассмотрена тоже для плавающей запятой и при более общем способе округления. В случае машины с плавающей запятой исследования, правда, несколько трудоемки, но не представляют никаких математических трудностей. Поэтому мы не приводим доказательства наших утверждений. Они основаны на простых преобразованиях формул с целыми и дробными частями.

С помощью полученных результатов, которые имеют значение даже для конструкции вычислительных машин, мы в конце нашей работы объясним парадокс показанного примера.

I. Машинные числа и округление вещественных чисел

Мы будем интересоваться работой цифровых вычислительных машин, в которых числа имеют вид

$$a = \operatorname{sgn} a \cdot q^{p} \sum_{k=1}^{M} \eta_k q^{k-1},$$

где η_k принимают значения $0, 1, \dots, q-1$. Этот способ представления чисел в машине, который является классической q-адической системой, применяется почти во всех машинах, чаще всего с $q = 2$ или $q = 10$, но этот факт не является существенным для теоретических исследований.

Мы будем пользоваться следующими обозначениями:

$[a]$ – есть целая часть вещественного числа a;

$\{a\}$ – его дробная часть;

$\langle a_1, a_2, \dots, a_n \rangle$ – упорядоченная совокупность чисел;

$\lg a$ – обозначает $\log_q a$ ($q > 1$ – основание используемой системы

счисления).

Теперь можно ввести понятие машинного числа следующим образом.

Определение 1. Пусть q , M , P_1 , P_2 - целые числа, для которых имеют место неравенства

$$q > 1 , \quad M > 1 , \quad P_1 \le P_2 .$$

Пару целых чисел $s = \langle p, m \rangle$, где

$$P_1 \le p \le P_2 , \qquad 1 - q^M \le m \le q^M - 1$$

мы будем называть м а ш и н н ы м ч и с л о м (в дальнейшем только м.ч.; точнее надо было бы писать q , M , P_1 , P_2 - м.ч.). Вместе с тем p будем называть его п о р я д к о м и m его м а н т и с с о й. Всякому м.ч. $s = \langle p, m \rangle$ сопоставим вещественное число, его з н а ч е н и е

$$H(s) = q^p \cdot m$$

Теперь надо определить функцию $S(a)$, которая, наоборот, вещественному числу сопоставляет м.ч. $s = S(a)$. При этом естественно требовать, чтобы, например, сложная функция $H(S(a))$ на множестве значений м.ч. являлась тождественной и монотонной на множестве всех вещественных чисел (точнее, на каком-то отрезке, содержащем значения м.ч.). Желательно также оценить или даже минимизировать $|a - H(S(a))|$.

Эта задача не совсем проста, т.к. изображение H не является в общем случае взаимнооднозначным. Из этого следует, что функций S , удовлетворяющих приведенным условиям, существует больше. Самая важная из них связана с понятием так называемых почти нормализованных м.ч.

Определение 2. Скажем, что м.ч. $s = \langle p, m \rangle$ нормализовано (дальше н.м.ч.), если $|m| \ge q^{M-1}$; если s является либо н.м.ч., либо $p = P_1$, то s будем называть почти нормализованным м.ч. (дальше только п.н.м.ч.).

Для $P_1 \le p \le P_2$ и $0 \le \theta < 1$ обозначим

$$S_{p,\theta}(a) = s = \langle p, \operatorname{sgn} a \cdot [\,|a| q^{-p} + \theta\,] \rangle . \qquad (I)$$

В случае $\theta < 1$ функция $S_{p,\theta}(a)$ удовлетворяет (на отрезке, где $S_{p,\theta}(a)$ является м.ч.) обоим условиям, требуемым от функции S .

Следующая теорема дает условия, при которых $S_{p,\theta}(a)$ будет м.ч. или п.н.м.ч. и дает оценку погрешности.

Теорема 1. При заданном вещественном a обозначим

$$p_0(a) = p_0 = \max\left(P_1 , 1 + \left[\lg \frac{|a|}{q^M - \theta}\right]\right)$$

$$p_0(0) = P_1 .$$

Если теперь

$$|a| < q^{P_2}(q^M - \theta), \tag{2}$$

то $p_0 \geq P_2$ и для каждого целого p, для которого

$$p_0 \leq p \leq P_2 \tag{3}$$

$S_{p,\theta}(a)$ является м.ч. Среди чисел $S_{p,\theta}(a)$ только $S_{p_0,\theta}(a)$ является п.н.м.ч. и если, кроме того, $|a| \geq q^{P_1}(q^M - \theta)$, то $S_{p_0,\theta}(a)$ является н.м.ч. Дальше для каждого p из отрезка (3) имеет место

$$|H(S_{p,\theta}(a))| - \theta q^p \leq |a| < |H(S_{p,\theta}(a))| + (1-\theta)q^p$$

и

$$|H(S_{p,\theta}(a)) - a| \geq q^p \max(\theta, 1-\theta). \tag{4}$$

Наконец,

$$|H(S_{p,0}(a))| \leq |a| < |H(S_{p,1}(a))|.$$

Из (4) следует, что наилучшая точность при замене вещественного числа машинным числом достигается при выборе p наименьшим (т.е. $p = p_0(a)$); из этого вытекает значение п.н.м.ч.) и $\theta = \frac{1}{2}$. Если, конечно, q нечетное число, то может оказаться практически невозможным взять $\theta = \frac{1}{2}$ и надо выбрать в качестве θ другое число, близкое $\frac{1}{2}$, но с конечным q-адическим рядом. Из-за этой причины желательно заниматься не только случаем $\theta = \frac{1}{2}$, но более общим $0 \leq \theta \leq 1$. Интересен также случай $\theta = 0$ – так называемое вычисление без округления или "отрезание".

Определение 3. Обозначим $k = [\log_q(|a| + \theta q^p)] + 1 - p$. Если $k > 0$, то k дает число значащих q-адических цифр вещественного числа a. Точнее, м.ч. s, определенное в (I), называется θ-округлением числа a на k значащих цифр.

По причинам, указанным выше, определим теперь

$$\tilde{S}(a) = \tilde{S}_\theta(a) = S_{p_0(a),\theta}(a)$$

и еще

$S(a) = S_\theta(a) = \tilde{S}(a)$, если $P_1 = P_2$ или $P_1 < P_2$ и $\tilde{S}(a)$ н.м.ч.,

$S(a) = \langle P_1, 0 \rangle$, если $P_1 < P_2$ и $\tilde{S}(a)$ является ненормализованным п.н.м.ч.

Будем различать два случая. Во-первых, $P_1 = P_2$ – т.наз. фиксированная запятая. Здесь обозначим $P_1 = P_2 = P$. Тогда (для чисел a, удовлетворяющих (I)) также $p_0 = P$, т.е. каждое машинное число является п.н.м.ч. Если мы хотим разумным способом (см.дальше) определить операции умножения и деления, то числа M, P должны удовлетворять неравенству

$$-2M < P < -M.$$

Чаще всего $P = -M$, $P = -M+1$, $P = -M+2$.

Второй случай. $P_1 < P_2$ называется плавающей запятой. Здесь естественно требовать, чтобы число $a = 1$ имело нормализованный машинный "образ", т.е. чтобы $S(1)$ было н.м.ч. Из этого вытекает условие

$$P_1 \leqq 1 - M \leqq P_2 .$$

Здесь чаще всего

$$M \cong 1 - \frac{P_1 + P_2}{2} ,$$

потому что при этом условии почти всегда $S\left(\frac{1}{a}\right)$ - н.м.ч., если $S(a)$ н.м.ч.

Объясним, почему мы в случае $P_1 < P_2$ различаем функции $S(a)$ и $\tilde{S}(a)$. У машин с плавающей запятой ненормализованных п.н.м.ч. настолько мало, что как с точки зрения громоздкости теоретических исследований, так именно и с точки зрения устройства машин, удобно их просто отбрасывать и вместо них пользоваться "наименьшим" числом $\langle P_1, 0 \rangle$, которое называется нормализованным нулем.

2. Псевдооперации

Если задана функция или операция над вещественными числами, то функции H и S, введенные в предыдущем параграфе, делают возможным естественным образом определить соответствующую функцию или операцию - так называемую псевдооперацию над м.ч.

Определение 4. Пусть $f(a_1, a_2, \ldots, a_n)$ есть вещественная функция n $(n \geq 1)$ переменных, определенная на каком-нибудь множестве D. Пусть $s_1, s_2, \ldots, s_n$ такие м.ч., что $\varphi = \langle H(s_1), H(s_2), \ldots, H(s_n) \rangle \in D$ и $a = f(\varphi)$ удовлетворяет неравенству (2). Тогда соответствующей $(\theta-)$ п с е в д о о п е р а ц и е й называется функция

$$\tilde{f}(s_1, s_2, \ldots, s_n) = S(f(\varphi)) \quad \left(= S_\theta(f(\varphi)) \right) .$$

Итак, мы желаем, чтобы результату операции соответствовало м.ч. с порядком наивыгоднейшим для точности аппроксимации (т.е. п.н.м.ч.). Это определение является немного компромиссным с теоретической точки зрения; казалось бы, лучше определить $\tilde{f}(s_1, \ldots, s_n) = \tilde{S}(f(\varphi))$ (см. конец предыдущего параграфа). Наоборот, (в случае $P_1 < P_2$) существующие машины выполняют некоторые псевдооперации по нашему определению только в случае, если и $s_1, s_2, \ldots, s_n$ - н.м.ч. Кроме того, бывают еще в машинах так называемые псевдооперации без нормализации, которые результату $f(\varphi)$ прибавляют м.ч. $S_{p,\theta}(f(\varphi))$, где p определяется по другому принципу, чем точность аппроксимации.

3. Операции произведения и деления

Рассмотрим подробнее псевдооперацию, определенную при помощи операций "умножения" и "деления" вещественных чисел. Для простоты мы ограничимся при этом только положительными числами. Для отрицательных чисел имеют место похожие результаты; число 0 ведет себя по-особому, но не представляет никаких трудностей.

Пусть $s_i = \langle p_i, m_i \rangle$ два машинные числа $(i = 1, 2)$ и пусть $s_3 = \langle p_3, m_3 \rangle$ является их произведением (соотв. частным), если оно существует. Для псевдоопераций произведения (будем обозначать $\times$) и деления (будем обозначать $:$) имеют место следующие формулы:

1) Произведение с фиксированной запятой.

Если $m_1 m_2 < q^{-P}(q^M - \Theta)$, то $s_1 \times s_2 = s_3 = \langle P, [m_1 m_2 q^P + \Theta] \rangle$.

2) Произведение с плавающей запятой.

Обозначим

$\gamma = 0$ если $m_1 m_2 < q^{M-1}(q^M - \Theta)$ и $\gamma = 1$ в противном случае.

Пусть $P_1 \le p_3 = p_1 + p_2 + \gamma + M - 1 \le P_2$.

Тогда $s_3 = \langle p_3, [m_1 m_2 q^{1-M-\gamma} + \Theta] \rangle$ является н.м.ч. и имеет место $s_1 \times s_2 = s_3$.

Если $p_3 < P_1$, то очевидно $s_3 = \langle P_1, 0 \rangle$.

3) Деление с фиксированной запятой.

Если $\frac{m_1}{m_2} < q^P(q^M - \Theta)$, то $s_1 : s_2 = s_3 = \langle P, [\frac{m_1}{m_2} q^{-P} + \Theta] \rangle$.

4) Деление с плавающей запятой.

Обозначим

$\beta = 0$, если $m_1 \ge m_2$ и $\beta = 1$, если $m_1 < m_2$.

Если

$$P_1 \le p_3 = p_1 - p_2 + 1 - M - \beta \le P_2$$

то

$s_3 = \langle p_3, [\frac{m_1}{m_2} q^{M-1-\beta} + \Theta] \rangle$ является н.м.ч. и имеет место

$s_1 : s_2 = s_3$.

В случае $p_3 < P_1$ опять $s_3 = \langle P_1, 0 \rangle$, а в случае $p_3 > P_2$ псевдооперация деления не является определенной.

Приведенные утверждения, именно для произведения и деления с плавающей запятой, имеют значение и для конструирования вычислительных машин. Очевидным способом выполнения псевдо-

операции является такой: прежде всего выполнить операцию над заданными м.ч. с достаточным количеством q -адических разрядов. Так как нормализованным м.ч. считается то число, первая цифра мантиссы которого не нулевая, мы должны полученный результат нормализовать, т.е. найти первую значащую цифру (и соответственно изменить порядок числа). Потом следует округление, которое вообще требует новой нормализации. Но именно нормализация отнимает очень много машинного времени, выгоднее найти такие свойства псевдоопераций, которые позволяют нам пользоваться только одной нормализацией. Это вполне возможно сделать, например, у деления (как следует из 4), потому что число β , которым по существу определен порядок частного (т.е. нормализация), не зависит от θ . В случае умножения утверждение 2 позволяет сделать упрощение, состоящее в том, что мы вычислим с достаточным количеством разрядов число

$$a = m_1 m_2 q^{1-M} + \theta$$

потом нормализуем, т.е. определим $\gamma = 0\ (1)$, если $a < q^M$ $(a \geq q^M)$, и тогда $m_3 = a\, q^{-r} + \theta\, \gamma \frac{q-1}{q}$.

Значит процесс, состоящий по существу из двух нормализаций и одного округления, переведен в процесс, часто более простой, состоящий из двух округлений и одной нормализации.

Однако рассмотрение отдельных операций не имеет такого значения, как изучение свойств целых цифровых вычислений, т.е. последовательностей этих операций.

В следующем параграфе мы изучим один простой пример одного вычисления такого процесса при помощи приведенных соображений.

4. Пример

Уже в введении мы упомянули о вычислении рекуррентной последовательности $y_n = \frac{y_{n-1}}{n} \cdot n$ или, если мы вычисляем ее на машине, то

$$\tilde{y}_n = (\tilde{y}_{n-1} : n) \times n \,. \qquad (5)$$

В качестве примера мы подробно рассмотрим результат постепенного выполнения псевдооперации деления и произведения, откуда получится обоснование ранее описанных свойств последовательности (5).

Обозначим $(a : b) \times b = a^*$. Покажем, что не всегда $a^* = a$. Пусть

$a = \langle p_1, m_1 \rangle$,

$b = \langle p_2, m_2 \rangle$,

$a^* = \langle p_3, m_3 \rangle$.

*) Фиксированная запятая.

Пусть

$$m_1 < m_2 q^P (q^M - \theta), \qquad (6)$$

$$m_1 < m_2 q^P \left(\left\{ \frac{m_1}{m_2} q^{-P} + \theta \right\} - \theta \right) + q^M - \theta . \qquad (7)$$

Тогда $a^X = (a : b) \times b$ является м.ч. с мантиссой $m_3 = m_1 + \varepsilon$, где для "ошибки" ε имеет место формула

$1 + [(\theta-1)(q^{P+M} - q^P + 1)] \le \varepsilon = [m_2 q^P (\theta - \{\frac{m_1}{m_2} q^{-P} + \theta\}) + \theta] \le [\theta(q^{P+M} - q^P + 1)]$.

Если считать, что $P \le -M$, как часто бывает, то для

$$1 - (q^{P+M} - q^P + 1)^{-1} \le \theta < (q^{P+M} - q^P + 1)^{-1} \qquad (8)$$

будет $\varepsilon = 0$. Это неравенство имеет место всегда, когда $\theta = \frac{1}{2}$.

Замечание. Условие (6) гарантирует существование машинного частного $a : b$. Если $P \le -M$ и для θ имеет место (8), потом $a^X = (a : b) \times b = a$ и условие (7) в этом случае тривиально выполнено. Однако если $P > M$, может случиться, что $a : b$ существует, но $(a : b) \times b$ является неопределенным.

Итак, возьмем $q = 2$, $M = 4$, $P = -3$, $m_1 = 15$, $m_2 = 14$, $\theta \ge \frac{3}{7}$, $a : b = <-3, 9>$ и $m_3 = 16$. Но так как $16 > q^{M-1} = 15$, то $<-3, 16>$ не является машинным числом.

2) Плавающая запятая.

Обозначим, как в параграфе 3, $\beta = 0$ или 1 в зависимости от того, будет ли $m_1 \ge m_2$ или $m_1 < m_2$, и $\delta = 0$ для $m_1 q^\beta + m_2 q^{1-M} (\theta - \{\frac{m_1}{m_2} q^{M-1+\beta} + \theta\}) < q^M - \theta$, в противоположном случае будет $\delta = 1$.

Пусть

$$P_1 \le p_3 - p_2 + 1 - M - \beta \le P_2 ,$$

$$P_1 \le p_3 - p_1 + \delta - \beta \le P_2 ,$$

что как раз является условиями для существования машинных выражений $a : b$ соотв. $(a : b) \times b$.

Потом

$a^X = <p_3, m_3>$, причем

$$m_3 = [m_1 q^{\beta-\delta} + m_2 q^{1-M-\delta} (\theta - \{\frac{m_1}{m_2} q^{M-1+\beta} + \theta\}) + \theta] .$$

Рассмотрение в этом случае является немного более сложным, т.к. для m_1, близкого максимальной или минимальной мантиссам, порядок p_3 может быть другой, чем порядок p_1. Следующие 6 лемм описывают различные возможности.

Лемма 1. Если $m_1 = m_2$ или $m_2 = q^{M-1}$, то $a^X = a$.

Лемма 2. Если $m_1 > q^{M-1}$ или, что то же самое (поскольку мы рассматриваем только нормализованные числа), если $m_1 \ne q^{M-1}$, то всегда $p_3 \le p_1$, точнее $p_3 = p_1$ или

$\rho_3 = \rho_1 + 1$.

Лемма 3. Если $m_1 = q^{M-1}$, то $\rho_3 = \rho_1 - 1$ имеет место тогда и только тогда, когда

$$\left\{ \frac{q^{2M-1}}{m_2} + \theta \right\} > \theta \left(1 + \frac{q^{M-1}}{m_2} \right) \qquad (9)$$

или, что то же самое, тогда и только тогда, когда

$$1 - \theta > \left\{ \frac{q^{2M-1}}{m_2} \right\} > \theta \frac{q^{M-1}}{m_2} .$$

В случае, когда $\rho_3 = \rho_1 - 1$ имеет место, то $m_3 = q^M - \varepsilon$, где

$$1 \leq \varepsilon = - \left[\theta - m_2 q^{M-1} \left\{ \frac{q^{2M-1}}{m_2} \right\} \right] < q - \theta(q+1) .$$

Лемма 4. Случай $m_1 < m_2$. Если, кроме того, $m_1 > q^{M-1}$, или хотя $m_1 = q^{M-1}$, но (9) не имеет места, то $\rho_3 = \rho_1$ и возможны три случая:

$$m_3 = m_1 \ , \ m_1 + 1 \ , \ m_1 - 1 .$$

При этом второй (соотв.третий) случай происходит тогда и только тогда, если

$$m_2 \left(\theta - \left\{ \frac{m_1}{m_2} q^M + \theta \right\} \right) q^{-M} + \theta \geq 1 \qquad (\text{соотв. } < 0).$$

Из этого далее вытекает, что последние два случая невозможны, если

$$\theta < \frac{1}{2 - q^{-M}} \ , \qquad \text{соотв.} \quad \theta \geq \frac{1 - q^{-M}}{2 - q^{-M}} . \qquad (10)$$

В частном случае $\theta = \frac{1}{2}$ всегда $m_3 = m_1$. Очевидно, что если $m_1 = q^{M-1}$ и (9) не имеет места, то третий случай невозможен.

Лемма 5. Пусть $m_1 > m_2$. Тогда существуют две возможности.

а) $\rho_3 = \rho_1 + 1$

и это имеет место тогда и только тогда, когда

$$1 - \theta \leq \left\{ \frac{m_1}{m_2} q^{M-1} \right\} \leq 1 - \frac{q^{M-1}}{m_2} (q^M - \theta - m_1) \qquad (11)$$

и тогда $m_3 = q^{M-1}$. Этот случай возможен только тогда, если

$$m_1 \geq q^M - \theta \frac{1+q}{1 + \theta q^{1-M}} .$$

б) $\rho_3 = \rho_1$ тогда и только тогда, когда (11) не имеет места. Если мы определим $\alpha = 0$ или 1 в зависимости от того, имеет ли место неравенство

$$\left\{ \frac{m_1}{m_2} q^{M-1} \right\} < 1 - \theta ,$$

то $m_3 = m_1 + i$, где

$$\max \left(q^{M-1} - m_1 - 1, (\theta+1)(q+1-2q^{1-M}) \right) < i = \left[\theta + m_2 q^{1-M} (\alpha - \left\{ \frac{m_1}{m_2} q^{M-1} \right\}) \right] \leq \min \left(q^M - m_1 - 1, \theta(q+1-2q^{1-M}) \right).$$

Лемма 6. Во всех случаях имеет место формула

$$- \theta(q+1) q^{\rho_1} < H(a) - H(a^*) < (1-\theta)(q+1) q^{\rho_1} .$$

Обратим теперь внимание на пример из введения. Последовательность y_n мы можем еще обобщить таким образом. Пусть дана последовательность чисел b_n и начальное условие

y_0 ; определим $y_n = \frac{y_{n-1}}{b_n} \cdot b_n$

(в нашем примере мы положили $b_n = n$). Машинное вычисление членов этой последовательности является в таком случае последовательным выполнением операции $(a \cdot b) \times b$ для различных a, b . Если выполняется для θ неравенство (8) (например, $\theta = \frac{1}{2}$), то при фиксированной запятой и $P \leq -M$ тождественно $\tilde{y}_n = \tilde{y}_0$. Если θ удовлетворяет неравенствам (10) (опять, например, $\theta = \frac{1}{2}$), то при плавающей запятой (и при выполнении соответствующих предположений о выборе P_1, P_2, M и y_0 и b_n) имеет место следующее:

а) Если $\tilde{y}_{n-1} = \langle p_1, q^{M-1} \rangle$, то $y_n \leq y_{n-1}$ (см. леммы 3 и 4).

б) Пусть $\tilde{b}_n = \langle p_2, m_2 \rangle$, $\tilde{y}_{n-1} = \langle p_1, m_1 \rangle$, $m_1 > q^{M-1}$

Тогда для $m_2 \geq m_1$ имеет место $y_n = y_{n-1}$ (см. леммы 1,4), т.е. "вероятность" ошибки вверх и вниз приблизительно одинакова. Этим разъяснено явление, приведенное в введении. При $\tilde{y}_0 = \langle p_1, q^{M-1} + 1 \rangle$ ошибка могла бы возникнуть из-за округления только для $m_2 < m_1$, но это возможно только в случае $m_2 = q^{M-1}$, который опять-таки противоречит лемме 1. Итак, $\tilde{y}_n = \tilde{y}_0$. При всех прочих y_0 вычисление y_n ,с точки зрения возникающих ошибок, является некоторым"случайным блужданием" с приблизительно одинаковой вероятностью вверх и вниз. При этом вероятность того, что число вообще изменится, является тем большей, чем больше число m_1 . (Вполне это справедливо при случайном m_2). Единственным исключением является $m_1 = q^{M-1}$. Здесь возможна только ошибка вниз, т.е. величина $y_n = q^k$ является барьером для этого случайного блуждания. Поэтому, если мы, например, рассматриваем $y_{100}, y_{200}, \ldots$, то кажется, что последовательность уменьшается. В заключение заметим еще, что несмотря на то, что из наших соображений вытекает, что в машинах с фиксированной запятой (где имеет место (8) и $P \leq -M$) ошибка не появляется, все-таки существуют машины (Урал-1, Киев), у которых последовательность $y_1, y_2, \ldots$ уменьшается. Однако эти машины не выполняют псевдооперации "наилучшим способом" - по определению в параграфе 2: из-за конструктивных упрощений они не работают с нужным количеством разрядов.

Рукопись поступила в издательский отдел
3 ноября 1964 г.

ЛИТЕРАТУРА

[1] J. Neumann, H. Goldstine : Numerical inverting of matrices of high order, Bull. Am. Math. Soc. 53 (1947), 1021 – 1099.