УДК [004.3 + 681.3]: 619.67

НАКОПЛЕНИЕ ПОГРЕШНОСТЕЙ ПРИ АППАРАТУРНОЙ РЕАЛИЗАЦИИ АЛГОРИТМА НОРМИРОВКИ

Бабенко В. Н.

ACCUMULATION OF ERRORS AT HARDWARE REALIZATION OF ALGORITHM OF NORMALIZATION

Babenko V. N.*

** Kuban State University, Krasnodar, Russia, e-mail: rnibvd@mail.ru*

Abstract. Development of algorithms, designing of devices for their realization are the important factor of increase of productivity of computing systems. Earlier the algorithm of inversion of a divider and application of algorithm for performance of normalization of a vector by the author have been submitted. Then the description of devices of normalization of a vector has been given. At performance of machine arithmetic operations because of limitation of a digit grid inevitably there are error of a founding. Process of accumulation of errors of algorithm of normalization of a vector at his realization on devices of normalization in this clause with the help of methods of the numerical analysis is For restraint of growth of errors of calculations in these devices on adders for a mantissa to m the basic bits is allocated in addition q younger bits. As a result of researches attitudes between numbers m and q, responding standard requirements of accuracy of the calculated result are established.

Keywords: inversion, machine number, algorithm, convergence, regularity, efficiency, unit vector normalization, computation error.

Введение

Одной из основных векторных операций в вычислительной математике является нормировка вектора , определяемая формулой

$$\mathbf{y} = \mathbf{x}/a^{\alpha}, \text{ где } \alpha = 1/n, \ n = 1, 2.$$

В [1] был предложен алгоритм инверсии делителя a^{α}, $n = 1, 2, \ldots$,

$$a_0 = a,$$

$$a_i = a_{i-1}(1 + \theta_{i-1}2^{t_{i-1}} f(u_{i-1}, z_{i-1}))^n,$$

где

$$\theta_{i-1} = \begin{cases} 1, \text{ если } a_{i-1} < 1, \\ 0, \text{ если } a_{i-1} = 1, \\ -1, \text{ если } a_{i-1} > 1, \end{cases}$$

$$\theta_{i-1}2^{t_{i-1}}u_{i-1} = \frac{1 - \sqrt[n]{u_{i-1}}}{\sqrt[n]{u_{i-1}}},$$

$$2^{-1} \leqslant u_{i-1} < 1,$$

$$f(u_{i-1}, z_{i-1}) = \begin{cases} 2^{-1}, \text{ если } u_{i-1} < z_{i-1}, \\ 1, \text{ если } z_{i-1} \leqslant u_{i-1}, \end{cases}$$

$$z_{i-1} = \frac{3 + \theta_{i-1}2^{t_{i-1}+1}}{4 + 3\theta_{i-1}2^{t_{i-1}}}, \quad i = 1, 2, \ldots,$$

предназначенный прежде всего для выполнения операции нормировки вектора.

В [2] была доказана оценка скорости сходимости алгоритма обращения. Пусть последовательность $\{c_i\}$, $i = 1, 2, \ldots$ определена следующим образом:

$$c_i = \prod_{j=1}^{i} (1 + \theta_{j-1}2^{-t_{j-1}} f(u_{i-1}, z_{i-1})).$$

Тогда последовательность $\{c_i\}$, $i = 1, 2, \ldots$ сходится к $a^{-\alpha}$, причем справедлива оценка

$$\left|c_i - a^{-\alpha}\right| < 2^{-2i}a^{-\alpha}. \qquad (1)$$

Изложенный алгоритм обладает регулярностью, что создает отличные предпосылки для его отображения на аппаратуру — допускает конвейеризацию вычислительного процесса.

Вторым достоинством алгоритма является его экономичность. Пусть 2^{-m} — допустимая (приемлемая) погрешность инверсии числа a, где m — четное положительное число, тогда, положив $i = m/2$ и подставив его в (1), получим оценку

$$\left|c_{m/2} - a^{-\alpha}\right| < 2^{-m}a^{-\alpha}. \qquad (2)$$

Последнее свойство обуславливает малую длину конвейера (число каскадов равно

Бабенко Виктор Николаевич, канд. физ.-мат. наук, доцент кафедры математических и компьютерных методов Кубанского государственного университета; e-mail: rnibvd@mail.ru.

$m/2$), что в свою очередь обеспечивает малую задержку сигнала ($m/2$ тактов) и, соответственно, его быструю заполняемость (время заполнения конвейера равно $m/2$ тактов, продолжительность такта равна времени выполнения операции сложения) [1, 2].

Здесь следует сказать, что число a в ЭВМ в формате с плавающей точкой (запятой) представляется в виде $a = \sigma\gamma^{k_a} m_a$, где σ — код знака числа a, ($\sigma \in \{0, 1\}$), γ — основание системы счисления (мы будем рассматривать $\gamma = 2$), k_a — порядок числа a, m_a — его мантисса, причем m_a удовлетворяет одному из неравенств $\gamma^{-1} \leqslant m_a < 1$ или же $1 \leqslant m_a < 2$.

При выполнении на арифметическом процессоре ЭВМ операции $*$ над числами a и b, представленными в формате с плавающей точкой, если ее результат не обратился в ноль, вместо $a * b$ получим машинный результат $(a * b)_{\text{м}}$, удовлетворяющий неравенству [3]

$$|(a * b)_{\text{м}} - a * b| < \varepsilon_1 |a * b|, \qquad (3)$$

где $\varepsilon_1 = 2^{-m+1}$, m — число разрядов, отводимых под мантиссу.

На устройства, реализующие нормировку вектора при $n = 1, 2$ получены патенты [4, 5]. В этих устройствах для обеспечения приемлемой точности вычисленного результата дополнительно к m основным разрядам, используемым в машинном представлении мантиссы числа в формате с плавающей точкой, были использованы q младших разрядов.

В этой работе ставится задача исследования накопления погрешностей при осуществлении вычислений на устройстве нормировки вектора и выявлению связи между величинами q и m.

1. Исследование процессов накопления погрешностей

Как отмечалось выше, при аппаратурном осуществлении вычислений из-за ограниченности разрядной сетки неизбежно возникают ошибки округления. Эти ошибки часто относят к эквивалентному возмущению исходной величины. Пусть, например, a — значение исходной величины и $\tilde{a}$ — ее возмущенное значение. Их связь можно описать соотношением

$$\tilde{a} = a(1 + \beta).$$

Очевидно,

$$\tilde{a}^{-\alpha} = a^{-\alpha}(1 + \beta)^{-\alpha}. \qquad (1.1)$$

Предполагая выполненным неравенство $|\beta| < 1$, отметим, что

$$(1 + \beta)^{-\alpha} = 1 + \sum_{i=1}^{\infty} \frac{(-1)^i}{i!} \prod_{j=1}^{i} (\alpha + (j - 1))\beta^i.$$

Если пренебречь малыми высших порядков, то из представленного ряда вытекает неравенство

$$\left|(1 + \beta)^{-\alpha} - 1\right| < \alpha |\beta|. \qquad (1.2)$$

Оценим теперь близость величин $\tilde{a}^{-\alpha}$ и $a^{-\alpha}$. Для этого, используя (1.1), запишем цепочку тождественных преобразований

$$\tilde{a}^{-\alpha} - a^{-\alpha} = a^{-\alpha}(1 + \beta)^{-\alpha} - a^{-\alpha} = \\ = a^{-\alpha}\left((1 + \beta)^{-\alpha} - 1\right).$$

Учитывая (1.2), отсюда следует искомая оценка близости

$$\left|\tilde{a}^{-\alpha} - a^{-\alpha}\right| < \alpha |\beta| \left|a^{-\alpha}\right|. \qquad (1.3)$$

При осуществлении вычислений в устройстве нормировки по предписаниям алгоритма

$$a_0 = a, y_0 = y,$$

$$a_i = \begin{cases} a_{i-1} + \theta_{i-1} 2^{-k_{i-1}} a_{i-1}, \text{ при } n = 1, \\ a_{i-1} + \theta_{i-1} 2^{-k_{i-1}+1} a_{i-1} + \\ \qquad + 2^{-2k_{i-1}} a_{i-1}, \text{ при } n = 2, \end{cases}$$

$$y_i = y_{i-1} + \theta_{i-1} 2^{-k_{i-1}} y_{i-1}, \qquad (1.4)$$

где

$$\theta_{i-1} = \begin{cases} 1, \text{ если } a_{i-1} < 1, \\ 0, \text{ если } a_{i-1} = 1, \\ -1, \text{ если } a_{i-1} > 1, \end{cases}$$

$$2^{-k_{i-1}} = 2^{t_{i-1}} f(u_{i-1}, z_{i-1}),$$

$$\theta_{i-1} 2^{t_{i-1}} u_{i-1} = \frac{1 - \sqrt[n]{a_{i-1}}}{\sqrt[n]{a_{i-1}}},$$

$$2^{-1} \leqslant u_{i-1} < 1,$$

$$f(u_{i-1}, z_{i-1}) = \begin{cases} 2^{-1}, \text{ если } u_{i-1} < z_{i-1}, \\ 1, \text{ если } z_{i-1} \leqslant u_{i-1}, \end{cases}$$

$$z_{i-1} = \frac{3 + \theta_{i-1} 2^{t_{i-1}+1}}{4 + 3\theta_{i-1} 2^{t_{i-1}}}, \quad i = 1, 2, \ldots,$$

вследствие погрешностей вычислений вместо последовательностей $\{a_i\}$ и $\{y_i\}$ мы получим иные последовательности $\{\tilde{a}_i\}$ и $\{\tilde{y}_i\}$ соответственно.

Определим источники погрешностей. Для этого обратимся к формулам (1.4). Рассмотрим сначала случай $n = 1$. Очевидно, при вычислении величины $\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1}$ $(\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{y}_{i-1})$, осуществляемом с помощью сдвига числа $\tilde{\theta}_{i-1}\tilde{a}_{i-1}$ $(\tilde{\theta}_{i-1}\tilde{y}_{i-1})$ на $\tilde{k}_{i-1}$ разрядов вправо, в вычисленную величину вносится погрешность α_{i-1} (δ_{i-1}), удовлетворяющая неравенству

$$|\alpha_{i-1}| < 2^{-(m+q)+1} \ (|\delta_{i-1}| < 2^{-(m+q)+1}).$$

Поэтому вместо

$$\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} \ (\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{y}_{i-1})$$

получим

$$\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} + \alpha_{i-1} \ (\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} + \delta_{i-1}).$$

При выполнении сложения $\tilde{a}_{i-1}(\tilde{y}_{i-1})$ с

$$\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} + \alpha_{i-1} \ (\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} + \delta_{i-1})$$

никаких погрешностей не вносится, откуда можно заключить, что

$$\tilde{a}_i = \tilde{a}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} + \alpha_{i-1}$$
$$(\tilde{y}_i = \tilde{y}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{y}_{i-1} + \delta_{i-1}).$$

Рассмотрим теперь второй случай $(n = 2)$. При вычислении величины $\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}+1}\tilde{a}_{i-1}$, осуществляемом с помощью сдвига числа $\tilde{\theta}_{i-1}\tilde{a}_{i-1}$ на $\tilde{k}_{i-1}-1$ разрядов вправо, в вычисленную величину вносится погрешность α_{i-1}, удовлетворяющая неравенству $|\alpha_{i-1}| < 2^{-(m+q)+1}$. Аналогично при вычислении величины $2^{-2\tilde{k}_{i-1}}\tilde{a}_{i-1}$, производимом также с помощью сдвига числа $\tilde{a}_{i-1}$ на $2\tilde{k}_{i-1}$ разрядов вправо, в вычисленную величину вносится погрешность β_{i-1}, удовлетворяющая неравенству $|\beta_{i-1}| < 2^{-(m+q)+1}$. Поэтому вместо $\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}+1}\tilde{a}_{i-1}$ получим $\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}+1}\tilde{a}_{i-1} + \alpha_{i-1}$, а вместо $2^{-2\tilde{k}_{i-1}}\tilde{a}_{i-1}$ получим $2^{-2\tilde{k}_{i-1}}\tilde{a}_{i-1} + \beta_{i-1}$.

При выполнении суммирования никаких погрешностей не вносится. Отсюда заключаем, что

$$\tilde{a}_i = \tilde{a}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}+1}\tilde{a}_{i-1} +$$
$$+ 2^{-2\tilde{k}_{i-1}}\tilde{a}_{i-1} + \alpha_{i-1} + \beta_{i-1}.$$

Заметим, что при $i > (m+q)/4$ значение величины $2^{-2\tilde{k}_{i-1}}\tilde{a}_{i-1}$ полностью выходит вправо за пределы разрядной сетки, поэтому вычисления осуществляются по формуле

$$a_i = a_{i-1} + \theta_{i-1}2^{-k_{i-1}+1}a_{i-1}.$$

Соответственно, формула, моделирующая погрешности вычислений, примет вид

$$\tilde{a}_i = \tilde{a}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}+1}\tilde{a}_{i-1} + \alpha_{i-1}.$$

Таким образом, следуя описанному выше алгоритму, из-за погрешностей вычислений на самом деле в вычислениях используются возмущенные величины, при этом связь между элементами последовательностей $\{\tilde{a}_i\}$ и $\{\tilde{y}_i\}$ описывается следующим образом:

$$\tilde{a}_0 = a, \tilde{y}_0 = y,$$

$$\tilde{a}_i = \begin{cases} \tilde{a}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} + \alpha_{i-1}, \\ \qquad\qquad \text{при } n = 1, \\ \begin{cases} \tilde{a}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}+1}a_{i-1} + \\ \quad + 2^{-\tilde{k}_{i-1}}\tilde{a}_{i-1} + \alpha_{i-1} + \beta_{i-1}, \\ \quad i \leqslant (m+q)/4, \\ \tilde{a}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}+1}\tilde{a}_{i-1} + \alpha_{i-1}, \\ \quad i > (m+q)/4, \end{cases} \\ \qquad\qquad \text{при } n = 2, \end{cases}$$

$$\tilde{y}_i = \tilde{y}_{i-1} + \tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\tilde{y}_{i-1} + \delta_{i-1},$$

где

$$\tilde{\theta}_{i-1} = \begin{cases} 1, \text{ если } \tilde{a}_{i-1} < 1, \\ 0, \text{ если } \tilde{a}_{i-1} = 1, \\ -1, \text{ если } \tilde{a}_{i-1} > 1, \end{cases}$$

$$2^{-\tilde{k}_{i-1}} = 2^{\tilde{t}_{i-1}}f(\tilde{u}_{i-1}, \tilde{z}_{i-1}),$$

$$\tilde{\theta}_{i-1}2^{\tilde{t}_{i-1}}\tilde{u}_{i-1} = \frac{1 - \sqrt[n]{\tilde{a}_{i-1}}}{\sqrt[n]{\tilde{a}_{i-1}}},$$

$$2^{-1} \leqslant \tilde{u}_{i-1} < 1,$$

$$f(\tilde{u}_{i-1}, \tilde{z}_{i-1}) = \begin{cases} 2^{-1}, \text{ если } \tilde{u}_{i-1} < \tilde{z}_{i-1}, \\ 1, \text{ если } \tilde{z}_{i-1} \leqslant \tilde{u}_{i-1}, \end{cases}$$

$$\tilde{z}_{i-1} = \frac{3 + \tilde{\theta}_{i-1}2^{\tilde{t}_{i-1}+1}}{4 + 3\tilde{\theta}_{i-1}2^{\tilde{t}_{i-1}}}, \quad i = 1, m/2. \quad (1.5)$$

Таким образом, может быть сформулирована теорема.

Теорема. Пусть величины y и a удовлетворяют неравенствам

$$2^{-1} \leqslant y < 1,$$

$$\begin{cases} 2^{-1} \leqslant a < 1, \text{ при } n = 1, \\ 2^{-2} \leqslant a < 1, \text{ при } n = 2, \end{cases}$$

элементы последовательностей $\{\tilde{y}_i\}$ и $\{\tilde{a}_i\}$ описываются соотношениями (1.5), а последовательность $\{\tilde{c}_i\}$ — формулой

$$\tilde{c}_i = \prod_{j=1}^{i} (1 + \tilde{\theta}_{j-1} 2^{-\tilde{k}_{j-1}}), \qquad (1.6)$$

причем для любого i выполнены неравенства

$$|\alpha_{i-1}|, |\beta_{i-1}|, |\delta_{i-1}| < 2^{-(m+q)+1}. \qquad (1.7)$$

Тогда справедливы оценки точности

$$\left|\tilde{c}_{m/2} - a^{-\alpha}\right| < 2^{-m/2} a^{-\alpha} + \\ + \begin{cases} m 2^{-(m+q)+1} a^{-\alpha}, \text{ при } n = 1, \\ \left(m + q + \dfrac{m-q}{2}\right) 2^{-(m+q)+1} a^{-\alpha}, \\ \qquad \text{при } n = 2, \end{cases}$$

$$\left|\tilde{y}_{m/2} - a^{-\alpha} y\right| < \\ < \frac{3m}{2} 2^{-(m+q)+1} a^{-\alpha} y + 2^{-m} a^{-\alpha} y + \\ + \begin{cases} m 2^{-(m+q)+1} a^{-\alpha} y, \text{ при } n = 1, \\ \left(m + q + \dfrac{m-q}{2}\right) \times \\ \quad \times 2^{-(m+q)+1} a^{-\alpha} y, \text{ при } n = 2. \end{cases} \qquad (1.8)$$

где $\alpha = 1/n$, $n = 1, 2$.

Доказательство. Обратимся к формуле для вычисления $\tilde{a}_i$ из списка формул (1.5). Пренебрегая разностью между $\tilde{a}_{i-1} + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}+1} a_{i-1} + 2^{-\tilde{k}_{i-1}} \tilde{a}_{i-1}$ и $\tilde{a}_{i-1} + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}+1} \tilde{a}_{i-1}$ при $i > (m+q)/2$, вынесем $\tilde{a}_{i-1}$ за скобки и представим формулу для вычисления $\tilde{a}_i$ в общем удобном для последующих выкладок виде

$$\tilde{a}_i = \left(1 + \tilde{\theta}_{i-1} 2^{\tilde{k}_{i-1}}\right)^n \tilde{a}_{i-1} + \phi_{i-1}, \quad i = 1, m/2,$$

$$\phi_{i-1} = \begin{cases} \alpha_{i-1}, \text{ при } n = 1, \\ \begin{cases} \alpha_{i-1} + \beta_{i-1}, \\ i \leqslant (m+q)/2, \\ \alpha_{i-1}, i > (m+q)/2, \end{cases} \text{при } n = 2, \end{cases}$$

Используя эту рекуррентную формулу, выразим $\tilde{a}_{m/2}$ через начальное значение величины a

$$\tilde{a}_{m/2} = \left(1 + \tilde{\theta}_{(m/2)-1} 2^{-\tilde{k}_{(m/2)-1}}\right)^n \tilde{a}_{(m/2)-1} + \\ + \phi_{(m/2)-1} = \ldots \\ \ldots = \prod_{i=1}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n a_0 + \\ + \prod_{i=2}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n \phi_0 + \ldots \\ \ldots + \prod_{i=(m/2)-1}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n \phi_{(m/2)-3} + \\ + \left(1 + \tilde{\theta}_{(m/2)-1} 2^{-\tilde{k}_{(m/2)-1}}\right)^n \phi_{(m/2)-2} + \\ + \phi_{(m/2)-1}. \qquad (1.9)$$

Умножив соотношение (1.9) на

$$\left(\prod_{i=1}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n\right)^{-1},$$

получим

$$\left(\prod_{i=1}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n\right)^{-1} a^{m/2} = \\ a + \left(\prod_{i=1}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n\right)^{-1} \times \\ \times \left[\prod_{i=2}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n \phi_0 + \ldots \right. \\ \ldots + \prod_{i=(m/2)-1}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n \phi_{(m/2)-3} + \\ + \left(1 + \tilde{\theta}_{(m/2)-1} 2^{-\tilde{k}_{(m/2)-1}}\right)^n \phi_{(m/2)-2} + \\ \left. + \phi_{(m/2)-1}\right].$$

Анализируя последнее соотношение, нетрудно убедиться, что произведение

$$\left(\prod_{i=1}^{m/2} \left(1 + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}}\right)^n\right)^{-1} a^{m/2}$$

есть не что иное, как $\tilde{a}$ — возмущенное значение величины a. Согласно сказанному можно записать

$$\tilde{a} = a + \left(\prod_{i=1}^{m/2}\left(1+\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\right)^n\right)^{-1} \times$$
$$\times \Big(\prod_{i=2}^{m/2}\left(1+\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\right)^n \phi_0 + \ldots$$
$$+ \prod_{i=(m/2)-1}^{m/2}\left(1+\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\right)^n \phi_{(m/2)-3} +$$
$$+ \left(1+\tilde{\theta}_{(m/2)-1}2^{-\tilde{k}_{(m/2)-1}}\right)^n \phi_{(m/2)-2} +$$
$$+ \phi_{(m/2)-1}\Big).$$

Раскрывая скобки в последнем выражении, получим

$$\tilde{a} = a + \left(\left(1+\tilde{\theta}_1 2^{-\tilde{k}_0}\right)^n\right)^{-1}\phi_0 +$$
$$+\left(\left(1+\tilde{\theta}_1 2^{-\tilde{k}_0}\right)^n\right)^{-1}\left(\left(1+\tilde{\theta}_2 2^{-\tilde{k}_1}\right)^n\right)^{-1}\phi_1 + \ldots$$
$$\ldots + \prod_{i=1}^{(m/2)}\left(\left(1+\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\right)^n\right)^{-1}\phi_{(m/2)-1}. \quad (1.10)$$

Замечание. Нетрудно также показать, что

$$\tilde{y} = y + \left(1+\tilde{\theta}_1 2^{-\tilde{k}_0}\right)^{-1}\delta_0 +$$
$$+\left(1+\tilde{\theta}_1 2^{-\tilde{k}_0}\right)^{-1}\left(1+\tilde{\theta}_2 2^{-\tilde{k}_1}\right)^{-1}\delta_1 + \ldots$$
$$+ \prod_{i=1}^{(m/2)}\left(1+\tilde{\theta}_{i-1}2^{-\tilde{k}_{i-1}}\right)^{-1}\delta_{(m/2)-1}).$$

Обращаясь к оценке (1), получим неравенство

$$\left|\tilde{c}_i - \tilde{a}^{-\alpha}\right| < 2^{-2i}\tilde{a}^{-\alpha},$$

которое также можно записать в виде

$$\tilde{c}_i = (1+o_i)\tilde{a}^{-\alpha}, \text{ где } |o_i| < 2^{-2i}. \quad (1.11)$$

Отсюда последует

$$\tilde{c}_i^{-n} = (1+o_i)^{-n}\tilde{a},$$

$$\tilde{c}_i^{-1} = (1+o_i)^{-1}\tilde{a}^{\alpha}.$$

Осуществляя подстановку (1.6) в два последних равенствах, получим

$$\prod_{j=1}^{i}\left(1+\tilde{\theta}_{j-1}2^{-\tilde{k}_{j-1}}\right)^{-n} = (1+o_i)^{-n}\tilde{a},$$

$$\prod_{j=1}^{i}\left(1+\tilde{\theta}_{j-1}2^{-\tilde{k}_{j-1}}\right)^{-1} = (1+o_i)^{-1}\tilde{a}^{\alpha}.$$

Используя эти два равенства в (1.10), с учетом сформулированного выше замечания, будем иметь

$$\tilde{a} = a + (1+o_1)^{-n}\tilde{a}\phi_0 + (1+o_2)^{-n}\tilde{a}\phi_1 + \ldots$$
$$\ldots + (1+o_{(m/2)})^{-n}\tilde{a}\phi_{(m/2)-1},$$

$$\tilde{y} = y + (1+o_1)^{-1}\tilde{a}^{\alpha}\delta_0 + (1+o_2)^{-1}\tilde{a}^{\alpha}\delta_1 + \ldots$$
$$\ldots + (1+o_{(m/2)})^{-1}\tilde{a}^{\alpha}\delta_{(m/2)-1}. \quad (1.12)$$

Из (1.11) следует цепочка отношений

$$(1+o_i)^{-n} = \frac{1}{(1+o_i)^n} < \frac{1}{(1-1/4)^n} =$$
$$= (4/3)^n = \begin{cases} 4/3, \text{ при } n=1, \\ 16/9, \text{ при } n=2. \end{cases}$$

Непосредственно из условий теоремы

$$\left(2^{-1} \leqslant y < 1, \begin{cases} 2^{-1} \leqslant a < 1, \text{ при } n=1, \\ 2^{-2} \leqslant a < 1, \text{ при } n=2 \end{cases}\right)$$

вытекает вторая цепочка отношений

$$a^{\alpha} < 1 = 2\cdot 2^{-1} \leqslant 2y.$$

Выпишем отдельно результирующие неравенства этих цепочек отношений:

$$(1+o_i)^{-n} < \begin{cases} 4/3, \text{ при } n=1, \\ 16/9, \text{ при } n=2, \end{cases}$$

$$a^{\alpha} < 2y.$$

Отсюда последуют оценки

$$(1+o_i)^{-n}a < 2a,$$

$$(1+o_i)^{-1}a^{\alpha} < 3y.$$

Используя последние неравенства, получим оценки близости величин $\tilde{a}$ к a и $\tilde{y}$ к y. Для этого обратимся к равенствам (1.12), откуда будем иметь

$$|\tilde{a}-a| \leqslant (1+o_1)^{-n}\tilde{a}\,|\phi_0| + (1+o_2)^{-n}\tilde{a}\,|\phi_1| + \ldots$$
$$\ldots + (1+o_{(m/2)})^{-n}\tilde{a}\left|\phi_{(m/2)-1}\right|,$$

$$|\tilde{y}-y| \leqslant (1+o_1)^{-1}\tilde{a}^{\alpha}\,|\delta_0| + \\ + (1+o_2)^{-1}\tilde{a}^{\alpha}\,|\delta_1| + \ldots \\ \ldots + (1+o_{(m/2)})^{-1}\tilde{a}^{\alpha}\,\left|\delta_{(m/2)-1}\right|.$$

Пренебрегая в последних двух неравенствах погрешностями высших порядков, запишем

$$|\tilde{a}-a| \leqslant 2a\,|\phi_0| + 2a\,|\phi_1| + \ldots + 2a\left|\phi_{(m/2)-3}\right| + \\ + 2a\left|\phi_{(m/2)-2}\right| + 2a\left|\phi_{(m/2)-1}\right|,$$

$$|\tilde{y}-y| \leqslant 3y\,|\delta_0| + 3y\,|\delta_1| + \ldots + 3y\left|\delta_{(m/2)-3}\right| + \\ + 3y\left|\delta_{(m/2)-2}\right| + 3y\left|\delta_{(m/2)-1}\right|.$$

Наконец, применяя (1.7) в двух последних неравенствах, получим окончательные оценки

$$|\tilde{y}-y| < \frac{3m}{2}2^{-(m+q)+1}y, \qquad (1.13)$$

$$|\tilde{a}-a| < \begin{cases} m2^{-(m+q)+1}a, & \text{при } n=1, \\ \left(m+q+\dfrac{m-q}{2}\right)2^{-(m+q)+1}a, & \\ \qquad\qquad \text{при } n=2. \end{cases}$$

Последнюю из этих оценок можно представить в эквивалентном виде

$$\tilde{a} = (1+\beta)a,$$

где

$$|\beta| < \begin{cases} m2^{-(m+q)+1}, & \text{при } n=1, \\ \left(m+q+\dfrac{m-q}{2}\right)2^{-(m+q)+1}, & \\ \qquad\qquad \text{при } n=2. \end{cases}$$

Из последнего соотношения и (1.3) следует неравенство

$$\left|\tilde{a}^{-\alpha}-a^{-\alpha}\right| < \\ < \begin{cases} m2^{-(m+q)+1}a^{-\alpha}, & \text{при } n=1, \\ \left(m+q+\dfrac{m-q}{2}\right)2^{-(m+q)+1}a^{-\alpha}, & \\ \qquad\qquad \text{при } n=2. \end{cases} \qquad (1.14)$$

Для завершения доказательства теоремы запишем тождество

$$\tilde{\mathrm{c}}_{m/2} - a^{-\alpha} = \tilde{\mathrm{c}}_{m/2} - \tilde{a}^{-\alpha} + \tilde{a}^{-\alpha} - a^{-\alpha}.$$

Используя неравенство треугольника, получим

$$\left|\tilde{\mathrm{c}}_{m/2} - a^{-\alpha}\right| \leqslant \left|\tilde{\mathrm{c}}_{m/2} - \tilde{a}^{-\alpha}\right| + \left|\tilde{a}^{-\alpha} - a^{-\alpha}\right|.$$

Применяя в последнем неравенстве соотношения (2) и (1.14), получим

$$\left|\tilde{\mathrm{c}}_{m/2} - a^{-\alpha}\right| < 2^{-m}\tilde{a}^{-\alpha} + \\ + \begin{cases} m2^{-(m+q)+1}a^{-\alpha}, & \text{при } n=1, \\ \left(m+q+\dfrac{m-q}{2}\right)2^{-(m+q)+1}a^{-\alpha}, & \\ \qquad\qquad \text{при } n=2. \end{cases}$$

. Пренебрегая малыми высших порядков, получим окончательную оценку точности алгоритма обращения

$$\left|\tilde{\mathrm{c}}_{m/2} - a^{-\alpha}\right| < 2^{-m}a^{-\alpha} + \\ + \begin{cases} m2^{-(m+q)+1}a^{-\alpha}, & \text{при } n=1, \\ \left(m+q+\dfrac{m-q}{2}\right)2^{-(m+q)+1}a^{-\alpha}, & \\ \qquad\qquad \text{при } n=2. \end{cases} \qquad (1.15)$$

Для оценки разности $\tilde{\mathrm{c}}_{m/2}\tilde{y} - a^{-\alpha}y$ запишем тождество

$$\tilde{\mathrm{c}}_{m/2}\tilde{y} - a^{-\alpha}y = \tilde{\mathrm{c}}_{m/2}\tilde{y} - a^{-\alpha}\tilde{y} + \\ + a^{-\alpha}\tilde{y} - a^{-\alpha}y,$$

из которого в соответствии с неравенством треугольника получаем

$$\left|\tilde{\mathrm{c}}_{m/2}\tilde{y} - a^{-\alpha}y\right| \leqslant \left|\tilde{\mathrm{c}}_{m/2}\tilde{y} - a^{-\alpha}\tilde{y}\right| + \\ + \left|a^{-\alpha}\tilde{y} - a^{-\alpha}y\right| = \left|\tilde{\mathrm{c}}_{m/2} - a^{-\alpha}\right|\tilde{y} + \\ + a^{-\alpha}\,|\tilde{y}-y|.$$

Из последней цепочки отношений и (1.13) и (1.15) следует

$$\left|\tilde{\mathrm{c}}_{m/2}\tilde{y} - a^{-\alpha}y\right| < \frac{3m}{2}2^{-(m+q)+1}a^{-\alpha}y + \\ + 2^{-m}a^{-\alpha}\tilde{y} + \\ + \begin{cases} m2^{-(m+q)+1}a^{-\alpha}\tilde{y}, & \text{при } n=1, \\ \left(m+q+\dfrac{m-q}{2}\right)2^{-(m+q)+1}a^{-\alpha}\tilde{y}, & \\ \qquad\qquad \text{при } n=2. \end{cases}$$

Пренебрегая погрешностями высших порядков, получим

$$\left|\tilde{y}_{m/2} - a^{-\alpha}y\right| < \frac{3m}{2}2^{-(m+q)+1}a^{\alpha}y + \\ + 2^{-m}a^{-\alpha}y + \\ + \begin{cases} m2^{-(m+q)+1}a^{-\alpha}y, & \text{при } n=1, \\ \left(m+q+\dfrac{m-q}{2}\right)2^{-(m+q)+1}a^{-\alpha}y, & \\ & \text{при } n=2. \end{cases}$$

Теорема доказана.

2. Обсуждение результатов

В неравенстве (3) указаны стандартные требования к точности арифметических операций, выполняемых на арифметическом процессоре. Произведем сопоставление точности устройства нормировки с точностью арифметического процессора при выполнении операции деления. Обращаясь к (1.8), можно записать

$$\left|\tilde{y}_{m/2} - a^{-1}y\right| < \left(\frac{3m}{2}2^{-(m+q)+1} + 2^{-m} + \right. \\ \left. + m2^{-(m+q)+1}\right)a^{-1}y.$$

Потребуем, чтобы

$$\frac{3m}{2}2^{-(m+q)+1} + 2^{-m} + m2^{-(m+q)+1} < 2^{-m+1}$$

и разрешим последнее неравенство относительно q. Получим

$$\frac{5m}{2}2^{-(m+q)+1} < 2^{-m},$$

следовательно $q > \log_2(5m)$. Таким образом, установлено, что при выполнении последнего неравенства можно ручаться, что выполнение операции деления на устройстве нормировки удовлетворит требованию

$$\left|\tilde{y}_{m/2} - a^{-1}y\right| < 2^{-m+1}a^{-1}y.$$

Пусть теперь $n = 2$. В этом случае при вычислении выражения $y/\sqrt{a}$ на арифметическом процессоре допускается погрешность, не превосходящая $2 \cdot 2^{-m+1}$ [3]. Сопоставим указанную погрешность с погрешностью, возникающей при осуществлении вычислений на устройстве нормировки вектора. Вновь обращаясь к (1.8), запишем

$$\left|\tilde{y}_{m/2} - a^{-\alpha}y\right| < \left(\frac{3m}{2}2^{-(m+q)+1} + 2^{-m} + \right. \\ \left. + \left(m+q+\frac{m-q}{2}\right)2^{-(m+q)+1}\right)a^{-\alpha}y$$

и выдвинем требование

$$\frac{3m}{2}2^{-(m+q)+1} + 2^{-m} + \\ + \left(m+q+\frac{m-q}{2}\right)2^{-(m+q)+1} < 2 \cdot 2^{-m+1}.$$

Выполним действия, аналогичные проделанным для $n = 1$, получим

$$\frac{6m+q}{2}2^{-(m+q)+1} < 3 \cdot 2^{-m},$$

следовательно $6m < 3 \cdot 2^q - q$.

Итак, выбор q, удовлетворяющего последнему неравенству, обеспечивает выполнение неравенства

$$\left|\tilde{y}_{m/2} - a^{-n}y\right| < 2 \cdot 2^{-m+1}a^{-n}y.$$

Наконец, считая y j-й компонентой вектора $\mathbf{x}$ ($y = x_j$, $j = 1, N$) и осуществляя вычисления по формулам (1.5) на устройстве нормировки вектора ($i = 1, m/2$), вместо вектора $\mathbf{z} = a^{-\alpha}\mathbf{x}$ мы получим вектор $\tilde{\mathbf{z}}$, при этом выполняются оценки близости

$$\|\tilde{\mathbf{z}} - \mathbf{z}\| < 2^{-m+1}\|\mathbf{z}\| \text{ при } n = 1,$$

$$\|\tilde{\mathbf{z}} - \mathbf{z}\| < 2 \cdot 2^{-m+1}\|\mathbf{z}\| \text{ при } n = 2.$$

В заключение автору хотелось бы выразить благодарность Е. А. Семенчину за постоянную поддержку и внимание к его работе.

Литература

1. *Бабенко В.Н.* Представление инверсии делителя в мультипликативной форме и ее применение // Известия вузов. Северо-Кавказский регион. Технические науки. 2010. №6. С. 33–37.
2. *Бабенко В.Н.* Алгоритм инверсии делителя // Экологический вестник научных центров Черноморского экологического сотрудничества. 2013. №4 .Т. 1. С. 19–25.
3. *Годунов С.К.* Решение систем линейных уравнений. Новосибирск: Наука, 1980. 177 с.
4. Пат. 2449357 RU, МПК G 06 F 17/16. Устройство нормировки вектора / Бабенко В.Н.
5. Пат. 2473961 RU, МПК G 06 F 17/16. Устройство нормировки вектора / Бабенко В.Н.

References

1. Babenko V.N. Predstavlenie inversii delitelja v mul'tiplikativnoj forme i ee primenenie [View inversion of the divisor in a multiplicative form and its application]. *Izvestija vuzov. Severo-Kavkazskij region. Tehnicheskie nauki* [Proc. of the Universities. North Caucasus region. Technical Sciences], 2010, no. 6, pp. 33–37. (In Russian)
2. Babenko V.N. Algoritm inversii delitelja [Inversion algorithm divider]. *Ekologicheskiy vestnik nauchnykh tsentrov Chernomorskogo ekologicheskogo sotrudnichestva* [Ecological Bulletin of Research Centers of the Black Sea Economic Cooperation], 2013, no. 4, vol. 1, pp. 19–25. (In Russian)
3. Godunov S.K. *Reshenie sistem linejnyh uravnenij* [Solution of systems of linear equations]. Novosibirsk, Nauka Publ., 1980, 177 p. (In Russian)
4. Babenko V.N. *Ustrojstvo normirovki vektora* [The device normalization of the vector]. Patient 2449357 RU, MPK G 06 F 17/16. (In Russian)
5. Babenko V.N. *Ustrojstvo normirovki vektora* [The device normalization of the vector]. Patient 2473961 RU, MPK G 06 F 17/16. (In Russian)

Статья поступила 12 февраля 2014 г.

Кубанский государственный университет, г. Краснодар