УДК [004.3+681.3]: 619.67

ТОЧНОСТЬ АППАРАТУРНОЙ РЕАЛИЗАЦИИ ПРЕОБРАЗОВАНИЯ ВРАЩЕНИЯ ВЕКТОРА

Бабенко В. Н.

THE ACCURACY OF HARDWARE REALIZATION
OF TRANSFORMATION VECTOR'S ROTATION

Babenko V. N.

Kuban State University, Krasnodar, 350040, Russia
e-mail: rnibvd@mail.ru

Abstract. Development of algorithms and designing of devices for their realization are the important factor of increase of productivity of computing systems. Earlier the author had been submitted algorithm of inversion of a divider and application of algorithm for performance of normalization of a vector. Then the description of devices of normalization of a vector has been given and research of their accuracy is made. The device of normalization of a vector is a component of the device of vector's rotation. The process of accumulation of errors in calculations investigated in the device of vector's rotation by methods of the numerical analysis. The addition younger bits allocated for restriction of growth of errors in this device on adders for a mantissa to the basic bits. As a result of investigations the attitudes of between numbers of the basic and additional bits are established.

Keywords: machine number, algorithm, convergence, regularity, profitability, device of vector's rotation, error of calculation

Устойчивость ортогональных преобразований (отражений и вращений) к погрешностям вычислений обусловливает их доминирующее положение в современной вычислительной математике. Одним из лучших аппаратно ориентированных алгоритмов, реализующих преобразование вращения, является метод *Cordic*, осуществляющий ортогональное аннулирование одной компоненты двумерного вектора за $2m$ итераций [1]. Автором предложена модификация *Cordic*-метода, позволяющая реализовать указанную операцию за m итераций [2, 3].

На практике вычисления по рассматриваемому методу можно выполнять в два этапа: 1) этап псевдовращений, 2) этап нормировки вектора. В [2] изложен алгоритм, реализующий этап псевдовращений за $m/2$ итераций. Соответственно, в [3, 4] дано описание алгоритма, реализующего этап нормировки за $m/2$ итераций. Наконец, в [5] проведено исследование процесса накопления погрешностей устройством нормировки вектора. В настоящее время работы по проектированию устройства вращения вектора находятся в стадии завершения. В его состав входят устройства псевдовращений и нормировки вектора, причем выход первого соединен с входом второго. Процесс накопления погрешностей в устройстве псевдовращений остался неизученным. Вследствие этого точность результата, полученного на выходе устройства вращения вектора, является неустановленной величиной. Эта статья призвана ликвидировать создавшийся пробел.

Чтобы ограничить рост указанных погрешностей и обеспечить приемлемую точность вычисленного результата, в рассматриваемых устройствах дополнительно к m основным разрядам были использованы младшие разряды: q_{ps} разрядов в устройстве псевдовращений и q — в устройстве нормировки.

1. Исследование процесса накопления погрешностей в устройстве вращения вектора

Будем предполагать, что на вход устройства вращения вектора подаются числа в формате с плавающей точкой. Этот формат для числа x определяется следующим образом: $x = \sigma\gamma^{k_x} m_x$, где σ — код знака числа x, ($\sigma \in \{0,1\}$), γ — основание системы счисления (мы будем рассматривать $\gamma = 2$), k_x — порядок числа x, m_x — его мантисса, удовлетворяющая неравенству $\gamma^{-1} \leqslant m_x < 1$

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

или $1 \leqslant m_x < \gamma$, а под мантиссу отведено m двоичных разрядов.

Далее предполагается, что непосредственно перед выполнением вычислений в устройстве вращения вектора по указанным ниже формулам компоненты векторов подвергаются операции выравнивания порядков. Согласно сказанному, в качестве исходных величин непосредственных вычислений можно принять не сами значения компонент вектора, а их сдвинутые мантиссы, представленные в дополнительном модифицированном коде.

Как отмечалось выше, при аппаратурном проведении 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,

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

Предполагая выполненным неравенство $|\beta| < 1$, получим

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

Из представленного ряда вытекает неравенство

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

Используя (1.1), запишем цепочку тождественных преобразований

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

Отсюда с учетом (1.2) следует оценка близости величин $\tilde{a}^{-\alpha}$ и $a^{-\alpha}$ [5]:

$$\left|\tilde{a}^{-\alpha} - a^{-\alpha}\right| < \alpha\,|\beta|\,\left|a^{-\alpha}\right| \tag{1.3}$$

Обращаясь к этапу псевдовращений [2], запишем алгоритм аннулирования одной компоненты двумерного вектора.

Теорема 1. Пусть для $\forall\ x \in R^2$ определена следующая последовательность:

$$\mathbf{x}^{(0)} = \mathbf{x}, \quad \mathbf{s}_0 = 1,$$

$$\begin{cases} \mathbf{x}^{(i)} = \mathbf{C}_{i-1}\mathbf{x}^{(i-1)}, i = 1,2,\ldots, \\ s_i = (1 + 2^{-2k_{i-1}})s_{i-1}, \end{cases} \tag{1.4}$$

где

$$\mathbf{C}_{i-1} = \begin{cases} \begin{pmatrix} 1 & w_{i-1} \\ -w_{i-1} & 1 \end{pmatrix}, \text{ если } \left|\dfrac{x_2^{(i-1)}}{x_1^{(i-1)}}\right| \leqslant 1, \\ \begin{pmatrix} w_{i-1} & 1 \\ 1 & -w_{i-1} \end{pmatrix}, \text{ если } \left|\dfrac{x_2^{(i-1)}}{x_1^{(i-1)}}\right| > 1, \end{cases}$$

$$w_{i-1} = \sigma_{i-1}2^{-k_{i-1}},$$

$$\sigma_{i-1}2^{-j_{i-1}}u_{t_{i-1}} = \begin{cases} \dfrac{x_2^{(i-1)}}{x_1^{(i-1)}}, \text{ если } \left|\dfrac{x_2^{(i-1)}}{x_1^{(i-1)}}\right| \leqslant 1, \\ \dfrac{x_1^{(i-1)}}{x_2^{(i-1)}}, \text{ если } \left|\dfrac{x_2^{(i-1)}}{x_1^{(i-1)}}\right| > 1, \end{cases}$$

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

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

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

$$z_{i-1} = \frac{\sqrt{1 + 2^{-2\left(j_{i-1}+1\right)}} + 2^{-1}\sqrt{1 + 2^{-2j_{i-1}}}}{\sqrt{1 + 2^{-2\left(j_{i-1}+1\right)}} + \sqrt{1 + 2^{-2j_{i-1}}}}.$$

Тогда последовательность $\{x_i\}$ сходится к линейному многообразию $\mathbf{R}(\mathbf{e}_1)$, причем справедлива следующая оценка скорости сходимости:

$$\frac{\left|x_2^{(i)}\right|}{\sqrt{\left(x_1^{(i)}\right)^2 + \left(x_2^{(i)}\right)^2}} < \frac{2^{-2i}}{\sqrt{1 + 2^{-4i}}}. \tag{1.5}$$

При осуществлении вычислений по формулам представленного алгоритма вследствие погрешностей округления вместо последовательностей $\left\{\mathbf{x}^{(i)}\right\}$ и $\{s_i\}$ получаются иные последовательности $\left\{\tilde{\mathbf{x}}^{(i)}\right\}$ и $\left\{\tilde{\tilde{s}}_i\right\}$.

Определим источники погрешностей. Для этого обратимся к формулам (1.4). Переходя от матричного обозначения к покомпонентной записи, из (1.4) получим итерационные формулы

$$x_1^{(i)} = x_1^{(i-1)} + \sigma_{i-1}2^{-k_{i-1}}x_2^{(i-1)}$$
$$(x_1^{(i)} = \sigma_{i-1}2^{-k_{i-1}}x_1^{(i-1)} + x_2^{(i-1)}),$$

$$x_2^{(i)} = -\sigma_{i-1}2^{-k_{i-1}}x_1^{(i-1)} + x_2^{(i-1)}$$
$$(x_2^{(i)} = x_1^{(i-1)} - \sigma_{i-1}2^{-k_{i-1}}x_2^{(i-1)}),$$
$$s_i = s_{i-1} + 2^{-2k_{i-1}}s_{i-1}.$$

Не нарушая общности исследования, из приведенных формул, будем рассматривать только те, которые не заключены в скобки. Так при вычислении величины

$$\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_2^{(i-1)}$$
$$(-\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_1^{(i-1)}, 2^{-2\tilde k_{i-1}}\tilde{\tilde s}_{i-1}),$$

осуществляемом с помощью сдвига числа $\tilde\sigma_{i-1}\tilde x_2^{(i-1)}$ $(-\tilde\sigma_{i-1}\tilde x_1^{(i-1)})$ на $\tilde k_{i-1}$, а $\tilde{\tilde s}_{i-1}$ — на $(2\tilde k_{i-1})$ разрядов вправо, в вычисленную величину вносится погрешность $\alpha_1^{(i-1)}$ $(\alpha_2^{(i-1)}, \delta_{i-1})$, удовлетворяющая неравенству

$$\left|\alpha_1^{(i-1)}\right| < 2^{-(m+q_{ps})+1}$$
$$(\left|\alpha_2^{i-1}\right| < 2^{-(m+q_{ps})+1}, |\delta_{i-1}| < 2^{-(m+q_{ps})+1}).$$

Поэтому вместо

$$\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_2^{(i-1)}$$
$$(-\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_1^{(i-1)}, 2^{-2\tilde k_{i-1}}\tilde{\tilde s}_{i-1})$$

мы получим

$$\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_2^{(i-1)} + \alpha_1^{(i-1)}$$
$$(-\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_1^{(i-1)} + \alpha_2^{(i-1)}, 2^{-2\tilde k_{i-1}}\tilde{\tilde s}_{i-1} + \delta_{i-1}).$$

При выполнении сложения $\tilde x_1^{(i-1)}(x_2^{(i-1)}, \tilde{\tilde s}_{i-1})$ с

$$\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_2^{(i-1)} + \alpha_1^{(i-1)}$$
$$(-\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_1^{(i-1)} + \alpha_2^{(i-1)}, 2^{-2k_{i-1}}\tilde{\tilde s}_{i-1} + \delta_{i-1})$$

никаких погрешностей не вносится, отсюда заключаем, что

$$\tilde x_1^{(i)} = \tilde x_1^{(i-1)} + \tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_2^{(i-1)} + \alpha_1^{(i-1)},$$
$$(\tilde x_2^{i} = -\tilde\sigma_{i-1}2^{-\tilde k_{i-1}}\tilde x_1^{(i-1)} + \tilde x_2^{(i-1)} + \alpha_2^{(i-1)},$$
$$\tilde{\tilde s}_i = \tilde{\tilde s}_{i-1} + 2^{-2\tilde k_{i-1}}\tilde{\tilde s}_{i-1} + \delta_{i-1}).$$

Таким образом, следуя предписаниям описанного выше алгоритма, вследствие погрешностей вычислений на самом деле производятся вычисления с возмущенными величинами, при этом связь между элементами последовательностей $\left\{\tilde{\mathbf{x}}^{(i)}\right\}$ и $\left\{\tilde{\tilde s}_i\right\}$ описывается следующим образом:

$$\tilde{\mathbf{x}}^{(0)} = \mathbf{x}, \quad \tilde{\tilde s}_0 = s = 1,$$

$$\tilde{\mathbf{x}}^{(i)} = \tilde{\mathbf{C}}_{i-1}\tilde{\mathbf{x}}^{(i-1)} + \boldsymbol{\alpha}^{(i-1)}, \quad i = 1, m/2,$$

$$\tilde{\tilde s}_i = \tilde{\tilde s}_{i-1} + 2^{-2\tilde k_{i-1}}\tilde{\tilde s}_{i-1} + \delta_{i-1}, \quad i = 1, m/4,$$

где

$$\tilde{\mathbf{C}}_{i-1} = \begin{cases} \begin{pmatrix} 1 & \tilde w_{i-1}) \\ -\tilde w_{i-1} & 1 \end{pmatrix}, \text{ если } \left|\dfrac{\tilde x_2^{(i-1)}}{\tilde x_1^{(i-1)}}\right| \leqslant 1, \\ \begin{pmatrix} \tilde w_{i-1} & 1 \\ 1 & -\tilde w_{i-1} \end{pmatrix}, \text{ если } \left|\dfrac{\tilde x_2^{(i-1)}}{\tilde x_1^{(i-1)}}\right| > 1, \end{cases}$$

$$\tilde w_{i-1} = \tilde\sigma_{i-1}2^{-\tilde k_{i-1}},$$

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

$$\tilde\sigma_{i-1}2^{-\tilde j_{i-1}}\tilde u_{t_{i-1}} = \begin{cases} \dfrac{\tilde x_2^{(i-1)}}{\tilde x_1^{(i-1)}}, \text{ если } \left|\dfrac{\tilde x_2^{(i-1)}}{\tilde x_1^{(i-1)}}\right| \leqslant 1, \\ \dfrac{\tilde x_1^{(i-1)}}{\tilde x_2^{(i-1)}}, \text{ если } \left|\dfrac{\tilde x_2^{(i-1)}}{\tilde x_1^{(i-1)}}\right| > 1, \end{cases}$$

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

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

$$\tilde z_{i-1} = \frac{\sqrt{1 + 2^{-2(\tilde k_{i-1}+1)}} + 2^{-1}\sqrt{1 + 2^{-2\tilde k_{i-1}}}}{\sqrt{1 + 2^{-2(\tilde k_{i-1}+1)}} + \sqrt{1 + 2^{-2\tilde k_{i-1}}}},$$

$$\boldsymbol{\alpha}^{(i-1)} = \begin{pmatrix} \alpha_1^{(i-1)} \\ \alpha_2^{(i-1)} \end{pmatrix}. \qquad (1.6)$$

Преобразования псевдовращения, применяемые на первом этапе, наряду с поворотом вектора также осуществляют его растяжение. Чтобы компенсировать это растяжение на втором этапе производится нормировка вектора $\tilde{\mathbf{x}}^{(m/2)}$. Операция нормировки осуществляется по формулам алгоритма [5], которые приводим в данной работе с учетом погрешностей, допускаемых при его реализации на устройстве нормировки вектора

$$\tilde a_0 = a, \quad \tilde y_0 = y,$$

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

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

где

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

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

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

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

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

$$\tilde{z}_{i-1} = \frac{3 + \tilde{\theta}_{i-1} 2^{\tilde{t}_{i-1}+1}}{4 + 3\tilde{\theta}_{i-1} 2^{\tilde{t}_{i-1}}}, \qquad (1.7)$$

$$i = 1, \ldots, m/2.$$

Ниже также представлена теорема [5], в которой приводятся оценки точности вычисленных результатов. Утверждения этой теоремы используются для вывода оценки точности результатов, полученных на выходе устройства вращения плоскости.

Теорема 2. Пусть величины y и a удовлетворяют неравенствам

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

элементы последовательностей $\{\tilde{y}_i\}$ и $\{\tilde{a}_i\}$ описываются соотношениями (1.7), а последовательность $\{\tilde{c}_i\}$ — формулой

$$\tilde{c}_i = \prod_{j=1}^{i} (1 + \tilde{\theta}_{j-1} 2^{-\tilde{k}_{j-1}}), \qquad (1.8)$$

причем для любого i выполнены неравенства

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

Тогда справедливы оценки точности

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

$$\left|\tilde{y}_{m/2} - a^{-\alpha} y\right| < \frac{3m}{2} 2^{-(m+q)+1} a^{-\alpha} y + 2^{-m} a^{-\alpha} y + \left(m + q + \frac{m-q}{2}\right) 2^{-(m+q)+1} a^{-\alpha}. \qquad (1.9)$$

где $\alpha = 1/n$, $n = 2$.

Замечание 1. В последнем неравенстве присутствие слагаемого $\frac{3m}{2} 2^{-(m+q)+1} a^{\alpha} y$, обусловлено накоплением погрешностей при непосредственном вычислении последовательности $\{y_i\}$ (из (1.7) $\tilde{y}_i = \tilde{y}_{i-1} + \tilde{\theta}_{i-1} 2^{-\tilde{k}_{i-1}} \tilde{y}_{i-1} + \delta_{i-1}$), которые отнесены к эквивалентному возмущению исходной величины y. Другими словами,

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

Замечание 2. Считая y j-й компонентой вектора $\tilde{x}^{(m/2)}$ ($y = \tilde{\mathbf{x}}_j^{(m/2)}$, $j = 1, 2$) и осуществляя вычисления по формулам (1.7) на устройстве нормировки вектора ($i = 1, m/2$), вместо вектора $a^{-\alpha} \tilde{\mathbf{x}}^{(m/2)}$ получим вектор $a^{-\alpha} \bar{\tilde{\mathbf{x}}}^{(m/2)}$. При этом выполняются оценки близости

$$\left\| a^{-\alpha} \bar{\tilde{\mathbf{x}}}^{(m/2)} - a^{-\alpha} \tilde{\mathbf{x}}^{(m/2)} \right\| < \frac{3m}{2} 2^{-(m+q)+1} \left\| a^{-\alpha} \tilde{\mathbf{x}}^{(m/2)} \right\|, \qquad (1.10)$$

где $\bar{\tilde{\mathbf{x}}}^{(m/2)}$ — вектор, полученный из $\tilde{\mathbf{x}}^{(m/2)}$ путем внесения в него возмущения, эквивалентного погрешностям вычислений (1.7).

Рассмотрим процессы накопления погрешностей, возникающих в устройстве вращения плоскости при реализации описанных выше алгоритмов.

Теорема 3. Пусть компоненты x_1 и x_2 произвольного вектора $\mathbf{x} \in R^2$ удовлетворяют неравенствам

$$0 \leqslant |x_2| \leqslant |x_1|, 2^{-1} \leqslant |x_1| < 1 \quad (0 \leqslant |x_1| \leqslant |x_2|, 2^{-1} \leqslant |x_2| < 1), \qquad (1.11)$$

элементы последовательностей $\left\{\tilde{\mathbf{x}}^i\right\}$ и $\left\{\tilde{\tilde{s}}_i\right\}$ описываются соотношениями (1.6) а элементы последовательности $\left\{\tilde{\mathbf{P}}_i\right\}$ имеют вид

$$\tilde{\mathbf{P}}_i = \tilde{\mathbf{C}}_{i-1} \ldots \tilde{\mathbf{C}}_0, \qquad (1.12)$$

где для любого i выполнены неравенства

$$\left|\alpha_1^{(i-1)}\right|, \left|\alpha_2^{(i-1)}\right|, \left|\delta_{(i-1)}\right| < \\ < 2^{-(m+q_{ps})+1}. \quad (1.13)$$

Тогда справедливы оценки точности

$$\left|\left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2} - \left(\tilde{s}_{m/4}\right)^{-1/2}\right| < \\ < \frac{1}{2}\left|\beta\right|\left(\tilde{s}_{m/4}\right)^{-1/2}, \quad (1.14)$$

$$\left\|\tilde{\mathbf{x}} - \mathbf{x}\right\| < 2\sqrt{2}m2^{-(m+q_{ps})+1}\left\|\mathbf{x}\right\|, \quad (1.15)$$

где

$$\tilde{s}_{m/4} = \prod_{i=1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right), \quad (1.16)$$

$$|\beta| < \frac{m}{4}2^{-(m+q_{ps})+1},$$

$\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}\mathbf{x}$, $\boldsymbol{\delta}\mathbf{x}$ — возмущение исходного вектора, эквивалентное погрешностям, описанным в (1.6).

Доказательство. Обратимся сначала к формулам для вычисления $\tilde{\mathbf{x}}^{(i)}$ и $\tilde{\tilde{s}}_i$ из (1.6). Используя эти рекуррентные формулы, выразим $\tilde{\mathbf{x}}^{(m/2)}$ и $\tilde{\tilde{s}}_{m/4}$ через входные данные алгоритма: вектор $\tilde{\mathbf{x}}^{(0)}$ и скаляр $\tilde{\tilde{s}}_0(s)$

$$\tilde{\mathbf{x}}^{(m/2)} = \tilde{\mathbf{C}}_{(m/2)-1}\tilde{\mathbf{x}}^{(m/2)-1} + \boldsymbol{\alpha}^{(m/2)-1} = \\ = \tilde{\mathbf{C}}_{(m/2)-1}\left(\tilde{\mathbf{C}}_{(m/2)-2}\tilde{\mathbf{x}}^{(m/2)-2} + \boldsymbol{\alpha}^{(m/2)-2}\right) + \\ + \boldsymbol{\alpha}^{(m/1)-1} = \dots \\ \dots = \tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_0\mathbf{x}^{(0)} + \tilde{\mathbf{C}}_{(m/2)-1}\dots \\ \dots\tilde{\mathbf{C}}_1\boldsymbol{\alpha}^{(0)} + \tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_2\boldsymbol{\alpha}^{(1)} + \dots \\ \dots + \tilde{\mathbf{C}}_{(m/2)-1}\tilde{\mathbf{C}}_{(m/2)-2}\boldsymbol{\alpha}^{(m/2)-3} + \\ + \tilde{\mathbf{C}}_{(m/2)-1}\boldsymbol{\alpha}^{(m/2)-2} + \boldsymbol{\alpha}^{(m/2)-1},$$

$$\tilde{\tilde{s}}_{m/4} = \left(1 + 2^{-2\tilde{k}_{(m/4)-1}}\right)\tilde{s}_{(m/2)-1} + \\ + \delta_{(m/4)-1} = \\ = \left(1 + 2^{-2\tilde{k}_{(m/4)-1}}\right) \times \\ \times\left(\left(1 + 2^{-2\tilde{k}_{(m/4)-2}}\right)\tilde{s}_{(m/4)-2} + \delta_{(m/4)-2}\right) + \\ + \delta_{(m/4)-1} = \dots \\ \dots = \prod_{i=1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\tilde{s}_0 + \prod_{i=2}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_0 + \\ + \prod_{i=3}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_1 + \dots \\ \dots + \prod_{i=(m/4)-1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_{(m/4)-3} + \\ + \left(1 + 2^{-2\tilde{k}_{(m/4)-1}}\right)\delta_{(m/4)-2} + \delta_{(m/4)-1}.$$

Перепишем отдельно полученные результаты

$$\tilde{\mathbf{x}}^{(m/2)} = \tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_0\mathbf{x}^{(0)} + \tilde{\mathbf{C}}_{(m/2)-1}\dots \\ \dots\tilde{\mathbf{C}}_{(1)}\boldsymbol{\alpha}^{(0)} + \tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_2\boldsymbol{\alpha}^{(1)} + \dots \\ \dots + \tilde{\mathbf{C}}_{(m/2)-1}\tilde{\mathbf{C}}_{(m/2)-2}\boldsymbol{\alpha}^{(m/2)-3} + \\ + \tilde{\mathbf{C}}_{(m/2)-1}\boldsymbol{\alpha}^{(m/2)-2} + \boldsymbol{\alpha}^{(m/2)-1}, \quad (1.17)$$

$$\tilde{\tilde{s}}_{m/4} = \prod_{i=1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\tilde{s}_0 + \\ + \prod_{i=2}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_0 + \prod_{i=3}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_1 + \dots \\ \dots + \prod_{i=(m/4)-1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_{(m/4)-3} + \\ + \left(1 + 2^{-2\tilde{k}_{(m/4)-1}}\right)\delta_{(m/4)-2} + \delta_{(m/4)-1}.$$

Умножив обе части последнего равенства на

$$\left(\prod_{i=1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\right)^{-1},$$

получим

$$\left(\prod_{i=1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\right)^{-1}\tilde{\tilde{s}}_{m/4} = \\ = s + \left(\prod_{i=1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\right)^{-1} \times \\ \times\left(\prod_{i=2}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_0 + \right. \\ \left. + \prod_{i=3}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_1 + \dots\right. \\ \left. \dots + \prod_{i=(m/4)-1}^{m/4}\left(1 + 2^{-2\tilde{k}_{i-1}}\right)\delta_{(m/4)-3} + \right.$$

$$+\left(1+2^{-2\tilde{k}_{(m/4)-1}}\right)\delta_{(m/4)-2}+\delta_{(m/4)-1}\Bigg).$$

Анализируя последнее соотношение, легко видеть, что произведение

$$\left(\prod_{i=1}^{m/4}\left(1+2^{-2\tilde{k}_{i-1}}\right)\right)^{-1}\tilde{\tilde{s}}_{m/4}$$

есть не что иное, как $\tilde{\tilde{s}}$ — возмущенное значение величины s ($\tilde{\tilde{s}}_0$, см. (1.6)). Согласно сказанному

$$\begin{aligned}\tilde{\tilde{s}}=1+\left(1+2^{-2\tilde{k}_0}\right)^{-1}\delta_0+\\ +\left(1+2^{-2\tilde{k}_0}\right)^{-1}\left(1+2^{-2\tilde{k}_1}\right)^{-1}\delta_1+\dots\\ \dots+\prod_{i=1}^{(m/4)-2}\left(1+2^{-2\tilde{k}_{i-1}}\right)^{-1}\delta_{(m/4)-3}+\\ +\prod_{i=1}^{(m/4)-1}\left(1+2^{-2\tilde{k}_{i-1}}\right)^{-1}\delta_{(m/4)-2}+\\ +\prod_{i=1}^{(m/4)}\left(1+2^{-2\tilde{k}_{i-1}}\right)^{-1}\delta_{(m/4)-1}.\end{aligned}$$

Учитывая условия (1.13) теоремы 3, а также тот факт, что для любого $j=1,\dots,m/4$ выполняется неравенство

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

последнее равенство можно записать в компактном виде

$$\tilde{\tilde{s}}=1+\beta,\ \text{где}\ |\beta|<\frac{m}{4}2^{-(m+q_{ps})+1}.$$

Обратно, умножая обе части последнего равенства на

$$\prod_{i=1}^{(m/4)}\left(1+2^{-2\tilde{k}_{i-1}}\right),$$

получим

$$\tilde{\tilde{s}}_{m/4}=(1+\beta)\prod_{i=1}^{(m/4)}\left(1+2^{-2\tilde{k}_{i-1}}\right).\qquad(1.18)$$

Обращаясь к (1.3), (1.17) и (1.18) получим (1.14).

Умножим обе части равенства (1.17) на $\left(\tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_0\right)^{-1}$. При этом получим

$$\begin{aligned}\left(\tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_0\right)^{-1}\tilde{\mathbf{x}}^{(m/2)}=\\ =\mathbf{x}+\left(\tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_0\right)^{-1}\Big(\tilde{\mathbf{C}}_{(m/2)-1}\dots\\ \dots\tilde{\mathbf{C}}_1\boldsymbol{\alpha}^{(0)}+\tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_2\boldsymbol{\alpha}^{(1)}+\dots\\ \dots+\tilde{\mathbf{C}}_{(m/2)-1}\tilde{\mathbf{C}}_{(m/2)-2}\boldsymbol{\alpha}^{(m/2)-3}+\\ +\tilde{\mathbf{C}}_{(m/2)-1}\boldsymbol{\alpha}^{(m/2)-2}+\boldsymbol{\alpha}^{(m/2)-1}\Big).\end{aligned}$$

Анализируя последнее соотношение, можно убедиться, что произведение

$$\left(\tilde{\mathbf{C}}_{(m/2)-1}\dots\tilde{\mathbf{C}}_0\right)^{-1}\tilde{\mathbf{x}}^{(m/2)}$$

представляет собой $\tilde{\mathbf{x}}$ — вектор, определенный в результативной части доказываемой теоремы. Согласно сказанному запишем

$$\begin{aligned}\tilde{\mathbf{x}}=x+\tilde{\mathbf{C}}_0^{-1}\boldsymbol{\alpha}^{(0)}+\tilde{\mathbf{C}}_0^{-1}\tilde{\mathbf{C}}_1^{-1}\boldsymbol{\alpha}^{(1)}+\dots\\ \dots+\tilde{\mathbf{C}}_0^{-1}\dots\left(\tilde{\mathbf{C}}_{(m/2)-3}\right)^{-1}\boldsymbol{\alpha}^{(m/2)-3}+\tilde{\mathbf{C}}_0^{-1}\dots\\ \dots\left(\tilde{\mathbf{C}}_{(m/2)-2}\right)^{-1}\boldsymbol{\alpha}^{(m/2)-2}+\\ +\tilde{\mathbf{C}}_0^{-1}\dots\left(\tilde{\mathbf{C}}_{(m/2)-1}\right)^{-1}\boldsymbol{\alpha}^{(m/2)-1}.\end{aligned}$$

Оценим близость векторов $\tilde{\mathbf{x}}$ и $\mathbf{x}$, учитывая, что для любого i $\left\|\tilde{\mathbf{C}}_i\right\|>1$ и выполняется неравенство (1.13)

$$\begin{aligned}\|\tilde{\mathbf{x}}-\mathbf{x}\|\leqslant\left\|\tilde{\mathbf{C}}_0^{-1}\right\|\left\|\boldsymbol{\alpha}^{(0)}\right\|+\\ +\left\|\tilde{\mathbf{C}}_0^{-1}\right\|\left\|\tilde{\mathbf{C}}_1^{-1}\right\|\left\|\boldsymbol{\alpha}^{(1)}\right\|+\dots\\ \dots+\left\|\tilde{\mathbf{C}}_0^{-1}\right\|\dots\left\|\left(\tilde{\mathbf{C}}_{(m/2)-3}\right)^{-1}\right\|\left\|\boldsymbol{\alpha}^{(m/2)-3}\right\|+\\ +\left\|\tilde{\mathbf{C}}_0^{-1}\right\|\dots\left\|\left(\tilde{\mathbf{C}}_{(m/2)-2}\right)^{-1}\right\|\left\|\boldsymbol{\alpha}^{(m/2)-2}\right\|+\\ +\left\|\tilde{\mathbf{C}}_0^{-1}\right\|\dots\left\|\left(\tilde{\mathbf{C}}_{(m/2)-1}\right)^{-1}\right\|\left\|\boldsymbol{\alpha}^{(m/2)-1}\right\|,\end{aligned}$$

то есть

$$\|\tilde{\mathbf{x}}-\mathbf{x}\|<m2^{-(m+q_{ps})+1}\sqrt{2}.\qquad(1.19)$$

Из (1.11) следует, что $\frac{1}{2}\leqslant\|\mathbf{x}\|$, но тогда

$$\sqrt{2}\leqslant 2\sqrt{2}\,\|\mathbf{x}\|.$$

Используя последнее неравенство в (1.19), получим оценку (1.15). □

Следствие. Пусть выполнены условия теорем 3 и 2, причем

$$a = 2^{-2}\tilde{s}_{m/4}, \quad y = \tilde{x}_j^{(m/2)}, \quad j = 1, 2,$$

элементы последовательностей $\{\tilde{y}_i\}$ и $\{\tilde{a}_i\}$ описываются соотношениями (1.7), а последовательности $\{\tilde{c}_i\}$ — формулой (1.8).

Тогда справедлива оценка точности

$$\left\|\tilde{\tilde{\mathbf{x}}}^{(m/2)} - \boldsymbol{Rt}\mathbf{x}\right\| < \\ < \left(\left(2^{-m} + \left(\frac{m+q}{2} + \frac{m-q}{4}\right)2^{-(m+q)+1}\right) + \right. \\ + \frac{1}{2}\cdot\frac{m}{4}2^{-(m+q_{ps})+1} + \frac{3m}{2}2^{-(m+q)+1} + \\ \left. + 2\sqrt{2}m2^{-(m+q_{ps})+1} + \frac{2^{-m}}{\sqrt{1+2^{-2m}}}\right)\|\mathbf{x}\|,$$

где $\tilde{\tilde{\mathbf{x}}}^{(m/2)}$ — вектор, полученный на выходе устройства вращения плоскости, а преобразование R таково, что

$$\mathbf{Rx} = \begin{pmatrix}\sigma\sqrt{x_1^2+x_2^2}\\ 0\end{pmatrix}, \text{ где } \sigma = \operatorname{sign}(x_1).$$

Доказательство. Согласно формулировке теоремы входными данными для алгоритма нормировки вектора являются: вектор $\tilde{\mathbf{x}}^{(m/2)}$ и значение нормирующей величины $\tilde{\tilde{s}}_{m/4}$. Обращаясь к неравенствам (1.9) (теорема 2) и (1.10) (замечание 2), можно записать

$$\left|\left(\tilde{c}_{m/2}\right)^{-1/2} - \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\right| < \\ \left(2^{-m} + \left(\frac{m+q}{2} + \frac{m-q}{4}\right)2^{-(m+q)+1}\right)\times \\ \times\left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}, \quad (1.20)$$

$$\left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\tilde{x}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{x}^{(m/2)}\right\| < \\ < \frac{3m}{2}2^{-(m+q)+1}\left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{x}^{(m/2)}\right\|. \quad (1.21)$$

Мысленно продолжим вычислительный процесс ортогонального аннулирования одной компоненты вектора по приведенным выше вычислительным процедурам. Очевидно, при этих вычислениях погрешности отсутствуют. В соответствии со сказанным, пользуясь обозначениями (1.12) $i = (m/2)+1, (m/2)+2, \ldots$ и (1.16) $i = (m/4)+1, (m/4)+2, \ldots$, получим

$$\lim_{i\to\infty}\tilde{s}_i^{-(1/2)}\tilde{\mathbf{P}}_i\mathbf{x} = \begin{pmatrix}\sigma\sqrt{x_1^2+x_2^2}\\ 0\end{pmatrix}.$$

Другими словами, последовательность $\{\tilde{s}_i^{-(1/2)}\tilde{\mathbf{P}}_i\}$ сходится к $\mathbf{R}$. С этой последовательностью мы свяжем последовательность углов $\{\varphi_i\}$, элементы которой определяются следующим образом:

$$\varphi_i = \angle\left(\mathbf{Rx}, \tilde{s}_i^{-(1/2)}\tilde{\mathbf{P}}_i\mathbf{x}\right).$$

Очевидно,

$$\left\|\tilde{s}_{m/2}^{-(1/2)}\tilde{\mathbf{P}}_{m/2}\mathbf{x} - \mathbf{Rx}\right\| = 2\sin\frac{\varphi_{m/2}}{2}\|\mathbf{x}\|,$$

кроме того $\lim\limits_{\varphi\to 0}\left(2\frac{\sin\frac{\varphi}{2}}{\sin\varphi}\right) = 1$. Поэтому, пренебрегая малыми высших порядков, примем

$$2\sin\frac{\varphi_{m/2}}{2} = \sin\varphi_{m/2}.$$

С другой стороны, обращаясь к неравенству (1.5) (теорема 1) мы можем записать

$$\sin\varphi_{m/2} = \frac{\left|\tilde{x}_2^{(m/2)}\right|}{\sqrt{\left(\tilde{x}_1^{(m/2)}\right)^2 + \left(\tilde{x}_2^{(m/2)}\right)^2}} < \\ < \frac{2^{-m}}{\sqrt{1+2^{-2m}}}.$$

Суммируя сказанное, приходим к неравенству

$$\left\|\tilde{s}_{m/2}^{-(1/2)}\tilde{\mathbf{P}}_{m/2}\mathbf{x} - \mathbf{Rx}\right\| < \\ < \frac{2^{-m}}{\sqrt{1+2^{-2m}}}\|\mathbf{x}\|. \quad (1.22)$$

Введем новое обозначение $\bar{\tilde{\mathbf{x}}}^{(m/2)}$ — вектор, полученный из $\tilde{\mathbf{x}}^{(m/2)}$, путем внесения в него возмущения, эквивалентного погрешностям

вычислений (1.7). Приступая к завершению доказательства, запишем цепочку тождеств

$$\begin{aligned}
&\tilde{\tilde{\mathbf{x}}}^{(m/2)} - \mathbf{R}\mathbf{x} = \\
&\quad = \tilde{c}_{m/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} + \\
&+ \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} + \\
&+ \left(\tilde{s}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{x}}^{(m/2)} + \\
&+ \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{x}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\mathbf{x} + \\
&\qquad + \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\mathbf{x} - \mathbf{R}\mathbf{x} = \\
&\quad = \tilde{c}_{m/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} + \\
&+ \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} + \\
&+ \left(\tilde{s}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{x}}^{(m/2)} + \\
&+ \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\tilde{\mathbf{x}} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\mathbf{x} + \\
&\qquad + \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\mathbf{x} - \mathbf{R}\mathbf{x}.
\end{aligned}$$

Пользуясь неравенством треугольника в последнем выражении, получим

$$\begin{aligned}
&\left\|\tilde{\tilde{\mathbf{x}}}^{(m/2)} - \mathbf{R}\mathbf{x}\right\| \leqslant \\
&\quad \leqslant \left\|\tilde{c}_{m/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)}\right\| + \\
&+ \left\|\left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)}\right\| + \\
&+ \left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{x}}^{(m/2)}\right\| + \\
&+ \left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\tilde{\mathbf{x}} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{P}_{m/2}\mathbf{x}\right\| + \\
&\qquad + \left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\mathbf{x} - \mathbf{R}\mathbf{x}\right\|.
\end{aligned}$$

Отсюда будем иметь

$$\begin{aligned}
&\left\|\tilde{\tilde{\mathbf{x}}}^{(m/2)} - \mathbf{R}\mathbf{x}\right\| \leqslant \\
&\quad \leqslant \left\|\tilde{c}_{m/2} - \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\right\| \left\|\bar{\tilde{\mathbf{x}}}^{(m/2)}\right\| + \\
&+ \left\|\left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2} - \left(\tilde{s}_{m/4}\right)^{-1/2}\right\| \left\|\bar{\tilde{\mathbf{x}}}^{(m/2)}\right\| + \\
&+ \left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\bar{\tilde{\mathbf{x}}}^{(m/2)} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{x}}^{(m/2)}\right\| + \\
&+ \left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\tilde{\mathbf{x}} - \left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\mathbf{x}\right\| + \\
&\qquad + \left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{P}}_{m/2}\mathbf{x} - \mathbf{R}\mathbf{x}\right\|.
\end{aligned}$$

Учитывая условия (1.20), (1.14), (1.21), (1.15), (1.22) в последнем неравенстве, запишем

$$\begin{aligned}
&\left\|\tilde{\tilde{\mathbf{x}}}^{(m/2)} - \mathbf{R}\mathbf{x}\right\| < \\
&< \left(2^{-m} + \left(\frac{m+q}{2} + \frac{m-q}{4}\right)2^{-(m+q)+1}\right) \times \\
&\qquad \times \left(\tilde{\tilde{s}}_{m/4}\right)^{-1/2}\left\|\bar{\tilde{\mathbf{x}}}^{(m/2)}\right\| + \\
&+ \frac{1}{2}\cdot\frac{m}{4}2^{-(m+q_{ps})+1}\left(\tilde{s}_{m/4}\right)^{-1/2}\left\|\bar{\tilde{\mathbf{x}}}^{(m/2)}\right\| + \\
&\quad + \frac{3m}{2}2^{-(m+q)+1}\left\|\left(\tilde{s}_{m/4}\right)^{-1/2}\tilde{\mathbf{x}}^{(m/2)}\right\| + \\
&+ 2\sqrt{2}m2^{-(m+q_{ps})+1}\|\mathbf{x}\| + \frac{2^{-m}}{\sqrt{1+2^{-2m}}}\|\mathbf{x}\|.
\end{aligned}$$

Пренебрегая погрешностями высших порядков, упростим последнее неравенство

$$\begin{aligned}
&\left\|\tilde{\tilde{\mathbf{x}}}^{(m/2)} - \mathbf{R}\mathbf{x}\right\| < \\
&< \left(\left(2^{-m} + \left(\frac{m+q}{2} + \frac{m-q}{4}\right)2^{-(m+q)+1}\right) + \right. \\
&\quad + \frac{1}{2}\cdot\frac{m}{4}2^{-(m+q_{ps})+1} + \frac{3m}{2}2^{-(m+q)+1} + \\
&\left. + 2\sqrt{2}m2^{-(m+q_{ps})+1} + \frac{2^{-m}}{\sqrt{1+2^{-2m}}}\right)\|\mathbf{x}\|.
\end{aligned} \tag{1.23}$$

Следствие доказано. □

Замечание 3. Вектор $\tilde{\tilde{\mathbf{x}}}^{(m/2)}$ можно рассматривать, как результат действия ортогонального отображения $\mathbf{R}$ на вектор $\mathbf{x}+\boldsymbol{\delta}\mathbf{x}$: $\tilde{\tilde{\mathbf{x}}}^{(m/2)} = \mathbf{R}(\mathbf{x}+\boldsymbol{\delta}\mathbf{x})$, где $\boldsymbol{\delta}\mathbf{x}$ моделирует возмущение вектора x, эквивалентное погрешностям, вносимым в вычисляемые величины при аппаратурной реализации описанного алгоритма. Пусть $\mathbf{u} \neq \mathbf{x}$. Поскольку в описанном алгоритме отображение $\mathbf{R}$ полностью определяется вектором $\mathbf{x}$, то для вектора $\tilde{\tilde{\mathbf{v}}}^{(m/2)}$, определенного как $\tilde{\tilde{\mathbf{v}}}^{(m/2)} = \mathbf{R}(\mathbf{u}+\boldsymbol{\delta}\mathbf{u})$, также справедлива оценка

$$\begin{aligned}
&\left\|\tilde{\tilde{\mathbf{v}}}^{(m/2)} - \mathbf{R}\mathbf{u}\right\| < \\
&< \left(\left(2^{-m} + \left(\frac{m+q}{2} + \frac{m-q}{4}\right)2^{-(m+q)+1}\right) + \right. \\
&\quad + \frac{1}{2}\cdot\frac{m}{4}2^{-(m+q_{ps})+1} + \frac{3m}{2}2^{-(m+q)+1} + \\
&\left. + 2\sqrt{2}m2^{-(m+q_{ps})+1} + \frac{2^{-m}}{\sqrt{1+2^{-2m}}}\right)\|\mathbf{u}\|.
\end{aligned}$$

2. Обсуждение результатов

К точности вычислений на арифметическом процессоре ЭВМ предъявляются вполне определенные требования, состоящие в том, что при выполнении арифметической 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 $(a*b)_m$, удовлетворяющий неравенству [6]

$$|(a*b)_m - a*b| < \varepsilon_1 |a*b|, \text{ где } \varepsilon_1 = 2^{-m+1}.$$

При вычислении выражения $y/\sqrt{a}$ на арифметическом процессоре допускается погрешность, не превосходящая $2\cdot 2^{-m+1}$ [6].

В [5] установлено, что для обеспечения аналогичной точности при выполнении вычислений на устройстве нормировки потребуется q дополнительных младших разрядов, причем q удовлетворяет неравенству

$$6m < 3\cdot 2^q - q. \qquad (2.1)$$

Будем предполагать, что в устройстве вращения плоскости используется устройство нормировки, отвечающее указанным требованиям.

Из правой части неравенства (1.23) видно, что слагаемые 2^{-m}, $\left(\frac{m+q}{2}+\frac{m-q}{4}\right)2^{-(m+q)+1}$ и $\frac{3m}{2}2^{-(m+q)+1}$, обусловлены погрешностями, допускаемыми в устройстве нормировки, остальные слагаемые обусловлены погрешностями вычислений в устройстве псевдовращений. Установим требование

$$\frac{1}{2}\cdot\frac{m}{4}2^{-(m+q_{ps})+1} + 2\sqrt{2}m2^{-(m+q_{ps})+1} + \frac{2^{-m}}{\sqrt{1+2^{-2m}}} < 2\cdot 2^{-m+1}. \qquad (2.2)$$

Очевидно, выполнение неравенства

$$\frac{1}{2}\cdot\frac{m}{4}2^{-(m+q_{ps})+1} + 2\sqrt{2}m2^{-(m+q_{ps})+1} < 3\cdot 2^{-m},$$

влечет за собой выполнение исходного неравенства.

В свою очередь из выполнения неравенства

$$\frac{1}{2}\cdot\frac{m}{4}2^{-(m+q_{ps})+1} + 3m2^{-(m+q_{ps})+1} < 3\cdot 2^{-m}$$

вытекает выполнение предыдущего неравенства.

Производя упрощение последнего неравенства, получим

$$\frac{25m}{8}2^{-(m+q_{ps})+1} < 3\cdot 2^{-m},$$

$$\frac{25m}{12}2^{-(m+q_{ps})} < 2^{-m}, \quad \frac{25m}{12}2^{-q_{ps}} < 1,$$

$$\frac{25m}{12} < 2^{q_{ps}}, \quad q_{ps} > \log_2\left(\frac{25m}{12}\right). \qquad (2.3)$$

Таким образом, установлено: при удоавлетворении последнего неравенства можно ручаться, что при осуществлении вычислений на устройстве псевдовращений будет также выполнено неравенство (2.2).

Подведем итог. Если в устройстве вращения вектора величины q и q_{ps} удовлетворяют неравенствам (2.1) и (2.3), соответственно, то из неравенства (1.23) вытекает оценка

$$\left\|\tilde{\tilde{\mathbf{x}}}^{(m/2)} - \mathbf{R}\mathbf{x}\right\| < 4\cdot 2^{-m+1}\|\mathbf{x}\|.$$

Литература

1. Сверхбольшие интегральные схемы и современная обработка сигналов / Под ред. С. Гуна, Х. Уайтхауса, Т. Кайлата. М.: Радио и связь, 1989. 345 с.
2. Бабенко В. Н. Способ повышения скорости сходимости процесса аннулирования одной компоненты двумерного вектора преобразованиями псевдовращения // Известия вузов. Сев.-Кавказ. регион. Технические науки. 2009. Спец. выпуск. С. 33–43.
3. Бабенко В. Н. Представление инверсии делителя в мультипликативной форме и ее применение // Известия вузов. Сев.-Кавказ. регион. Технические науки. 2010. № 6. С. 33–37.
4. Бабенко В. Н. Алгоритм инверсии делителя // Экологический вестник научных центров Черноморского экологического сотрудничества. 2013. №4. Т. 1. С. 19–25.
5. Бабенко В. Н. Накопление погрешностей при аппаратурной реализации алгоритма нормировки // Экологический вестник научных центров Черноморского экологического сотрудничества. 2014. № 2. С. 5–12.
6. Годунов С. К. Решение систем линейных уравнений. Новосибирск: Наука, 1980. 177 с.

References

1. Gun S., Uaythaus H., Kaylat T. (Eds.) *Sverhbolshye integralnye shemy i sovremennaja obrabotka signalov* [VSLI and Modern Signal Processing]. Moscow, Radio i svjaz', 1989, 345 p. (In Russian)
2. Babenko V. N. Sposob povyshenija skorosti shodimosti processa annulirovanija odnoj komponenty dvumernogo vektora preobrazovanijami psevdovrashenija [The Way of increase speed of convergence process of nulling one component of binariate vector by transformation of psevdorotation]. *Izvestija vuzov. Severo-Kavkazskij region. Tehnicheskie nauki* [Proc. of the Universities. North-Caucasian region. Engineering], 2009, Spets. vypusk [Special Iss.], pp. 33–39. (In Russian)
3. 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-Caucasian region. Engineering], 2010, no. 6, pp. 33–37. (In Russian)
4. Babenko V. N. Algoritm inversii delitelja [Inversion algorithm ivider]. *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)
5. Babenko V. N. Nakoplenie pogreshnostej pri apparaturnoj realizacii algoritma normirovki [Accumulation of errors at hardware realization of algorithm of normalization]. *Ekologicheskiy vestnik nauchnykh tsentrov Chernomorskogo ekologicheskogo sotrudnichestva* [Ecological Bulletin of Research Centers of the Black Sea Economic Cooperation], 2014, № 2, pp. 5–12. (In Russian)
6. Godunov S. K. *Reshenie sistem linejnyh uravnenij* [Solution of systems of linear equations]. Novosibirsk, Nauka Publ., 1980, 177 p. (In Russian)

Статья поступила 2 июля 2014 г.