

# Об одной реализации метода Ньютона на ЭЦВМ

В. Е. Шаманский

Для решения систем нелинейных уравнений

$$\mathbf{g}(\mathbf{x}) = \mathbf{d} \quad (1)$$

где  $\mathbf{g}(\mathbf{x})$  — вектор-функция векторного аргумента  $\mathbf{x}$  с  $n$  компонентами  $g_i(\mathbf{x})$ , а  $\mathbf{d}$  — заданный вектор, часто применяют метод Ньютона. Если обозначить через  $\Gamma(\mathbf{y})$  матрицу Якоби

$$\Gamma(\mathbf{y}) = \left( \frac{\partial \mathbf{g}}{\partial y_1}, \frac{\partial \mathbf{g}}{\partial y_2}, \dots, \frac{\partial \mathbf{g}}{\partial y_n} \right),$$

то вычислительную схему метода Ньютона можно записать так:

$$\begin{aligned} \Gamma(\mathbf{y}_k) \mathbf{z}_k &= \mathbf{g}(\mathbf{y}_k) - \mathbf{d}, \\ \mathbf{y}_{k+1} &= \mathbf{y}_k - \mathbf{z}_k, \\ (k &= 0, 1, 2, \dots), \end{aligned} \quad (2)$$

где вектор  $\mathbf{y}_0$  — произволен. Метод Ньютона в общем виде исследован в [1], где показано, что если начальное приближение  $\mathbf{y}_0$  выбрано достаточно близким к решению системы (1) и в окрестности решения существует обратная матрица  $\Gamma^{-1}(\mathbf{y})$ , то для достаточно гладких вектор-функций  $\mathbf{g}(\mathbf{x})$  итерационный процесс (2) обладает квадратичной сходимостью. Однако, во многих случаях вычисление производных от компонент вектор-функции  $\mathbf{g}(\mathbf{x})$  в аналитическом виде приводит к громоздким выражениям, кроме того, в ряде случаев может быть даже неизвестным явное выражение для  $\mathbf{g}(\mathbf{x})$ , а известен только алгоритм, позволяющий вычислять значение  $\mathbf{g}(\mathbf{x})$  при заданных значениях аргумента  $\mathbf{x}$ . Последняя ситуация имеет место, например, при решении нелинейных краевых задач для систем обыкновенных дифференциальных уравнений. Поэтому вместо вычислительной схемы (2) применяют несколько иную вычислительную схему метода Ньютона. Для записи этой вычислительной схемы введем матрицу

$$R(h, \mathbf{y}) = \left( \frac{\mathbf{g}(\mathbf{y} + h\mathbf{e}_1) - \mathbf{g}(\mathbf{y})}{h}, \frac{\mathbf{g}(\mathbf{y} + h\mathbf{e}_2) - \mathbf{g}(\mathbf{y})}{h}, \dots, \frac{\mathbf{g}(\mathbf{y} + h\mathbf{e}_n) - \mathbf{g}(\mathbf{y})}{h} \right),$$

где  $\mathbf{e}_i$   $i$ -тый координатный орт,  $h$  — вещественный параметр. Матрица  $R(h, \mathbf{y})$  представляет собой разностный аналог матрицы Якоби  $\Gamma(\mathbf{y})$ . С помощью матрицы  $R(h, \mathbf{y})$  указанную модификацию итерационного процесса Ньютона можно записать в следующем виде:

$$R(h_k, \mathbf{y}_k) \mathbf{z}_k = \mathbf{g}(\mathbf{y}_k) - \mathbf{d}, \quad \mathbf{y}_{k+1} = \mathbf{y}_k - \mathbf{z}_k, \quad (k = 0, 1, 2, \dots). \quad (3)$$

В (3)  $h_k$  представляет собой элемент заданной сходящейся последовательности вещественных чисел, в частности,  $h_k = h = \text{const}$ . Применение вычислительной схемы (3) вместо (2), естественно, как-то повлияет на быстроту сходимости итерационного процесса Ньютона, а при некоторых наборах  $h_k$  он вообще может не сходить. Принимая это во внимание, в настоящей статье мы рассмотрим вопросы сходимости итерационного процесса (3) и дадим рекомендации относительно выбора последовательности значений  $h_k$ .

В дальнейшем будем рассматривать следующую норму векторов  $\mathbf{x} = \{x_i\}$  и матриц  $A = \{a_{ij}\}$  (см[2]):

$$\|\mathbf{x}\| = \max_i |x_i|, \quad \|A\| = \max_i \sum_{j=1}^n |a_{ij}|.$$

Теорема 1. Пусть  $S$  — сфера  $\|y - y_0\| \leq \varrho$  и пусть выполняются условия:

1) в  $S$  компоненты вектор-функции  $g(x)$  имеют абсолютно непрерывные первые производные по всем аргументам;

$$2) \|R^{-1}(h, y)\| \leq a \text{ при } y \in S \text{ и } h \in (0, h_0];$$

$$3) \max_i \sum_{j=1}^n \left[ \sum_{m=1}^n \int_0^1 \left( \int_0^1 \frac{\partial^2 g_i}{\partial x_j \partial x_m} dt \right)^2 dt \right]^{1/2} < b \sqrt{3}, \quad (4)$$

где  $y \in S, z \in S, h \in (0, h_0], a$  аргумент второй производной имеет вид:  $w_{t\tau j} = y + \tau t(z - y) + (1 - \tau) t h e_j, z \neq y;$

$$4) \|h_k\| \leq c \|y_{k+1} - y_k\|, c - \text{произвольно, } k = 0, 1, \dots \quad (5)$$

Тогда, если начальное приближение  $y_0$  выбрано так, что

$$\beta = (1 + c) ab \|R^{-1}(h_0, y_0) [g(y_0) - d]\| < 1,$$

$$\omega(\beta) = \sum_{i=0}^{\infty} \beta^{2^i} \leq ab \varrho,$$

то в сфере  $S$  существует решение  $x$  системы  $g(x) = d$ , к которому сходится последовательность векторов  $y_k$ , получаемых по вычислительной схеме (3), причем быстрота сходимости оценивается неравенством:

$$\|y_k - x\| \leq \frac{1}{(1 + c) ab} \cdot \frac{\beta^{2^k}}{1 - \beta^{2^k}}. \quad (6)$$

Доказательство. Обозначим  $R_k = R(h_k, y_k), g_k = g(y_k)$ . Из (3) находим:

$$\begin{aligned} y_{k+1} - y_k &= R_k^{-1} [R_{k-1}(y_k - y_{k-1}) - (g_k - g_{k-1})] = \\ &= R_k^{-1} [R_{k-1} - \int_0^1 \Gamma(u_i^{(k-1)}) dt] (y_k - y_{k-1}), \end{aligned} \quad (7)$$

где  $u_i^{(k-1)} = y_{k-1} + t(y_k - y_{k-1})$ . Введем матрицу

$$A_{k-1} = \{a_{ij}^{(k-1)}\} = R_{k-1} - \int_0^1 \Gamma(u_i^{(k-1)}) dt$$

и рассмотрим ее элемент  $a_{ij}^{(k-1)}$ . Имеем:

$$\begin{aligned} a_{ij}^{(k-1)} &= \frac{g_i(y_{k-1} + h_{k-1} e_j) - g_i(y_{k-1})}{h_{k-1}} - \int_0^1 \left( \frac{\partial g_i}{\partial x_j} \right)_{u_i^{(k-1)}} dt = \\ &= \int_0^1 \left[ \left( \frac{\partial g_i}{\partial x_j} \right)_{v_{ij}^{(k-1)}} - \left( \frac{\partial g_i}{\partial x_j} \right)_{u_i^{(k-1)}} \right] dt. \end{aligned} \quad (8)$$

Здесь обозначено  $v_{ij}^{(k-1)} = y_{k-1} + t h_{k-1} e_j$ . Пусть  $s_{ij}$  — направление отрезка прямой, соединяющего точки  $u_i^{(k-1)}$  и  $v_{ij}^{(k-1)}$ .

Тогда

$$\left( \frac{\partial g_i}{\partial x_j} \right)_{v_{ij}^{(k-1)}} - \left( \frac{\partial g_i}{\partial x_j} \right)_{u_i^{(k-1)}} = t \|y_k - y_{k-1} - h_{k-1} e_j\| \int_0^1 \frac{\partial^2 g_i}{\partial x_j \partial s_{ij}} dt.$$

Производная под знаком интеграла в последнем выражении вычисляется при значении аргумента

$$\mathbf{w}_{\tau j}^{(k-1)} = \mathbf{y}_{k-1} + \tau t (\mathbf{y}_k - \mathbf{y}_{k-1}) + (1 - \tau) t h_{k-1} \mathbf{e}_j. \quad (9)$$

Таким образом,

$$\begin{aligned} a_{ij}^{(k-1)} &= \|\mathbf{y}_k - \mathbf{y}_{k-1} - h_{k-1} \mathbf{e}_j\| \int_0^1 t \int_0^1 \frac{\partial^2 g_i}{\partial x_i \partial s_{ij}} d\tau dt = \\ &= \|\mathbf{y}_k - \mathbf{y}_{k-1} - h_{k-1} \mathbf{e}_j\| \int_0^1 t \sum_{m=1}^n \left( \int_0^1 \frac{\partial^2 g_i}{\partial x_i \partial x_m} d\tau \right) \cos(x_m, s_{ij}) dt. \end{aligned}$$

Оценим норму матрицы  $A_{k-1}$ . Находим, используя неравенство Буняковского

$$\begin{aligned} \|A_{k-1}\| &= \max_i \sum_{j=1}^n |a_{ij}^{(k-1)}| \leq (\|\mathbf{y}_k - \mathbf{y}_{k-1}\| + |h_{k-1}|) \max_i \times \\ &\times \sum_{j=1}^n \int_0^1 t \left[ \sum_{m=1}^n \left( \int_0^1 \frac{\partial^2 g_i}{\partial x_i \partial x_m} d\tau \right)^2 \right]^{1/2} dt \leq (\|\mathbf{y}_k - \mathbf{y}_{k-1}\| + |h_{k-1}|) \max_i \times \\ &\times \sum_{j=1}^n \left[ \int_0^1 t^2 dt \cdot \sum_{m=1}^n \left( \int_0^1 \frac{\partial^2 g_i}{\partial x_i \partial x_m} d\tau \right)^2 \right]^{1/2} = (\|\mathbf{y}_k - \mathbf{y}_{k-1}\| + |h_{k-1}|) \times \\ &\times \frac{1}{\sqrt{3}} \max_i \sum_{j=1}^n \left[ \sum_{m=1}^n \int_0^1 \left( \int_0^1 \frac{\partial^2 g_i}{\partial x_i \partial x_m} d\tau \right)^2 dt \right]^{1/2}. \quad (10) \end{aligned}$$

Предположим пока, что выполняется оценка (4) из условия теоремы. Тогда из (7) и (10) получаем:

$$\|\mathbf{y}_{k+1} - \mathbf{y}_k\| \leq ab \|\mathbf{y}_k - \mathbf{y}_{k-1}\| (\|\mathbf{y}_k - \mathbf{y}_{k-1}\| + |h_{k-1}|). \quad (11)$$

Если еще выполняется и условие  $|h_{k-1}| \leq c \|\mathbf{y}_k - \mathbf{y}_{k-1}\|$ , то неравенство (11) можно заменить следующим:

$$\|\mathbf{y}_{k+1} - \mathbf{y}_k\| \leq (1 + c) ab \|\mathbf{y}_k - \mathbf{y}_{k-1}\|^2. \quad (12)$$

Чтобы установить справедливость неравенства (12), необходимо показать, что аргумент  $\mathbf{w}_{\tau j}^{(k-1)}$  второй производной в (10) и аргумент  $\mathbf{y}_k$  матрицы  $R_k^{-1}$  в (7) принадлежат сфере  $S$  при каждом  $k$ . Доказательство будем вести по индукции. При  $k=1$  имеем:

$$\|\mathbf{y}_1 - \mathbf{y}_0\| = \frac{\beta}{(1+c)ab} < \varrho,$$

$$\|\mathbf{w}_{\tau j}^{(0)} - \mathbf{y}_0\| \leq \|\mathbf{y}_1 - \mathbf{y}_0\| + |h_0| \leq (1+c)\|\mathbf{y}_1 - \mathbf{y}_0\| = \frac{\beta}{ab} < \varrho,$$

т. е.  $\mathbf{y}_1 \in S$ ,  $\mathbf{w}_{\tau j}^{(0)} \in S$ , поэтому для  $k=1$  неравенство (12) выполняется:

$$\|\mathbf{y}_2 - \mathbf{y}_1\| \leq (1+c)ab \|\mathbf{y}_1 - \mathbf{y}_0\|^2 = \frac{\beta^2}{(1+c)ab}.$$

Предположим теперь, что неравенство

$$\|y_{i+1} - y_i\| \leq \frac{\beta^{2^i}}{(1+c)ab} \quad (13)$$

имеет место для всех  $i = 1, 2, \dots, k-1$  и покажем справедливость этого неравенства при  $i = k$ . Находим оценки:

$$\|y_k - y_0\| \leq \sum_{i=0}^{k-1} \|y_{i+1} - y_i\| \leq \frac{1}{(1+c)ab} \sum_{i=0}^{k-1} \beta^{2^i} \leq \frac{\omega(\beta)}{(1+c)ab} \leq \varrho, \quad (14)$$

$$\|w_{t\tau j}^{(k-1)} - y_0\| \leq \|w_{t\tau j}^{(k-1)} - y_{k-1}\| + \|y_{k-1} - y_0\| \leq \frac{\omega(\beta)}{ab} \leq \varrho,$$

из которых следует, что  $y_k \in S$ ,  $w_{t\tau j}^{(k-1)} \in S$ . Поэтому выполняется неравенство (12), а следовательно и (13) при  $i = k$ . Из (13) заключаем, что последовательность векторов  $y_k$  сходится к некоторому пределу  $x$ , а переход к пределу при  $k \rightarrow \infty$  в равенстве

$$R_k(y_{k+1} - y_k) = g(y_k) - d \quad (15)$$

позволяет установить, что  $x$  является решением системы  $g(x) = d$ . Переход к пределу при  $m \rightarrow \infty$  в неравенстве

$$\|y_{k+m} - y_k\| \leq \sum_{i=k}^{k+m-1} \|y_{i+1} - y_i\| \leq \frac{\beta^{2^k}}{(1+c)ab} \cdot \frac{1}{1-\beta^{2^k}}$$

дает оценку быстроты сходимости, приведенную в условии теоремы.

**Замечание.** Согласно доказанной теореме, квадратическая сходимость итерационного процесса (3) будет обеспечена, если параметр  $h_k$  в (3) выбирать из условия  $|h_k| \leq c \|y_{k+1} - y_k\|$ . Этому условию можно удовлетворить только с помощью подбора, так как  $y_{k+1}$  зависит от  $h_k$ . Поэтому при практической реализации процесса (3) удобнее пользоваться условием  $|h_k| \leq c_1 \|g(y_k) - d\|$ , где  $c_1$  — произвольная константа. Из этого условия следует предыдущее, если положить  $c = c_1 \rho$ . Здесь  $\rho$  определяется неравенством  $\|R(h, y)\| \leq \rho$  при  $y \in S$ ,  $h \in (0, h_0]$ . Действительно,

$$|h_k| \leq c_1 \|g(y_k) - d\| = c_1 \|R_k R_k^{-1}(g_k - d)\| \leq c_1 \rho \|y_{k+1} - y_k\|.$$

**Теорема 2.** Если  $h_k = h = \text{const}$  и выполняются условия (1) — (3) теоремы 1, а начальное приближение  $y_0$  и параметр  $h$  выбраны так, что

$$\beta = 2ab \|R^{-1}(h, y_0)[g(y_0) - d]\| < 1,$$

$$\delta = 2ab |h| < 1,$$

$$|h| \left(1 + \frac{\beta}{1-\delta}\right) + \frac{\omega(\beta)}{ab} \leq \varrho,$$

то в сфере  $S$  существует решение  $x$  системы  $g(x) = d$ , к которому сходится последовательность векторов  $y_k$ , получаемых по вычислительной схеме (3).

*Быстрота сходимости при  $k \leq m$ , где  $m$  определяется соотношением*

$$\|y_{m+1} - y_m\| \leq |h| \leq \|y_m - y_{m-1}\|,$$

оценивается неравенством

$$\|y_k - x\| \leq \frac{1}{2ab} \left( \frac{1}{1 - \beta^{2k}} + \frac{\delta \beta^{2m-2k}}{1 - \delta} \right) \beta^{2k},$$

а при  $k \geq m + 1$  — неравенством

$$\|y_k - x\| \leq \frac{\beta^{2m}}{2ab(1 - \delta)} \delta^{k-m}.$$

Доказательство. Пусть выполняется неравенство  $|h| \leq \|y_1 - y_0\|$ . Тогда из теоремы 1 следует, что найдется такое число  $m$ , для которого имеет место соотношение

$$\|y_{m+1} - y_m\| \leq |h| \leq \|y_m - y_{m-1}\|,$$

причем при  $k = 0, 1, \dots, m$  справедлива оценка

$$\|y_{k+1} - y_k\| \leq \frac{1}{2ab} \beta^{2k}.$$

Если неравенство  $|h| \leq \|y_1 - y_0\|$  не выполняется, то доказательство теоремы необходимо начать со следующего предложения, положив  $m = 0$ .

Пусть  $k \geq m + 1$ . Покажем, что для таких  $k$  выполняется оценка

$$\|y_{k+1} - y_k\| \leq \delta \|y_k - y_{k-1}\|, \quad (16)$$

где  $\delta < 1$ . Проверим сначала выполнение неравенства (16) для  $k = m + 1$ . Для этого находим оценки:

$$\|y_{m+1} - y_0\| \leq \sum_{i=0}^m \|y_{i+1} - y_i\| \leq \frac{1}{2ab} \sum_{i=0}^m \beta^{2i} \leq \frac{\omega(\beta)}{2ab} < \varrho,$$

$$\begin{aligned} \|w_{\tau j}^{(m)} - y_0\| &\leq \|w_{\tau j}^{(m)} - y_m\| + \sum_{i=0}^{m-1} \|y_{i+1} - y_i\| \leq |h| + \sum_{i=0}^m \|y_{i+1} - y_i\| \leq \\ &\leq |h| + \frac{\omega(\beta)}{2ab} \leq \varrho, \end{aligned}$$

из которых следует, что при  $k = m + 1$  выполняется неравенство (11):

$$\|y_{m+2} - y_{m+1}\| \leq ab \|y_{m+1} - y_m\| (\|y_{m+1} - y_m\| + |h|).$$

Отсюда

$$\|y_{m+2} - y_{m+1}\| \leq 2ab |h| \cdot \|y_{m+1} - y_m\| = \delta \|y_{m+1} - y_m\|.$$

Аналогично устанавливается справедливость неравенства (16) и для  $k > m + 1$  на основе оценок:

$$\begin{aligned} \|y_k - y_0\| &\leq \sum_{i=m+1}^{k-1} \|y_{i+1} - y_i\| + \sum_{i=0}^m \|y_{i+1} - y_i\| \leq \frac{\beta^{2m}}{2ab} \sum_{i=m+1}^{k-1} \delta^{i-m} + \\ &+ \frac{1}{2ab} \sum_{i=0}^m \beta^{2i} \leq \frac{\beta^{2m}}{2ab} \cdot \frac{\delta}{1 - \delta} + \frac{\omega(\beta)}{2ab} < \varrho, \end{aligned}$$

$$\|w_{\tau j}^{(k-1)} - y_0\| \leq |h| + \sum_{i=0}^{k-1} \|y_{i+1} - y_i\| \leq |h| + \frac{\beta^{2m}}{2ab} \cdot \frac{\delta}{1 - \delta} + \frac{\omega(\beta)}{2ab} < \varrho.$$

Из (16) находим:

$$\|y_{h+s} - y_h\| \leq \sum_{i=h}^{h+s-1} \|y_{i+1} - y_i\| \leq \frac{\beta^{2m}}{2ab} \sum_{i=h}^{h+s-1} \delta^{i-m} = \frac{\beta^{2m}}{2ab} \cdot \frac{1 - \delta^s}{1 - \delta} \delta^{h-m}. \quad (17)$$

Последнее неравенство позволяет заключить, что последовательность векторов  $y_k$  сходится к некоторому пределу  $x$ . Предельный переход в (15) показывает, что  $x$  является решением системы уравнений  $g(x) = d$ . Оценку скорости сходимости итерационного процесса (3) можно получить из (17) для  $k \geq m+1$  путем предельного перехода при  $s \rightarrow \infty$ :

$$\|y_k - x\| \leq \frac{\beta^{2m}}{2ab(1-\delta)} \delta^{k-m}.$$

Для  $1 \leq k \leq m$  оценка скорости сходимости получается так:

$$\begin{aligned} \|y_k - x\| &\leq \|y_k - y_{m+1}\| + \|y_{m+1} - x\| \leq \sum_{i=k}^m \|y_{i+1} - y_i\| + \\ &+ \frac{\beta^{2m}}{2ab(1-\delta)} \delta \leq \frac{1}{2ab} \sum_{i=k}^m \beta^{2i} + \frac{\beta^{2m}}{2ab(1-\delta)} \delta \leq \frac{1}{2ab} \left( \frac{\beta^{2k}}{1 - \beta^{2k}} + \frac{\delta}{1 - \delta} \beta^{2m} \right). \end{aligned}$$

Теорема 2 показывает, что при постоянном  $h$  скорость сходимости итерационного процесса (3) хуже, чем при переменном  $h_k$ , удовлетворяющем неравенству (5). Однако, при достаточно малых  $h$ , как показывают оценки скорости сходимости, это ухудшение практически мало влияет на продолжительность итерационного процесса при получении приближенного решения с определенной точностью.

#### ЛИТЕРАТУРА

1. Л. В. Канторович, Г. П. Акилов, Функциональный анализ в нормированных пространствах, Физматгиз, 1959.
2. Д. К. Фаддеев, В. Н. Фаддеева, Вычислительные методы линейной алгебры, Физматгиз, 1960.

Поступила 28.IX 1965 г.  
Киев