Известия СПбГЭТУ «ЛЭТИ». 2022. Т. 15, № 10. С. 35–44
LETI Transactions on Electrical Engineering & Computer Science. 2022. Vol. 15, no. 10. P. 35–44

УДК 519.61 Научная статья

https://doi.org/10.32603/2071-8985-2022-15-10-35-44

Анализ точности и формальное обоснование численного решения нелинейных уравнений методами Ньютона и Хэйли при использовании арифметики с плавающей запятой

А. А. Чусов✉, Ю. И. Ефимова

Дальневосточный федеральный университет, Владивосток, Россия
✉ chusov.and@dvfu.ru

Аннотация. Представлено оригинальное исследование методов численного решения линейных и нелинейных алгебраических уравнений в контексте их реализуемости с использованием арифметики с плавающей запятой для нахождения решений нелинейных уравнений с помощью численной аппроксимации.

Распространение вычислительных устройств, которые реализуют некоторое подмножество действительных (и, таким образом, комплексных) чисел с помощью арифметики с плавающей запятой (точкой) требует учета специальных для этого случая особенностей отсутствия ассоциативности, катастрофических отмен точности, невозможности получения численных значений с произвольным числом разрядов.

Анализ научной литературы показывает скудость работ, изучающих методы численных решений нелинейных алгебраических уравнений при наличии ограничений и особенностей, налагаемых на числа с плавающей запятой, несмотря на общеизвестную важность этих аспектов, когда речь идет об общей точности, предсказуемости и удобстве использования методов решения нелинейных уравнений.

Проведенное исследование показало интересные результаты теоретического и экспериментального изучения хорошо известных методов Ньютона и Галлея с точки зрения их реализуемости и проблем, возникающих при использовании арифметики с плавающей запятой. С одной стороны, эксперименты демонстрируют соответствие между прогнозируемыми коэффициентами эффективности и теми, которые измеряются, однако, с другой стороны, эти измерения противоречат интуитивно предсказанному поведению методов, если не учитывать специфику с плавающей запятой двойной точности распространенного сегодня стандарта IEEE-754.

Ключевые слова: методы Хаусхолдера, метод Ньютона, метод Хэйли, численное решение нелинейных уравнений, плавающая запятая, плавающая точка, IEEE-754

Для цитирования: Чусов А. А., Ефимова Ю. И. Усечение падающего поля в задаче рассеяния электромагнитных волн на случайных поверхностях конечной длины // Изв. СПбГЭТУ «ЛЭТИ». 2022. Т. 15, № 10. С. 35–44. doi: 10.32603/2071-8985-2022-15-10-35-44.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Original article

Accuracy of Numerical Solvers of Non-Linear Equations Based on Newton's and Halley's Approximation Using Floating-Point Arithmetics

A. A. Chusov✉, Yu. I. Efimova

Far-Eastern Federal University, Vladivostok, Russia
✉ chusov.and@dvfu.ru

Abstract. The paper describes particularities of using floating-point arithmetics for finding solutions of non-linear equations by the means of numerical approximation.

Analysis of scientific literature shows scarcity of works studying these methods of numerical solvers in presence of limitations and particularities imposed by algebras of floating-point numbers despite well-known significance of these aspects when it comes to overall accuracy, predictability and usability of numerical solvers.

Therefore, the paper describes some interesting results of theoretical and experimental study of these well-known Newton's and Halley's methods from the point of view of their implementability and problems that arise when floating-point arithmetic is used. The analysis is conducted both theoretically and experimentally using floating-point machines. On one hand, the experiments demonstrate correspondence between the predicted efficiency factors and ones that are measured, but on the other hand these measurements contradict intuitively predicted behavior of solvers if no floating-point specifics are taken into account.

Keywords: Householder methods, Newton's method, Halley's method, numerical solvers of non-linear equations, floating-point arithmetics, IEEE-754

For citation: Chusov A. A., Efimova Yu. I. Accuracy of Numerical Solvers of Non-Linear Equations Based on Newton's and Halley's Approximation Using Floating-Point Arithmetics // LETI Transactions on Electrical Engineering & Computer Science. 2022. Vol. 15, no. 10. P. 35–44. doi: 10.32603/2071-8985-2022-15-10-35-44.

Conflict of interest. The authors declare no conflicts of interest.

Введение. Актуальность методов приближенного вычисления корней алгебраических уравнений, задаваемых дифференцируемыми функциями, обусловлена широтой области их применения. Например, они используются при численном решении систем дифференциальных уравнений, математическом и компьютерном моделировании физических процессов, математические модели которых заданы системами дифференциальных уравнений, при минимаксном приближении функций и информационных сигналов с помощью полиномов с гарантированной для наихудшего случая точностью.

Вместе с тем, распространение вычислительных устройств, которые реализуют некоторое подмножество действительных (и таким образом комплексных) чисел с помощью арифметики с плавающей запятой (точкой), требует учета специальных для этого случая особенностей – отсутствия ассоциативности, катастрофических отмен точности, невозможности получения численных значений с произвольным числом разрядов. Поэтому многие методы численной обработки цифровых сигналов и данных исследуются в наиболее современных научных публикациях применительно к давно известным классическим алгоритмам и численным методам (преобразование Фурье, численное решение матричных и линейных уравнений, аспекты приближения дифференциальных и интегральных уравнений, свертки), а результаты этих исследований показывают определяющую зависимость результативности численного метода от особенностей реализации над алгеброй чисел с плавающей точкой.

Вместе с тем, анализ источников показывает, что рассмотрение методов численного решения алгебраических уравнений и их систем методами Хаусхолдера (включая методы Ньютона и Хэйли) в литературе представлено слабо и не проработано в степени, достаточной для их практической реализации в общем случае.

Постановка задачи. В настоящей работе выполняется такой анализ как с теоретических позиций, но с учетом вышеуказанных особенностей алгебры чисел с плавающей запятой, так и на основе выполненных экспериментально измерений результативности приближенного вычисления корня на основе чисел с плавающей запятой двойной точности распространенного сегодня стандарта IEEE-754.

Методы численного решения алгебраических уравнений в поле действительных чисел. Метод Ньютона (метод касательных, метод секущих) – алгоритм нахождения корней дифференцируемых функций, который использует итеративную функцию для аппроксимации ее корней [1]. Это простой и эффективный алгоритм для приближенного нахождения корней действительнозначных функций, т. е. решения уравнений вида $f(x) = 0$. Единственные требования, накладываемые на функцию, – чтобы у нее был хотя бы один корень и чтобы она была непрерывна и дифференцируема на интервале поиска, а ее производная не принимала нулевое значение на этом интервале.

Алгоритм начинается с начального приближения, которое задается вблизи предположительного корня, после чего строится касательная к графику исследуемой функции в точке приближения, для которой находится пересечение с осью абсцисс. Эта точка берется в качестве следующего приближения. И так далее, пока не будет достигнута необходимая точность. В случае нахождения численного решения уравнения $f(x) = 0$ формула сводится к итерационной процедуре вычисления:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \qquad (1)$$

Считается, что метод Ньютона обладает квадратичной сходимостью, т. е. на каждой итерации абсолютное отклонение приближенного значения корня от истинного возводится в квадрат, т. е. число верных знаков удваивается [2]. Метод Ньютона крайне важен в вычислительной математике: в большинстве случаев именно он используется для нахождения численных решений уравнений.

Метод Хэйли (Halley's method, tangent hyperbolas method, метод касательных гипербол) представляет собой алгоритм нахождения корня, используемый для дважды дифференцируемых функций одной действительной переменной с непрерывной второй производной [3]. Алгоритм является вторым в классе методов Хаусхолдера после метода Ньютона. Метод Хэйли для решения нелинейных уравнений рассматривается некоторыми как расширение метода Ньютона.

Он состоит из последовательности итераций

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)}. \qquad (2)$$

Преимущество метода Хэйли по сравнению с методом Ньютона заключается в том, что он сходится быстрее (у метода Хэйли кубическая сходимость), однако для этого требуется вычисление второй производной функции *f*, как видно из (2). Поскольку эти два метода часто взаимозаменяемы (т. е. для данного корня условия сходимости выполняются для обоих методов), было бы предпочтительнее использовать метод Хэйли, если вторая производная от функции существует и ее несложно вычислить.

Особенности реализации арифметики с плавающей запятой. Термин «число с плавающей запятой» представляет собой экспоненциальную форму представления вещественных (действительных) чисел. Число с плавающей запятой состоит из частей: знак мантиссы, который показывает отрицательность или положительность числа, мантисса, которая выражает значение числа без учета порядка, знак порядка и сам порядок, который выражает степень основания системы счисления, на которую умножается мантисса.

В научной литературе асимптотический анализ точности применительно к различным унарным функциям представлен широко, однако использование для аппроксимации технических средств, в большинстве случаев опирающихся на представление множества действительных значений с помощью чисел с плавающей запятой, требует рассмотрения еще одного источника ошибок машинных вычислений, связанных с ограниченностью числа разрядов мантиссы, и этот анализ применительно к методам численного решения уравнений в литературе представлен слабо, поэтому в настоящей статье выполнено теоретическое и экспериментальное исследование методов Ньютона и Хэйли при их реализации числами с плавающей запятой.

Также стоит заметить, что если для реализации вычислений над действительными или комплексными значениями используется представление с плавающей запятой с фиксированной точностью мантиссы и в алгоритме присутствует вычитание или сложение, как в (1) и (2), то можно столкнуться с катастрофической отменой точности. После таких вычислений происходит аннулирование старших разрядов операндов и результат таких вычислений будет далеким от истинного, так как на выход возвращается число, которое сформировано только младшими разрядами аргументов, которые могут быть потеряны в результате промежуточных вычислений. Это – серьезная проблема для десятично-двоичных вычислений [4].

Для снижения остроты данной проблемы с числами с плавающей запятой некоторые реализации арифметики с плавающей запятой поддерживают совмещение мультипликативной и последующей аддитивной операций с последующим однократным вместо двукратного округлением. Такие операции называются fused multiply-add (FMA). Это позволяет более эффективно реализовать операции деления и извлечения квадратного корня (при отсутствии аппаратной реализации), умножение векторов и матриц, вычисление полиномов по схеме Горнера.

Поскольку наиболее распространенные реализации арифметики с плавающей запятой регламентированы стандартом IEEE-754 с описанными в нем форматами binary_32 (известным также как тип данных float) и binary_64 (double), представленные далее результаты соответствуют одному из них – формату binary_64 с точностью 53 двоичных разряда и с разрядностью экспоненты 11 бит. Класс решаемых задач не предполагает получения чрезмерно большой или малой экспоненты, поэтому, предполагая, что магнитуда численных значений на всех этапах алгоритмов аппроксимации всегда принадлежит интервалу $\left[2^{(-2^{11}+1)}; 2^{(2^{11}-1)}\right]$, разрядность экспоненты считается неограниченной, а число – всегда нормализованным.

При численном анализе точности представления и операций над числами с плавающей запятой используется понятие ULP (Unit-In-the-Last-Place), которое для заданного $x = m\beta^e$ (где m – мантисса из интервала [1, 2), e – экспонента и β – основание системы счисления) представляет собой значение $1\text{ulp} = \beta^{e-(p-1)}$ и показывает магнитуду денормализованного числа с минимальной мантиссой 0.0…01 и экспонентой e, равной экспоненте числа x, где p – это число двоичных защитных разрядов. В большинстве случаев – кроме тех, когда действительное значение, аппроксимируемое числом с плавающей точкой, принадлежит интервалу $(\beta^E - \beta^{E-p}, \beta^E)$ для некоторого E, – единица ULP равна расстоянию между двумя последовательными числами с плавающей запятой с экспонентой e, т. е. значение, которое представляет наименьшая значащая (крайняя правая) единица. Единица ULP используется в качестве меры точности при числовых вычислениях. Например, проверка на равенство двух значений с плавающей запятой, имеющих равные во всех разрядах экспоненты E, с допустимой ошибкой в 16 ulp, предполагает, что младшие 4 бита (значение, равное логарифму числа 16 по основанию 2) могут не совпадать, однако числа все равно будут считаться равными [5].

Спецификация IEEE-754, которой следуют все современные аппаратные средства с плавающей запятой, требует, чтобы результат элементарной арифметической операции (сложение, вычитание, умножение, деление и квадратный корень) был правильно округлен, что подразумевает, что при округлении до ближайшего округленный результат находится в пределах 0.5 ulp. Авторитетные числовые библиотеки вычисляют основные трансцендентные функции с точностью от 0.5 до примерно 1 ulp, только несколько библиотек вычисляют их в пределах 0.5 ulp, потому что это – сложная задача.

Теоретическое обоснование численной аппроксимации исследуемых уравнений. В настоящем разделе выполняется анализ эффективности двух вышеуказанных методов численного решения уравнений, заданных унарной дифференцируемой функцией, определенной на алгебре чисел с плавающей запятой. В качестве функций выбраны отображения $f(x) = x^2 - 2$ и $f(x) = x^2 - 2$ $g(x) = \sin e^x - 1,$ истинные значения корней которых равны, соответственно, алгебраическому $\sqrt{2}$ и одному из трансцендентных значений из $\{\ln(\sin^{-1}(1 + 2\pi k)) \mid k \in \mathbb{N}_0\}$ в зависимости от выбора начального значения x_0 в (1). Для оценки и сравнения результативности решения этих двух уравнений измеряются следующие показатели эффективности: относительная ошибка в процентах и в единицах на последнем месте ULP.

Сначала проанализируем применение двух вышеописанных методов численной аппроксимации решения квадратного уравнения $x^2 - 2 = 0.$ Единственное в области неотрицательных действительных чисел «истинное» значение корня этого уравнения во множестве неотрицательных действительных чисел приведено в [6] с точностью до 20 000 десятичных разрядов. Для его отображения на число формата binary_64 с ошибкой округления, не превосходящей 0.5 ulp, необходимо получение p = 53 бит точности плюс некоторое малое количество ε бит, в литературе называемых защитными, для определения ближайшего к «истинному» значения с плавающей запятой. Для обеих функций достаточным оказывается использование трех защитных двоичных разрядов, т. е. $\lceil (p + \varepsilon)\log_{10}\beta \rceil = \lceil (53 + 3)\log_{10} 2 \rceil = 17$ десятичных разрядов, поэтому на основе [6] в качестве «истинного» значения корня уравнения $f(x)$ = 0 принимается число 1.4142135623730950.

Для перевода мантиссы числа с плавающей запятой из одной позиционной системы счисления, по основанию β_1, в иную, по основанию β_2, необходимо сопоставление запятой с учетом значений экспоненты. Действительно, без такого сопоставления позиция (значимость) разрядов системы счисления β_2 будет определяться целочисленной значимостью разрядов β_1 тогда и только тогда, когда β_1 является целой степенью β_2. В противном же случае перевод мантисс возможен только в отношение целочисленной части мантиссы, выраженной в системе счисления β_1, в целом и соответствующей целочисленной поправкой экспоненты.

Более формально, число для любых ненулевых целочисленных β_1 и β_2, любых $d_i \in \mathbb{Z}_{\beta_1}$ и $e_i \in \mathbb{Z}_{\beta_2}$, а также действительного

$$x = \beta_1^{E_1} \sum_{i=-\infty}^{\infty} d_i \beta_1^i = \sum_{i=0}^{\infty} d_i \beta_1^{i+E_1} + \sum_{i=-\infty}^{-1} d_i \beta_1^{i+E_1} =$$

$$= \sum_{i=0}^{\infty} e_i \beta_2^{i+E_2} + \sum_{i=-\infty}^{-1} e_i \beta_1^{i+E_2},$$

значение выбранной части мантиссы в левой части, $\sum_{i=n_1}^{n_2} d_i \beta_1^{i+E_1}$, может быть равно некоторой части $\sum_{i=m_1}^{m_2} d_i \beta_2^{i+E_2}$ мантиссы справа, только если $\beta_1^{n+E_1} = \beta_2^{m_2+E_2}$. Поэтому в общем случае, когда $\beta_1^{n+E_1}$ не является целой степенью $\beta_2^{m_2+E_2}$ (как рассматриваемый случай с $\beta_1 = 10$ и $\beta_2 = 2$), необходимо выбрать такие части x, т. е. n_2, m_2 и E_2, чтобы оба вышеуказанных равенства выполнялись, и это будет всегда справедливо, если $n_1 + E_1 = m_1 + E_2 = 0$.

Поэтому, если с учетом защитных разрядов требуется точность $p_\varepsilon = p + \varepsilon$ выражения x в системе счисления β_2, используется преобразование (в предположении, что n наиболее старший ненулевой разряд, $\forall nN(d_n \neq 0 \wedge N > n \Rightarrow d_n = 0)$)

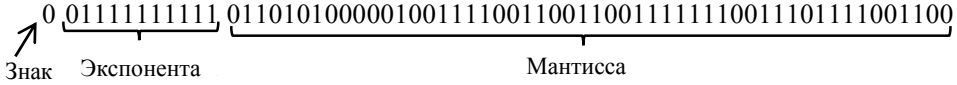
$$x \approx \beta_1^{E_1} \sum_{i=-0}^{n} d_i \beta_1^i = \beta_2^{\left|E_1 \log_{\beta_2} \beta_1\right| - p_\varepsilon} \times \times \sum_{i=0}^{n} d_i \beta_1^{i+E_1} \beta_2^{p_\varepsilon - \left|E_1 \log_{\beta_2} \beta_1\right|}, \quad (3)$$

в котором значение справа от знака суммы оказывается приближенным значением мантиссы, а показатель степени при β_2 слева – экспонентой E_2.

Тогда двоичное представление десятеричного приближения числа $\sqrt{2}$ в формате с плавающей запятой двойной точности может быть получено выделением как минимум 56-битной целочисленной части с соответствующей поправкой экспоненты следующим образом (здесь индекс с правой стороны мантиссы показывает основание используемой для записи числа системы счисления):

$$1.4142135623730950 = (1.4142135623730950 \cdot 2^{56}) \cdot 2^{-56} =$$
$$= 10190482676041235.51744875528192_{10} \cdot 2^{-56} \approx$$
$$\approx 10110101000001001111001100110011111111001110111001100011 0_2 \cdot 2^{-56} \approx$$
$$\approx 10110101000001001111001100110011111111001110111001100.0110_2 \cdot 2^{-56} \approx$$
$$\approx 10110101000001001111001100110011111111001110111001100 \cdot 2^0.$$

В формате binary_64 стандарта IEEE-754 число будет выглядеть так:



Приведенное значение в шестнадцатеричной системе счисления равно 0x3FF6A09E 667F3BCC, которое является «истинным», т. е. с 53 корректными двоичными разрядами мантиссы.

Реализация методов Ньютона и Хэйли для численного решения уравнения $f(x) = x^2 - 2 = 0$ осуществляется следующим образом. Правая часть формулы (1) метода Ньютона равна

$$f_{\mathrm{N}}(x) = -\frac{f(x)}{f'(x)} = -\frac{x^2 - 2}{2x}. \quad (4)$$

Вычисление (4) может приводить к катастрофической отмене арифметики с плавающей запятой вследствие аддитивной операции над значениями с разными знаками в знаменателе. Однако подстановка (4) в (1) после упрощений исключает условие отмены:

$$x_{n+1} = \frac{x_n^2 + 2}{2x_n}.$$

Аналогичен источник катастрофической потери точности при аппроксимации методом Хэйли: подставляя определение $f(x)$ в (2), имеем

$$f_{\mathrm{H}}(x) = -\frac{2f(x)f'(x)}{2(f'(x)^2 - f(x)f''(x)} = -\frac{4x^3 - 8x}{6x^2 + 4}.$$

Аналогичным же образом этот источник можно исключить:

$$x_{n+1} = \frac{2x_n^3 + 12x_n}{6x_n^2 + 4}.$$

Графически уравнение и его решения методами Ньютона и Хэйли представлены на рис. 1.

Теперь рассмотрим уравнение $g(x) = \sin e^x - 1 = 0$ и его решение двумя методами. Для этой функции правая часть (1) метода Ньютона имеет вид

$$g_{\mathrm{N}}(x) = -\frac{\sin e^x - 1}{e^x \cos e^x}.$$

С помощью (2) для метода Хэйли уравнение запишется следующим образом:

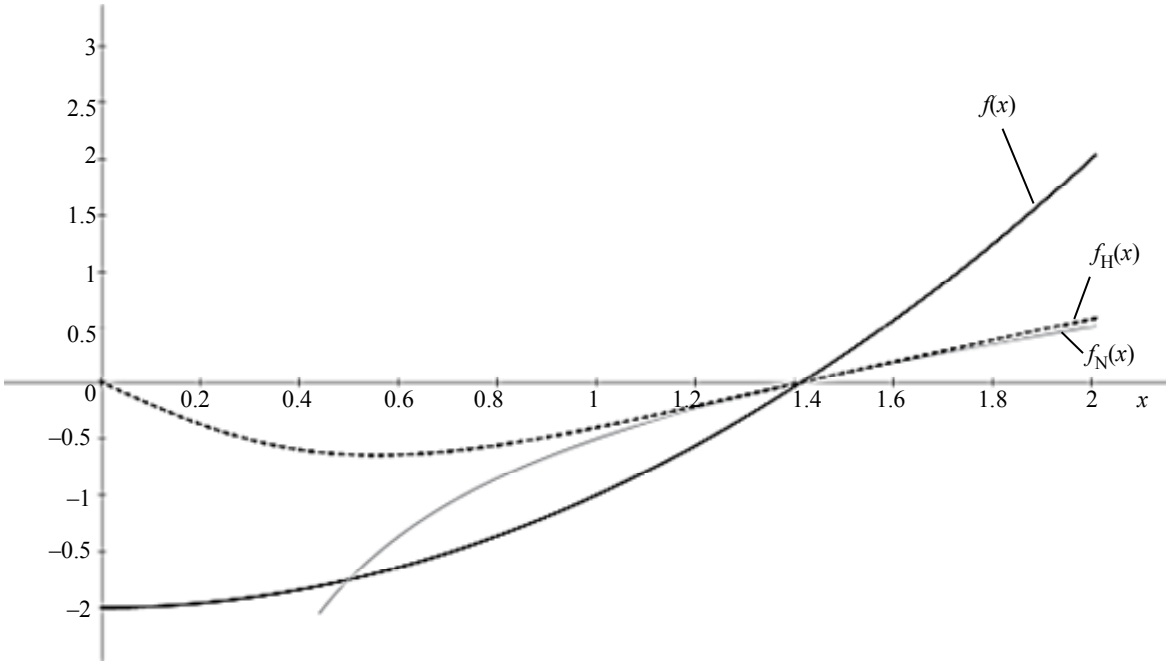


Рис. 1. Графическое представление $f(x)$, а также поправок $f_N(x)$ и $f_H(x)$, используемых соответственно в методах Ньютона и Хэйли

Fig. 1. Plots of $f(x)$ together with adjustments $f_N(x)$) and $f_H(x)$, used as part of the Newton's and Halley's methods

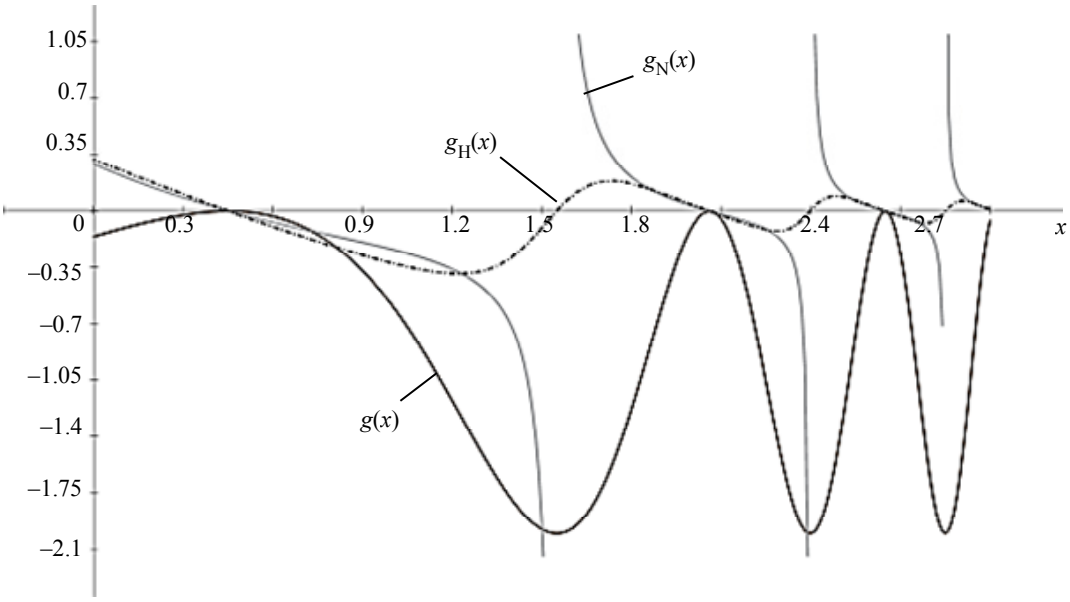


Рис. 2. Графическое представление функции $g(x)$, поправок в реализациях методов Ньютона $g_N(x)$ и Хэйли $g_H(x)$

Fig. 2. Visualization of the original function $g(x)$ together with the Newton's $g_N(x)$ and Halley's $g_H(x)$ adjustments

$$g_H(x) = -\frac{2e^x \cos e^x \left(\sin e^x - 1\right)}{2\left(e^x \cos e^x\right)^2 - \left(\sin e^x - 1\right)\left(e^x \cos e^x - e^{2x} \sin e^x\right)}.$$

Графически функция $g(x)$ и решения методами Ньютона и Хэйли представлены на рис. 2.

На рис. 3 видно, что функция $g(x)$ и функции двух методов имеют ноль в одной точке.

Для нахождения «истинного» значения сначала решается уравнение

$$\sin e^x - 1 = 0 \Rightarrow e^x = \frac{\pi}{2} + 2k\pi,\ k \in \mathbb{Z},$$

$$\Rightarrow x = \ln\left(\frac{\pi}{2} + 2k\pi\right)\Big|_{k=0} = \ln\left(\frac{\pi}{2}\right).$$

Аналогично реализации $g(x)$ для получения корректного приближения с плавающей точкой,

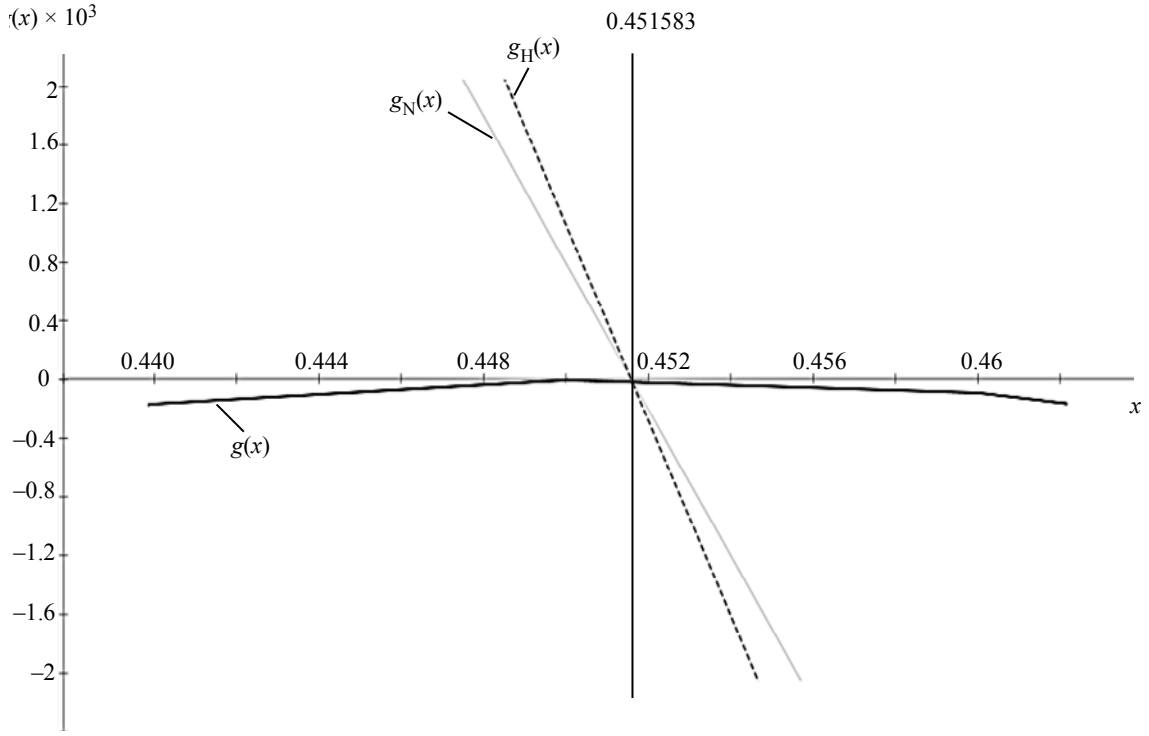


Рис. 3. Графическое представление первого нуля функции $g(x)$, а также аппроксимированных решений методами Ньютона $g_N(x)$ и Хэйли $g_H(x)$

Fig. 3. Visualization of the first zero of the function $g(x)$ as well as approximate solutions which result from employment of the Newton's $g_N(x)$ and Halley's $g_H(x)$ methods

отличающегося от истинного не более чем на 0.5 ulp, требуется 17 верных десятичных разрядов истинного корня, полученного вычитанием $\ln(\pi) - \ln(2)$ и последующим округлением разницы значений логарифмов, приведенных в [6] с точностью в 2000 десятичных разрядов:

$$x = 0.45158270528945486.$$

Отсюда с учетом (3):

$$0.45158270528945486_{10} = \left(0.45158270528945486_{10} \cdot 10^{-1} \cdot 2^{56-\lfloor -\log_2 10 \rfloor}\right) \cdot 2^{56-\lfloor -\log_2 10 \rfloor - 56} \approx$$

$$\approx 260319706018374325_{10} \cdot 2^{-59} =$$

$$= 1110011100111010111011001001011010101000010011000101 01.10101_2 \cdot 2^{-59} \approx$$

$$\approx 1110011100111010111011001001011010101000010011000101 01_2 \cdot 2^{-59} =$$

$$= 1.1100111001110101110110010010110101010000100110001010 1_2 \cdot 2^{-2} = CE6BB25AA1316.$$

Т. е. 0x3FDCE6BB25AA1316 – решение уравнения $g(x) = 0$ с в формате binary_64 с 53 корректными двоичными разрядами мантиссы.

Эксперименты. На рис. 4 и 5 отображены результаты проведенного экспериментального исследования. Из рисунков видно, что ошибка, которая вытекает из применения метода Хэйли с увеличением количества итераций снижается быстрее, чем у метода Ньютона.

Более подробно результаты измерений результативности представлены в табл. 1–4. В табл. 1 и 2 видно, что аппроксимация методом Ньютона, реализованным над числами двойной точности, приводит к значению, корректному во всех разрядах, т. е. решение и «истинное» значение совпали полностью, а минимальная полученная ошибка вычислений методом Хэйли в наилучшем случае равна 1 ulp, что вызвано бóльшим числом операций и, следовательно, округлений на каждой итерации.

В табл. 3 и 4 видно, что в двух методах достигается маленькая относительная погрешность, а вот ulp слишком большой.

Обсуждение результатов. Эксперименты показывают интересные результаты: например, несмотря на соответствие предсказанной сходимости методов фактической, более быстро сходящийся метод Хэйли в некоторых случаях оказывается менее точным по сравнению с методом Ньютона, достигая меньшего предела точности,

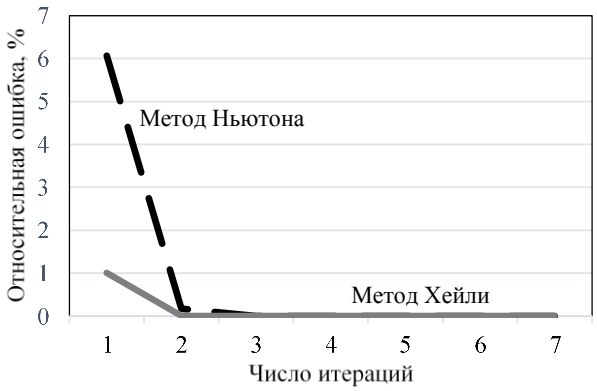


Рис. 4. Зависимость относительной ошибки от числа итераций для аппроксимации функции $f(x) = x^2 - 2$

Fig. 4. Relative error with respect to a number of iterations employed to approximate $f(x) = x^2 - 2$

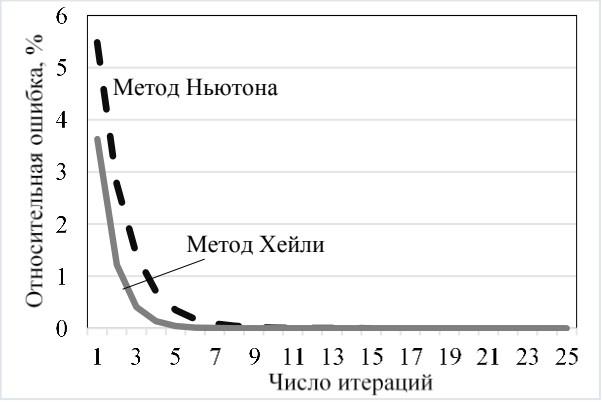


Рис. 5. Зависимость относительной ошибки от числа итераций для аппроксимации функции $g(x) = \sin e^x - 1$

Fig. 5. Relative error with respect to a number of iterations employed to approximate $g(x) = \sin e^x - 1$

Табл. 1. Решения уравнения $f(x) = x^2 - 2$, полученные методом Ньютона

Tab. 1. Solutions for $f(x) = x^2 - 2$ obtained by the implementation of the Newton's method

Номер итерации	Решение в двоичной системе счисления	Решение в шестнадцатеричной системе счисления	«Истинное значение»	ULP	Относительная ошибка, %
1	1.5	3ff8000000000000	3ff6a09e667f3bcc	3.86E+14	6.06602
2	1.41667	3ff6aaaaaaaaaaa	3ff6a09e667f3bcc	1.10E+13	0.173461
3	1.41422	3ff6a0a0a0a0a0a0	3ff6a09e667f3bcc	9.57E+09	0.000150183
4	1.41421	3ff6a09e667f3bcc	3ff6a09e667f3bcc	7183	1.1278E-10
5	1.41421	3ff6a09e667f3bcc	3ff6a09e667f3bcc	0	0
6	1.41421	3ff6a09e667f3bcc	3ff6a09e667f3bcc	0	0

Табл. 2. Решения уравнения $f(x) = x^2 - 2$ полученные методом Хэйли

Tab. 2. Solutions for $f(x) = x^2 - 2$ obtained by the implementation of the Halley's method

Номер итерации	Решение в двоичной системе счисления	Решение в шестнадцатеричной системе счисления	«Истинное значение»	ULP	Относительная ошибка, %
1	1.4	3ff6666666666666	3ff6a09e667f3bcc	6.40E+13	1.00505
2	1.41421	3ff6a09e04ad9cb8	3ff6a09e667f3bcc	1.64E+09	2.57672E-05
3	1.41421	3ff6a09e667f3bcd	3ff6a09e667f3bcc	1	1.57009E-14
4	1.41421	3ff6a09e667f3bcd	3ff6a09e667f3bcc	1	1.57009E-14

которая достаточно быстро оказывается ограничена разрядностью мантиссы числа в противоположность количеству итераций выбранного метода. Таким образом показано, что увеличение порядка выбранного метода Хаусхолдера для численной аппроксимации, хотя и увеличивает скорость сходимости, может приводить к увеличению минимальной ошибки, обусловленной увеличенным количеством вычислений.

Также показано, что вид аппроксимируемого уравнения также определяет минимально достижимую ошибку численного приближения вследствие большего числа аппроксимаций из-за большего числа элементарных операций над числами с плавающей запятой. В частности, шаг, реализующий аппроксимацию второго уравнения, допускает возникновение катастрофической отмены, при которой степень близости истинных значений с одинаковым знаком, которые подвергаются вычитанию с предварительной аппроксимацией числами с плавающей запятой, линейно увеличивает относительную ошибку вычислений.

Выводы и заключение. В статье описываются особенности использования арифметики с плавающей запятой для нахождения решений нелинейных уравнений с помощью численной аппроксимации.

Табл. 3. Решения уравнения $g(x) = \sin e^x - 1$ методом Ньютона
Tab. 3. Solutions for $g(x) = \sin e^x - 1$ obtained by the implementation of the Newton's method

Номер итерации	Решение в двоичной системе исчисления	Решение в шестнадцатеричной системе счисления	«Истинное значение»	ULP	Относительная ошибка, %
1	0.476356	3fde7c9e4c4da29c	3fdce6bb25aa1315	4.46E+14	5.4859
2	0.46412	3fddb42455c692dc	3fdce6bb25aa1315	2.26E+14	2.7763
3	0.45789	3fdd4e1304a7381e	3fdce6bb25aa1315	1.14E+14	1.39677
4	0.454746	3fdd1a909ac6a4bd	3fdce6bb25aa1315	5.70E+13	0.700577
5	0.453167	3fdd00b058a1fa9a	3fdce6bb25aa1315	2.85E+13	0.350841
6	0.452376	3fdcf3b860294cff	3fdce6bb25aa1315	1.43E+13	0.175559
7	0.451979	3fdced3a6b882a16	3fdce6bb25aa1315	7.14E+12	0.087815
8	0.451781	3fdce9faf2cc79fb	3fdce6bb25aa1315	3.57E+12	0.043916
9	0.451682	3fdce85b16c9953e	3fdce6bb25aa1315	1.79E+12	0.02196
10	0.451632	3fdce78b20dd923e	3fdce6bb25aa1315	8.93E+11	0.010981
11	0.451607	3fdce72323ecab62	3fdce6bb25aa1315	4.47E+11	0.00549
12	0.451595	3fdce6ef24f55e6f	3fdce6bb25aa1315	2.23E+11	0.002745
13	0.451589	3fdce6d5255a9513	3fdce6bb25aa1315	1.12E+11	0.00137264
14	0.451586	3fdce6c82585b5c8	3fdce6bb25aa1315	5.58E+10	0.000686322
15	0.451584	3fdce6c1a597fdd4	3fdce6bb25aa1315	2.79E+10	0.000343161
16	0.451583	3fdce6be65a2ae9e	3fdce6bb25aa1315	1.40E+10	0.000171582
17	0.451583	3fdce6bcc5acb802	3fdce6bb25aa1315	6.98E+09	8.58E-05
18	0.451583	3fdce6bbf5ace52f	3fdce6bb25aa1315	3.49E+09	4.29E-05
19	0.451583	3fdce6bb8dadb999	3fdce6bb25aa1315	1.757E+09	2.15E-05
20	0.451583	3fdce6bb59cf2d68	3fdce6bb25aa1315	8.757E+08	1.08E-05
21	0.451583	3fdce6bb3ff172da	3fdce6bb25aa1315	4.41E+08	5.42E-06
22	0.451583	3fdce6bb321fe60a	3fdce6bb25aa1315	2.09E+08	2.57E-06
23	0.451583	3fdce6bb2df61467	3fdce6bb25aa1315	1.39E+08	1.71E-06
24	0.451583	3fdce6bb27b56fb4	3fdce6bb25aa1315	3.43E+07	4.21E-07
25	0.451583	3fdce6bb27b56fb4	3fdce6bb25aa1315	3.43E+07	4.21E-07

Табл. 4. Решения уравнения $g(x) = \sin e^x - 1$ методом Хэйли
Tab. 4. Solutions for $g(x) = \sin e^x - 1$ obtained by the implementation of the Halley's method

Номер итерации	Решение в двоичной системе исчисления	Решение в шестнадцатеричной системе счисления	«Истинное значение»	ULP	Относительная ошибка, %
1	0.467986	3fddf37d39f78965	3fdce6bb25aa1315	2.96E+14	3.63249
2	0.457081	3fdd40cf416dc9a5	3fdce6bb25aa1315	9.90E+13	1.21749
3	0.453419	3fdd04cff6776ec4	3fdce6bb25aa1315	3.31E+13	0.406574
4	0.452195	3fdcf0c3a8595401	3fdce6bb25aa1315	1.10E+13	0.135608
5	0.451787	3fdcea137df9b839	3fdce6bb25aa1315	3.68E+12	0.045212
6	0.451651	3fdce7d89d688422	3fdce6bb25aa1315	1.23E+12	0.015072
7	0.451605	3fdce71a4e21cbf1	3fdce6bb25aa1315	4.09E+11	0.005024
8	0.45159	3fdce6dadde142ce	3fdce6bb25aa1315	1.36E+11	0.001675
9	0.451585	3fdce6c5b8693e59	3fdce6bb25aa1315	4.54E+10	0.000558
10	0.451584	3fdce6beabe8d044	3fdce6bb25aa1315	1.51E+10	0.000186
11	0.451583	3fdce6bc526a189e	3fdce6bb25aa1315	5.05E+09	6.20E-05
12	0.451583	3fdce6bb89d9867e	3fdce6bb25aa1315	1.68E+09	2.07E-05
13	0.451583	3fdce6bb46d20417	3fdce6bb25aa1315	5.56E+08	6.84E-06
14	0.451583	3fdce6bb2f93b2bc	3fdce6bb25aa1315	1.66E+08	2.05E-06
15	0.451583	3fdce6bb28777254	3fdce6bb25aa1315	4.70E+07	5.78E-07
16	0.451583	3fdce6bb28777254	3fdce6bb25aa1315	4.70E+07	5.78E-07

Для этого выполнены теоретическое и экспериментальное исследования и обоснования методов Ньютона и Галлея с точки зрения их реализуемости и проблем, возникающих при использовании арифметики с плавающей запятой. С одной стороны, эксперименты демонстрируют соответствие между прогнозируемыми показателями эффективности и теми, которые измеряются, но, с другой стороны, эти измерения противоречат интуитивно предсказанному поведению методов, если не учитывать специфику с плавающей запятой двойной точности стандарта IEEE-754.

Результаты статьи можно использовать при выборе методов численной обработки цифровых сигналов и в их реализации, так как исследование показывает определяющую зависимость результативности численного метода от особенностей реализации над алгеброй чисел с плавающей точкой.

Список литературы

1. Chun-Hua G. On Newton's method and Halley's method for the principal root of a matrix // Regina, Linear Algebra and its Appl. 2009. Vol. 432, no. 8. P. 1905–1922.

2. Eagan N., Hauser G., Flaherty T. Newton's method on a system of nonlinear equations. Pittsburgh: Carnegie Mellon University, 2014. 14 p.

3. Naseem A., Rehman M. A., Abdeljawad T. Numerical methods with engineering applications and their visual analysis via polynomiography // IEEE. 2021. Vol. 9. P. 99287–99298. doi: 10.1109/ACCESS.2021.3095941.

4. Harrison J. A machine-checked theory of floating-point arithmetic // Proc. of the 1999 Intern. Conf. on Theorem Proving in Higher Order Logics. Nice, 1999. P. 113–130. doi: 10.1007/3-540-48256-3_9.

5. Quinnell E., Swartzlander E. E., Lemonds C. Floating-point fused multiply-add architectures // Conf. Record of the Forty-First Asilomar Conf. on Signals, Systems and Computers. Pacific Grove, 2008. P. 331–337. doi: 10.1109/ACSSC.2007.4487224.

6. The On-Line Encyclopedia of Integer Sequences, Decimal expansion of square root of 2. URL: https://oeis.org/A002193 (дата обращения 28.09.2022).

Информация об авторах

Чусов Андрей Александрович – канд. техн. наук, доцент департамента электроники, телекоммуникации и приборостроения Политехнического института, Дальневосточный федеральный университет, г. Владивосток, о. Русский, п. Аякс, 10, Россия.
E-mail: chusov.aa@dvfu.ru
http://orcid.org/0000-0002-7931-5368

Ефимова Юлия Игоревна – магистрант департамента электроники, телекоммуникации и приборостроения Политехнического института, Дальневосточный федеральный университет, г. Владивосток, о. Русский, п. Аякс, 10, Россия.
E-mail: efimova.iui@students.dvfu.ru
http://orcid.org/0000-0002-7694-0694

References

1. Chun-Hua G. On Newton's method and Halley's method for the principal root of a matrix // Regina, Linear Algebra and its Appl. 2009. Vol. 432, no. 8. P. 1905–1922.

2. Eagan N., Hauser G., Flaherty T. Newton's method on a system of nonlinear equations. Pittsburgh: Carnegie Mellon University, 2014. 14 p.

3. Naseem A., Rehman M. A., Abdeljawad T. Numerical methods with engineering applications and their visual analysis via polynomiography // IEEE. 2021. Vol. 9. P. 99287–99298. doi: 10.1109/ACCESS.2021.3095941.

4. Harrison J. A machine-checked theory of floating-point arithmetic // Proc. of the 1999 Intern. Conf. on Theorem Proving in Higher Order Logics. Nice, 1999. P. 113–130. doi: 10.1007/3-540-48256-3_9.

5. Quinnell E., Swartzlander E. E., Lemonds C. Floating-point fused multiply-add architectures // Conf. Record of the Forty-First Asilomar Conf. on Signals, Systems and Computers. Pacific Grove, 2008. P. 331–337. doi: 10.1109/ACSSC.2007.4487224.

6. The On-Line Encyclopedia of Integer Sequences, Decimal expansion of square root of 2. URL: https://oeis.org/A002193 (data obrashcheniya 28.09.2022).

Information about the authors

Andrey A. Chusov – Cand. Sci. (Eng.), Associate Professor of the Department of Electronics, Telecommunications and Instrumentation of the Polytechnic Institute, Far Eastern Federal University, Island Russkiy, Ajax, 10, Russia.
E-mail: chusov.aa@dvfu.ru
http://orcid.org/0000-0002-7931-5368

Yulia I. Efimova – master student of the Department of Electronics, Telecommunications and Instrumentation of the Polytechnic Institute, Far Eastern Federal University, Island Russkiy, Ajax, 10, Russia.
E-mail: efimova.iui@students.dvfu.ru
http://orcid.org/0000-0002-7694-0694

Статья поступила в редакцию 11.10.2022; принята к публикации после рецензирования 01.11.2022; опубликована онлайн 25.12.2022.

Submitted 11.10.2022; accepted 01.11.2022; published online 25.12.2022.