АСТРОНОМИЧЕСКИЙ ЖУРНАЛ, 2020, том 97, № 12, с. 1022–1029

УДК 521.3

РЕШЕНИЕ УРАВНЕНИЯ КЕПЛЕРА С МАШИННОЙ ТОЧНОСТЬЮ

© 2020 г. М. К. Абубекеров[1, *], Н. Ю. Гостев[1, **]

[1] Московский государственный университет им. М.В. Ломоносова, Государственный астрономический институт им. П.К. Штернберга, Москва, Россия

**E-mail: marat@sai.msu.ru*

***E-mail: ngostev@mail.ru*

Поступила в редакцию 31.05.2020 г.
После доработки 08.07.2020 г.
Принята к публикации 15.08.2020 г.

Представлен алгоритм численного решения уравнения Кеплера с машинной точностью. Доказана сходимость итерационной последовательности метода Ньютона при указанном начальном приближении. Сформулирована задача нахождения численного решения уравнения Кеплера как числа с плавающей запятой. Учтены аспекты, связанные с вычислениями вблизи машинного нуля. Проанализирована точность возможного результата. Выявлена проблема, возникающая при стремлении к максимально возможной точности, и предложено ее решение. Дана оценка машинного времени, необходимого для решения уравнения Кеплера данным методом.

DOI: 10.31857/S0004629920120014

1. ВВЕДЕНИЕ

Как правило, в задачах, связанных с движением звезд (а также других космических тел) по эллиптическим орбитам, в том числе в задаче интерпретации кривых блеска двойных систем (см., напр., [1–4]), возникает необходимость решения уравнения Кеплера:

$$x - e\sin x = \mathcal{M} \qquad (1)$$

относительно x при заданных величинах e (эксцентриситете) и $\mathcal{M}$ (средней аномалии).

Данное уравнение является трансцендентным уравнением, которое при $0 < e < 1$ имеет единственное решение. На данный момент известны различные методы решения данного уравнения. Например, часто используется решение методом разложения по степеням e [5]. Этот метод преимущественно полезен при аналитическом исследовании движения по эксцентрической орбите, в предположении малых значений e, не превышающих предела Лапласа 0.6627.... Однако точность численных результатов, получаемых методом степенных рядов, ухудшается при увеличении значений e. В то же время во многих задачах желательно получать решение уравнения Кеплера с максимальной точностью, которая возможна, исходя из используемого компьютерного формата представления действительных чисел. Хотя современные компьютеры могут оперировать с числами, точность которых намного превосходит точность современных наблюдений (например, точность 80-битных чисел расширенной точности с 64-битной мантиссой соответствует 19 десятичным знакам), мы полагаем, что получение результата с максимально возможной точностью по-прежнему актуально по следующим причинам.

Во-первых, иногда для увеличения скорости вычислений целесообразно использовать машинное представление меньшей разрядности, чем максимально возможное, поскольку те же самые операции с числами меньшей разрядности выполняются быстрее. Например, для ускорения вычислений иногда имеет смысл использовать не 80-битные числа расширенной точности, а 64-битные числа двойной точности (с 52-битной мантиссой, соответствующей 15–16 десятичным знакам). Или может даже 32-битные числа одинарной точности с 24-битной мантиссой, соответствующей семи десятичным знакам. И чем ниже точность, обусловленная используемым компьютерным форматом представления действительного числа, тем важнее не допускать ее потерь, связанных с особенностями решения данной задачи.

Во-вторых, решение уравнения Кеплера обычно является промежуточным результатом, используемым для дальнейших вычислений, порой весьма сложных. И в ходе этих вычислений также возможна некоторая потеря точности (см. напр., [6]). При этом не всегда можно надежно оценить и правильно учесть эти потери точности. Таким образом, полезно иметь максимально возможный запас точности вычисления решения уравнения Кеплера.

В-третьих, даже если ошибка, связанная с вычислением теоретического значения физической величины, значительно меньше ошибки наблюдения, в некоторых случаях такая вычислительная ошибка может заметно искажать результаты, связанные со статистическим анализом наблюдений. Так, в [7] на примере интерпретации транзитной кривой блеска системы HD 209458 показано, как ошибка при вычислении модельных значений кривой блеска влечет статистически значимое изменение невязки χ^2.

Также необходимость высокой точности может проявиться независимо от точности наблюдений при решении задачи минимизации невязки χ^2 в ходе интерпретации транзитной кривой блеска. Во многих методах нелинейной оптимизации (напр., методе Левенберга-Марквардта) используются производные минимизируемого выражения (или входящих в него функций) и на сходимость последовательности метода может существенно влиять точность вычисления этих производных, в выражение для которых входит решение уравнения Кеплера для различных значений $\mathcal{M}$ и e и при фиксированных значениях наблюдательных данных (точность измерения которых не имеет значения для процесса минимизации).

На данный момент существуют различные алгоритмы решения уравнения Кеплера, основанные на быстро сходящихся итерационных последовательностях действительных чисел. Однако числа с плавающей запятой отличаются от действительных чисел, а разница между расчетами с числами с плавающей запятой и соответствующими расчетами с действительными числами может становиться значительной, если в эти расчеты входит вычисление выражений, близких к машинному нулю. И в некоторых случаях ошибки округления промежуточных результатов вызывают заметную потерю точности в конечном результате.

Мы рассмотрим данную проблему на примере из работ [8, 9], где описан алгоритм решения уравнения Кеплера, основанный на итерационной последовательности действительных чисел, которая быстро сходится к требуемому результату. Тут следует заметить, что данное утверждение о сходимости верно для $0 < \mathcal{M} < \pi$. Элементарной численной проверкой можно убедиться в том, что последовательность расходится при $e = 0.93$ и $-0.41 \le \mathcal{M} \le -0.39$. Возможно, эта расходимость устраняется выбором правильных начальных приближений для соответствующих диапазонов значений параметров. Однако и при $0 < \mathcal{M} < \pi$, если реализовать этот алгоритм с 64-битными числами двойной точности, можно обнаружить, что при некоторых значениях исходных данных соответствующая последовательность чисел с плавающей запятой не приближается к требуемому результату ближе некоторого значения, заметно большего ошибки в последнем разряде. Можно указать такие $\mathcal{M}$ и e, что построенная по упомянутому алгоритму последовательность компьютерных чисел не приближается к решению уравнения Кеплера ближе чем на $10^3\epsilon$, где $\epsilon = 2^{-52}$ – ошибка в последнем разряде чисел двойной точности. Следовательно, решение уравнения Кеплера при таких значениях параметров по данному алгоритму возможно с ошибкой, не меньшей, чем $10^3\epsilon$. Если же в качестве условия прерывания итерационного цикла в программе установить достижение большей точности, то цикл станет бесконечным.

Следует отметить, что подобный “сбой” в сходимости компьютерной последовательности не является результатом особенности и/или неустойчивости задачи решения уравнения Кеплера, сформулированной на множестве действительных чисел. Он обусловлен именно данным компьютерным представлением чисел. При переходе к другому представлению, например, к 80-битным числам расширенной точности, при данных параметрах значения итерационной последовательности приблизятся к значению решения уравнения Кеплера с точностью порядка соответствующего машинного ϵ. Также подобных “сбоев” сходимости уже не будет при незначительном отклонении от вышеприведенных значений $\mathcal{M}$ и e, например, при изменении лишь их последних цифр. Более того, наличие такого сбоя может зависеть и от выбора компилятора, используемого для того, чтобы запрограммировать алгоритм. Поэтому подобные “сбои” сходимости трудно выявить простым тестированием программы, т.е. с испытанием при небольшом количестве вариантов исходных данных. Однако данный сбой может проявиться при обработке больших массивов данных.

В настоящей работе построен алгоритм решения уравнения Кеплера, обеспечивающий достижение машинной точности результата путем учета особенностей вычислений вблизи машинного нуля. Для этого используется метод касательных (метод Ньютона), погрешность которого уменьшается до машинного нуля в среднем примерно за 5 итераций при использовании чисел расширенной точности (точное количество зависит от значений e и $\mathcal{M}$). Построение алгоритма на основе именно метода Ньютона представляется наиболее подходящим для наших целей, поскольку именно так удобнее контролировать влияние особенностей вычислений вблизи машинного нуля.

Существенным моментом в использовании метода Ньютона является выбор начального при-

ближения, при котором итерационная последовательность заведомо сходится к решению уравнения Кеплера. Отметим, что произвольно выбранное начальное приближение может не обеспечить оптимальную сходимость соответствующей итерационной последовательности. Итерационная последовательность может даже оказаться расходящейся при некоторых значениях начального приближения. Поэтому для решения уравнения Кеплера важен выбор начального приближения. При этом сформулирована задача численного решения уравнения Кеплера именно как задача нахождения соответствующего числа с плавающей запятой. Также описано тестирование точности и скорости алгоритма.

2. ПОСТРОЕНИЕ ИТЕРАЦИОННОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

При $0 \le e < 1$ и любом действительном $\mathcal{M}$ запишем уравнение Кеплера как

$$f(x) = 0, \tag{2}$$

где

$$f(x) = x - e\sin x - \mathcal{M}. \tag{3}$$

Нетрудно видеть, что $f(\mathcal{M} - e) = -e(1 + \sin(\mathcal{M} - e)) \le 0$, $f(\mathcal{M} + e) = e(1 - \sin(\mathcal{M} - e)) \ge 0$, т.е. функция f меняет знак на отрезке $[\mathcal{M} - e, \mathcal{M} + e]$. Поскольку функция $f(x)$ непрерывная и возрастающая при любых x, уравнение (2) всегда имеет единственное решение.

В тривиальном случае $\mathcal{M} = \pi k$, где k – целое число, его решением является $x = \mathcal{M}$, т.е., $\mathcal{K}(\pi k, e) = \pi k$.

Для решения данного уравнения методом касательных необходимо существование интервала D, содержащего искомое решение, причем такого, что на нем вторая производная $f''(x)$ не меняет знак, и один из концов этого интервала z удовлетворяет условию

$$f''(x)f(z) > 0, \quad \forall x \in D. \tag{4}$$

В таком случае, согласно известному утверждению [10], последовательность метода касательных, определяемая итерационным выражением

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{5}$$

сойдется к искомому решению, если взять z в качестве начального приближения.

Возьмем в качестве z ближайшее к $\mathcal{M}$ число вида $\pi(2m + 1)$, где m – целое число. Если $z > \mathcal{M}$, то

$$f''(x) = e\sin x > 0 \quad \forall x \in (2\pi m, z),$$
$$f(z) = z - \mathcal{M} > 0,$$

таким образом условие (4) выполняется, если за D взять интервал $(2\pi m, z)$. Если $z < \mathcal{M}$, то

$$f''(x) = e\sin x < 0 \quad \forall x \in (2\pi(m + 1), z),$$
$$f(x_0) = z - \mathcal{M} < 0,$$

таким образом условие (4) выполняется, если за D взять интервал $(x_0, 2\pi(m + 1))$.

Таким образом, если в качестве начального приближения выбрать $x_0 = \pi(2m + 1)$, последовательность метода касательных сходится к искомому решению уравнения (1). Отметим, что это начальное условие можно улучшить, отметив, что если x является нетривиальным решением уравнения (1), то

$$\mathcal{M} - e < x < \mathcal{M} + e,$$

т.е. решение уравнения принадлежит и интервалу $[\mathcal{M} - e, \mathcal{M} + e]$. Поэтому если $z > \mathcal{M}$, можно в качестве улучшенного начального приближения x_0 взять $\min(z, \mathcal{M} + e)$, а если $z < \mathcal{M}$, то взять $\max(z, \mathcal{M} - e)$. Условие (4) будет выполняться и при замене в нем z на x_0.

Подставив (3) в (5), получим

$$x_{n+1} = x_n - \frac{x_n - e\sin x_n - \mathcal{M}}{1 - e\cos x_n} = \frac{e(\sin x_n - x_n\cos x_n) - \mathcal{M}}{1 - e\cos x_n}. \tag{6}$$

Таким образом, последовательность (6) сходится к искомому решению уравнения Кеплера, если взять в качестве начального приближения

$$x_0 = \begin{cases} \max(\pi(2m + 1), \mathcal{M} - e), & \pi(2m + 1) < \mathcal{M} \\ \min(\pi(2m + 1), \mathcal{M} + e), & \pi(2m + 1) > \mathcal{M}, \end{cases}$$

где m – такое целое число, при котором величина $|\pi(2m + 1) - \mathcal{M}|$ принимает минимальное значение. Построенная выше итерационная последовательность $\{x_n\}$ является монотонной, и каждый ее член ближе к искомому решению, чем предыдущие.

Следует отметить, что задача решения уравнения Кеплера при произвольном значении средней аномалии может быть сведена к задаче решения уравнения Кеплера в пределах первого оборота. Мы рассматриваем общий случай для удобства при практической реализации метода, в том числе для того, чтобы исключить применение к значению средней аномалии операции получения остатка от деления на 2π, учитывая, что архитектура современных ЭВМ позволяет вычислять значения тригонометрических функций для любых значений аргумента на аппаратном уровне.

Подчеркнем, что существенную роль играет именно выбор начального приближения x_0. Указанного поведения итерационной последователь-

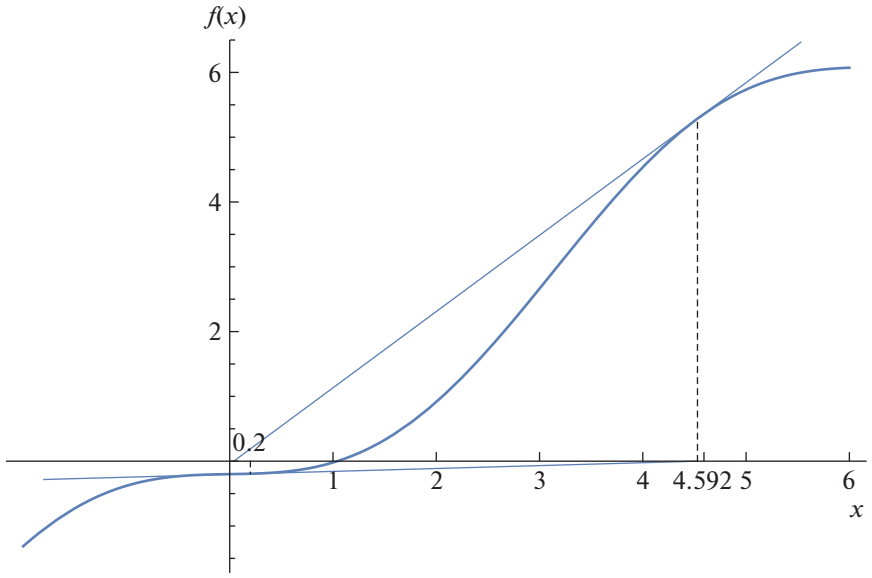

Рис. 1. Пример использования начального приближения $x_0 = \mathcal{M}$, приводящего к расходящейся последовательности решений уравнения (3), при значениях средней аномалии $\mathcal{M} = 0.2$ и эксцентриситета $e = 0.9747$.

ности не может гарантировать произвольный выбор. Например, если выбрать в качестве начального приближения $x_0 = \mathcal{M}$, это во многих случаях также даст сходящуюся последовательность, и некоторые авторы предлагают такой выбор (см., напр., [11]). Однако могут быть значения $\mathcal{M}$ и e, при которых члены последовательности с $x_0 = \mathcal{M}$ значительно удаляются от искомого решения, например, если $\mathcal{M} = 0.2$, $e = 0.9747$ (см. рис. 1). В работе [11] также указывалось на существование значений $\mathcal{M}$ и e, при которых итерационная последовательность расходится.

3. ТОЧНОСТЬ ВЫЧИСЛЕНИЙ ВБЛИЗИ МАШИННОГО НУЛЯ

Приведенное в предыдущем разделе построение итерационной последовательности, сходящейся к решению уравнения Кеплера, получено применительно к абстрактным действительным числам. В реальных компьютерных вычислениях мы оперируем числами с плавающей запятой, которые рассматриваются как приближенные значения действительных чисел. И в случае замены в вычислениях действительных чисел на числа с плавающей запятой соответствующие результаты также верны лишь приближенно. При этом величина ошибки при непосредственной замене действительного числа числом с плавающей запятой соответствует размеру мантиссы. Если действительное число аппроксимируется числом расширенной точности, мантисса которого содержит 64 бит (что гарантирует 19 десятичных значащих цифр), то величина относительной ошибки аппроксимации будет 2^{-64}, или $\sim 10^{-19}$.

Однако относительная ошибка результата вычислений с числами с плавающей запятой может быть значительно больше. Прежде всего это относится к случаю сложения (вычитания) двух чисел, когда результат по модулю значительно меньше, чем слагаемые. Если два действительных числа настолько близки, что соответствующие им числа с плавающей запятой не различаются, то результатом компьютерного вычисления их разности будет ноль. В то время как конечный результат решения задачи для действительных чисел зависит от этой разности как от промежуточного результата и обращение этой разности в ноль существенно искажает конечный результат.

Если два числа с плавающей запятой отличаются в нескольких последних знаках, относительная точность их разности будет соответствовать именно этому количеству знаков (скажем, при отличии чисел в двух последних десятичных знаках относительная погрешность вычисления их разности будет 1%). Поэтому замена абстрактных действительных чисел компьютерными числами с плавающей запятой может существенно ухудшить точность вычисления итогового результата, несмотря на то, что формальная задача поставлена корректно и устойчива к малым изменениям входных данных в виде абстрактных действительных чисел (изменения в 2^{-64} в исходных данных в виде действительных чисел влечет изменения того же порядка в итоговом результате). Для получения результата с заданной точностью ϵ путем

формальной реализации алгоритма, гарантирующего такую точность применительно к абстрактным действительным числам, необходимо использовать числа с плавающей запятой, точность которых значительно выше ε.

Разумеется, можно получить результат с заранее заданной точностью ε путем формальной реализации алгоритма с числами произвольной точности. Однако использование этого способа нежелательно по причине значительного возрастания времени вычислений, поскольку эмуляция элементарных операций с числами произвольной точности требует затрат машинного времени, значительно больших, чем выполнение реализованных на машинном уровне элементарных операций с 80-битными числами, не говоря уже о 64-битных.

Ввиду сказанного выше, для получения решения уравнения Кеплера с максимально возможной точностью, в рамках использования операций с числами с плавающей запятой, представляется целесообразным ставить задачу не как задачу нахождения числа, отличающегося от истинного решения уравнения не более чем на некоторое заданное ε, а как нахождение машинного числа с плавающей запятой, подстановка которого в уравнение (1) дает минимальное по модулю значение левой части, т.е. функции $f(x) = x - e\sin x - \mathcal{M}$.

Отметим, что одним из результатов вычисления (1) с использованием чисел с плавающей запятой может быть машинный ноль. Хотя возможна и ситуация, когда при некоторых машинных значениях $\mathcal{M}$ и e машинный ноль не будет достигаться ни при каком машинном значении x.

Для решения такой задачи сначала вычисляется последовательность, заданная выражением (5) или (6). Применительно к абстрактным числам разность между x_n и истинным решением уравнения Кеплера, равно как и величина $f(x_n)$, при любом n не меняют свой знак. В случае если $f(x_n)$ не меняет знак при любом n применительно к машинному x_n, его абсолютная величина будет уменьшаться по мере увеличения n, и на некотором шаге N (исходя из вида (5)) станет равным нулю. В таком случае x_n можно считать искомым решением.

Однако ввиду неточности, связанной с конечным представлением действительных чисел как чисел с плавающей запятой, машинное значение $f(x_n)$ может изменить знак на некотором шаге. В таком случае про дальнейшие элементы последовательности x_n нельзя утверждать, что они являются наилучшим приближением уравнения Кеплера. Однако они достаточно близки к истинному уравнению Кеплера, поэтому имеет смысл зафиксировать некоторые два из них, при которых функция $f(x_n)$ имеет разные знаки, и, начиная с образуемого ими отрезка, решать уравнение (1) методом половинного деления. Данный метод сходится значительно медленнее, чем метод касательных, и поэтому нецелесообразно использовать его на больших интервалах. В то же время этот метод основан исключительно на вычислении значения функции $f(x_n)$ и сравнении его с нулем, поэтому не чувствителен к ошибкам, которые возникают при операциях со значением $f(x_n)$. Поэтому имеет смысл использовать его для уточнения результата, полученного с помощью быстро сходящегося метода касательных.

Подчеркнем, что данный подход к нахождению численного значения решения уравнения Кеплера устраняет ошибку, связанную с промежуточными вычислениями (с конечным представлением используемых в нем чисел), но в любом случае конечный результат может содержать ошибку, вызванную ошибкой в исходных данных, которая, в свою очередь, связана с конечным представлением этих исходных данных (значений M и e) и соответствует значению последнего машинного разряда (2^{-64} для чисел расширенной точности с 64-битной мантиссой). Такая ошибка (превосходящая ошибку округления) появится, если производная $f'(x) < 1$. Можно сказать, что эта ошибка обусловлена чувствительностью результата к исходным данным.

Для сравнения мы реализовали алгоритм, описанный в [8, 9] с 64-битными числами двойной точности, относительная ошибка округления которых составляет $\epsilon = 2^{-52}$. Данный алгоритм основан на использовании сходящейся к решению уравнения Кеплера итерационной последовательности

$$E_{n+1} = E_n - \frac{(\mathcal{M} + e\sin E_n - E_n)^2}{E_n - 2(\mathcal{M} + e\sin E_n) + \mathcal{M} + e\sin(\mathcal{M} + e\sin E_n)}$$

при $E_0 = M + 0.85e$. Такую последовательность предлагается вычислять, пока $|E_{n+1} - E_n| > \varepsilon$, где ε – требуемая точность вычислений. Еще раз напомним, что при указанном начальном приближении эта последовательность не сходится для некоторых диапазонов значений $\mathcal{M}$ вне $0 < \mathcal{M} < \pi$. Однако и при $0 < \mathcal{M} < \pi$ могут возникать сложности с построением соответствующей последовательности компьютерных чисел для максимально точного вычисления. Аналогичные сложности могут возникнуть и с другими алгоритмами решения уравнения Кеплера, мы рассмотрим их на примере алгоритма [8, 9].

Для реализации с 64-битными числами двойной точности, как уже указывалось выше, при не-

которых значениях $\mathcal{M}$ и e, значения данной итерационной последовательности не приближаются к решению уравнения Кеплера ближе, чем на величину, существенно большую ошибки округления до последнего разряда 64-битного машинного числа ϵ. Напомним, что такие числа зависят от конкретной компьютерной реализации алгоритма. Например, при реализации на одном из компиляторов языка C при $\mathcal{M} = 0.09912109375$ и $e = 0.70849609375$ значения относительной разности $|(E_{n+1} - E_n)/E_{n+1}|$ не становятся меньше, чем $\epsilon 10^3$, где ϵ – ошибка округления до последнего разряда 64-битного машинного числа. Для реализации на другом компиляторе такими значениями оказываются, например, $\mathcal{M} = 0.00653076171875$, $e = 0.9605560302734375$. В конце работы мы даем ссылку на программу, которая выявляет такие значения $\mathcal{M}$ и e. Если в качестве условия прерывания итерационного цикла устанавливается уменьшение модуля относительной разности ($|(E_{n+1} - E_n)/E_{n+1}|$) между элементами итерационной последовательности менее чем на $\epsilon 10^3$, такой цикл становится бесконечным (программа зацикливается). Чтобы гарантированно избежать такого рода ситуации для любых значений исходных параметров, в приведенном алгоритме придется выбирать достаточно большое ε. При этом возникает еще и довольно нетривиальный вопрос о том, каким именно должно быть такое ε. Но уже из приведенного примера понятно, что это значение ε должно как минимум в 10^3 раз превышать ошибку округления до последнего разряда машинного представления числа.

Реализация нашего алгоритма с числами двойной точности позволяет вычислить решение уравнения Кеплера при вышеупомянутых значениях $\mathcal{M}$ и e с точностью до 15-десятичного знака, т.е. на уровне машинной точности, соответствующей такому представлению числа.

Далее мы провели численное тестирование точности значений решения уравнения Кеплера, полученных с помощью описанного алгоритма. Для этого мы реализовали алгоритм с числами расширенной точности (80-битными числами с 64-битной мантиссой, что в десятичной форме дает 19 значащих цифр) и с числами более высокой (эмулированной) точности. Как уже упоминалось выше, для практических вычислений использование таких чисел во многих случаях неэффективно, поскольку сильно увеличивает время вычислений. Однако с их помощью удобно проверить вычисления, сделанные с обычными машинными числами.

Мы осуществили вычисления с помощью описанного алгоритма как с эмулированными числами, содержащими значительно больше знаков, чем машинное 80-битное число, так и с 80-битными числами (с которыми элементарные операции выполняются на машинном уровне). Результат с эмулированными числами при этом имел гарантированно большую точность, чем результат с 80-битными числами. Далее мы вычисляли разность результата x, полученного с использованием 80-битных чисел и с помощью эмулированных чисел повышенной точности при одинаковых входных значениях $\mathcal{M}$ и e. Всего мы осуществили вычисления такой разности при 10^8 значениях пары $\mathcal{M}$ и e, в качестве значения $\mathcal{M}$ бралось псевдослучайное число с равномерным распределением на интервале $[0, \pi]$, а в качестве значения e бралось псевдослучайное число с равномерным распределением на интервале $[0,1]$. Такие же вычисления были произведены со значениями $\mathcal{M}$ и e, расположенными в виде равномерной сетки из 10^4 значений на соответствующем интервале. В случае $f'(x) < 1$ мы умножали полученную разность на $f'(x)$, чтобы учесть неизбежную ошибку, вызванную чувствительностью результата к ошибке округления входных значений. В каждом случае результат оказывался менее 10^{-19}, что позволяет сделать вывод о том, что 80-битное значение решения уравнения Кеплера получено с точностью, максимально возможной для 80-битного представления числа, и практически не искажено в результате промежуточных вычислений.

Для таких же 10^8 пар входных значений M и e (случайных и расположенных на равномерной сетке) мы протестировали количество итераций, необходимое для достижения максимальной точности. Среднее количество таких итераций, с учетом итераций методом половинного деления, оказалось равным примерно 5.51 (оно может возрастать при значениях e, близких к единице). Отметим, что если считать только итерации методом касательных, среднее их количество оказывается незначительно меньше, примерно 5.28. Таким образом, можно сделать вывод, что необходимость в уточнении результата методом половинного деления возникает, хотя общий объем таких вычислений сравнительно небольшой, около 4% от общего числа итераций. Однако при массовой обработке наблюдений мы практически неизбежно сталкиваемся с такими ситуациями.

4. ЗАКЛЮЧЕНИЕ

Часто решение уравнения Кеплера является промежуточным результатом, используемым для дальнейших, нередко весьма сложных вычислений, в которых неизбежно происходит существенная потеря точности. Примером таких вычислений может являться интерпретация кривой

блеска. При этом для статистического анализа наблюдательных данных иногда может требоваться точность, существенно превышающая точность наблюдений [7]. Равным образом высокая точность вычисления выражений, содержащих решение уравнения Кеплера, независимо от точности наблюдательных данных, желательна при решении задачи минимизации невязки χ^2 (см., напр., [1–4]). В сложных вычислениях представляется полезным использовать любую возможность для увеличения точности вычислений, в том числе потому, что иногда весьма сложно оценить необходимую точность, особенно, если речь идет о вычислительных задачах, которые могут возникнуть в будущем. Поэтому и при решении уравнения Кеплера важно стремиться к максимально точному результату.

Авторами в работе предложен алгоритм для быстрого вычисления решения уравнения Кеплера с точностью, наилучшей при данном машинном представлении действительных чисел. Предложенный алгоритм состоит из двух этапов.

Первым этапом является использование быстро сходящегося метода касательных для получения результата с точностью, которую данный метод может обеспечить с учетом конечности машинного представления чисел. При этом существенное значение имеет эффективный выбор начального приближения такой, что последовательные приближения представляют собой сходящуюся монотонную последовательность, каждый член которой ближе к искомому значению, чем предыдущие.

Вторым этапом является использование метода половинного деления для окончательного уточнения полученного значения, которое в некоторых случаях может содержать неточность в последних разрядах. При этом метод Ньютона обеспечивает простой критерий, по которому можно эффективно определить необходимость специального уточнения решения (изменение знака $f(x_n)$, невозможное в итерационной последовательности применительно к абстрактным действительным числам).

При этом задача решения уравнения Кеплера формулируется применительно к машинному представлению чисел как задача нахождения машинного числа, представляющего наилучшее возможное приближение решения уравнения Кеплера при заданных значениях средней аномалии и эксцентриситета.

Проведено сравнение эффективности предложенного авторами алгоритма решения уравнения Кеплера и другого, во многом сходного с методом Ньютона, алгоритма. Отмечено, что для некоторых значений $\mathcal{M}$ и e итерационная последовательность, построенная в машинном представлении действительных чисел, может перестать сходиться, отличаясь от минимально возможной для данного представления ошибки не менее чем в 10^3 раз. При этом такая ситуация возникает в сравнительно небольшом проценте возможных значений $\mathcal{M}$ и e, она не связана с какими-либо особенностями решения уравнения Кеплера применительно к абстрактным действительным числам. Поэтому возможность такой ситуации вполне может быть не выявлена простым тестированием формальной реализации алгоритма (которая не учитывает специфики машинного представления чисел), но может проявиться в дальнейшем при использовании такой реализации при обработке большого массива данных. При этом учесть возможность подобной ситуации, не выходя за рамки описанного в [8, 9] алгоритма, можно лишь существенным завышением ошибки для конечного результата ε.

Предложенный авторами алгоритм позволяет получить решение уравнения Кеплера с точностью на уровне ошибки округления для используемого компьютерного представления действительного числа (машинной точности) абсолютно для любых входных значений $\mathcal{M}$ и e. Эффективность использования такого алгоритма проявляется прежде всего при сложной обработке больших массивов данных, например, в задаче интерпретации современных наблюдаемых транзитных кривых блеска, которые могут содержать до нескольких десятков тысяч точек наблюдений. Таким образом, предложенный авторами алгоритм имеет определенное преимущество перед описанным в [8, 9] алгоритмом, даже если для большинства входных значений $\mathcal{M}$ и e для этих алгоритмов нет существенной разницы в скорости сходимости и достигаемой точности.

Путем численного эксперимента на большом массиве синтетических исходных данных проверена точность, обеспечиваемая алгоритмом, показано, что она является максимально достижимой для используемого представления действительного числа. Также оценено среднее количество итераций, от которого напрямую зависит время вычислений применительно к 80-битным числам с плавающей запятой (которые во многих современных ЭВМ обеспечивают максимальную машинную точность).

Разработанный авторами алгоритм общедоступен. Его программная реализация на языке C расположена на сайте Государственного астрономического института им. П.К. Штернберга[1]. Здесь расположен файл с реализацией алгоритма для 64- и 80-битного представления. Отдельно находится реализация для 128-битного представления, используемого в некоторых компиляторах C/C++. Хотя для большинства современных за-

[1] http://lnfm1.sai.msu.su/ngostev/Files/Kepler.zip

дач такая точность скорее всего будет избыточной и не оправдывающей увеличения временных затрат на операции со столь длинным числами, данная реализация полезна для тестирования точности. Также по приведенной ссылке расположена программа с реализацией метода Danby [8, 9] для 64-битных чисел, с контролем достижения заданной относительной точности путем подсчета числа итераций. Она демонстрирует упомянутое в статье отсутствие сходимости последовательности E_n при некоторых значениях $\mathcal{M}$ и e, которые находятся случайным перебором. При этом за критерий того, что последовательность не сходится, берется условие, что при $n > 500$, т.е. после 500 шагов, $|(E_{n+1} - E_n)/E_{n+1}| > N\varepsilon$, где N – заданное число (оно изначально задано как 1000).

Примеры использования разработанного авторами алгоритма решения уравнения Кеплера из работы [7] также расположены на сайте Государственного астрономического института им. П.К. Штернберга[2], в программных комплексах OccultationPack3 и DemoPack1.

[2] http://lnfm1.sai.msu.su/ngostev/algorithm.html; http://lnfm1.sai.msu.su/~ngostev/algorithm.html

СПИСОК ЛИТЕРАТУРЫ

1. *М. К. Абубекеров, Н. Ю. Гостев, А. М. Черепащук*, Астрон. журн. **85**, 121 (2008).
2. *М. К. Абубекеров, Н. Ю. Гостев, А. М. Черепащук*, Астрон. журн. **86**, 778 (2009).
3. *М. К. Абубекеров, Н. Ю. Гостев, А. М. Черепащук*, Астрон. журн. **87**, 1199 (2010).
4. *Н. Ю. Гостев*, Астрон. журн. **88**, 704 (2011).
5. *Г. Н. Дубошин*, *Небесная механика. Основные задачи и методы* (М.: Наука, 1968).
6. *М. К. Абубекеров, Н. Ю. Гостев*, Астрон. журн. **96**, 70 (2019).
7. *M. K. Abubekerov and N. Yu. Gostev*, Astron. and Astrophys. **633**, id. A96 (2020).
8. *J. M. A. Danby*, *Fundamentals of Celestial Mechanics. Second Edition* (Willmann-Bell, Inc., USA, 1995).
9. *Н. В. Емельянов*, *Динамика естественных спутников планет на основе наблюдений* (Фрязино: Век-2, 2019).
10. *А. Н. Колмогоров, С. В. Фомин*, *Элементы теории функций и функционального анализа* (М.: Наука, 1976).
11. *A. W. Odell and R. H. Gooding*, Celestial Mechanics and Dynamical Astronomy **38**, 307 (1986).