Math-Net.Ru

Общероссийский математический портал

С. В. Яхонтов, *LINSPACE* конструктивный аналог функции $(1+x)^h$, *Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. науки*, 2008, выпуск 2(), 239–249

DOI: <https://doi.org/10.14498/vsgtu612>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением

<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 52.6.47.48

26 апреля 2023 г., 16:05:59

Информатика

УДК 519.688

LINSPACE КОНСТРУКТИВНЫЙ АНАЛОГ ФУНКЦИИ $(1 + x)^h$

С. В. Яхонтов

Санкт-Петербургский государственный университет,
математико-механический факультет
198504, г. Санкт-Петербург, Старый Петергоф, Университетский пр-т., 28
E-mails: sergey_home_mail@inbox.ru, ysv@tercom.ru

Проводится построение алгоритмов класса FLINSPACE на основе разложения в ряд Тейлора для вычисления приближенных значений функции $(1 + x)^h$ с произвольной точностью. Данная функция рассматривается на множестве конструктивных вещественных чисел, т.е. в качестве аргумента x берутся алгоритмы, вычисляющие рациональные приближения.

Ключевые слова: рациональные приближения, алгоритмы, точность.

Введение. В данной статье предлагаются алгоритмы для вычисления в пределах LINSPACE и с полиномиальным временем приближенных значений вещественной функции

$$f(x) = (1 + x)^h \quad (1)$$

для x из интервала $(-1, +\infty)$ и рациональных h , $|h| < 1$. Работа выполнена в рамках построения LINSPACE вычислимого математического анализа аналогично построению P вычислимого математического анализа [1].

Основные сведения о классах вычислительной сложности можно найти в [2]. FLINSPACE является стандартным обозначением для класса алгоритмов, использующих в процессе вычислений промежуточную память, объем которой ограничен сверху линейной функцией от длины входных данных (т.е. Cn). Под вычислениями в пределах LINSPACE будем понимать те, которые применяют алгоритмы из класса FLINSPACE.

В данной работе функция (1) вычисляется на основе разложения в ряд Тейлора. Отличительной особенностью алгоритмов расчета элементарных функций, использующих разложения в ряды, является относительная простота, так как при этом обычно производятся только арифметические операции над рациональными числами без обращения к дополнительным алгоритмам. К тому же для вычислений с небольшим количеством знаков после двоичной точки ряды эффективнее других способов в связи с небольшими константами в оценках вычислительной сложности. Поэтому такие алгоритмы важны в информатике для практических приложений.

Выражение $(1 + x)^h$ рассматривается на множестве конструктивных вещественных чисел. Такие числа задаются алгоритмами, вычисляющими рациональные приближения, а построение алгоритма, с помощью которого рассчитываются приближения функции по приближениям аргумента, позволяет

Яхонтов Сергей Викторович — старший преподаватель каф. информатики Санкт-Петербургского государственного университета.

говорить о конструктивной функции, являющейся конструктивным аналогом исходной функции.

Для практических приложений важно уметь вычислять значения конструктивных элементарных функций на множестве конструктивных вещественных чисел с помощью алгоритмов из некоторого емкостного класса сложности, которые одновременно были бы полиномиальными по времени. В качестве такого класса емкостной сложности представляется разумным взять класс *LINSPACE*. По аналогии с классом временной сложности *P* его можно назвать классом осуществимых вычислений с точки зрения требуемой для вычислений памяти. Подтверждением этой неформальной характеристики может служить то, что в настоящее время в работах по теоретической информатике для вычислений в пределах *LINSPACE* утвердился термин «space-efficient evaluation».

На данный момент нами не найдены публикации других авторов, в которых содержался бы анализ емкостной вычислительной сложности функции (1).

1. Конструктивные вещественные числа и функции. В данной статье основу представления конструктивных объектов составляет понятие вычислимой последовательности ϕ , сходящейся по Коши [1]. Такая последовательность определяется на множестве всех натуральных чисел N , включая 0, а областью аппроксимирующих значений является всюду плотное в R естественное подмножество множества рациональных чисел. Для последовательности, сходящейся по Коши и задающей вещественное число x , требуют, чтобы выполнялось $|\phi(n) - x| \leq 2^{-n}$ для любого натурального n .

В качестве множества аппроксимирующих значений берется множество двоично-рациональных чисел D [1]. Рациональное число d называется двоично-рациональным, если $d = \frac{m}{2^n}$ для некоторого целого m и натурального n . Двоично-рациональные числа имеют конечное двоичное представление: строка s , равная $\pm u_p u_{p-1} \dots u_0 . v_1 v_2 \dots v_r$, обозначает число

$$d = \pm \sum_{i=0}^p u_i 2^i \pm \sum_{j=1}^r v_j 2^{-j}.$$

Длина представления двоично-рационального числа определяется как количество символов в строке s , равное (с учётом знака и двоичной точки) $p+r+3$, и обозначается $l(s)$. Под точностью представления $\text{грес}(s)$ понимается число битов справа от двоичной точки, т.е. r . С точки зрения изучения вычислительной сложности двоично-рациональные числа удобны тем, что для любого n двоично-рациональные числа с точностью n равномерно распределены на вещественной прямой [1].

Последовательность $\phi : N \rightarrow D$ двоично-рационально сходится к вещественному числу x , если для любого $n \in N$ выполняется $\text{грес}(\phi(n)) = n + 1$ и $|\phi(n) - x| \leq 2^{-n}$. Множество всех функций, двоично-рационально сходящихся к вещественному числу x , обозначается CF_x . Вещественное число x называется *CF* конструктивным [1], если CF_x содержит вычислимую функцию ϕ .

Вещественная функция f , заданная на отрезке $[a, b]$, называется конструктивной функцией [1] на этом отрезке, если существует машина Тьюринга M с оракульной функцией такая, что для любого $x \in [a, b]$ и любой вычислимой функции $\phi \in CF_x$ функция ψ , вычисляемая M с оракульной функцией ϕ , принадлежит $CF_{f(x)}$. Фактически это означает, что для любой вычислимой функции $\phi \in CF_x$ и любого $n \in N$ машина M последовательно вычисляет

$m \in N$ и $d \in D$ такие, что

$$|\phi(m) - x| \leq 2^{-m}, \quad |d - f(x)| \leq 2^{-n}.$$

Сложность расчёта рациональных приближений конструктивных чисел и функций определяется в [1] на основе длины двоичного представления точности вычисления. Память ленты запроса и ленты ответа оракульной функции, как и время вычисления этой функции, при оценке вычислительных ресурсов машины Тьюринга не учитывается. Обращение к оракульной функции $\phi \in CF_x$ аргумента x конструктивной функции осуществляется следующим образом:

- на ленту запроса оракульной функции записывается точность вычисления аргумента 2^{-m} ;
- рассчитывается значение $\phi(m)$ оракульной функции, и результат записывается на ленту ответа;
- значение $\phi(m)$ считывается с ленты ответа в промежуточную память.

При описании алгоритмов данные шаги в явном виде здесь приводиться не будут; просто будет подразумеваться, что процесс вычислений содержит обращения к оракульной функции.

Вещественное число x называется *LINSPACE* конструктивным вещественным числом, если существует *LINSPACE* вычислимая функция $\phi \in CF_x$. Вещественная функция f , заданная на отрезке $[a, b]$, называется *LINSPACE* конструктивной вещественной функцией на отрезке $[a, b]$, если для любого $x \in [a, b]$ существует *LINSPACE* вычислимая функция ψ из $CF_{f(x)}$. Множества *LINSPACE* конструктивных вещественных чисел и функций будем обозначать $LINSPACE_{CF}$ и $LINSPACE_{C[a,b]}$ соответственно. Индекс $C[a, b]$ в обозначении $LINSPACE_{C[a,b]}$ обусловлен тем, что конструктивные функции являются непрерывными на всей области определения [1]. Построение конструктивного аналога вещественной функции f на отрезке $[a, b]$ означает описание алгоритма, вычисляющего двоично-рациональные приближения с произвольной точностью значений $f(x)$ для $x \in [a, b]$.

Под приближением t с точностью 2^{-n} будем понимать двоично-рациональное число t^* такое, что $|t^* - t| \leq 2^{-n}$. Для модуля t^* выполняются следующие неравенства:

$$\begin{aligned} |t^*| - |t| &\leq |t^* - t| \leq 2^{-n}, & \text{т. е.} \quad |t^*| &\leq |t| + 2^{-n}, \\ |t| - |t^*| &\leq |t - t^*| \leq 2^{-n}, & \text{т. е.} \quad |t^*| &\geq |t| - 2^{-n}. \end{aligned} \quad (2)$$

Далее, пусть двоично-рациональное число t^{**} таково, что $|t^{**} - t| \leq 2^{-(n+1)}$. Отбросим все биты t^{**} после двоичной точки, начиная с $(n+2)$ -го, а новую величину обозначим t^* . Тогда $prec(t^*) = n+1$ и $|\varepsilon| = |t^* - t^{**}| < 2^{-(n+1)}$. Отсюда получаем

$$|t^* - t| \leq |t^* - t^{**}| + |t^{**} - t| < 2^{-(n+1)} + 2^{-(n+1)} = 2^{-n}. \quad (3)$$

То есть, чтобы вычислить t с точностью 2^{-n} , можно рассчитать t с точностью $2^{-(n+1)}$, а затем отбросить биты полученной величины t^{**} после двоичной точки, начиная с $(n+2)$ -го.

2. Вычисление значений степенного ряда. Рассмотрим степенной ряд

$$S(x) = \sum_{k=0}^{\infty} a_k x^k = b_0 + b_0 b_1 x + b_0 b_1 b_2 x^2 + \dots + b_0 \dots b_{k-1} b_k x^k + \dots \quad (4)$$

на отрезке $[-\varrho, \varrho]$, $0 < \varrho \leq \rho$, где ρ — радиус сходимости. Пусть b_i — *LINSPACE* конструктивные вещественные числа, x — произвольное *LINSPACE* конструктивное вещественное число из отрезка $[-\varrho, \varrho]$, β_i — *LINSPACE* вычислимые функции из CF_{b_i} , ϕ — *LINSPACE* вычисляемая функция из CF_x . Пусть также m — некоторое натуральное число, $\beta_i(m)$ — двоично-рациональные приближения коэффициентов b_i с точностью 2^{-m} , $\phi(m)$ — двоично-рациональные приближения аргумента x с точностью 2^{-m} .

Рассмотрим частичные суммы $S_k(x) = \sum_{i=0}^k a_i x^i$ ряда (4) с подстановкой $\beta_i(m)$ вместо b_i и $\phi(m)$ вместо аргумента x :

$$S_k(m) = \sum_{i=0}^k \left[\prod_{j=0}^i \beta_j(m) \right] \phi(m)^i.$$

Коэффициенты ряда (4) получаются последовательным умножением на величины b_i с ростом номера коэффициента, т. е. $a_i = a_{i-1} b_i$ с начальным значением $a_0 = b_0$. При вычислении частичных сумм такого ряда удобно не пересчитывать коэффициенты a_i , а накапливать их значение вместе с вычислением промежуточных значений. Поэтому воспользуемся следующей схемой с округлением промежуточных результатов для расчета приближения $S_k^*(m)$ величины $S_k(m)$:

$$\begin{aligned} h_0(m) &= \beta_k(m), \\ \vartheta_r(m) &= \beta_{k-r}(m)[1 + h_{r-1}(m)\phi(m)], \quad r = 1, 2, \dots, k, \\ h_r(m) &= \vartheta_r(m) + \varepsilon_r \end{aligned} \quad (5)$$

при $r = k$ полагаем $S_k^*(m) = h_k(m)$. Величины $h_r(m)$ получаются отбрасыванием битов $q_{m+1}q_{m+2} \dots$ чисел $\vartheta_r(m)$ после двоичной точки, начиная с $(m+1)$ -го, т. е.

$$|\varepsilon_r| = |h_r(m) - \vartheta_r(m)| = 0.0 \dots 0q_{m+1}q_{m+2} \dots, \quad (6)$$

а знак совпадает со знаком $\vartheta_r(m)$ (ясно, что $|\varepsilon_r| < 2^{-m}$). Оценим величину погрешности вычислений

$$\Delta(k, m) = |S_k^*(m) - S_k(x)|$$

при следующих ограничениях на модуль b_i и ϱ :

$$|b_i| \leq 1 \text{ при всех } i, \quad \varrho \leq 2^{-1} + 2^{-5}. \quad (7)$$

Предварительно заметим, используя неравенства (2), что $|\beta_i(m)| \leq |b_i| + 2^{-m}$ и $|\phi(m)| \leq |x| + 2^{-m}$. Для вывода основной оценки понадобится вспомогательное утверждение.

ЛЕММА 1. *Для любого фиксированного $r \geq 0$ выполняется следующее: если $m \geq 5$, то справедлива оценка*

$$|h_r(m)| < 5. \quad (8)$$

Доказательство. Применим математическую индукцию по i для $h_i(m)$. База индукции при i , равном 0:

$$|h_0(m)| = |\beta_k(m)| \leq |b_k| + 2^{-m} \leq 1 + 2^{-5} < 5.$$

Индукционный переход для $(i+1) \geq 1$:

$$\begin{aligned}
 |h_{i+1}(m)| &= |\beta_{k-(i+1)}(m)[1 + h_i(m)\phi(m)] + \varepsilon_{i+1}| \leq \\
 &\leq (|b_{k-(i+1)}| + 2^{-m}) [1 + |h_i(m)|(|x| + 2^{-m})] + 2^{-m} < \\
 &< (1 + 2^{-m}) [1 + 5 \cdot (2^{-1} + 2^{-5} + 2^{-m})] + 2^{-m} \leq \\
 &\leq (1 + 2^{-5}) [1 + 5 \cdot (2^{-1} + 2^{-5} + 2^{-5})] + 2^{-5} = \\
 &= (1 + 2^{-5}) [1 + 90 \cdot 2^{-5}] + 2^{-5} < 5. \quad \square
 \end{aligned}$$

Теперь можно сформулировать утверждение относительно $\Delta(k, m)$.

ЛЕММА 2. При выполнении условий (7) погрешность вычислений по схеме (5) оценивается как

$$\Delta(k, m) < 2^{-m} 2^5 (k + 1). \quad (9)$$

Доказательство. Обозначим

$$\begin{aligned}
 \zeta &= \max_{i=0}^k |\beta_i(m) - b_i|, \quad \xi = |\phi(m) - x|, \\
 H_r(x) &= b_{k-r}[1 + H_{r-1}(x)x], \quad H_0(x) = b_k, \quad \eta(r, m) = |h_r(m) - H_r(x)|.
 \end{aligned}$$

Ясно, что $\zeta \leq 2^{-m}$, $\xi \leq 2^{-m}$. Воспользуемся методом математической индукции для $\eta(r, m)$ по r , предполагая, что $\eta(r, m) < 2^{-m} 2^5 (r + 1)$. База индукции при r , равном 0:

$$\eta(0, m) = |h_0(m) - H_0(x)| = |\beta_k(m) - b_k| \leq 2^{-m} < 2^{-m} 2^5 (0 + 1).$$

Оценим сверху выражение $\eta(r+1, m)$ для $(r+1) \geq 1$:

$$\begin{aligned}
 \eta(r+1, m) &= |\beta_{k-(r+1)}(m)[1 + h_r(m)\phi(m)] + \varepsilon_{r+1} - b_{k-(r+1)}[1 + H_r(x)x]| \leq \\
 &\leq |(\beta_{k-(r+1)}(m) - b_{k-(r+1)}) + \varepsilon_{r+1}| + \\
 &+ |\beta_{k-(r+1)}(m)h_r(m)\phi(m) - b_{k-(r+1)}H_r(x)x|.
 \end{aligned}$$

Рассмотрим последнее слагаемое под модулем, которое обозначим $w(r, m)$:

$$\begin{aligned}
 w(r, m) &= \beta_{k-(r+1)}(m)h_r(m)\phi(m) - b_{k-(r+1)}H_r(x)x = \\
 &= \beta_{k-(r+1)}(m)h_r(m)\phi(m) - b_{k-(r+1)}h_r(m)\phi(m) + \\
 &+ b_{k-(r+1)}h_r(m)\phi(m) - b_{k-(r+1)}H_r(x)x = \\
 &= [\beta_{k-(r+1)}(m) - b_{k-(r+1)}] h_r(m)\phi(m) + \\
 &+ b_{k-(r+1)} [h_r(m)\phi(m) - H_r(x)x] = \\
 &= [\beta_{k-(r+1)}(m) - b_{k-(r+1)}] h_r(m)\phi(m) + \\
 &+ b_{k-(r+1)} [h_r(m)\phi(m) - h_r(m)x + h_r(m)x - H_r(x)x].
 \end{aligned}$$

В силу $|b_i| \leq 1$ и $|\phi(m)| < 1$ справедлива оценка

$$|w(r, m)| < \zeta |h_r(m)| + |h_r(m)|\xi + |\eta(r, m)|(2^{-1} + 2^{-5}).$$

Так как $|h_r(m)| < 5$ (из (8)) и по индукционному предположению $\eta(r, m) < 2^{-m}2^5(r+1)$, то

$$\begin{aligned}\eta(r+1, m) &< \zeta + |\varepsilon_{r+1}| + \zeta \cdot 5 + 5 \cdot \xi + 2^{-m}2^5(r+1) \cdot 17 \cdot 2^{-5} < \\ &< 2^{-m}(12 + 17(r+1)) < 2^{-m}2^5(r+1).\end{aligned}$$

Из $S_k^*(m) = h_k(m)$, $S_k(x) = H_k(x)$, $\Delta(k, m) = \eta(k, m)$ получаем искомое неравенство. $\square$

Пусть требуется вычислить частичную сумму $S_{\mu(k)}(x)$ степенного ряда (4) с точностью 2^{-k} для некоторого натурального k , где μ — линейная функция от k , $\mu(k) \geq 0$. Положим

$$m = m(k_1, k_2) = 5 + \lceil \log_2(k_1 + 1) \rceil + k_2, \quad (10)$$

где $k_1 = \mu(k)$, $k_2 = k$. Тогда $m > 5$ и, следовательно, погрешность вычисления приближенного значения $S_{\mu(k)}(x)$ будет оцениваться в соответствии с (9):

$$\Delta(\mu(k), m) < 2^{-m}2^5(\mu(k) + 1) \leq 2^{-k_2} = 2^{-k}. \quad (11)$$

На основании приведенных оценок можно сформулировать следующий алгоритм.

Алгоритм PartialSum. Приближенное значение частичной суммы ряда (4).

Вход: Запись точности вычисления 2^{-k} .

Оракульная функция: ϕ для аргумента x .

Выход: Приближенное значение $S_{\mu(k)}(x)$ с точностью 2^{-k} .

Описание:

- 1) вычисляем $k_1 := \mu(k)$, $k_2 := k$;
- 2) находим m по формуле (10);
- 3) рассчитываем $\phi := \phi(m)$;
- 4) инициализируем цикл: $h := \beta(k_1, m)$;
- 5) выполняем цикл по r от 1 до k_1 :
 - а) рассчитываем $b := \beta(k_1 - r, m)$,
 - б) вычисляем $\vartheta = b \cdot [1 + h \cdot \phi]$,
 - в) h присваиваем величину ϑ , округленную в соответствии с (6);
- 6) на выход записываем двоичное представление h (в этом представлении будет m битов после двоичной точки).

Пусть степенной ряд (4) линейно сходится, т.е. для остаточного члена выполняется

$$|r_{\mu(n+1)}(x)| = |S_{\mu(n+1)}(x) - S(x)| \leq 2^{-(n+1)}$$

для любого $x \in [-\varrho, \varrho]$, где μ — линейная функция. Тогда

$$\begin{aligned}|S_{\mu(n+1)}^*(m) - S(x)| &\leq |S_{\mu(n+1)}^*(m) - S_{\mu(n+1)}(x)| + |S_{\mu(n+1)}(x) - S(x)| \leq \\ &\leq \Delta(\mu(n+1), m) + |r_{\mu(n+1)}(x)| \leq \Delta(\mu(n+1), m) + 2^{-(n+1)}.\end{aligned}$$

Возьмём $m = m(k_1, k_2)$ по формуле (10) для $k_1 = \mu(n+1)$, $k_2 = n+1$. Тогда, учитывая неравенство (11), получаем

$$|S_{\mu(n+1)}^*(m) - S(x)| < 2^{-(n+1)} + 2^{-(n+1)} = 2^{-n}.$$

Отсюда видно, что линейной зависимости m от n достаточно для вычисления суммы степенного ряда (4) на отрезке $[\varrho, \varrho]$ с требуемой точностью. По условию величины b_i являются *LINSPACE* вычислимыми, поэтому алгоритм *PartialSum*, определяющий $h_r(m)$ по схеме (5), будет оперировать с числами, длины которых линейно зависят от n , т.е. данный алгоритм будет принадлежать классу *FLINSPACE*. Если величины b_i являются также полиномиально вычислимыми по времени, то алгоритм *PartialSum* будет в классе $FLINSPACE \cap FP$ (через *FP* обозначается класс алгоритмов, полиномиальных по времени).

3. Вычисление приближенных значений функции. Рядом Тейлора функции (1) является биномиальный ряд [3]:

$$(1+x)^h = 1 + hx + \frac{h(h-1)}{1 \cdot 2}x^2 + \dots + \frac{h(h-1) \dots (h-n+1)}{1 \cdot 2 \cdot \dots \cdot n}x^n + \dots \quad (12)$$

Остаточный член в форме Коши для $|x| < 1$ имеет вид [3]

$$r_n(x) = \frac{f^{(n+1)}(\theta x)}{n!}(1-\theta)^n x^{n+1} = \nu_1 \nu_2 \nu_3,$$

где

$$\nu_1 = \frac{h(h-1)(h-2) \dots (h-n)}{1 \cdot 2 \cdot \dots \cdot n}, \quad \nu_2 = x^{n+1}(1+\theta x)^{h-1}, \quad \nu_3 = \left(\frac{1-\theta}{1+\theta x} \right)^n$$

для $\theta \in (0, 1)$. Очевидно, что $|\nu_1| < n+1$. Так как $1+\theta x > 1-\theta$ при $x > -1$, то $|\nu_3| < 1$ для таких x . Рассмотрим выражение ν_2 для x из отрезка $[-\frac{1}{2} - 2^{-5}, 2^{-5}]$. Второй множитель из ν_2 преобразуем к виду

$$(1+\theta x)^{h-1} = (1+\theta x)(1+\theta x)^{h-2} = \frac{(1+\theta x)}{(1+\theta x)^{2-h}}.$$

Так как $2-h > 1$, то выражение $(1+\theta x)^{2-h}$ больше величины $(1-17 \cdot 2^{-5})^3$. Поэтому

$$(1+\theta x)^{h-1} < \frac{33}{2^5} \cdot \left[\frac{2^5}{15} \right]^3 < 2^4$$

и для модуля ν_2 верна оценка сверху:

$$|\nu_2| < \left[\frac{17}{2^5} \right]^{n+1} \cdot 2^4 < 2^{(n+1)(4.1-5)} 2^4 = 2^{-0.9(n+1)+4}.$$

Следовательно, для остаточного члена выполняется соотношение $|r_n(x)| < 2^{-\zeta(n)}$, где $\zeta(n) = 0.9(n+1) - 4 - \log_2(n+1)$, т.е. ряд (12) линейно сходится на отрезке $[-\frac{1}{2} - 2^{-5}, 2^{-5}]$. Если в качестве b_i ряда (4) взять $\frac{h-i+1}{i}$, а $b_0 = 1$, то будут верны условия (7). Поэтому приближения биномиального ряда можно вычислять в пределах *LINSPACE* на этом отрезке с помощью алгоритма *PartialSum*. При этом $\mu(n+1)$ нужно брать по формуле

$$\mu(n+1) = 2(n+5), \quad (13)$$

так как тогда

$$\begin{aligned}\zeta(\mu(n+1)) &= 0.9(2(n+5)+1) - 4 - \log_2(2(n+5)+1) = \\ &= 1.8n + 9.9 - 4 - \log_2(2n+11) = \\ &= (n+1) + (0.8n + 4.9 - \log_2(2n+11)) > n+1,\end{aligned}$$

и поэтому $|r_{\mu(n+1)}| < 2^{-(n+1)}$.

Для расчёта ряда (12) на произвольном отрезке воспользуемся мультипликативной редукцией интервала. Для этого введём величину u' , равную $\frac{u}{2^r}$, где u — обозначение для $1+x$, r — целое число. Значение r подберем так, чтобы $u' \in [\frac{1}{2} - 2^{-5}, 1 + 2^{-5}]$; далее будет видно, что такое r подобрать можно. Тогда значение x' , равное $u' - 1$, будет в интервале $[-\frac{1}{2} - 2^{-5}, 2^{-5}]$, который обозначим через $\sigma(5)$. Выражение $(1+x)^h$ сведем к двум множителям:

$$u^h = (u'2^r)^h = (u')^h(2^r)^h = (u')^h(2^h)^r = (u')^h \left[(2^{-1})^{-h} \right]^r. \quad (14)$$

Второй множитель в последнем выражении удобно вычислять для положительных r . Поэтому если $r > 0$, то положим $t = (2^{-1})^{-h}$, а если $r < 0$, то положим $t = (2^{-1})^{+h}$ ($t = 1$ при $r = 0$). Тогда

$$u^h = w \cdot \left[(2^{-1})^{-h} \right]^{+r} = w \cdot t^{+r} \quad \text{или} \quad u^h = w \cdot \left[(2^{-1})^{+h} \right]^{-r} = w \cdot t^{-r}$$

в зависимости от знака r , где $w = (u')^h$. Величину w будем вычислять как значение биномиального ряда для $x' \in \sigma(5)$, а величину t — как значение биномиального ряда для $x = -\frac{1}{2}$ и затем возводить в степень $|r|$.

Найдём точность 2^{-v} вычисления аргумента x и точность 2^{-m_3} вычисления t для расчёта значения ряда (12) по формуле (14) с точностью $2^{-(n+1)}$ и последующим округлением до точности 2^{-n} в соответствии с (3).

Точность вычисления аргументов x, y операции умножения, достаточная для получения произведения с точностью $2^{-(n+1)}$, равна $2^{-(n+q+2)}$ [4], где q таково, что $|x| + 1 < 2^q$, $|y| + 1 < 2^q$. Так как $w < 3$ для $x' \in \sigma(5)$, $t^{|r|} < 2^{|r|}$, то для расчёта выражения (14) с точностью $2^{-(n+1)}$ достаточно знать множители $w, t^{|r|}$ с точностью 2^{-m_1} , где

$$m_1 = n + q + 2, \quad q = \max(2, |r| + 1). \quad (15)$$

Следовательно, значение x' нужно вычислять с точностью 2^{-m_2} , где m_2 рассчитывается по формуле (10) для $k_1 = \mu(m_1 + 1)$, $k_2 = m_1 + 1$, а $\mu(m_1 + 1)$ находится по формуле (13). Так как при сложении с константой ($x' = u' - 1$) точность не уменьшается, то u' нужно вычислять также с точностью 2^{-m_2} .

Пусть x^* — приближение x , u^* — величина $1+x^*$, $(u')^*$ — величина $\frac{u^*}{2^r}$. Введём обозначения:

$$v_{<} = m_2 + |r|, \quad v_{>} = \max(m_2 - |r|, 0).$$

Рассмотрим отдельно случаи, когда $r < 0$ и $r > 0$ ($(u')^* = u^*$ для $r = 0$). Если $r < 0$, то x достаточно знать с точностью $2^{-v_{<}}$, так как

$$|(u')^* - u'| = \left| \frac{1+x^*}{2^r} - \frac{1+x}{2^r} \right| = \frac{1}{2^r} |x^* - x| \leq 2^{-m_2}.$$

Аналогично, если $r > 0$, то x достаточно знать с точностью 2^{-v} . Для обоих случаев знака r подойдет значение $v = \max(v_<, v_>) = m_2 + |r|$.

Приближенные значения аргумента конструктивной функции являются двоично-рациональными числами, а такие числа имеют конечное двоичное представление. Поэтому для расчета $(u')^*$ будем сдвигать битовое представление u^* так, чтобы старший значащий бит оказался сразу после двоичной точки (т.е. $(u')^* = 0.1\dots$). Если $u^* \geq 1$, то r будет такое, что $r > 0$, иначе $r \leq 0$. При этом $(u')^* \in [\frac{1}{2}, 1)$, а величина u' будет в интервале $[\frac{1}{2} - 2^{-5}, 1 + 2^{-5})$, так как точность расчета приближения $(u')^*$ не ниже 2^{-5} (из формулы для v).

Здесь возникает некоторая сложность, состоящая в том, что для вычисления x^* нужна величина сдвига, которую можно узнать только после расчета $u^* = 1 + x^*$. Поэтому предположим, что

$$x \geq -1 + 2^{-p} \quad \text{и} \quad x \leq 2^p - 2$$

для некоторого натурального $p \geq 1$. Тогда для $x^* = \phi(v)$ из (2) получаем следующие соотношения, которые выполняются при $v \geq p + 1$:

$$\phi(v) \geq -1 + 2^{-p} - 2^{-(p+1)} = -1 + 2^{-(p+1)} \quad \text{и} \quad \phi(v) \leq 2^p - 2 + 2^{-(p+1)} < 2^p - 1,$$

поэтому $u^* \geq 2^{-(p+1)}$ и $u^* < 2^p$. Величина $(u')^*$ получается сдвигом u^* на r битов. Если $r < 0$, то $1 > (u')^* \geq 2^{-(p+1)+|r|}$, и для r будет выполняться $(p+1) - |r| > 0$, т.е. $|r| < p+1$. В случае $r > 0$ будет $2^{-1} \leq (u')^* < 2^{p-r}$, и тогда $p-r > -1$, т.е. $r < p+1$. Следовательно, всегда r такое, что $|r| < p+1$ для $p \geq 1$, и в качестве v достаточно взять $m_2 + p$. В этом случае количество битов после двоичной точки величины $(u')^*$ не будет превышать $m_2 + 2p + C$. Из (15) также заметим, что m_1 достаточно рассчитывать по формуле $n+p+3$.

Далее, покажем, что для расчета выражения $t^{|r|}$ с точностью 2^{-m_1} при последовательном умножении и округлении достаточно вычислить m_3 по формуле $m_1 + 3|r|$. Обозначим через z_i приближение величины t^i . Значение z_i будем рассчитывать по формулам

$$z_1 = t^*, \quad z_{i+1} = z_i t^* + \kappa_{i+1} \quad \text{для} \quad i = 1, \dots, |r| - 1, \quad (16)$$

где κ_{i+1} получаются округлением результата с точностью 2^{-m_3} аналогично (6), т.е. $|\kappa_{i+1}| < 2^{-m_3}$. Предварительно заметим, что верна оценка $|z_i| < 2^{3i}$. И действительно, выполняются неравенства

$$|z_1| \leq |t| + 2^{-m_3}, \quad |z_{i+1}| \leq |z_i| (|t| + 2^{-m_3}) + |\kappa_{i+1}|,$$

то есть

$$|z_1| < 2^3, \quad |z_{i+1}| < 2^{3i}(2 + 2^{-m_3}) + 2^{-m_3} < 2^{3(i+1)},$$

откуда и следует по принципу математической индукции, что $|z_i| < 2^{3i}$ для произвольного i . Теперь рассмотрим погрешность расчета t^i , которую обозначим через δ_i , т.е. $\delta_i = z_i - t^i$. Предположим, что $|\delta_i| \leq 2^{-m_3+3(i-1)}$. Для $i = 1$ данное соотношение выполняется: $\delta_1 \leq 2^{-m_3}$. Рассмотрим погрешность

δ_{i+1} для $(i+1) \geq 2$:

$$\begin{aligned}
 |z_{i+1} - t^{i+1}| &= |z^i t^* + \kappa_{i+1} - t^i t| = \\
 &\leq |z^i t^* - z_i t| + |z_i t - t^i t| + |\kappa_{i+1}| = \\
 &= |z_i| \cdot |t^* - t| + |z_i - t^i| \cdot |t| + |\kappa_{i+1}| < \\
 &< 2^{3i} 2^{-m_3} + 2^{-m_3+3(i-1)} 2 + 2^{-m_3} = \\
 &= 2^{-m_3} (2^{3i} + 2^{3(i-1)+1} + 1) < 2^{-m_3} 2^{3i} 4 < 2^{-m_3} 2^{3(i+1)},
 \end{aligned}$$

т. е. индукционное предположение подтверждается. Следовательно, если взять $m_3 = m_1 + 3|r|$, то $|z_r - t^{|r|}| < 2^{-m_1}$, и поэтому приближение $t^{|r|}$ с точностью 2^{-m_1} по формулам (16) будет искомым.

В результате получаем следующий алгоритм.

Алгоритм BinomSeriesValue. Приближенное значение $(1+x)^h$.

Вход: Запись точности вычисления 2^{-n} .

Оракульная функция: ϕ для аргумента x .

Параметры: Константа p .

Выход: Приближенное значение $(1+x)^h$ с точностью 2^{-n} .

Описание:

- 1) вычисляем $m_1 := n + p + 3$;
- 2) вычисляем $k_1 := 2(m_1 + 5)$ и $k_2 := m_1 + 1$;
- 3) находим $m_2 := m(k_1, k_2)$ по формуле (10);
- 4) вычисляем $v := m_2 + p$;
- 5) рассчитываем $\phi := \phi(v)$;
- 6) вычисляем $u := 1 + \phi$;
- 7) определяем r :
 - а) если $u < 1$, то $r \leq 0$,
 - б) если $u \geq 1$, то $r > 0$;
- 8) сдвигая битовое представление u на r позиций влево или вправо в зависимости от знака r , получаем u' ;
- 9) вычисляем $x' := u' - 1$;
- 10) используем алгоритм *PartialSum* для вычисления приближенного значения w с точностью 2^{-m_1} ;
- 11) находим $m_3 := m_1 + 3|r|$;
- 12) используем алгоритм *PartialSum* для расчета приближенного значения t с точностью 2^{-m_3} ;
- 13) вычисляем $f_1 := t^{|r|}$ с точностью 2^{-m_1} ;
- 14) вычисляем $f_2 := w \cdot f_1$;
- 15) на выход записываем двоичное представление f_2 , округленное в соответствии с (3).

Учитывая свойства алгоритма *PartialSum*, а также линейную зависимость m_1 , m_2 и m_3 от n , получаем, что объем промежуточной памяти будет ограничен сверху линейной функцией от n . Очевидно также, что все вычисления будут полиномиальными по времени.

ТЕОРЕМА. Вещественная функция $f(x) = (1+x)^h$ для рационального h , $|h| < 1$, является *LINSPACE* конструктивной на любом отрезке $[-1 + 2^{-p}, 2^p - 2]$, где p — натуральное число, $p \geq 1$.

Заключение. Алгоритм *BinomSeriesValue* (вместе с алгоритмом *PartialSum*) можно применять в информатике как основу *LINSPACE* конструктивной вещественной функции $f(x) = (1+x)^h$, заданной на множестве *LINSPACE*

конструктивных вещественных чисел. Предложенные алгоритмы не являются оптимальными по времени выполнения, но для построения *LINSPACE* вычислимого математического анализа важна принципиальная возможность построения линейных по памяти алгоритмов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Ko, K.* Complexity Theory of Real Functions [Text] / K. Ko — Boston: Birkhauser, 1991. — 309 p.
2. *Du, D.* Theory of Computational Complexity [Text] / D. Du, K. Ko — New York: John Wiley & Sons, 2000. — 491 p.
3. *Фихтенгольц, Г. М.* Курс дифференциального и интегрального исчисления [Текст] / Г. М. Фихтенгольц — М.: Физматлит, 2003. Т. 2. — 864 с.
4. *Шурыгин, В. А.* Основы конструктивного математического анализа [Текст] / В. А. Шурыгин. — М.: Едиториал УРСС, 2004. — 328 с.

Поступила в редакцию 01/III/2008;
в окончательном варианте — 10/VII/2008.

MSC: 68Q25

LINSPACE CONSTRUCTIVE ANALOGUE OF $(1+x)^h$ FUNCTION

S. V. Yakhontov

St. Petersburg State University, Dept. of Mathematics and Mechanics
198504, St. Petersburg, Staryy Petergoff, Universitetskii prosp., 28.

E-mails: sergey_home_mail@inbox.ru, ysv@tercom.ru

Algorithms of class FLINSPACE are constructed in order to calculate the arbitrary precision of $(1+x)^h$ function with the help of Taylor expansion. This function is analyzed with the multitude of constructive real numbers, that is, the algorithms calculating rational approximations are used as x argument.

Key words: *rational approximations, algorithms, closeness.*

Original article submitted 01/III/2008;
revision submitted 10/VII/2008.

Yakhontov Sergey Viktorovich, Senior Lecturer, Dept. of Mathematics and Mechanics (Computer Science chair) of St. Petersburg State University.