УДК 004.222.3

ОЦЕНКА ПАРАМЕТРОВ МНОГОРАЗРЯДНЫХ ЧИСЕЛ С ПЛАВАЮЩЕЙ ТОЧКОЙ ДЛЯ ВЫПОЛНЕНИЯ ОПЕРАЦИЙ ВЫСОКОЙ ТОЧНОСТИ

Пехотин Е.В., Бабков В.С.
Донецкий национальный технический университет

В работе предлагается альтернативный подход к оценке параметров многоразрядных чисел с плавающей точкой. Результатом работы является система связующих критериев для выбора эффективно-представляющего формата многоразрядных чисел с плавающей точкой для выполнения операций высокой точности.

На сегодняшний день в математике существует множество методов решения задач в различных областях человеческой деятельности. Однако непосредственное применение этих методов затруднительно, ибо на практике можно использовать только числа, имеющие конечное представление, в то время, как большинство этих методов работают с произвольными вещественными числами. Очевидно, что подавляющее большинство вещественных чисел невозможно выразить конечным представлением. А т.к. данное ограничение касается как входных, так и выходных и промежуточных данных, то в области реальных вычислений встает задача как повышения точности расчетов, так и разработки устойчивых методов расчета [1-2].

Данная работа является продолжением одноименной работы автора, опубликованной в [3].

1 Особенности чисел с плавающей точкой

Рис. 1 демонстрирует общее представление чисел с плавающей точкой. Ясно, что 1+w+t = N. Для обычной экспоненты значения порядка лежат в диапазоне: $D[E] = [-2^{w-1}; 2^{w-1}-1]$. Обычно используется смещенная экспонента. Пусть p_{min}, p_{max} – минимальное

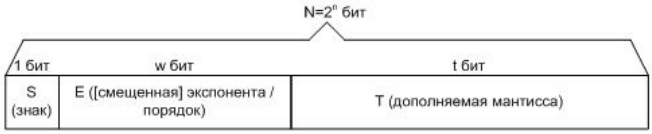

Рисунок 1 – Общий вид числа с плавающей точкой

и максимальное значения порядка, w – размерность порядка, bias – прибавляемое смещение. Для получения смещения была в [3, с. 13] составлена и решена система уравнений аффинного преобразования одномерного пространства x2 = b0 + b1 · x1, в результате чего получено равенство

$$bias = -p_{\min} = 2^w - p_{\max} - 1.$$

Назовем **характеристикой** величину q = p + bias. Будем представлять значение экспоненты числа с плавающей точкой с помощью характеристики. Следует отметить, что в стандарте на кодирование чисел с плавающей точкой IEEE-754 [4] используют именно смещенную экспоненту на интервале $[1–2^{w-1}; 2^{w-1}]$.

Постулат 1. Любое число с фиксированной точкой может быть представлено в виде числа с плавающей точкой в данном формате абсолютно без потери точности при условии, что разность между номерами самого старшего и самого младшего единичных бит будет меньше размера мантиссы, а показатель степени самого старшего единичного разряда числа лежит в диапазоне $[p_{min}, p_{max}]$.

Доказательство. Пусть L[M] – размер используемой мантиссы (а через M будем обозначать любую допустимую мантиссу). Представляем число X в двоичном виде в формате с фиксированной точкой. У любого двоичного числа (кроме 0) будет хотя бы одна значащая 1 в представлении. Пускай номер самого старшего единичного разряда числа при отсчет с 0 от его самого младшего единичного бита равен n (т.е. в числе всего n+1 разрядов), а его показатель степени – z.

Для 0: значение мантиссы возьмем 0.0, значение порядка может быть любое (в том числе и 0). Для чисел, отличных от 0: номер разряда самой старшей значащей 1 целой части обозначен через z.

По условию $z \in [p_{min}, p_{max}]$, поэтому может быть допустимым для данного формата порядком представляемого числа. Разделим число на 2^z, или, что то же самое, сдвинем значение представление числа вправо через двоичную точку на z разрядов. Мы получим $m = \frac{X}{2^z}$, у которой целая часть будет равна 1 – старшему единичному биту X, а дробная – остальным битам X. Дополнив m справа нулями до L[M] (по условию их количество не превзойдет длины мантиссы, а т.к. один единичный бит будет у нас всегда, то количество дополняющих нулей будет лежать в интервале [0,L[M] – 1]), получим M – мантиссу числа. Итак,

$$X = \frac{X}{2^z} 2^z = M * 2^p,$$

где $M = \frac{X}{2^z}$, дополненное справа 0 до L[M], – мантисса представляемого числа, p = z – порядок представляемого числа.

5

Примечание. Данный постулат расширяет эквивалентный постулат в [3, с. 13].

Следствие 1. Любое отличное от 0 представимое число может быть представлено в виде 1.0 + ε, где ε < 1.

Из следствия 1 вытекает, что мы можем сэкономить 1 разряд, подразумевая 1.0 и не отображая его в представлении числа. Таким образом, размер мантиссы увеличится на 1 разряд, что приведет к увеличению точности на 1 разряд.

Однако, встает проблема – как представить бесконечности, 0 и ненормализованные числа (которые в данном FP–формате нельзя представить в виде 1.0 + ε, а только 0.0 + ε из-за ограничений хранения порядка)? Можно пойти двумя путями:

- хранить полную мантиссу;
- использовать 2 граничных значения порядка для кодирования особых случаев.

Теперь опишем форматы 0.0 + ε и 1.0 + ε для дальнейшей работы с ними.

Лемма 0: вне зависимости от способа интерпретации значение

результирующего числа $a_{q=0}$ для любой наперед заданной мантиссы γ будет одним и тем же.

Доказательство. Пускай N – размерность мантиссы данного формата. Пускай

$$\gamma = \sum_{i=-1}^{-N} \gamma_i \cdot 2^{i+\psi} = 2^{\psi} \cdot \sum_{i=-1}^{-N} \gamma_i \cdot 2^{i}, \gamma_i \in \{0,1\} \quad (3)$$

Из такой записи видно, что γ здесь представлена комбинированно – как произведение представления числа с фиксированной точкой и некоторого коэффициента. Естественно, что количество членов суммы представленного ряда в точности равно размерности используемой мантиссы. Осталось выяснить значение коэффициента 2^{ψ}, точнее, значение ψ. Пускай $\psi = 0$. Тогда возможные значения γ будут лежать в пределах $\gamma \in [0;1)$. И действительно, γ при $\psi = 0$ представляет собой ряд вида:

$$\gamma_{-1}\frac{1}{2^1} + \gamma_{-2}\frac{1}{2^2} + \ldots + \gamma_{-N+1}\frac{1}{2^{N-1}} + \gamma_{-N}\frac{1}{2^N}, \quad (*)$$

что также можно назвать физическим представлением мантиссы γ; но тогда (3) можно назвать логическим представлением γ.

Если $\forall i : \gamma_i = 0$, то и сумма данного ряда становится равной 0, следовательно, $\gamma = 0$. С другой стороны, это наименьшее значение мантиссы в данном формате.

Если же $\forall i : \gamma_i = 1$, то ряд превращается в известный в математике отрезок непрерывного ряда по степеням $\frac{1}{2}$, и его сумма также известна

$$\sum_{i=-1}^{-N} 2^i = \sum_{i=1}^{N} 2^{-i} = \sum_{i=1}^{N} \frac{1}{2^i} = \left(\sum_{i=0}^{N} \frac{1}{2^i}\right) - 1 = \frac{1-\left(\frac{1}{2}\right)^{N+1}}{1-\frac{1}{2}} - 1 = 1 - 2 \cdot \frac{1}{2^{N+1}} = 1 - \frac{1}{2^N} \quad (4)$$

следовательно, $\gamma = 1 - \frac{1}{2^N}$, и очевидно, что $\gamma < 1$. Очевидно также, что все остальные значения γ лежат в этом интервале, следовательно, $\gamma \in [0;1)$.

Однако, данное условие как раз и является условием на ε в случае первой интерпретации. Соответственно, мантисса γ при

$\psi = 0$ является подходящей для первой интерпретации q = 0 (т.е. $\varepsilon = \gamma_{\psi=0}$) и значение числа в этой интерпретации равно

$$a_{q=0} = (0.0 + \varepsilon) \cdot 2^{p_{min}+1} = 2^{p_{min}+1} \sum_{i=-1}^{-N} \varepsilon_i \cdot 2^i = 2^{p_{min}} \cdot 2^1 \cdot \sum_{i=-1}^{-N} \gamma_i \cdot 2^i$$

Теперь пускай $\psi = 1$. Можно сказать, что $\psi = 1$ – фиксированный коэффициент особого случая рассматриваемого нами формата – денормализованных чисел, а фиксированные коэффициенты можно хранить «в уме», следовательно, физически мантисса γ и здесь будет представлена в виде (*), из чего следует один важный вывод: размерность используемой мантиссы при $\psi = 1$ осталась прежней, т.е. относительно представления при $\psi = 0$ не потерялось ни одного бита представления – и ни одного бита точности ⇒ точности физических представлений обоих форматов совпадают.

Однако для получения истинного (логического) результата нам необходимо умножать все получаемые значения на $2_{\psi=1}^{\psi} = 2^1 = 2$. Следовательно, результаты всех вышеприведенных расчетов надо просто умножить на 2 и мы получим:

– левый граничный случай $\gamma_{\psi=1}^{\gamma_i=0} = 0$ – наименьшее логическое значение мантиссы.

– правый граничный случай

$2 \cdot \gamma_{\psi=0}^{\gamma_i=1} = 2 \cdot \left(1 - \frac{1}{2^N}\right) = 2 - \frac{1}{2^{N-1}}$, из чего ясно, что $\gamma_{\psi=1}^{\gamma_i=1} < 2$ – наибольшее логическое значение мантиссы.

Итак, $\gamma_{\psi=1} \in [0; 2)$, следовательно, логическое значение мантиссы при таком представлении является подходящим для второй интерпретации q = 0, тогда $\varepsilon = \gamma_{\psi=1}$ и фактическое значение числа с данной мантиссой равно

$$a_{q=0} = (0.0 + \varepsilon) \cdot 2^{p_{min}} = 2^{p_{min}} \cdot \gamma_{\psi=1} = 2^{p_{min}} \cdot 2^1 \cdot \gamma_{\psi=0}.$$

Однако,

$$2^{p_{min}} \cdot 2^1 \cdot \gamma_{\psi=0} = 2^{p_{min}} \cdot 2^1 \cdot \sum_{i=-1}^{-N} \gamma_i \cdot 2^i = 2^{p_{min}+1} \sum_{i=-1}^{-N} \varepsilon_i \cdot 2^i = (0.0 + \varepsilon) \cdot 2^{p_{min}+1}$$

– это первый способ интерпретации мантиссы денормализованного числа, откуда

$$a_{q=0}^{\psi=0} = (0.0 + \varepsilon_{\psi=0}) \cdot 2^{p_{min}+1} = 2^{p_{min}} \cdot 2^1 \cdot \gamma_{\psi=0} = 2^{p_{min}} \cdot \gamma_{\psi=1} = (0.0 + \varepsilon_{\psi=1}) \cdot 2^{p_{min}} = a_{q=0}^{\psi=1}, \quad (3)$$

значит, $a_{q=0}^{\psi=0} = a_{q=0}^{\psi=1} = a_{q=0}$, что означает, что логически оба способа интерпретации из одной и той же физической мантиссы γ (*) получают одно и то же число $a_{q=0}$. А т.к. и значения чисел, и их погрешность для одной и той же мантиссы γ вне зависимости от используемого способа интерпретации совпадают, то и эти способы интерпретации являются эквивалентными и, следовательно, представляют эквивалентные числа. ■

Формат 0.0 + ε, называемый полный или свободный, определяет представление числа a в виде $a = (0.0 + \varepsilon_a) \cdot 2^{p_a}$, при этом $p_a \in [p_{min}, p_{max}]$ – целое. Для наложения ограничения на ε необходимо определить положение двоичной точки в представлении числа. В [3, с. 14] было показано, что рациональнее всего использовать представление числа в виде $1.xxxx \cdot 2^{yy}$ для двоичной с.с., т.е. в формате 1.0 + ε. Поэтому будем располагать двоичную точку между старшим и предыдущим разрядом мантиссы, т.е. между 2^{t-1} и 2^{t-2} разрядами (отсчет по рисунку 1 от нуля справа налево).

Также, имеется некоторое различие между полным и свободным форматом 0.0 + ε. Для полного формата обязательно наличие лидирующей 1 в первом разряде, т.е. мантисса числа $\varepsilon = 1.0 + \varepsilon^*, 0 \le \varepsilon^* < 1$ – это формат 1.0 + ε с явно присутствующей в мантиссе лидирующей 1. У свободного формата условие такого типа отсутствует.

2 Эффективно-представляющие форматы чисел с плавающей точкой

В работе [3] вводится понятие **эффективно-представляющего** формата. Формат является эффективно-представляющим, если он эффективно использует все w_0 битов порядка и t_0 битов мантиссы, т.е. имеет такое свойство: количество представимых форматом **различных** чисел равно общему числу двоичных чисел

разрядностью N_0 (булеану от N_0).

У эффективно–представляющих форматов имеется исходящее из их определения преимущество перед другими: они представляют максимальное количество различных чисел, которые можно представить имеющейся разрядностью.

В [3, с. 14-15] была доказана **лемма 1**: все числа, представимые в формате 1.0 + ε, является различными или (что то же) любое число, представимое в формате 1.0 + ε, представимо в нем единственным образом.

Утверждение 1. Формат 1.0 + ε является эффективно-представляющим.

Данное утверждение является заменой эквивалентному в [3, с. 15] – его доказательство дополнено анализом влияния денормализованных чисел.

Доказательство. По лемме 1 все нормализованные числа формата 1.0 + ε являются различными. Для денормализованных же используем факты леммы 0:

1) обсуждаемые в ней способы интерпретации денормализованных чисел эквивалентны, следовательно, они имеют идентичную уникальность представимых ими чисел ⇒ достаточно проверить уникальность только одной интерпретации;

2) очевидно, что каждое денормализованное число – уникально, т.к. является двоичной комбинацией мантиссы с одним и тем же порядком (при q = 0).

3) через t_0 определена размерность мантиссы, откуда из (3) и (4) наибольшее значение денормализованного числа равно

$$a_{q=0}^{max} = 2^{p_{min}} \cdot 2^1 \cdot \gamma_{\psi=0}^{\gamma_i=1} = 2^{p_{min}} \cdot 2^1 \cdot \left(1 - \frac{1}{2^{t_0}}\right),$$

а т.к. очевидно, что раз характеристика q = 0 отводится для представления денормализованных чисел, то минимальная характеристика для представления нормализованных чисел – q = 1, следовательно, значение минимального нормализованного числа (из определения формата 1.0 + ε) равно

$$a_{norm}^{min} = a_{q=1}^{min\varepsilon} = a_{q=1}^{\varepsilon=0} = (1.0+0)\cdot 2^{p_{min}+1} = 2^{p_{min}} \cdot 2^1 \cdot 1.$$

Очевидно, что

$$\forall t_0 \geq 0 : 1 - \frac{1}{2^{t_0}} < 1 \Rightarrow 2^{p_{min}} \cdot 2^1 \cdot \left(1 - \frac{1}{2^{t_0}}\right) < 2^{p_{min}} \cdot 2^1 \cdot 1 \Rightarrow a_{q=0}^{max} < a_{norm}^{min}.$$

Теперь ясно, что т.к. размерность мантиссы физически не может быть меньше 0, то по (3) для любого формата 1.0 + ε всегда будет $a_{q=0}^{max} < a_{norm}^{min}$, т.е. максимальное значение денормализованного числа в любом формате типа 1.0 + ε будет меньше значения минимального нормализованного числа этого формата.

Теперь, т.к. формат не накладывает никаких ограничений ни на мантиссу, ни на порядок, ни на знак числа, это значит, что:

1. Для представления дробной части значения числа используются все биты мантиссы (а значит, оная используется эффективно), а количество различных значений мантиссы равно булеану от ее длины, т.е. 2^{t_0}.
2. Для представления порядка числа используются все биты порядка (а значит, и порядок используется эффективно), а количество различных значений порядка равно соответственно булеану его длины, т.е. 2^{w_0}.
3. Знак – это элемент множества {+, −}, мощность которого равна 2, при этом каждое представляемое число имеет знак.

Соответственно, всего в формате могут быть представлено $2 \cdot 2^{w_0} \cdot 2^{t_0} = 2^{1+w_0+t_0} = 2^{N_0}$ (булеан от N_0) различных чисел.

Утверждение 2. Формат 0.0 + ε не является эффективно-представляющим, что доказано в [6, с. 15-16].

Итак, в качестве базового формата представления чисел с плавающей точкой рекомендуется использовать эффективно-представляющий формат 1.0 + ε. Задача выбора конкретного для заданных условий эффективно-представляющего формата решается с помощью системы связующих критериев, которые получены в ранее проведенных исследованиях и представлены в [3].

Выводы

В результате исследования получена система, выражающая зависимость граничных значений и точности формата представления 1.0 + ε от длин его составляющих – мантиссы и порядка. Полученные зависимости могут быть использованы на практике для реализации программных библиотек для математических расчетов, ориентированных на получение достоверных результатов с максимально эффективным представлением чисел заданной разрядности.

Литература

[1] Alefeld G., Herzberger J. Introduction to interval computations. — New York etc.; Academic Press, 1983. – XVIII, 333 p.; Рус. перев.; Алефельд Г., Херцбергер Ю. Введение в интервальные вычисления: Пер. с англ. – М.: Мир, 1987. – 356 с.

[2] Nickel K. Can we trust the results of our computing? // Mathematics for Computer Science; Proc Symposium held in Paris, March 16–18, 1982. –S. l.; Association francaise pour la cybernetique et technique (AFCET), 1982. – P. 167–175.

[3] Бабков В.С., Пехотин Е.В. Оценка параметров многоразрядных чисел с плавающей точкой для выполнения операций высокой точности // Научные труды Донецкого национального технического университета. Серия «Информатика, кибернетика и вычислительная техника» (ИКВТ-2010). Выпуск 12 (165) – Донецк: ГВУЗ «ДонНТУ». 2010 – с. 12-17.

[4] IEEE Standard for Floating-Point Arithmetic (Revision of IEEE Std 754-1985) // IEEE Computer Society. – 2008. – P. 70.