Числа: точность неотрицательные системы счисления дополнительный код

Д. В. Луцив

Кафедра системого программирования СПбГУ

CS103

Содержание

Предмет

- Целых, рациональных, а *тем более* вещественных чисел бесконечное количество.
- Машинное представление обычно конечно и адекватно решаемой задаче.
- Представление с произвольной точностью и/или рациональное бывает, но аппаратно не поддерживается, да и программно условно.

Абсолютная погрешность — отклонение числа, оптимально представленного машиной, от его действительного значения.

Для $I^{(m)}$ абсолютная погрешность не превышает 0,5, когда исходное число лежит в множестве I, для представления элементов которого предназначены машинные целые из множества $I^{(m)}$.

Для наилучшего представления –

$$x^{(m)} \in I^{(m)}, x \in I : \Delta_a^{(m)}(x) = \min_{x_i^{(m)} \in I^{(m)}} |x_i^{(m)} - x|$$

$$\text{справедливо } \Delta_a^{(m)}(I^{(m)}) = \max_{x \in I} \Delta_a^{(m)}(x)$$

Постоянная абсолютная погрешность и у представления с фиксированной точкой.

Абсолютная погрешность — отклонение числа, оптимально представленного машиной, от его действительного значения. Для $I^{(m)}$ абсолютная погрешность не превышает 0,5, когда исходное число лежит в множестве I, для представления элементов которого предназначены машинные целые из множества $I^{(m)}$.

Для наилучшего представления –

$$x^{(m)} \in I^{(m)}, x \in I : \Delta_a^{(m)}(x) = \min_{x_i^{(m)} \in I^{(m)}} |x_i^{(m)} - x|$$

справедливо $\Delta_a^{(m)}(I^{(m)}) = \max_{x \in I} \Delta_a^{(m)}(x)$

Постоянная абсолютная погрешность и у представления с фиксированной точкой.

Абсолютная погрешность — отклонение числа, оптимально представленного машиной, от его действительного значения.
Для $I^{(m)}$ абсолютная погрешность не превышает 0,5, когда исходное число лежит в множестве I, для представления элементов которого предназначены машинные целые из множества $I^{(m)}$.
Для наилучшего представления –

$$x^{(m)} \in I^{(m)}, x \in I : \Delta_a^{(m)}(x) = \min_{x_i^{(m)} \in I^{(m)}} |x_i^{(m)} - x|$$

$$\text{справедливо } \Delta_a^{(m)}(I^{(m)}) = \max_{x \in I} \Delta_a^{(m)}(x)$$

Постоянная абсолютная погрешность и у представления с фиксированной точкой.

Относительная погрешность — это абсолютная погрешность, деленная на абсолютную величину конкретного представляемого числа:

$$\Delta_r^{(m)}(x) = \frac{\Delta_a^{(m)}(x)}{|x|}$$

В общем же случае,

$$\Delta_r^{(m)}(I_f^{(m)}) = \max_{x \in I_f} \frac{\Delta_a^{(m)}(I_f^{(m)})}{|x|}$$

С постоянной относительной погрешностью представляют значения машинные числа с плавающей запятой.
При представлении с постоянной погрешностью бóльших чисел, относительная погрешность будет падать.

Относительная погрешность — это абсолютная погрешность, деленная на абсолютную величину конкретного представляемого числа:

$$\Delta_r^{(m)}(x) = \frac{\Delta_a^{(m)}(x)}{|x|}$$

В общем же случае,

$$\Delta_r^{(m)}(I_f^{(m)}) = \max_{x \in I_f} \frac{\Delta_a^{(m)}(I_f^{(m)})}{|x|}$$

С постоянной относительной погрешностью представляют значения машинные числа с плавающей запятой.

При представлении с постоянной погрешностью бóльших чисел, относительная погрешность будет падать.

Относительная погрешность — это абсолютная погрешность, деленная на абсолютную величину конкретного представляемого числа:

$$\Delta_r^{(m)}(x) = \frac{\Delta_a^{(m)}(x)}{|x|}$$

В общем же случае,

$$\Delta_r^{(m)}(I_f^{(m)}) = \max_{x \in I_f} \frac{\Delta_a^{(m)}(I_f^{(m)})}{|x|}$$

С постоянной относительной погрешностью представляют значения машинные числа с плавающей запятой.
При представлении с постоянной погрешностью бóльших чисел, относительная погрешность будет падать.

||

Для многих вычислений (особенно физических) важна именно относительная погрешность:
Например, у «результата с тремя значащими цифрами», относительная погрешность всегда $\frac{1}{2000}$.
Биологическая рецепторная функция $f(x) = \ln(x)$.

$$f'(x) = \ln' x = 1/x$$

При $\Delta \to 0$ справедливо $\ln(x+\Delta) - \ln(x) = O(\frac{\Delta}{x})$.
Если рецептор различает $f_1 = f(x)$ и $f_2 = f(x) + \Delta_f = f(x + \Delta_x)$, то при $\Delta_f = \text{const}$, $\Delta_x = O(x)$.

||

Для многих вычислений (особенно физических) важна именно относительная погрешность:
Например, у «результата с тремя значащими цифрами», относительная погрешность всегда $\frac{1}{2000}$.
Биологическая рецепторная функция $f(x) = \ln(x)$.

$$f'(x) = \ln' x = 1/x$$

При $\Delta \to 0$ справедливо $\ln(x+\Delta) - \ln(x) = O(\frac{\Delta}{x})$.
Если рецептор различает $f_1 = f(x)$ и $f_2 = f(x) + \Delta_f = f(x + \Delta_x)$, то при $\Delta_f = \text{const}$, $\Delta_x = O(x)$.

||

Для многих вычислений (особенно физических) важна именно относительная погрешность:
Например, у «результата с тремя значащими цифрами», относительная погрешность всегда $\frac{1}{2000}$.
Биологическая рецепторная функция $f(x) = \ln(x)$.

$$f'(x) = \ln' x = 1/x$$

При $\Delta \to 0$ справедливо $\ln(x+\Delta) - \ln(x) = O(\frac{\Delta}{x})$.
Если рецептор различает $f_1 = f(x)$ и $f_2 = f(x) + \Delta_f = f(x + \Delta_x)$, то при $\Delta_f = \text{const}$, $\Delta_x = O(x)$.

Диапазон — разность между наименьшим и наибольшим представляемыми числами.

Фиксированная запятая

Важны не значения, как таковые, а, скорее, их количество.

Плавающая запятая

Информативны границы диапазона, даже для чисел одного знака.

$\min_{x \in I_f} |x|$ — максимальная точность для малых чисел (но не относительная погрешность, т.к. в самом маленьком числе м.б. единица в конце мантиссы, о длине которой мы наперед не знаем, см. ниже).

$\max_{x \in I_f} |x|$ — диапазон в его исходном смысле.

Диапазон — разность между наименьшим и наибольшим представляемыми числами.

Фиксированная запятая

Важны не значения, как таковые, а, скорее, их количество.

Плавающая запятая

Информативны границы диапазона, даже для чисел одного знака.

$\min_{x \in l_f} |x|$ — максимальная точность для малых чисел (но не относительная погрешность, т.к. в самом маленьком числе м.б. единица в конце мантиссы, о длине которой мы наперед не знаем, см. ниже).

$\max_{x \in l_f} |x|$ — диапазон в его исходном смысле.

Диапазон — разность между наименьшим и наибольшим представляемыми числами.

Фиксированная запятая

Важны не значения, как таковые, а, скорее, их количество.

Плавающая запятая

Информативны границы диапазона, даже для чисел одного знака.

$\min_{x \in l_f} |x|$ — максимальная точность для малых чисел (но не относительная погрешность, т.к. в самом маленьком числе м.б. единица в конце мантиссы, о длине которой мы наперед не знаем, см. ниже).

$\max_{x \in l_f} |x|$ — диапазон в его исходном смысле.

В самом общем виде неотрицательная позиционная система обладает переменными основаниями $B_i : i \in \mathbb{N}, B_0 = 1$. Число записывается при помощи цифр $c_i : i \in [1, N], c_i \in [0, B_i)$. Значение же числа, записанного N цифрами[1] вычисляется по формуле:

$$v = \sum_{k=1}^{N} c_k \prod_{i=0}^{k-1} B_i$$

На практике, обычно, используется постоянное основание системы счисления B, совпадающее с B_1, так что:

$$v = \sum_{k=1}^{N} c_k B_1^{k-1}$$

[1] Слово *цифра* арабское, по происхождению такое же, как слово *шифр*

В самом общем виде неотрицательная позиционная система обладает переменными основаниями $B_i : i \in \mathbb{N}, B_0 = 1$. Число записывается при помощи цифр $c_i : i \in [1, N], c_i \in [0, B_i)$. Значение же числа, записанного N цифрами[1] вычисляется по формуле:

$$v = \sum_{k=1}^{N} c_k \prod_{i=0}^{k-1} B_i$$

На практике, обычно, используется постоянное основание системы счисления B, совпадающее с B_1, так что:

$$v = \sum_{k=1}^{N} c_k B_1^{k-1}$$

[1] Слово *цифра* арабское, по происхождению такое же, как слово *шифр*

Однозначность I

- Сначала покажем, что $1 \times B^N > \sum_{k=1}^{N} (B-1) \times B^{k-1}$.
- Для геометрической прогрессии

$$S_n = \sum_{i=1}^{n} b_i = \frac{b_1 - b_{n+1}}{1-q} = b_1 \frac{q^n - 1}{q-1},$$

соответственно, $$\sum_{k=1}^{N} B^{k-1} = 1 \times \frac{B^N - 1}{B-1},$$

А $$\sum_{k=1}^{N} (B-1) \times B^{k-1} = (B-1) \frac{B^N - 1}{B-1} = B^N - 1.$$

- Таким образом, N максимальными цифрами можно записать число, на 1 меньшее, чем единицей и N нулями.

Однозначность I

- Сначала покажем, что $1 \times B^N > \sum_{k=1}^{N} (B-1) \times B^{k-1}$.
- Для геометрической прогрессии

$$S_n = \sum_{i=1}^{n} b_i = \frac{b_1 - b_{n+1}}{1-q} = b_1 \frac{q^n - 1}{q-1},$$

$$\text{соответственно, } \sum_{k=1}^{N} B^{k-1} = 1 \times \frac{B^N - 1}{B-1},$$

$$\text{А } \sum_{k=1}^{N} (B-1) \times B^{k-1} = (B-1) \frac{B^N - 1}{B-1} = B^N - 1.$$

- Таким образом, N максимальными цифрами можно записать число, на 1 меньшее, чем единицей и N нулями.

Однозначность II

- Пусть две записи числа отличаются, отличия проявляются вплоть до k-й цифры. Отличия в k-й цифре должны компенсироваться младшими.

$$x = B^{N-1}c_N + \ldots + B^{k-1}c_k + \sum_{i=1}^{k-1} B^{i-1}c_i$$

и

$$x = B^{N-1}c_N + \ldots + B^{k-1}c'_k + \sum_{i=1}^{k-1} B^{i-1}c'_i$$

- Но это невозможно, т.к.

$$\forall c'_i \in [0, B) \quad B^{k-1} \times 1 > \sum_{i=1}^{k-1} B^{i-1}c'_i$$

Однозначность II

- Пусть две записи числа отличаются, отличия проявляются вплоть до k-й цифры. Отличия в k-й цифре должны компенсироваться младшими.

$$x = B^{N-1}c_N + \ldots + B^{k-1}c_k + \sum_{i=1}^{k-1} B^{i-1}c_i$$

и

$$x = B^{N-1}c_N + \ldots + B^{k-1}c'_k + \sum_{i=1}^{k-1} B^{i-1}c'_i$$

- Но это невозможно, т.к.

$$\forall c'_i \in [0, B) \quad B^{k-1} \times 1 > \sum_{i=1}^{k-1} B^{i-1}c'_i$$

Выбор

Выбор основания системы счисления — задача важная сама по себе. Например, двоичную арифметику при финансовых расчётах не используют, т.к. $\frac{1}{10}$ в двоичной системе — бесконечная периодическая дробь, а по сему точного конечного представления не имеет.

Упражнение

Придумать алгориѳм преобразования периодической дроби в простую. Можно запрограммировать.

Упражнение

Придумать алгориѳм преобразования простой дроби в периодическую. Можно запрограммировать.

Выбор

Выбор основания системы счисления — задача важная сама по себе. Например, двоичную арифметику при финансовых расчётах не используют, т.к. $\frac{1}{10}$ в двоичной системе — бесконечная периодическая дробь, а по сему точного конечного представления не имеет.

Упражнение

Придумать алгориѳм преобразования периодической дроби в простую. Можно запрограммировать.

Упражнение

Придумать алгориѳм преобразования простой дроби в периодическую. Можно запрограммировать.

осуществляется по простому алгоритму (использован язык Scheme, диалект ▸ LISP):

```
(define (val->posn value base)
  (if (= value 0)
      '() #| no digits for 0 |#
      (let ((digit (modulo value base)))
        (append
         (val->posn (/ (- value digit) base) base)
         (list digit)
         )
        )
      )
  )
```

Применим только к числам фиксированного размера. Беззнаковое число $X^{(m)} \in [0, P), P = B^N$. Машина хранит только N цифр.

$$[x]_P = [x + nP]_P = \{x + Pk | k \in \mathbb{Z}\}$$

$$[-1]_P = [P-1]_P, [-2]_P = [P-2]_P, \ldots$$

— мы сами вольны рассматривать числа, как положительные, или как отрицательные.

Применим только к числам фиксированного размера. Беззнаковое число $X^{(m)} \in [0, P), P = B^N$. Машина хранит только N цифр.

$$[x]_P = [x + nP]_P = \{x + Pk | k \in \mathbb{Z}\}$$

$$[-1]_P = [P - 1]_P, [-2]_P = [P - 2]_P, \ldots$$

— мы сами вольны рассматривать числа, как положительные, или как отрицательные.

- Для представления отрицательных чисел берется число $s \in (0, P-1)$ и дальше принимается $[-s]_P = [P-s]_P$.
- Обычно для четных P берут $s = P/2$. Например, для $P = 256$ — $s = 128$.
- Беззнаковый байт — $[0, 256)$, а знаковый — $[-128, 0) \cup [0, 128) = [-128, 128)$.
- Если машина работает с двоичными числами, то установленный старший бит соответствует строго отрицательному числу.

- Для представления отрицательных чисел берется число $s \in (0, P-1)$ и дальше принимается $[-s]_P = [P-s]_P$.
- Обычно для четных P берут $s = P/2$. Например, для $P = 256$ — $s = 128$.
- Беззнаковый байт — $[0, 256)$, а знаковый — $[-128, 0) \cup [0, 128) = [-128, 128)$.
- Если машина работает с двоичными числами, то установленный старший бит соответствует строго отрицательному числу.

Вопросы

EDU.DLUCIV.NAME