Алгоритмічні мови та програмування

Лабораторна робота № 3

«Похибки обчислень з плаваючою точкою.»

Теоретичні відомості та рекомендації:

Для виконання цієї роботи, необхідно повторити тему «Двоїчний запис реальних чисел» та «Похибки обчислень з реальними числами» на конспектом та літературою.

Число з плаваючою комою (точкою) - форма представлення дійсних чисел, в якій число зберігається у формі мантиси і показника ступеня. При цьому число з плаваючою комою має фіксовану відносну точність і мінливу абсолютну. Найбільш часто використовуване уявлення затверджено в стандарті IEEE 754. Реалізація математичних операцій з числами з плаваючою комою в обчислювальних системах може бути як апаратна, так і програмна.

В деяких, переважно англомовних та англофіцірованних, країнах при записі чисел ціла частина відділяється від дробової точкою, то в термінології цих країн фігурує назва "плаваюча крапка" (floating point). Так як в Росії ціла частина числа від дробової традиційно відокремлюється комою, то для позначення того ж поняття історично використовується термін "плаваюча кома", проте в даний час в російськомовній літературі і технічної документації можна зустріти обидва варіанти.

Назва "плаваюча кома" походить від того, що кома в позиційному поданні числа (десяткова кома, або, для комп'ютерів, двійкова кома - далі за текстом просто кома) може бути поміщена де завгодно щодо цифр в рядку. Це положення коми вказується окремо у внутрішньому поданні. Таким чином, подання числа у формі з плаваючою комою може розглядатися як комп'ютерна реалізація експоненційної запису чисел.

Перевага використання представлення чисел у форматі з плаваючою комою над виставою у форматі з фіксованою комою (і цілими числами) полягає в тому, що можна використовувати істотно більший діапазон значень при незмінній відносної точності. Наприклад, у формі з фіксованою комою число, що займає 8 розрядів в цілій частині і 2 розряду після коми, може бути представлено у вигляді 123456,78; 8765,43; 123,00 і так далі. У свою чергу, у форматі з плаваючою комою (у тих же 8 розрядах) можна записати числа

1,2345678; 1234567,8; 0,000012345678; 12345678000000000 і так далі, але для цього необхідно дворозрядне додаткове поле для запису показників ступеня 10 від 0 до 16 10, при цьому загальне число розрядів складе 8 +2 = 10.

Швидкість виконання комп'ютером операцій з числами, представленими у формі з плаваючою комою, вимірюється в мегафлопсах (від англ. *FLOPS - число операцій з плаваючою комою в секунду*), гігафлопсах і так далі, і є однією з основних одиниць вимірювання швидкодії обчислювальних систем.

Структура числа

Число з плаваючою комою складається з:

- Мантиси (що виражає значення числа без урахування порядку)
- Знака мантиси (що вказує на негативні чи позитивні числа)
- Порядку (виражає ступінь підстави числа, на яке множиться мантиса)
- Знака порядку

Нормальна форма і нормалізована форма

Нормальною формою числа з плаваючою комою називається така форма, в якій мантиса (без урахування знака) знаходиться на полуінтервалі [0; 1) ($0 \le a < 1$). Число з плаваючою комою, що знаходиться не в *нормальній формі,* втрачає точність у порівнянні з *нормальною формою.* Така форма запису має недолік: деякі числа записуються неоднозначно (наприклад, 0,0001 можна записати у 4 формах - 0,0001 10^{0}, 0,001 10^{-1}, 0,01 10^{-2}, 0,1 10 -3), тому поширена (особливо в інформатиці) також інша форма запису - *нормалізована,* в якій мантиса десяткового числа приймає значення від 1 (включно) до 10 (не включно), а мантиса двійкового числа приймає значення від 1 (включно) до 2 (не включно) ($1 \le a < 2$). У такій формі будь-яке число (крім 0) записується єдиним чином. Недолік полягає в тому, що в такому вигляді неможливо уявити 0, тому представлення чисел в інформатиці передбачає спеціальний ознака (біт) для числа 0.

Так як старший розряд (ціла частина числа) мантиси двійкового числа (крім 0) в *нормалізованому* вигляді дорівнює "1", то при записі мантиси числа в еом старший розряд можна не записувати, що і використовується в стандарті IEEE 754. В позиційних системах числення з підставою більшим, ніж 2 (в троичной, четверичной та ін), цієї властивості немає.

В обчислювальних машинах показник ступеня прийнято відокремлювати від мантиси буквою "E" (exponent). Наприклад, число 1,528535047 10^{-25} в більшості мов програмування високого рівня записується як 1.528535047E-25.

Способи машинної реалізації.

Існує кілька способів того, як рядки з цифр можуть представляти числа:

Найбільш поширений шлях подання значення числа з рядка з цифрами - у вигляді цілого числа - кома (radix point) за замовчуванням знаходиться в кінці рядка.

Загалом математичному уявленні рядок з цифр може бути як завгодно довгою, а положення коми позначається шляхом явною запису символу комою (або, на Заході, точки) в потрібному місці.

У системах з поданням чисел у форматі з фіксованою комою існує певна умова щодо положення коми. Наприклад, у рядку з 8 цифр умова може наказувати положення коми в середині запису (між 4-ю і 5-ю цифрою). Таким чином, рядок "00012345" означає число 1,2345 (нулі зліва завжди можна відкинути).

У експоненційної запису використовують стандартний *(нормалізований)* вид представлення чисел. Число вважається записаним в стандартному *(нормалізованому)* вигляді, якщо воно записане у вигляді aq^n , Де a , зване мантисою, таке, що , $1 \le a < q$ n - ціле, називається показник ступеня та q - ціле, основа системи числення (на практиці це зазвичай 10). Тобто у мантиси кома поміщається відразу після першої значущої (не дорівнює нулю) цифри, рахуючи зліва направо, а подальша запис дає інформацію про дійсному значенні числа. Наприклад, період обігу (на орбіті) супутника планети Юпітера Іо, який дорівнює 152853,5047 с, в стандартному вигляді можна записати як 1,528535047 10^5с. Побічним ефектом обмеження на значення мантиси є те, що в такого запису неможливо зобразити число 0.

Запис у формі з плаваючою комою схожа на запис чисел у стандартному вигляді, але мантиса і експонента записуються роздільно. Мантиса записується в *нормалізованому* форматі - з фіксованою комою, подразумеваемой після першої значущої цифри. Повертаючись до прикладу з Іо, запис у формі з плаваючою комою буде 1528535047 з показником 5. Це означає, що записане число в 10^5 разів більше числа 1,528535047, тобто для отримання подразумеваемого числа кома зсувається на 5 розрядів вправо. Однак, запис у формі з плаваючою комою використовується в основному в електронному поданні чисел, при якому використовується основа системи числення 2, а не 10. Крім того, в двійковій запису мантиса зазвичай денормалізована, тобто кома мається на увазі до першої цифри, а не після, і цілої частини взагалі не мається на увазі - так з'являється можливість і значення 0 зберегти природним чином. Таким чином, десяткова 9 в двійковому поданні з плаваючою комою буде записана як мантиса +1001000 ... 0 і показник +0 ... 0100. Звідси, наприклад, біди з двійковим поданням чисел типу однієї десятої (0,1), для якої двійкове подання мантиси виявляється періодичної двійковій

дробом - за аналогією з 1/3, яку не можна кінцевим кількістю цифр записати в десятковій системі числення.

Запис числа у формі з плаваючою комою дозволяє робити обчислення над широким діапазоном величин, поєднуючи фіксована кількість розрядів і точність. Наприклад, у десятковій системі надання чисел з плаваючою комою (3 розряду) операцію множення, яку ми б записали як

0,12 0,12 = 0,0144

в нормальній формі представляється у вигляді

(1,20 10 $^{-1}$) (1,20 10 $^{-1}$) = (1,44 10 -2).

У форматі з фіксованою комою ми б отримали вимушене округлення

0,120 0,120 = 0,014.

Ми втратили крайній правий розряд числа, так як даний формат не дозволяє коми "плавати" по запису числа.

Діапазон чисел, представимих у форматі з плаваючою комою

Діапазон чисел, які можна записати даними способом, залежить від кількості біт, відведених для представлення мантиси і показника. На звичайній

32-бітної обчислювальній машині, що використовує подвійну точність (64 біта), мантиса становить 1 біт знак + 52 біта, показник - 1 біт знак + 10 біт. Таким чином отримуємо діапазон точності приблизно від 4,94 10 -^{324} до 1.79 10 ^{308} (від 2 $^{-52}$ 2 $^{-1022}$ до ~ 1 2 ^{1024}). Пара значень показника зарезервована для забезпечення

можливості подання спеціальних чисел. До них відносяться значення NaN (Not a Number, не число) і + /-INF (Infinity, нескінченність), які утворюються в результаті операцій типу поділу на нуль нуля, позитивних і негативних чисел. Також сюди потрапляють денормалізованние числа, у яких мантиса менше одиниці. У спеціалізованих пристроях (наприклад GPU) підтримка спеціальних чисел часто відсутня. Існують програмні пакети, в яких обсяг пам'яті виділений під мантиссу і показник задається програмно, і обмежується лише обсягом доступної пам'яті ЕОМ.

Точність	**Одинарна**	**Подвійна**	**Розширена**
Розмір (байти)	4	8	10
Число десяткових знаків	7	15	19
Найменше значення (> 0), denorm	$1{,}4\ 10^{-45}$	$5{,}0\ 10^{-324}$	$1{,}9\ 10^{-4\ 951}$
Найменше значення (> 0), normal	1,2 10 -38	$2{,}3\ 10^{-308}$	$3{,}4\ 10^{-4\ 932}$
Найбільше значення	$3{,}4\ 10^{+38}$	$1{,}7\ 10^{+308}$	$1{,}1\ 10^{+4932}$
Поля	SEF	SEF	SEIF
Розміри полів	1-8-23	1-11-52	1-15-1-63

S - знак, E - показник ступеня, I - ціла частина, F - дробова частина

Так само, як і для цілих, знаковий біт - старший.

Машинний іпсилон

На відміну від чисел з фіксованою комою, сітка чисел, які здатна відобразити арифметика з плаваючою комою, нерівномірна: вона густіша для чисел з малими порядками і більш рідкісна - для чисел з великими порядками. Але відносна похибка запису чисел однакова і для малих чисел, і для великих. Тому можна ввести поняття машинної епсилон.

Машинним іпсилон називається найменше позитивне число ε таке, що $1 \oplus \varepsilon \neq 1$ (знаком $\oplus$ позначено машинне складання). На практиці це означає, що машинна арифметика для даного типу даних не розрізняє числа a та b такі, що $1 < a/b < 1 + \varepsilon$

Завдання:

1. Написати програму, яка обчислює у типі такі *float* дробі:
    a. 1/3
    b. 1/33
    c. 1/333
    d. 1/3333
    e. 1/33333
2. Обчислити ті самі дробі у типі *long double* та обчислити похибку округлення. Порівняти та пояснити результат.
3. Написати програму, яка виконує такі дії із змінною типу *float*:
    a. Додає 0.1 десять разів,
    b. Додає 0.01 сто разів,

c. Додає 0.001 тисячу разів

d. Додає 0.000001 мільйон разів

4. Обчислити похибку додавання. Порівняти та пояснити результат.
5. Написати програму, яка обчислює такі послідовності (лімітом є log 2), використовуючи змінні типу *float*:

а) $1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\cdots-\frac{1}{10000}$

б) $-\frac{1}{10000}+\frac{1}{9999}-\frac{1}{9998}+\cdots-\frac{1}{2}+1$

в) $\left(1+\frac{1}{3}+\frac{1}{5}+\cdots\frac{1}{9999}\right)-\left(\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\cdots\frac{1}{10000}\right)$

г) $\left(\frac{1}{9999}+\frac{1}{9997}+\cdots\frac{1}{3}+1\right)-\left(\frac{1}{10000}+\frac{1}{9998}+\cdots\frac{1}{4}+\frac{1}{2}\right)$

6. Порівняти та пояснити результат.
7. Написати функцію для обчислення машинного іпсилон. Обчислити машинний іпсилон для типу *float*.

Примітки:

Результати роботи подати письмово. **Обов`язково зробити висновки.**