МИНОБРНАУКИ РОССИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ «ЛЭТ И» ИМ. В. И. УЛЬЯНОВА (ЛЕНИНА)

Кафедра МО ЭВМ

ОТЧЕТ

по практической работе №1

по дисциплине «Вычислительная математика»

Тема: Особенность машинной арифметики, точность вычисления на ЭВМ

Студент гр. 0304 Крицын Д. Р.

Преподаватель Попова Е. В.

Санкт-Петербург

2021

Вариант 12.

Цель работы.

Изучить особенности вычислений с плавающей точкой.

Основные теоретические положения.

В фундаменте математического анализа прочно утвердилась система действительных чисел. Однако, как бы она не упрощала анализ, практические вычисления вынуждены обходиться без нее.

Обычным способом аппроксимации системы действительных чисел в ЭВМ посредством конкретных математических представлений являются числа с плавающей точкой. Множество F чисел с плавающей точкой характеризуется четырьмя параметрами: основанием b, точностью t и интервалом показателей [L,M]. Каждое число с плавающей точкой, принадлежащее F, имеет значение

$$x=\pm\left(\frac{d_1}{b}+\frac{d_2}{b^2}+\ldots+\frac{d_t}{b^t}\right)b^n,$$

где целые числа d_1, d_2, …, d_t удовлетворяют неравенствам $0\le d_j<b\left(j=\overline{1,t}\right)$ $L\le n\le M$. Если для каждого ненулевого x из F справедливо $d_1 \neq 0$, то система F называется нормализованной. Целое число n называется показателем, а число $f=\sum_{j=1}^{t} d_j/b^j$ — дробной частью. Обычно целое число b_n хранится по той или иной схеме представления, принятой для целых чисел, например, величины со знаком, дополнения до единицы или дополнения до двух. Если принять $-N\le n<N$, $N = 2^{m-1}$ то переходим к общепринятой

терминологии, при которой t — разрядность мантиссы, m — разрядность порядка.

Действительность машинная реализация представлений чисел с плавающей точкой может отличатся в деталях от рассматриваемой идеальной, однако различия несущественны, и на практике их почти всегда можно игнорировать, анализируя основные проблемы ошибок округления. Величина b^{-t} является оценкой относительной точности плавающей арифметики, которая характеризуется посредством машинного эпсилон, т. е. наименьшего числа с плавающей точкой ε, такого, что $1 + \varepsilon > 1$. Точное значение машинного эпсилон зависит не только от указанных выше параметров, но и от принятого способа округления. В вычислительных машинах используются различные системы чисел с плавающей точкой, причём в некоторых ЭВМ несколько систем. Так, для современных ПЭВМ характерно применение двух систем, которые называются обычной точностью и удвоенной точностью.

Рассматриваемое множество F не является котинуумом или даже бесконечным множеством. Оно содержит ровно $2(b - 1)b^{t}(M - L + 1) + 1$ чисел, которые расположены неравномерно (равномерность расположения имеет место лишь при фиксированном показателе). В силу того, что F — конечное множество не представляется возможным сколь-нибудь детально отобразить континуум действительных чисел. Например, действительные числа модулей, большим максимального элемента из F, вообще не могут быть отображены, причём последнее справедливо также в отношении ненулевых действительных чисел, меньших по абсолютной величине по сравнению с наименьшим положительным числом из F, и, наконец, каждое число из F должно представлять целый интервал действительных чисел, для которой, как и для любой модели, присущи допущения и ограничения.

На множестве F определены арифметические операции в соответствии с тем, как они выполняются ЭВМ. Эти операции, в свою очередь моделируются в машине посредством приближений, называемых плавающими операциями. Для плавающих операций сложения, вычитания, умножения и деления существует возможность возникновения ошибок округления, переполнения и появления машинного нуля. Следует отметить, что операции плавающего сложения и умножения коммутативны, но не ассоциативны, и дистрибутивный закон для них также не выполняется. Невыполнение указанных алгебраических законов, имеющих фундаментальное значение для математического анализа, приводит к сложности анализа плавающих вычислений и возникающих при этом ошибок.

Постановка задачи.

Используя готовые программы, выполнить исследования машинной арифметики и точности вычислений на ПЭВМ. Порядок выполнения работы следующий:

1) Исследования распредления нормализованных чисел с плавающей точкой на вещественной оси для различных значений параметров b, m, t.

2) Вычисление значения величины машинного эпсилон при различных значениях константы ***c*.**

3) Исследование абсолютных и относительных ошибок округления при вычислениях с плавающей точкой сумм чисел при различных значениях шага суммирования.

4) Исследование проявления ошибок округления, возникающих при вычислении показательной функции e^x для чисел с плавающей точкой для двух вариантов алгоритма вычислений, а также скорости сходимости обоих вариантов

Выполнение работы.

1. Были проведены исследования распределения нормализованных чисел с плавающей точкой на вещественной оси для различных значений параметров $b = 2$, $m = 3$, $t = 3$. Результаты расчётов см. в табл. 1.

Таблица 1 — Числа, сгенерированные программой с разными значениями параметров.

```
x[ 0]=0.000000
x[ 1]=0.003906
x[ 2]=0.004883
x[ 3]=0.005859
x[ 4]=0.006836
x[ 5]=0.007812
x[ 6]=0.009766
x[ 7]=0.011719
x[ 8]=0.013672
x[ 9]=0.015625
x[10]=0.019531
x[11]=0.023437
x[12]=0.027344
x[13]=0.031250
x[14]=0.039062
x[15]=0.046875
x[16]=0.054687
```

```
x[16]=0.054687
x[17]=0.062500
x[18]=0.078125
x[19]=0.093750
x[20]=0.109375
x[21]=0.125000
x[22]=0.156250
x[23]=0.187500
x[24]=0.218750
x[25]=0.250000
x[26]=0.312500
x[27]=0.375000
x[28]=0.437500
x[29]=0.500000
x[30]=0.625000
x[31]=0.750000
x[32]=0.875000
```

```
x[32]=0.875000
x[33]=1.000000
x[34]=1.250000
x[35]=1.500000
x[36]=1.750000
x[37]=2.000000
x[38]=2.500000
x[39]=3.000000
x[40]=3.500000
x[41]=4.000000
x[42]=5.000000
x[43]=6.000000
x[44]=7.000000
x[45]=8.000000
x[46]=10.000000
x[47]=12.000000
x[48]=14.000000
```

```
x[45]=8.000000
x[46]=10.000000
x[47]=12.000000
x[48]=14.000000
x[49]=16.000000
x[50]=20.000000
x[51]=24.000000
x[52]=28.000000
x[53]=32.000000
x[54]=40.000000
x[55]=48.000000
x[56]=56.000000
x[57]=64.000000
x[58]=80.000000
x[59]=96.000000
x[60]=112.00000
```

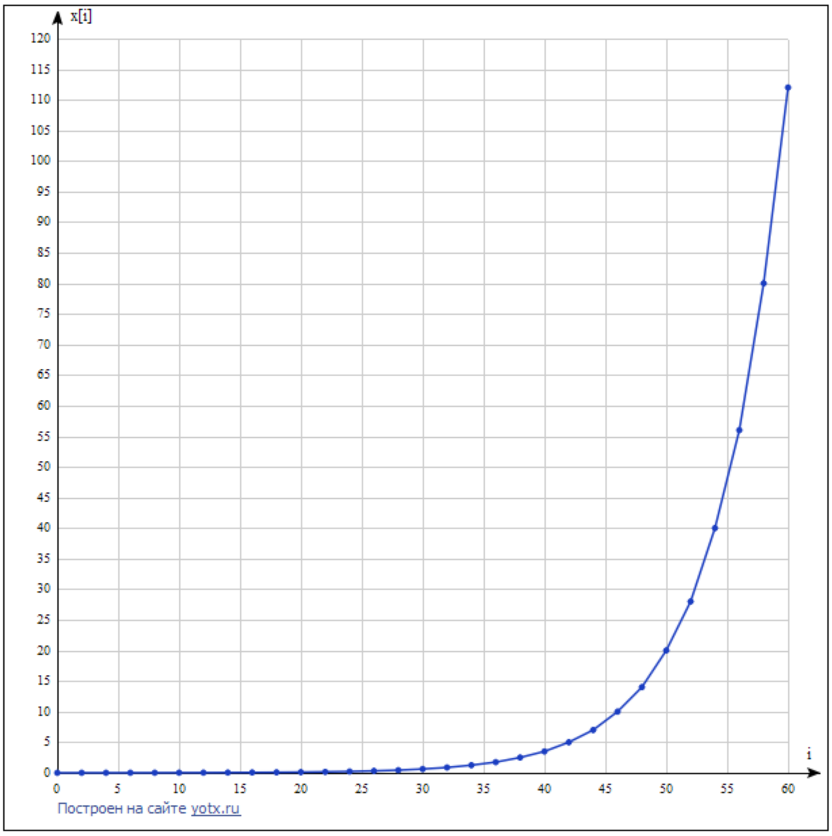

Рисунок 1 – график распределения чисел с плавающей точкой

Вывод: распредление нормализованных чисел с плавающей точкой на вещественной оси неравномерно. Плотность распределения велика при малых значениях параметров *b*, *m*, *t* и мала при значениях, близких к верхней границе диапазона множества чисел (x_i, когда *i* близок к $2(b — 1)b^t(M — L + 1) + 1)$.

2. Были вычислены значения ε, при разных значениях аргумента с (наименьшее значение ε, такое что $c+\varepsilon(c)>c$). Результаты вычислений см. в табл. 2.

Таблица 2 — Результаты вычисления ε при разных значениях с

Значение *c*	Значение *ε*	Шаг итерации
5	$888 * 10^{-19}$	50
10	$1776 * 10^{-19}$	49
25	$3552 * 10^{-19}$	48
125	$14211 * 10^{-19}$	46
512	$113687 * 10^{-19}$	43

Вывод: при маленьком значении *c* значение *ε* значительно меньше, чем при больших значениях *c*. При увеличении *c* в два раза *ε* так же увеличивается в 2 раза (с точностью до 10^{-19}). График зависимости *ε* от *c* показан на рис. 1.

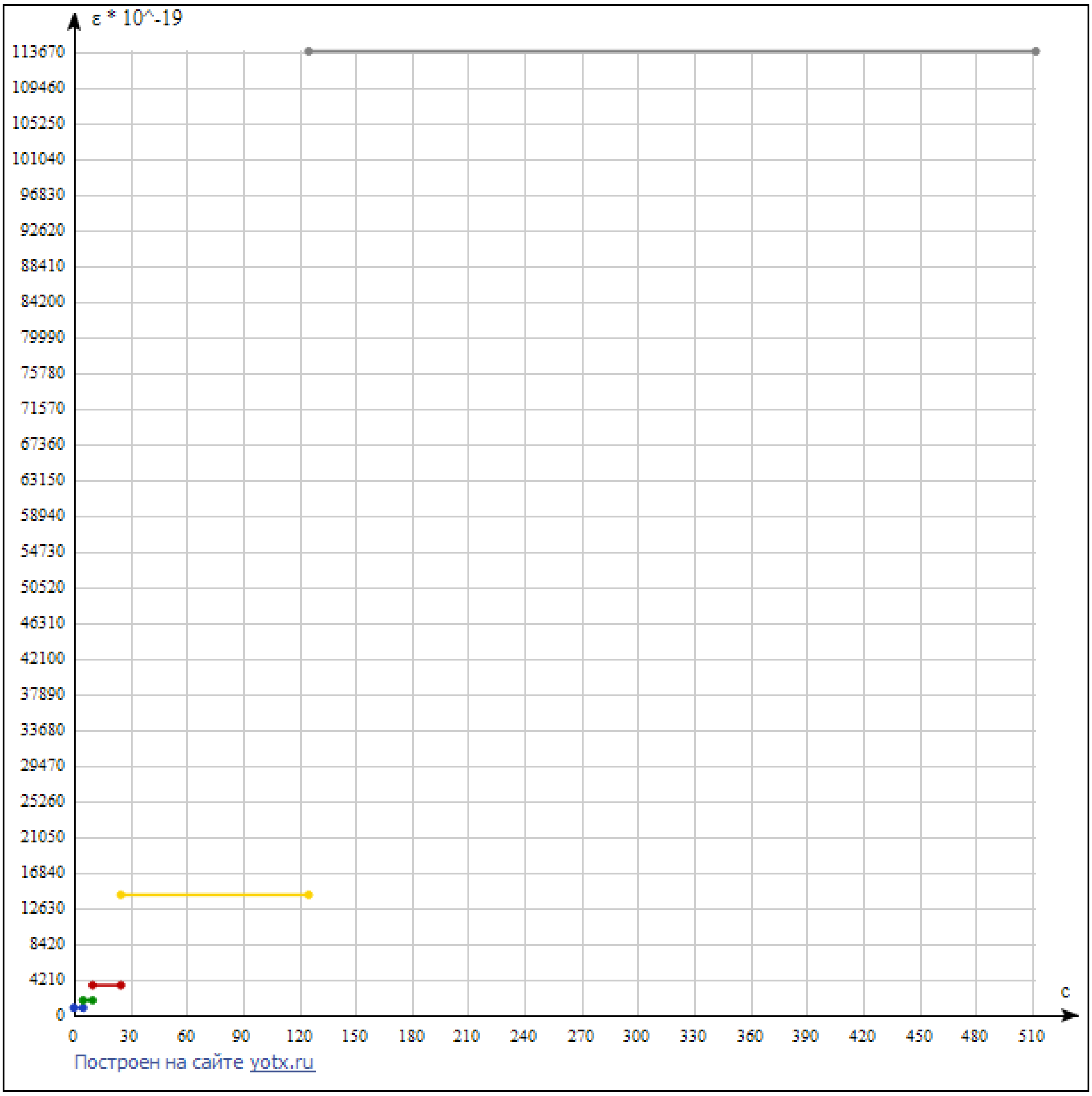

Рисунок 1 – График зависимости машинного ε от c

3. Было проведено исследование абсолютных и относительных ошибок округления при вычислениях с плавающей точкой сумм чисел ($\sum_{i=1}^{N}\frac{1}{N}, x_i = x_{i-1} + \frac{1}{N}$) при различных значениях шага суммирования. Результаты вычислений см. в табл. 3.

Таблица 3 — Результаты исследования абсолютных и относительных ошибок округления (*N* — шаг суммирования, *x — dx* — абсолютная погрешность, *(x-dx)/x* — относительная погрешность).

N	*x - dx*	*(x - dx)/x*
5	0.0000000149	0.000001%
30	0.0000000522	0.000005%
180	0.0000000242	0.000002%
400	0.0000000224	0.000002%
700	0.0000000107	0.000001%
1800	0.0000000242	0.000002%

Вывод: абсолютная погрешность, ровно как и относительная, колебалась относительно небольших значений в каждом шаге суммирования, поэтому накопление абсолютной погрешности суммы происходило относительно равномерно.

4. Было проведено исследование проявления ошибок округления, возникающих при вычислении показательной функции e^x для чисел с плавающей точкой для двух вариантов алгоритма вычислений: 1) $e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!}$; 2) $x = m + f$, $e^x = e^m * e^f$, где m — целая часть числа, f — дробная. Также были найдены скорости сходимости обоих вариантов. Результаты обработки введённых данных при помощи программы приведены в таблице 4.

Таблица 4 — Исследование проявление проявления ошибок округления, возникающих при вычислении функции e^x для двух алгоритмов.

Введенное значение x	Введенное значение ε	Разложение Тейлора	Улучшенный алгоритм	Абсолютная погрешность	Относительная погрешность
0.4	0.001	1.491818675558090 i = 5	1.491818675558090 i = 5	0	0%
3.7	0.001	40.447231220944140 i = 15	40.446946970066763 i = 6	0.000284250876511	0.000703%
6.55	0.001	699.243993004821 i = 22	699.2274893587799 i = 5	0.016503646041087	0.00236%
20.34	0.001	681631776.3248589 i = 60	681612312.7969973 i = 4	19463.52786171436	0.002856%
70.56	0.001	$4.403698556129254*10^{30}$ i = 196	$4.4035815481196796*10^{30}$ i = 5	117007932458083960000000000	0.002657%
130.33	0.001	$3.995771285902443*10^{56}$ i = 358	$3.995672191897767*10^{56}$ i = 4	$9.909400467662330*10^{51}$	0.00248%
255.77	0.001	$1.2008851983223055*10^{111}$ i = 698	$1.200865630516735*10^{111}$ i = 6	$1.956780631967093*10^{106}$	0.001629%

Вывод: при увеличении аргумента абсолютная погрешность возрастает, в то время как относительная погрешность мала и при больших значениях аргумента выходит за пределы машинного эпсилон и не может быть посчитана. Сходимость ряда Тейлора вычисляется медленно. Так, для аргумента 20.34 и порядка 0.001 требуется 60 итераций, для аргумента 70.56 — 196, и т.д. Улучшенный алгоритм является более рациональным вариантом, так как на целых числах дает сходимость за 1 итерацию, а на вещественных — за куда менее быстро возрастающее количество итераций.

Выводы.

В ходе выполнения заданий лабораторной работы, были исследованы машинная арифметика, точность вычислений на ПЭВМ, распределение нормализованных чисел на вещественной оси, абсолютные и относительные ошибки округления при вычислениях с плавающей точкой, зависимость машинного эпсилон от значения константы и проявление ошибок округления при вычислении показательной функции e^x. Все результаты исследований были занесены в таблицы, для некоторых из них были построены графики.