计算物理第二周作业

王潇卫 515072910032

1. Consider the 32-bit single-precision floating-point number A

	s	e	f
Bit position	31	30...23	22...0
value	0	0000 1110	1010 0000 0000 0000 0000 000

Determine the full value of A.

s = 0

e = 0000 1110 = $2^3 + 2^2 + 2 =$ （14）$_{10}$

p = e – 127 = -113

f = 1.1010 0000 0000 0000 0000 000 = $(1+0.5+0.125)_{10}$ = 1.625

$(-1)^s \times 1.625 \times 2^p = 1.565 \times 10^{-34}$

2. Sometimes the loss of significance error can be avoided by rearranging terms in the function using a known identity from trigonometry or algebra. Find an equivalent formula for the following functions that avoids a loss of significance.

   (a) $\ln(x+1) - \ln(x)$ for large x

   (b) $\sqrt{x^2+1} - x$ for large x

   (c) $\cos^2(x) - \sin^2(x)$ for $x \approx \pi/4$

   (d) $\sqrt{\dfrac{1+\cos(x)}{2}}$ for $x \approx \pi$

（a） ln(x+1) – ln(x) = ln (1 + 1/x)

（b） $\sqrt{x^2+1} - x = \frac{1}{\sqrt{x^2+1}+x}$

（c） $cos^2(x) - sin^2(x) = \cos(2x)$

（d） $\sqrt{\frac{1+\cos(x)}{2}} = \cos(\frac{x}{2})$

3. Write a program to determine the under- and overflow limits.

答案：over-flow: 8.9885e+307 ; under-flow: 4.9407e-324

程序：

```
%Write a programme to determine the under and over-flow limits.

under = 1;
over = 1;
a = 0;
b = 0;
```

```
while a ~= under
    a = under;
    under = under/2
end

while b~=over
    b = over;
    over = over*2
end
```

4. Write a program to determine your machine precision for single-precision floats and double-precision floats.

答案：single-precision is: 5.9605e-08

double-precision is: 1.1102e-16

程序：

```
clc
clear all

one = single(0);
sing_eps = single(1);
while one ~= 1
    sing_eps = single(sing_eps/2);
    one = single(1 + sing_eps);
end

fprintf (1,'single-precision is: \n')
single_eps = sing_eps

dou_one = 0;
dou_eps = 1;
while dou_one ~= 1
    dou_eps = dou_eps/2;
    dou_one = 1 + dou_eps;
end
fprintf (1,'double-precision is: \n')
double_eps = dou_eps
```

5. The value of π can be calculated with the series:

$$\pi = 4\sum_{n=1}^{\infty}(-1)^{n-1}\frac{1}{2n-1} = 4\left(1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\cdots\right)$$

- Describe your algorithm that calculates the value of π by using n terms of the series.
- Write a program to implement your algorithm and calculate the corresponding true relative error.
- Use the program to calculate π and the true relative error for: (a) $n = 10$, (b) $n = 20$, (c) $n = 40$ (d) $n > 50$ of your choice.
- Comment on your results of relative errors.

1. a = 0
2. 当 N 为偶数时，n =N-1；当 N 为奇数时，n=N-2
3. a = a + $\frac{1}{2n-1} - \frac{1}{2n+1} = \frac{2}{4n^2-1}$
4. n = n-2
5. 重复直至 n = 1
6. N 为偶数时输出 4a；N 为奇数时，输出 a =4(a + $\frac{1}{2N-1}$)

（2）

```
clear all
clc

fprintf(1,' 输入级数 N:   \n ');
N = input('N = ');
a = 0;

if mod(N,2) == 0
    n = N-1;
    while n>=1
        a = a +2/(4*n^2-1);
        n = n - 2;
    end
else
    n = N-2;
    while n>=1
        a = a +2/(4*n^2-1);
        n = n - 2;
    end
    a = a + 1/(2*N-1);
end
```

```
a = 4*a;
relative_error = 100 * (a-pi)/pi;
fprintf(1,' pi 的估计值为  %f \n ',a);
fprintf(1,' pi 的相对误差为  %f%%\n ',relative_error);
```

（3）

n	π 估计值	相对误差
10	3.041840	-3.175238%
20	3.091624	-1.590558%
40	3.116597	-0.795650%
100	3.131593	-0.318302%

（4）绘制 N=10、20、40、50、100、200 级数——误差双对数图，可以看到，随着 n 的增大，相对误差会逐渐减小，双对数图斜率拟合为-0.994，误差主要是 approximation errors