数值分析第二次大作业

林嘉成 2016011498
(自动化系 自 66)

目录

1 需求分析

本次大作业的目标是求 π^x。

为了计算得到 π、$\ln\pi$ 以及 π^x 的值，首先使用数值方法求取 π(如泰勒展开等等)，再使用得到的结果计算 $\ln\pi$(使用数值积分、常微分方程或牛顿迭代法)，最后使用得到的结果计算 π^x。其中，为了保证精度，需要选择合适的数值方法。

2 方案设计及相关说明

2.1 方案设计

求 π 使用了泰勒展开以及 BBP 公式。

求 $\ln\pi$ 使用了复化辛普生公式、复化柯特斯公式、牛顿迭代法[1]。

求 π^x 使用了四阶龙格-库塔公式和泰勒展开公式。

2.2 相关说明

2.2.1 存储误差

由于 double 的二进制有效数字为 52 位，且由于 $\log_2 50$ 已经达到了十进制的 15 位，所以如果只是按照 10 的指数次幂来分析舍入误差的话，误差最大会有 4 倍 (即有两位二进制被忽略)，所以本报告的舍入误差分析均是在 2 的指数次幂的基础上进行分析。

2.2.2 累积误差

对于累积误差的分析，除了求 π 的泰勒展开法、四阶龙格-库塔是在方法误差和舍入误差分别分析之外，其他的累积误差均在舍入误差中分析。

2.2.3 贴入代码

由于在分析舍入误差时，有的方法需要参考代码，故贴出了部分相关的代码。

3 求取 π

3.1 泰勒展开求 π

3.1.1 原理简述

将 $y = \arctan x$ 泰勒展开并代入 $x = \dfrac{\sqrt{3}}{3}$，可以得到

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - ... + \frac{(-1)^n x^{2n+1}}{2n+1} + ...$$

其中，$x = \dfrac{\sqrt{3}}{3}$。

使用牛顿迭代法求取 $\sqrt{3}$，选取初值为 2，即

$$f(x) = x^2 - 3$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

[1]代码中还写了复化龙贝格公式，但是没有在本报告中作误差分析

即有

$$x_{n+1} = x_n - \frac{x_n^2 - 3}{2x_n}$$

3.1.2 收敛速度

先考虑牛顿法求取 $\sqrt{3}$。由于牛顿法为二阶收敛，选取初值为 2。选取初始区间为 $[1.5, 2]$，记其上 $max|\phi^{''}(x)| = M$，则有

$$\begin{aligned} e_{n+1} &= \frac{1}{2}\phi^{''}(\xi_n)e_n^2 \\ |e_{n+1}| &\le \frac{M}{2}|e_n|^2 \\ \frac{M}{2}|e_{n+1}| &\le \left[\frac{M}{2}|e_n|^2\right]^2 \le ... \le \left[\frac{M}{2}|e_0|\right]^{2^{n+1}} \end{aligned}$$

即有

$$|e_{n+1}| \le \frac{2}{M}\left[\frac{M}{2}|e_0|\right]^{2^{n+1}}$$

计算可以得到，$M = \dfrac{8}{9}, |e_0| = |2 - \sqrt{3}| \le 0.3$，则代入可得

$$|e_{n+1}| < \frac{9}{4}\left(\frac{2}{15}\right)^{2^{n+1}}$$

令 $e_{n+1} < \epsilon$，则有

$$n \ge \log_2 P - 1,$$

其中，

$$P = \log_{\frac{2}{15}} \frac{4}{9}\epsilon$$

由此可以得知，牛顿法求取 $\sqrt{3}$ 的值的收敛速度为 $\log\log(1/\epsilon)$。

下面分析利用 $y = \arctan x$ 的泰勒展开求取 π 的收敛速度。对于第 $\arctan x$ 泰勒展开的第 n 项 (代入 $x = \sqrt{3}/3$)，有

$$\frac{(-1)^n x^{2n+1}}{2n+1} = \frac{1}{\sqrt{3}}\frac{(-1)^n}{(2n+1)3^n}$$

对于 n 项之后的余项放缩成等比数列，则有

$$R \le \frac{1}{\sqrt{3}(2n+1)}\frac{\left(\dfrac{1}{3}\right)^n}{1 - \dfrac{1}{3}}$$

即得到

$$R \le \frac{\sqrt{3}}{4n+2} \cdot \frac{1}{3^n} \le \epsilon$$

经过放缩后有

$$n \ge \log_3 (\sqrt{3}/(2\epsilon))$$

由此可以得知，将 $\arctan x (x = \sqrt{3}/3)$ 泰勒展开求取 π 的收敛速度为 $\log(1/\epsilon)$，即 $o(3^n)$。

3.1.3 计算代价

使用牛顿迭代法求取 $\sqrt{3}$ 时，迭代 6 次即已经收敛至 15 位精度。而计算代价主要取决于选取的 $\arctan x$ 展开选取的余项个数 n，且时间复杂度为 o(n)。

3.2 BBP 求 π

3.2.1 原理简述

BBP 公式 (Bailey-Borwein-Plouffe 公式) 提供了一个计算圆周率 π 的第 n 位二进制数的 spigot 算法。这个求和公式于 1995 年提出，公式的形式为

$$\pi = \sum_{k=0}^{\infty} \left[\frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right) \right]$$

这个公式也可以使用下述两个多项式的商来表示：

$$\pi = \sum_{k=0}^{\infty} \left[\frac{1}{16^k} \left(\frac{120k^2+151k+47}{512k^4+1024k^3+712k^2+194k+15} \right) \right]$$

3.2.2 收敛速度

设选取前 n 项，将余项放缩为等比数列，即有

$$R = \sum_{k=n+1}^{\infty} \left[\frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right) \right] < \sum_{k=n+1}^{\infty} \frac{1}{16^k}$$

可以得到

$$R < \frac{1}{15 \times 16^n} < \epsilon$$

可以解得

$$n > \log_{16}(1/(15\epsilon))$$

由此可以得出，BBP 的收敛速度为 $\log(1/\epsilon)$，即 $o(16^n)$。

3.2.3 计算代价

根据所写代码，BBP 算法求取 π 的时间复杂度为 o(n^2)。

4 求取 $\ln \pi$

4.1 复化辛普生公式求 $\ln \pi$

4.1.1 原理简述

考虑使用数值积分的方法进行求解，表达式如下。之后的科特斯公式、龙贝格公式均为数值积分的求解方法。

$$\ln \pi = \int_1^{\pi} \frac{1}{x} dx$$

复化辛普生公式形式如下

$$I = \frac{h}{6} \left[f(a) + 2 \sum_{k=1}^{n-1} f(x_k) + 4 \sum_{k=0}^{n-1} f(x_k + \frac{h}{2}) + f(b) \right]$$

4.1.2 收敛速度

对于复化辛普生公式，有

$$I - S_n = -\frac{h^4}{2880}\sum_{k=0}^{n-1} h \cdot f^{(4)}(\eta_k)$$

而由于当 n 趋近于无穷时，h 趋近于 0

$$\sum_{k=0}^{n-1} h \cdot f^{(4)}(\eta_k) \to \int_a^b f^{(4)}(x)dx = f^{(3)}(b) - f^{(3)}(a)$$

即复化辛普生公式的收敛速度为

$$I - S_n = o(\frac{1}{n^4})$$

4.1.3 计算代价

根据所写的代码，复化辛普生公式求取 $\ln \pi$ 的时间复杂度为 o(n)。

4.2 复化柯特斯公式求 $\ln \pi$

4.2.1 原理简述

由龙贝格算法，由复化梯形公式进行推导，可以得到复化辛普生公式、科特斯公式、龙贝格公式分别为

$$S_n = \frac{4}{3}T_{2n} - \frac{1}{3}T_n$$

$$Cn = \frac{16}{15}S_{2n} - \frac{1}{15}S_n$$

$$R_n = \frac{64}{63}C_{2n} - \frac{1}{63}C_n$$

其中，复化梯形公式为

$$T_n = \frac{h}{2}\left[f(a) + 2\sum_{k=1}^{n-1} f(x_k) + f(b)\right]$$

4.2.2 收敛速度

对于复化柯特斯公式，有[2]

$$I - C_n = -\frac{2h^6}{945 \times 4^6}\sum_{k=0}^{n-1} hf^{(6)}(\eta_k)$$

而由于当 n 趋于无穷时，h 趋近于 0

$$\sum_{k=0}^{n-1} hf^{(6)}(\eta_k) \to f^{(5)}(b) - f^{(5)}(a)$$

即复化柯特斯公式的收敛速度为

$$I - C_n = o(\frac{1}{n^6})$$

[2]这部分的柯特斯公式的余项来自于数值分析教材上的公式，在误差分析中，由另一种推导方式，但是不影响收敛速度的判断。

4.2.3 计算代价

由所写的代码，可以得到复化柯特斯的计算代价为 o(n)。

4.3 牛顿迭代法求 $\ln\pi$

4.3.1 原理简述

考虑选取 $f(x)=e^x-\pi$，选取初值为 1.5，则由牛顿迭代法有

$$x_{n+1}=x_n+\frac{\pi}{e^{x_n}}-1$$

而对于 e^x 的求取，考虑使用泰勒公式，即

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+...+\frac{x^n}{n!}+...$$

4.3.2 收敛速度

对于 e^x 的求取，设选取展开式的前 n 项，则后项与前项的比值小于 1 时收敛

$$\frac{x^{n+1}}{(n+1)!}\frac{n!}{x^n}=\frac{x}{n+1}<1$$

则有

$$n>x-1$$

则 n 取 20 可以满足条件。而 e^x 泰勒展开的拉格朗日余项为

$$R=\frac{e^\xi}{n!}x^n<\epsilon$$

则可以得出，e^x 的收敛速度为 $o(n!)$。

而牛顿法求根的收敛速度为 $o(P^{2^n})$，故这里不再赘述。

4.3.3 计算代价

由于牛顿迭代法的计算代价仍是线性，且其收敛极快，故牛顿迭代法的计算代价极小，接近常数。

5 求取 π^x

对于 π^x 的求取，为了尽可能减小误差，考虑将 x 拆分为整数部分 x_i 与小数部分 x_j。则对于整数部分，直接计算 π^{x_i}；而对于小数部分，计算 $\exp\{x_j\ln\pi\}$。最后将二者相乘即为得到的最终结果。以下方法均是为了求取 $\exp\{x_j\ln\pi\}$。

5.1 四阶龙格-库塔求取 π^x

5.1.1 原理简述

经典的四阶龙格-库塔公式如下：

$$
\begin{aligned}
&y_{n+1} = y_n + \frac{h}{6}[K1 + 2K2 + 2K3 + K4] \\
&K1 = f(x_n, y_n) \\
&K2 = f(x_n + \frac{h}{2}, y_n + \frac{h}{2}K_1) \\
&K3 = f(x_n + \frac{h}{2}, y_n + \frac{h}{2}K2) \\
&K4 = f(x_{n+1}, y_n + hK_3)
\end{aligned}
$$

为了求取 $y = e^x$，构造方程如下，其中 $y(0) = 1$

$$y^{'} = e^x$$

则由于 $y = y^{'}$，故可以将四阶 RK 公式整理成如下形式，即

$$y_{n+1} = \left(1 + h + \frac{h^2}{2} + \frac{h^3}{6} + \frac{h^4}{24}\right) y_n$$

5.1.2 收敛速度

由数值分析课件可以得知，四阶龙格-库塔为四阶精度，

$$y(x_{n+1}) - y_{n+1} = o(h^5)$$

即整体收敛速度为 $o(\frac{1}{n^4})$。

5.1.3 计算代价

由所写代码，可以得到四阶龙格-库塔公式的计算代价为 $o(n)$。

5.2 泰勒展开求 π^x

5.2.1 原理简述

见牛顿迭代法求 $\ln\pi$，故不再赘述。

5.2.2 收敛速度

由于 x_j 的区间为 (0,1)，故一定收敛，且由之前所证，收敛速度为 $o(n!)$。

5.2.3 计算代价

e^x 泰勒展开计算代价为线性，但是由于收敛较快，所以计算代价接近常数。

6 误差分析

在求取 π 和 $\ln\pi$ 时，使精度达到 14 位小数[3]，记误差 $\delta_0 = \dfrac{1}{2} \times 2^{-50}$; 在求取 π^x 时，使精度达到 6 位小数，记误差 $\delta_1 = \dfrac{1}{2} \times 2^{-20}$。

6.1 double 类型运算带来的误差

C++ 中的 double 数据类型是 64 位浮点数的基本数据结构。其中包括 1 位二进制符号位，11 位二进制指数位和 52 位二进制尾数位，其表示方式为二进制科学计数法。即 double 的二进制有效数字为 52 位，转换为 10 进制后可以认为 double 可以提供 15 位有效数字。

double 类型在进行加减乘除运算时是在二进制下进行，且运算的过程中会产生存储误差。假设一次运算后得到的二进制结果的最高位为 n，则该运算过程产生的舍入误差为 $\frac{1}{2} \times 2^{-52+n}$。

6.2 求取 π

6.2.1 泰勒展开求 π

方法误差

首先分析用牛顿法求取 $\sqrt{3}$ 的方法误差。取区间为 [1.5,2]，则由之前原理部分可知：

$$\phi(x) = \frac{1}{2}x + \frac{3}{2x}$$

$$|e_n| \le \frac{2}{M}\left[\frac{M}{2}|e_0|\right]^{2^n}$$

其中，M 为 $|\phi''(x)|$ 的最大值，为 $\dfrac{8}{9}$。所以有

$$|e_n| < \frac{9}{4}\left(\frac{2}{15}\right)^{2^n} < \delta_0$$

解得 $n \ge 4.17$，可以取 n=6。

$$|e_6| < 2.23 \times 10^{-56}$$

故由牛顿法求取 $\sqrt{3}$ 的值对后续操作的误差基本上由舍入误差决定。

下面分析用泰勒展开求取 π 的方法误差。先假定 $\sqrt{3}$ 的值是准确的，则由原理简述部分可以得到，余项为

$$R \le \frac{\sqrt{3}}{4n+2} \cdot \frac{1}{3^n}$$

考虑到前一步得到 $\sqrt{3}$ 造成的影响，则总的方法误差为

$$|E| < (1 - x^2 + x^4 - ... + (-1)^n x^{2n})|\Delta_x| + \frac{\sqrt{3}}{4n+2} \cdot \frac{1}{3^n}$$

进一步放缩

$$|E| < \frac{1}{1+x^2}|\Delta_x| + \frac{\sqrt{3}}{4n+2} \cdot \frac{1}{3^n} < \delta_0$$

其中，$x = \sqrt{3}/3, |\Delta_x| < 7.43 \times 10^{-57}$，代入可以解得

$$n > 29$$

[3] 由于在 C++ 内部 double 类型实现是通过二进制，故本报告的误差分析多数建立在二进制的基础之上。

故 n 可以取 30。此时，总的方法误差为

$$|E| < 6.90 \times 10^{-17} \approx 2^{-54}$$

舍入误差

先考虑使用牛顿法求取 $\sqrt{3}$ 值的舍入误差。根据所写代码，递推式如下，

$$x_{n+1} = x_n - \frac{x_n \times x_n - 3}{2x_n}$$

其中，$2x_n$ 为移位操作故精度不损失，$x_n \times x_n$ 数值范围在 [1,4]，二进制条件下整数部分最多有两位，故单步具有舍入误差为 $\frac{1}{2} \times 2^{-50}$，再考虑上两次减法和一次除法，可以得到一步操作的舍入误差为 $\frac{1}{2} \times 2^{-50}$。则整体的舍入误差为

$$\Delta_{n+1} \leq M\Delta_n + \frac{1}{2} \times 2^{-50}$$

其中

$$M = \frac{1}{2} - \frac{3}{2x^2} < \frac{1}{2}$$

则有

$$|\Delta_{n+1}| \leq \frac{1}{2}|\Delta_n| + \frac{1}{2} \times 2^{-50}$$

$$|\Delta_{n+1}| - 2^{-50} \leq \frac{1}{2}(|\Delta_n| - 2^{-50}) \leq ...$$

$$|\Delta_{n+1}| - 2^{-50} \leq \left(\frac{1}{2}\right)^{n+1} (|\Delta_0| - 2^{-50})$$

代入 $\Delta_0 = \frac{1}{2} \times 2^{-50}$，即有

$$|\Delta_n| \leq 2^{-51}\left(2 - \left(\frac{1}{2}\right)^n\right)$$

由上述方法误差取 n=6，则可以得到

$$|\Delta_6| \leq 8.81 \times 10^{-16} \approx 2^{-50}$$

下面分析用泰勒展开求取 π 的方法误差。首先假定 x 不存在舍入误差，由于各项整数部分均为 0，考虑最差的情况，设每一项的舍入误差为 $\frac{1}{2} \times 2^{-52}$。又由于 n=30，所以所有项相加的舍入误差为 $6.66 \times 10^{-15} \approx 2^{-47}$。故整体的舍入误差为

$$|E| < \frac{1}{1+x^2}|\Delta_x| + 2^{-47}$$

其中，$x = \sqrt{3}/3, |\Delta_x| < |\Delta_6| = 8.81 \times 10^{-16} \approx 2^{-50}$，代入后可以得到整体舍入误差为

$$|E| < 7.32 \times 10^{-15} \approx 2^{-47}$$

这里的舍入误差考虑的是最差情况。

故总的误差为

$$|E| < 7.389 \times 10^{-15} \approx 2^{-47}$$

由于舍入误差的上界在放缩计算时放大较多，所以我认为这里只需要根据方法误差确定 n 的值，且其实这里所算的总误差没有太大的价值，因为在精度较高的情况下，估计的舍入误差会相对非常大，而这时只需要考虑方法误差。[4]后续的误差分析均遵循该原则。

[4]在精度较低的情况下，舍入误差和方法误差都需要考虑进去。

6.2.2 BBP 求 π

```
//k=15 k为位数精度
double Cal_Pi::BBPPi(int k)
{
    if (k < 0)
        k = 0;
    double result = 0;
    for (int i = 0; i < k; i++)
    {
        double temp = 1.0 * (120 * i * i + 151 * i + 47) /
        (512 * i*i*i*i + 1024 * i*i*i + 712 * i*i + 194 * i + 15);
        for (int j = 0; j < i; j++)
            temp /= 16.0;
        result += temp;
    }
    return result;
}
```

方法误差

$$R = \sum_{k=n+1}^{\infty}\left[\frac{1}{16^k}\left(\frac{4}{8k+1}-\frac{2}{8k+4}-\frac{1}{8k+5}-\frac{1}{8k+6}\right)\right] < \sum_{k=n+1}^{\infty}\frac{1}{16^k}$$

$$|E| < \frac{1}{15\times 16^n} < \delta_0$$

可以解得 $n > 11.7$，故这里取 $n = 12$。此时 BBP 的方法误差为

$$|E| < 2.37\times 10^{-16} \approx 2^{-51}$$

舍入误差

根据所写的代码，逐项分析舍入误差。

$$\pi \approx \sum_{k=0}^{n}\left[\frac{1}{16^k}\left(\frac{120k^2+151k+47}{512k^4+1024k^3+712k^2+194k+15}\right)\right]$$

分子和分母均为整数，所以在计算的过程中无误差，而在分子与分母相除之后，会产生误差，误差大小放大至 $\frac{1}{2}\times 2^{-50}$。又 double 类型除以 16^k 时，相当于移位操作，几乎无误差。故单步的误差只有 $\frac{1}{2}\times 2^{-50}$。所以整体舍入误差为

$$|E| < \frac{n}{2}\times 2^{-50}$$

由方法误差得到 $n = 12$，代入可以得到最差情况下的舍入误差为

$$|E| < 5.33\times 10^{-15} \approx 2^{-47}$$

故整体的误差为

$$|E| < 2.37\times 10^{-16} + 5.33\times 10^{-15} = 5.57\times 10^{-15} \approx 2^{-47}$$

6.3 求取 $\ln\pi$

6.3.1 复化辛普生公式求 $\ln\pi$

方法误差

先假设所代入的 π 值 (即 b) 是准确的，则由原理简述得到以下等式

$$\ln\pi = \int_1^\pi \frac{1}{x}dx$$

$$f(x) = \frac{1}{x}$$

$$f^{(4)}(x) = \frac{24}{x^5}$$

$$R[f] = -\frac{(b-a)^5}{2880n^4}f^{(4)}(\eta)$$

在区间 [1,π] 上，$f^{(4)}(x)$ 的最大值为 24，则有

$$|R[f]| < \frac{(\pi-1)^5}{2880n^4} \times 24 = \frac{(\pi-1)^5}{120n^4} < \delta_0$$

可以解得 $n > 5392$，但实际上 $n = 2000$ 时，能到达 14 位小数精度，因为四阶导数放缩的时候导致偏大。这里按照计算所得结果，取 $n = 8192$。故此时的方法误差为

$$|R[f]| \leq 8.34 \times 10^{-17} \approx 2^{-53}$$

舍入误差及累积误差

$$S_n = \frac{h}{6}\left[f(a) + 2\sum_{k=1}^{n-1} f(x_k) + 4\sum_{k=0}^{n-1} f(x_k + \frac{h}{2}) + f(b)\right]$$

先考虑加法产生的误差，中括号内的值近似 $\frac{6\ln\pi}{h}$，即

$$\frac{6\ln\pi}{\pi-1}n$$

代入 $n = 8192$ 则可以计算出这个结果的二进制下的最高位为 15，则加法所产生的误差为 $\dfrac{1}{2}\times 2^{-37}$，即十进制下保留至 11 位小数。则舍入误差为

$$|\Delta S_n| \leq \left|\frac{\partial S_n}{\partial h}\right| |\Delta h| + \sum_{k=1}^{n-1}\left|\frac{\partial S_n}{\partial f(x_k)}\right| |\Delta f(x_k)| + \sum_{k=0}^{n-1}\left|\frac{\partial S_n}{\partial f(x_k+\frac{h}{2})}\right| \left|\Delta f(x_k + \frac{h}{2})\right| + \left|\frac{\partial S_n}{\partial f(a)}\right| |\Delta f(a)| + \left|\frac{\partial S_n}{\partial f(b)}\right| |\Delta f(b)| + \frac{1}{2}\times 2^{-37}$$

逐项进行分析，对于第一项，有

$$h = \frac{\pi-1}{n}$$

根据 h 的表达式，由于 n 为 2 的整数次幂，故除以 n 只是在进行移位，即没有误差，所以有

$$\Delta h = \frac{\Delta\pi}{n}$$

所以

$$\left|\frac{\partial S_n}{\partial h}\right||\Delta h| = \left|\frac{\partial S_n}{\partial \pi}\right||\Delta \pi| < \frac{1}{\pi}|\Delta \pi|$$

对于第二项

$$|\Delta f(x_k)| < \left|\frac{1}{x_k^2}\right||\Delta x_k| + \frac{1}{2}\times 2^{-50}$$

其中

$$x_k = 1 + kh = 1 + k\frac{\pi - 1}{n}$$

由于 n 为 2 的整数次幂，所以除以 n 不产生误差，故乘法会产生舍入误差，即

$$|\Delta x_k| < \frac{k}{n}|\Delta \pi| + \frac{1}{2}\times 2^{-50}$$

所以带回 $|\Delta f(x_k)|$，有

$$|\Delta f(x_k)| < \left|\frac{1}{x_k^2}\right||\Delta x_k| + \frac{1}{2}\times 2^{-50} < \frac{k}{n}|\Delta \pi| + 2^{-50}$$

即

$$\sum_{k=1}^{n-1}\left|\frac{\partial S_n}{\partial f(x_k)}\right||\Delta f(x_k)| \leq \frac{h}{3}\left(\sum_{k=1}^{n-1}\frac{k}{n}|\Delta\pi| + (n-1)\times 2^{-50}\right) = \frac{(n-1)(\pi-1)}{6n}\left(|\Delta\pi| + 2^{-49}\right)$$

对于下一项, 与上述同理

$$\sum_{k=0}^{n-1}\left|\frac{\partial S_n}{\partial f(x_k + \frac{h}{2})}\right|\left|\Delta f(x_k + \frac{h}{2})\right| \leq \frac{(\pi - 1)}{3}(|\Delta\pi| + 2^{-49})$$

再下一项，由于 $f(a) = 1$，即 $|\Delta f(a)| = 0$，故

$$\left|\frac{\partial S_n}{\partial f(a)}\right||\Delta f(a)| = 0$$

再下一项

$$|\Delta f(\pi)| < \left|\frac{1}{\pi^2}\right||\Delta \pi| + \frac{1}{2}\times 2^{-50}$$

故代入有

$$\left|\frac{\partial S_n}{\partial f(\pi)}\right||\Delta f(\pi)| < \frac{h}{6}|\Delta f(\pi)| + \frac{1}{2}\times 2^{-50} < \frac{\pi - 1}{6n}\left[\frac{1}{\pi^2}|\Delta\pi| + \frac{1}{2}\times 2^{-50}\right] + \frac{1}{2}\times 2^{-37}$$

所以整体的舍入误差为

$$\begin{aligned}|\Delta S_n| \leq \left(\frac{1}{\pi} + \frac{(n-1)}{6n} + \frac{\pi - 1}{3} + \frac{\pi - 1}{6n}\cdot\frac{1}{\pi^2}\right)|\Delta\pi| + (\frac{(n-1)(\pi-1)}{6n}\times 2^{-49} \\ + \frac{\pi - 1}{3}\times 2^{-49} + \frac{\pi - 1}{12n}\times 2^{-50} + \frac{1}{2}\times 2^{-37})\end{aligned}$$

若代入之前所计算出的 π 的误差，则得到舍入误差最大上限为

$$|\Delta S_n| < 2^{-36}$$

显然复化辛普生的舍入误差非常大。

6.3.2 复化柯特斯公式求 $\ln\pi$

方法误差

复化柯特斯公式如下，其中 $h=\frac{\pi-1}{n}$ $x_k=1+kh$

$$C_n=\frac{h}{90}\sum_{k=0}^{n-1}\left[7f(x_k)+12f(x_k+\frac{h}{4})+32f(x_k+\frac{h}{2})+12f(x_k+\frac{3h}{4})+7f(x_k+h)\right]$$

首先求取柯特斯公式在 [a,b] 上的余项。因为柯特斯公式具有 5 阶代数精度，故可以构造 H(x) 满足如下条件

$$H(a)=f(a)$$
$$H(\frac{3a+b}{4})=f(\frac{3a+b}{4})$$
$$H(\frac{a+b}{2})=f(\frac{a+b}{2})$$
$$H(\frac{a+3b}{4})=f(\frac{a+3b}{4})$$
$$H(b)=f(b)$$
$$H^{'}(\frac{a+b}{2})=f^{'}(\frac{a+b}{2})$$

这样可以得到

$$f(x)-H(x)=\frac{f^{(6)}(\eta)}{6!}(x-a)\left(x-\frac{3a+b}{4}\right)\left(x-\frac{a+b}{2}\right)^2\left(x-\frac{a+3b}{4}\right)(x-b)$$

则可以得到柯特斯公式余项为

$$\begin{aligned}|R_c|&=\left|\int_a^b\frac{f^{(6)}(\eta)}{6!}(x-a)\left(x-\frac{3a+b}{4}\right)\left(x-\frac{a+b}{2}\right)^2\left(x-\frac{a+3b}{4}\right)(x-b)\right|\\&\leq\left|\frac{f^{(6)(\eta)}}{6!}\right|\int_a^b\left|(x-a)\left(x-\frac{3a+b}{4}\right)\left(x-\frac{a+b}{2}\right)^2\left(x-\frac{a+3b}{4}\right)(x-b)\right|dx\\&=\frac{|f^{(6)}(\eta)|}{720\times2048}(b-a)^7\end{aligned}$$

所以复化柯特斯公式的余项为

$$|R[f]|=\left|\sum_{k=0}^{n-1}|R_c|\right|=\frac{(\pi-1)^7}{720\times2048n^6}\left|f^{(6)}(\eta)\right|<\delta_0$$

又 $f^{(6)}(\eta)$ 的最大值为则可以解得 $n>247$，为了减少舍入误差，选取 $n=256$。此时，方法误差为

$$R[f]<3.58\times10^{-16}\approx2^{-51}$$

舍入误差及累积误差

代码中，复化柯特斯公式是使用了龙贝格算法从梯形公式外推出。即

$$S_n=\frac{4}{3}T_{2n}-\frac{1}{3}T_n$$

$$Cn=\frac{16}{15}S_{2n}-\frac{1}{15}S_n$$

下面对复化梯形公式进行分析，复化梯形公式如下

$$T_n = \frac{h}{2}\left[f(a) + 2\sum_{k=1}^{n-1} f(x_k) + f(b)\right]$$

具体的舍入误差分析与复化辛普生公式相似，即

$$|\Delta T_n| \le \left|\frac{\partial T}{\partial h}\right| |\Delta h| + \sum_{k=1}^{n-1}\left|\frac{\partial T}{\partial f(x_k)}\right| |\Delta f(x_k)| + \left|\frac{\partial T}{\partial f(a)}\right| |\Delta f(a)| + \left|\frac{\partial T}{\partial f(b)}\right| |\Delta f(b)| + \frac{1}{2}\times 2^{-43}$$

分别求取每一项，中间过程省略，可以得到

$$\left|\frac{\partial T}{\partial h}\right| |\Delta h| < \frac{1}{\pi}|\Delta\pi|$$

$$\sum_{k=1}^{n-1}\left|\frac{\partial T}{\partial f(x_k)}\right| |\Delta f(x_k)| < \frac{(n-1)(\pi-1)}{2n}(|\Delta\pi| + x^{-49})$$

$$\left|\frac{\partial T}{\partial f(a)}\right| |\Delta f(a)| = 0$$

$$\left|\frac{\partial T}{\partial f(b)}\right| |\Delta f(b)| < \frac{\pi-1}{2n}\left(\frac{1}{\pi^2}|\Delta\pi| + \frac{1}{2}\times 2^{-50}\right) + \frac{1}{2}\times 2^{-50}$$

故复化梯形公式的整体舍入误差为

$$|\Delta T_n| < \left(\frac{1}{\pi} + \frac{(n-1)(\pi-1)}{2n} + \frac{\pi-1}{2n\pi^2}\right)|\Delta\pi| + \left(\frac{(n-1)(\pi-1)}{n}2^{-50} + \frac{\pi-1}{n}2^{-52} + 2^{-44}\right)$$

则对于 S_n，有 S_n 的整体舍入误差为

$$|\Delta S_n| \le \frac{4}{3}|T_{2n}| + \frac{1}{3}|\Delta T_n| + \frac{1}{2}\times 2^{-50}$$

同理

$$|\Delta C_n| \le \frac{16}{15}|\Delta S_{2n}| + \frac{1}{15}|\Delta S_n| + \frac{1}{2}\times 2^{-50}$$

代入之后即为复化柯特斯的整体舍入误差，代入推出的最差情况下的 π 的整体误差进行数值计算后，得到舍入误差最大上限为[5]

$$|\Delta C_n| < 2^{-43}$$

6.3.3 牛顿迭代法求 $\ln\pi$

```
//计算e^x 泰勒展开
double Myexp(double x, int nCount)
{
    double result = 1;
    double Factorial = 1;
    double xinput = x;
    for (int i = 1; i <= nCount; i++)
```

[5] 此时，如果代入 $|\Delta\pi|$ 的实际的误差值和所取的 n 进行计算，则可以得到舍入误差的上限。但是由于在题目的背景下，无法得知实际的误差值，且在计算得到 π 时的舍入误差时，由于放缩较大，所以得到的舍入误差的上界非常大，所以若代入该值计算则没有任何意义而言。所以只需要考虑方法误差即可。

```
    {
        Factorial *= i;
        result += 1.0 * xinput / Factorial;
        xinput *= x;
    }
    return result;
}

//牛顿法 e^x = pi
double Cal_lnPi::Newton(double x_unit, double pi, int nCount)
{
    double result = x_unit;
    for (int i = 0; i < nCount; i++)
        result += pi / Myexp(result,20) - 1;
    return result;
}
```

方法误差

先考虑使用泰勒展开求取 e^x。x 的初值选取为 1.5，有

$$R = \frac{e^\xi}{n!}x^n < \frac{e^{1.5}}{n!}1.5^n < \delta_0$$

故可以解得 $n > 20$，所以取 $n = 21$。此时，e^x 泰勒展开所得到的方法误差为

$$R < 4.38 \times 10^{-16} \approx 2^{-51}$$

再考虑使用牛顿法求取 $e^x = \pi$。先假设求得的 e^x 为准确值。初值选取为 1.5，相关等式如下

$$\phi(x) = x + \frac{\pi}{e^x} - 1$$

故 $M = \max|\phi^{''}(x)| = 1$，$|e_0| < 0.4$，且有

$$|e_n| \leq \frac{2}{M}\left[\frac{M}{2}|e_0|\right]^{2^n}$$

即

$$|e_n| < 2 \times \left[\frac{1}{5}\right]^{2^n} < \delta_0$$

则可以解得 $n > 4.4$，即取 $n = 5$。此时，牛顿迭代法的方法误差为

$$|e_n| < 8.6 \times 10^{-23} \approx 2^{-73}$$

舍入误差及累积误差

首先考虑使用泰勒展开求取 e^x 造成的舍入误差。根据所写代码，由于使用 double 类型存储阶层，且 double 最大可以存储 17!，所以从第 19 项开始，对于阶乘会产生一定的舍入误差。又由于在做除法之后，该部分的值微乎其微，故可以考虑忽略该部分所造成的误差。对于前 18 项，由于所求结果最大值为 $e^{1.5}$，二进制下的最高位为 3，所以泰勒展开求取 e^x 整体的舍入误差为

$$|E| < \frac{1}{2} \times 2^{-49}$$

所以，泰勒公式求取 e^x 的总误差为

$$|E| < 1.33 \times 10^{-15} \approx 2^{-49}$$

下面分析牛顿迭代法的整体舍入误差。由于牛顿迭代的过程中，最大值不超过 2，所以在假设 e^x 的求取准确的情况下的舍入误差为 $\frac{1}{2}\times 2^{-51}$，考虑 e^x 求取的误差对牛顿迭代法造成的影响，分别考虑 x 和 e^x 产生的累积误差。则牛顿迭代法的整体舍入误差为

$$|\Delta_{n+1}| \leq |\Delta_n| + \pi e^{-x_n}|\Delta e^{-x_n}| + \frac{1}{2}\times 2^{-51} < |\Delta_n| + |\Delta e^{-x}| + \frac{1}{2}\times 2^{-51}$$

递推即有

$$|\Delta_{n+1}| \leq |\Delta_0| + (n+1)(|\Delta e^{-x}| + \frac{1}{2}\times 2^{-51})$$

其中

$$|\Delta e^{-x}| < \frac{1}{e^{2x}}|\Delta e^x| + \frac{1}{2}\times 2^{-52} < \frac{1}{\pi^2}|\Delta e^x| + \frac{1}{2}\times 2^{-52}$$

将其代入可以得到下式，其中 $\Delta_0 = 0$

$$\Delta_{n+1} \leq (n+1)2^{-50}$$

这里取 n=5，代入可以计算得到牛顿迭代法整体舍入误差为

$$|\Delta_5| < 4.44\times 10^{-15} \approx 2^{-47}$$

6.4 求取 π^x

为了尽可能减小误差，在这里将 x 拆成整数部分 x_i 和小数部分 x_j。整数部分使用之前算出 π 的值进行迭代相乘，用以下方法求取 π^{x_j}，由于其最大值不超过 4，所以求取 π^{x_j} 的单步舍入误差最大为 $\frac{1}{2}\times 2^{-50}$。

先分析由 x 的整数部分带来的误差 (只有舍入累积误差)。由题目要求，x_i 的取值范围为 [1,10]。

$$y_1 = \pi^x$$

由于 π^{10} 在二进制下最高位为第 17 位，则该部分误差为

$$|\Delta y_1| < x\pi^{x-1}|\Delta\pi| + \frac{1}{2}\times 2^{-35} < 3\times 10^5|\Delta\pi| + \frac{1}{2}\times 2^{-35}$$

若代入所计算的 π 的总误差，则得到 π^{x_i} 的最大舍入误差为

$$|\Delta y_1| < 2.15\times 10^{-9} \approx 2^{-28}$$

6.4.1 四阶龙格-库塔求取 π^x

由之前的原理简述可以得到关系式如下

$$y_{n+1} = \left(1 + h + \frac{h^2}{2} + \frac{h^3}{6} + \frac{h^4}{24}\right) y_n$$

方法误差

假设 y_n 为准确值，则将 $y(x_{n+1})$ 在 x_n 处泰勒展开，则有

$$|R| = \frac{h^5}{120}\cdot y^{(5)}(\xi_n)$$

考虑方法累积误差，则有

$$|\Delta_{n+1}| < \left(1 + h + \frac{h^2}{2} + \frac{h^3}{6} + \frac{h^4}{24}\right)|\Delta_n| + \frac{L\cdot h^5}{120}$$

其中 $L=\max y^{(5)}(x)=\pi$。则递推后，有

$$|\Delta_{n+1}|+A<\left(1+h+\frac{h^2}{2}+\frac{h^3}{6}+\frac{h^4}{24}\right)^{n+1}A$$

其中

$$A=\frac{\frac{Lh^4}{120}}{1+\frac{h}{2}+\frac{h^2}{6}+\frac{h^3}{24}}$$

当 $n\to\infty$ 时，$h\to 0$，且 $\max nh=\ln\pi$，则有

$$|\Delta_{n+1}|<(\pi-1)\frac{Lh^4}{120}<\frac{1}{2}\times 2^{-50}$$

可以解得 $n>3837$，即取 $n=4069$，此时方法误差可以达到 15 位精度。但实际上，若只需要实现 6 位小数精度的话，可以取 $n=1024$，此时的方法累积误差为

$$|\Delta_{n+1}|<8.76\times 10^{-14}\approx 2^{-43}$$

又因为求取整数部分 π^{x_i} 无方法误差，所以此时使用四阶 RK 公式可以达到 13 位小数精度。

舍入误差

考虑整体舍入误差，有

$$|\Delta_{n+1}|<\left(1+h+\frac{h^2}{2}+\frac{h^3}{6}+\frac{h^4}{24}\right)|\Delta_n|+\max|y_n|\left(1+h+\frac{1}{2}h^2+\frac{1}{6}h^3\right)|\Delta h|+\frac{1}{2}\times 2^{-50}$$

由于区间为 $(0,\ln\pi)$，故有

$$|\Delta_{n+1}|<\left(1+h+\frac{h^2}{2}+\frac{h^3}{6}+\frac{h^4}{24}\right)|\Delta_n|+\pi\left(1+h+\frac{1}{2}h^2+\frac{1}{6}h^3\right)|\Delta h|+\frac{1}{2}\times 2^{-50}$$

而又有 $h<\frac{\ln\pi}{n}$，且 n 为 2 的整数次幂，故没有舍入误差，则

$$|\Delta h|<\frac{|\Delta\ln\pi|}{n}<|\Delta\ln\pi|$$

递推后可以得到

$$|\Delta_{n+1}|<\frac{\pi-1}{h}\left[\pi\left(1+h+\frac{1}{2}h^2+\frac{1}{6}h^3\right)|\Delta\ln\pi|+\frac{1}{2}\times 2^{-50}\right]$$

设区间长度为 l，则 $h=\dfrac{l}{n}$，故整体舍入误差的上限受区间长度的影响。再考虑到整数部分 π^x 的最大舍入误差上限，可以得到四阶 RK 法的舍入误差上限的最小值为 $2.15\times 10^{-19}\approx 2^{-28}$。经过实际程序进行测试，满足六位小数精度的要求。

6.4.2 泰勒展开求 π^x

方法误差

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+...+\frac{x^n}{n!}+...$$

其中，x 的取值范围为 $[0,\ln\pi]$，在假定 $\ln\pi$ 值准确的条件下，有

$$R=\frac{e^{\xi}}{n!}x^n<\frac{\pi}{n!}(\ln\pi)^n<\frac{1}{2}\times 2^{-50}$$

解得 $n > 18$，故取 $n = 19$。此时的方法误差为

$$R < 3.37 \times 10^{-16} \approx 2^{-51}$$

在假设 π 的值准确的情况下，整数部分没有方法误差。故将整数部分和小数部分的方法误差整合，可以得到使用泰勒展开求解 π^x 总体的方法误差为

$$R < 3.37 \times 10^{-16} \approx 2^{-51}$$

舍入误差及累积误差

在 $\ln\pi$ 的值准确的条件下，舍入误差由结果的位数决定，又计算的结果的最大值在二进制下的最高位为第 2 位，所以舍入误差为 $\frac{1}{2} \times 2^{-50}$。则考虑 $\ln\pi$ 的误差的影响有

$$|\Delta y_2| < |e^x||\Delta x| + \frac{1}{2} \times 2^{-50} < \pi|\Delta x| + \frac{1}{2} \times 2^{-50}$$

其中 $x = x_j \ln\pi$，故有 $\Delta x < |\Delta \ln\pi|$，即有

$$|\Delta y_2| < \pi|\Delta \ln\pi| + \frac{1}{2} \times 2^{-50}$$

代入所计算的 $\ln\pi$ 的误差 (选取复化柯特斯或牛顿迭代法[6]，这里选取舍入误差相对大一些的复化柯特斯进行计算分析)，有

$$|\Delta y_2| < 3.58 \times 10^{-13} \approx 2^{-41}$$

而所求的 $\pi^x = y_1 y_2$，所以

$$|\Delta\pi^x| < \max|y_2||\Delta y_1| + \max|y_1||\Delta y_2| + \frac{1}{2} \times 2^{-35} < \pi|\Delta y_1| + \pi^{10}|\Delta y_2| + \frac{1}{2} \times 2^{-35}$$

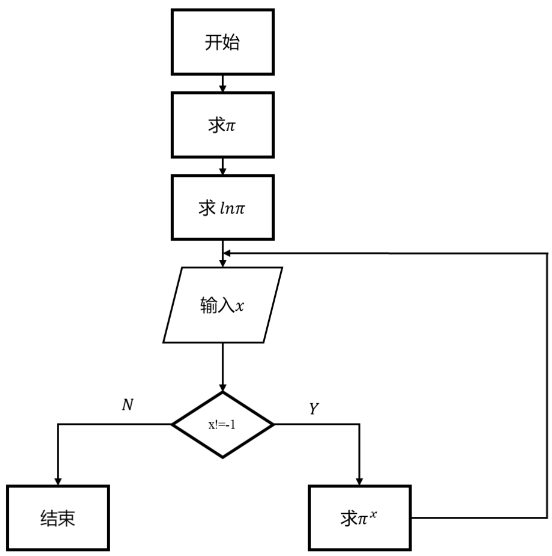
代入之前计算的数值，可以得到舍入误差最大上限为

$$|\pi^x| < 5.43 \times 10^{-8} \approx 2^{-24}$$

即使用复化柯特斯或牛顿迭代法求取 $\ln\pi$，再使用泰勒公式求取 π^x 满足 6 位小数精度。

7 实验结果与分析

7.1 程序框图



[6]复化辛普生的舍入误差较大，故不考虑使用复化辛普生进行理论分析

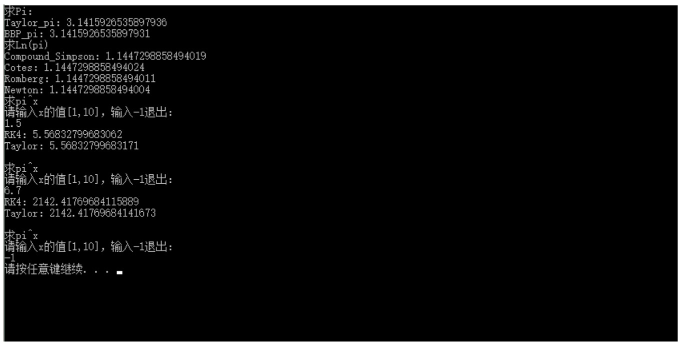
图 1: 程序框图

求 π 使用了泰勒展开以及 BBP 公式。

求 $\ln\pi$ 使用了复化辛普生公式、复化柯特斯公式、牛顿迭代法。

求 π^x 使用了四阶龙格-库塔公式和泰勒展开公式。

7.2 实验结果



```
求Pi:
Taylor_pi: 3.1415926535897936
BBP_pi: 3.1415926535897931
求Ln(pi)
Compound_Simpson: 1.1447298858494019
Cotes: 1.1447298858494024
Romberg: 1.1447298858494011
Newton: 1.1447298858494004
求pi^x
请输入x的值[1,10]，输入-1退出:
1.5
RK4: 5.56832799683062
Taylor: 5.56832799683171

求pi^x
请输入x的值[1,10]，输入-1退出:
6.7
RK4: 2142.41769684115889
Taylor: 2142.41769684141673

求pi^x
请输入x的值[1,10]，输入-1退出:
-1
请按任意键继续. . .
```

图 2: 实验结果

具体实验结果的分析见方法比较。

8 方法比较

通过 wolfram alpha 计算所得结果与实际结果的偏差，则方法比较表格如下[7]

	方法	实际误差	理论误差	n的取值
求π	泰勒展开	3.62×10^{-16}	10^{-15}	30
	BBP	1.38×10^{-16}	10^{-15}	12
求$ln\pi$	复化辛普生	1.73×10^{-15}	10^{-11}	8192
	复化柯特斯	2.23×10^{-15}	10^{-13}	256
	复化龙贝格	1.26×10^{-16}	\	128
	牛顿迭代法	2.26×10^{-16}	10^{-14}	5
求π^x $(x = 9.0009)$	四阶RK	3.15×10^{-10}	10^{-8}	1024
	泰勒展开	2.70×10^{-11}	10^{-8}	20
求π^x $(x = 9.9)$	四阶RK	1.03×10^{-8}	10^{-8}	1024
	泰勒展开	2.38×10^{-11}	10^{-8}	20

图 3: 方法比较表格，其中 n 的取值可以看出计算代价和收敛速度

由表格可以看出来，在求 π 时，两种方法的实际误差数量级相同，但是 BBP 需要迭代的次数更少。

在求 $\ln\pi$ 时，复化柯特斯和复化辛普生的实际误差数量级几乎相同，但是复化柯特斯的 n 取值更小；而牛顿迭代法的误差最小，且迭代次数只有 5 次，收敛特别快。

在求 π^x 时，显然 x 越大，则可能产生的误差越大，故尝试 $x = 9.9, 9.0009$。四阶 RK 只能达到 7 位小数精度，而泰勒公式可以达到 10 位小数精度。其原因显而易见，即由于我们所求比较简

[7]只对复化龙贝格公式列出结果对比，但是不进行分析

单，用龙格库塔公式可能相当于小题大做，而泰勒公式是最基本的，故其又快又准。

而对于四阶 RK 的理论误差的计算，实验时取的这两个值的小数部分并不是很大，所以计算得到的理论误差均在 10^{-8}，而当小数部分非常非常小时，理论误差可能会达到 10^{-1}，但是实际误差却仍在 $10^{-10} \sim 10^{-8}$ 之间。理论计算偏大的原因可能是由于 h 作为分母，且所求到的值只是舍入误差的最大上限。

至于各方法的收敛速度和计算代价，从 n 的取值就能够看出来 (虽然对于不同的方法，n 的含义可能不同，即迭代次数或区间等分数)。显而易见最好的方法是牛顿迭代法，最差的方法是复化辛普生公式。

若按照题目要求，求取每个值时使用不同的方法，则可以找到求取最大精度、收敛最快、计算代价最小的 π^x 的方法为：先使用 BBP 计算 π，在使用牛顿迭代法计算 $\ln\pi$，最后使用泰勒展开计算 π^x。

9 总结与反思

本次数值大作业让我对常用的一些数值计算方法进行了复习，锻炼了我推导公式的能力，且极大地提升了我使用 Latex 编辑公式的熟练度。除此之外，我还感受到方法误差和舍入误差之间的权重衡量，以及 n 值的确定与这二者的关系。对于精度较低的要求，可以同时考虑方法误差和舍入误差对 n 进行确定；但对于精度较高的要求，比如这里的 10^{-15} 或 2^{-50}，舍入误差的计算会由于不等式放缩等等原因会计算得偏大，所以我认为可以先通过方法误差进行对 n 的确定，再带入舍入误差进行检验。由于我们最终的目的是为了得到具有六位小数精度的 π^x，所以计算 π、$\ln\pi$ 时得到的舍入误差大于我们的预计误差并不是什么大问题，毕竟估计的是上限，只需要验证最后我们要精度符合题意即可。