Chapter 1. Preliminaries

Numerical Analysis, a branch of Applied Mathematics situated at the border between Mathematics and Computer Science, produces methods and procedures for finding numerical solutions of various problems, with a given precision. Standard topics in a Numerical Analysis course are the approximation of problems by simpler problems, the construction of algorithms, iteration methods, error analysis, stability, asymptotic error formulas, the effects of machine arithmetic, etc.

The study will focus on issues such as:

- Problems modeled by functions that do not have an analytical expression, whose values are known only at a discrete set of points. Based on these, we want to approximate values of the function at new points, values of the derivatives or integrals of the function, etc. Such problems lead to finite and divided differences, interpolation formulas, numerical differentiation and integration schemes and many others.

- In many practical situations it is necessary to solve various types of equations or systems of equations, such as: algebraic, transcendental, differential, or integral equations, whose exact solutions cannot be found. They need to be approximated numerically.

An approximating procedure must satisfy several properties:

1. To be *convergent*, meaning the sequence of iterations (successive approximations) must converge to the exact solution, in order to produce "better and better" approximations.

2. To be *stable* (to have stability), meaning that small variations of the input data should lead to small variations in the results (the approximating solutions).

3. From its structure and properties, to be able to estimate the *error* and the *speed of convergence* of the approximating method.

Many times, the conditions that ensure stability of a numerical method coincide with the ones that guarantee its convergence.

1 Taylor Polynomials

We start with a very useful tool from Calculus, Taylor's theorem. This will be needed for both the development and understanding of many of the numerical methods discussed later on.

In fact, Taylor polynomials give the very first example of a numerical method, being used as a way to evaluate other functions approximately.

Theorem 1.1. [Taylor's Theorem] *Let $f : [a, b] \to \mathbb{R}$ be a function with $n+1$ continuous derivatives on $[a, b]$, for some $n \geq 0$ and let $x, x_0 \in [a, b]$. Then*

$$f(x) \quad = \quad p_n(x) + R_{n+1}(x), \tag{1.1}$$

where

$$p_n(x) \quad = \quad f(x_0) + \frac{(x - x_0)}{1!} f'(x_0) + \cdots + \frac{(x - x_0)^n}{n!} f^{(n)}(x_0) \tag{1.2}$$

*is **Taylor's polynomial of degree** n **of** f **at** x_0 and*

$$R_{n+1}(x) \quad = \quad \frac{1}{n!} \int_{x_0}^{x} (x - t)^n f^{(n+1)}(t) dt \quad = \quad \frac{(x - x_0)^{n+1}}{(n + 1)!} f^{(n+1)}(\xi),\ \xi \text{ between } x \text{ and } x_0, \tag{1.3}$$

*is the **error** of the Taylor approximation.*

Example 1.2. Using Taylor's theorem with $x_0 = 0$, we obtain the following well-known formulas:

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n + 1)!} e^{\xi_x}, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + (-1)^{n+1} \frac{x^{2n+2}}{(2n + 2)!} \cos \xi_x, \\ \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^{n-1} \frac{x^{2n-1}}{(2n - 1)!} + (-1)^n \frac{x^{2n+1}}{(2n + 1)!} \sin \xi_x, \\ (1 + x)^a &= 1 + \binom{a}{1} x + \binom{a}{2} x^2 + \cdots + \binom{a}{n} x^n + \binom{a}{n + 1} \frac{x^{n+1}}{(1 + \xi_x)^{n+1-a}}, \end{aligned}$$

with

$$\binom{a}{k} \quad = \quad \frac{a(a - 1) \cdots (a - k + 1)}{k!},\ k = 1, 2, 3, \ldots,\ a \in \mathbb{R}.$$

An important special case of the last formula is $a = -1$, with x replaced by $-x$:

$$\frac{1}{1-x} \quad = \quad 1 + x + x^2 + \cdots + x^n + \frac{x^{n+1}}{1-x},$$

where the remainder has a simpler form than before. This is easily proved by multiplying both sides by $1 - x$ and then simplifying. Rearranging the terms, we obtain the familiar formula for a partial sum of the Geometric series:

$$1 + x + x^2 + \cdots + x^n \quad = \quad \frac{1 - x^{n+1}}{1-x}, \; x \neq 1.$$

Series representations for the functions in Example 1.2 can be obtained by letting $n \to \infty$. Recall that the series for the first three functions converge for all $x \in \mathbb{R}$, while those for the last two converge for $|x| < 1$. So, by taking Taylor polynomials of higher and higher degree, we can obtain better and better approximations of the functions above.

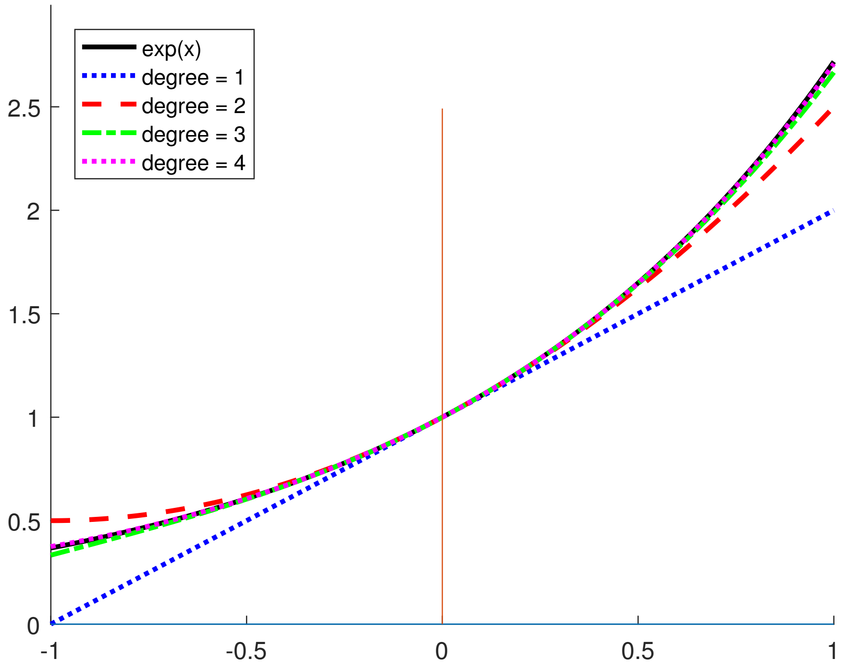

Fig. 1: Taylor approximations of e^x

Example 1.3. Consider the function $f(x) = e^x$. Figure 1 shows the approximations of f with

Taylor polynomials of degree $1, 2, 3$ and 4, for $x \in [-1, 1]$. The errors of these approximations, $\max\limits_{x\in[-1,1]}\{e^x - p_n(x)\}$, are graphed in Figure 2.

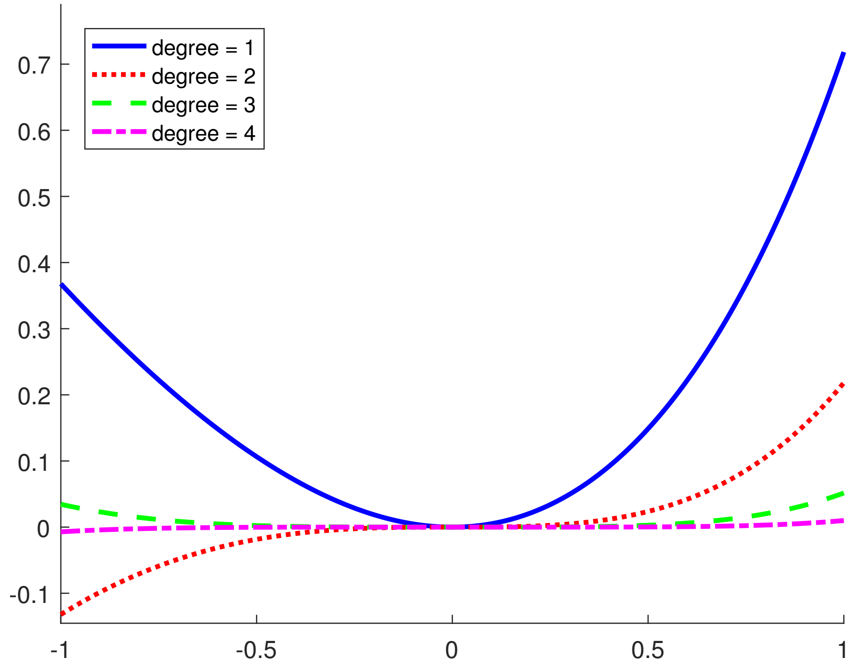

Fig. 2: Errors in Taylor approximations of e^x

Taylor's formula in two dimensions

Theorem 1.4. *Let $f : D \subset \mathbb{R}^2 \to \mathbb{R}^2$ be a function which is $n + 1$ times continuously differentiable on D, for some $n \geq 0$ and let $(x, y), (x_0, y_0) \in D$. Then*

$$f(x, y) \quad = \quad p_n(x, y) + R_{n+1}(x, y), \tag{1.4}$$

where

$$\begin{aligned} p_n(x,y) &= f(x_0,y_0) + \frac{x-x_0}{1!} f'_x(x_0,y_0) + \frac{y-y_0}{1!} f'_y(x_0,y_0) \\ &+ \frac{1}{2!}\Big[(x-x_0)^2 f''_{x^2}(x_0,y_0) + 2(x-x_0)(y-y_0) f''_{xy}(x_0,y_0) + (y-y_0)^2 f''_{y^2}(x_0,y_0)\Big] \\ &+ \sum_{j=1}^{n} \frac{1}{j!}\left[(x-x_0)\frac{\partial}{\partial x} + (y-y_0)\frac{\partial}{\partial y}\right]^j f(x,y)\Bigg|_{\substack{x=x_0\\ y=y_0}} \end{aligned} \tag{1.5}$$

and

$$R_{n+1}(x,y) = \frac{1}{(n+1)!}\left[(x-x_0)\frac{\partial}{\partial x} + (y-y_0)\frac{\partial}{\partial y}\right]^{n+1} f(x,y)\Bigg|_{\substack{x=\xi\\ y=\eta}}, \tag{1.6}$$

with (ξ,η) *a point on the line segment determined by the points* (x,y) *and* (x_0,y_0).

2 Errors: Sources, Propagation, Analysis

2.1 Types of Numerical Problems

Let us first consider the following simple examples:

Example 2.1. Compute

$$\int\limits_1^3 \frac{1}{x} dx.$$

Solution Its exact value is

$$\int\limits_1^3 \frac{1}{x} dx = \ln x\Big|_1^3 = \ln 3 - \ln 1 = \ln 3$$

and its approximation is

$$\int\limits_1^3 \frac{1}{x} dx = 1.0986...$$

■

Example 2.2. Solve the equation

$$x^2 = 3, x > 0.$$

Solution The true solution is

$$x = \sqrt{3},$$

with approximating value

$$x = 1.732...$$

■

Example 2.3. Consider the data

x_i	1	2	3	4	5	6	7
y_i	5	13	16	23	33	38	40

Discuss the nature of the relationship between x and y.

Solution We plot the data (see Figure 3).

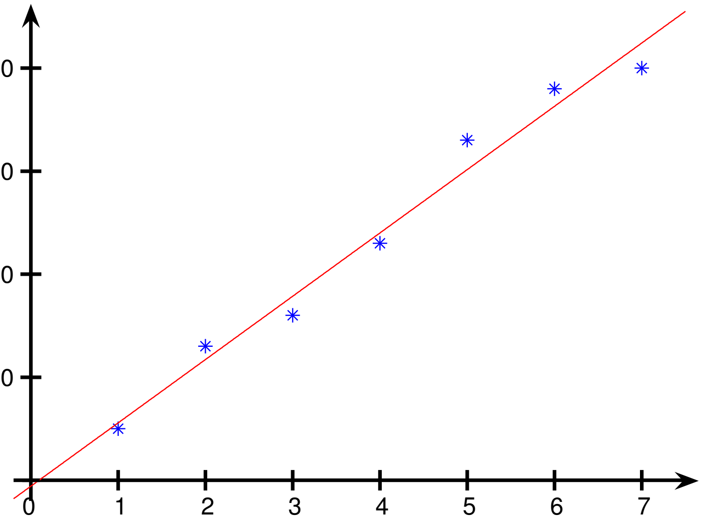

Fig. 3: Scatterplot and linear function

Notice that a straight line "best" approximates the data. So based on these values, we seek a relationship between x and y of the form

$$f(x) = ax + b.$$

In the next chapters, we will analyze several rigorous procedures for approximating scattered data.

■

From these examples, we see that a numerical problem is, in general, of the form

$$f(x) = y.$$

- If x and f are given, and we seek y, then this is called a **direct problem** or an **evaluation problem** (Example 2.1).
- If y and f are given, and we want to find x, then it is called an **inverse problem** (Example 2.2).
- If x and y are given, and f must be determined, then we have an **identifying problem** (Example 2.3).

For each type of problem, we obtain an *approximating* value, which is affected by *errors*, perturbations from the true value. The study of errors and their propagation is an important task in Numerical Analysis. In practical problems, it is important that the approximations obtained have *small (negligible)* errors, that do not affect the overall precision of the numerical procedure.

2.2 Sources and Propagation of Errors

Let $\mathcal{A} : \mathbb{R} \to \mathcal{P}(\mathbb{R})$ be a mapping that assigns a set $\mathcal{A}(x) \subseteq \mathbb{R}$ to each real number $x \in \mathbb{R}$.

Definition 2.4. *Let $x \in \mathbb{R}$. The number x^* is called an* ***approximation*** *of $x \in \mathbb{R}$, if $x^* \in \mathcal{A}(x)$ (notation $x \approx x^*$. The mapping $\mathcal{A}$ is called an* ***approximating procedure (method)***

In practice, $\mathcal{A}(x)$ consists of numbers that vary "little" from x.

Let $x \in \mathbb{R}$ and $x^* \in \mathcal{A}(x)$ be an approximation of the true (exact) value x.

Definition 2.5. *The number*

$$\Delta x = x - x^* \tag{2.1}$$

is called the ***error*** *of the approximation x^*.*

If $\Delta x > 0$, then x^* is an **under-approximation**, and if $\Delta x < 0$, x^* is an **over-approximation**.

For example, for the number π = 3.141592..., the number $x^* = 3.14$ is an under-approximation, while $x^* = 3.142$ is an over-approximation.

Definition 2.6. *The value*

$$|\Delta x| = |x - x^*| \tag{2.2}$$

is called **absolute error** *and the quantity*

$$\delta x = \frac{|\Delta x|}{|x|}, x \neq 0 \tag{2.3}$$

is called **relative error** .

Since in practice x is unknown, we use instead

$$\delta x = \frac{|\Delta x|}{|x^*|}. \tag{2.4}$$

For the relative error in $\mathbb{R}$, one can use

$$\delta x = \frac{\Delta x}{x}, \tag{2.5}$$

from which we get

$$\Delta x = x\delta x \Longrightarrow x^* - x = x\delta x,$$

or

$$x^* = (1 + \delta x)x, \tag{2.6}$$

which is widely used in applications.
For the absolute and relative error, one can also use the notations Δx^* and δx^*.

Sources of error

- *Mathematical modeling of a physical problem.* A mathematical model for a physical situation is an attempt to give mathematical relationships between certain quantities of physical interest. Because of the complexity of physical reality, a variety of simplifying assumptions are used to construct a more tractable mathematical model. The resulting model has limitations on its accuracy as a consequence of these assumptions, and these limitations may or may not be troublesome, depending on the uses of the model.

- *Blunders (human errors).* In pre-computer times, chance arithmetic errors were always a serious problem. With the introduction of digital computers, the type of blunder has changed. Chance arithmetic errors (e.g., computer malfunctioning) are now relatively rare, and programming errors are currently the main difficulty. Often a program error will be repeated many times in the course of executing the program, and its existence becomes obvious because of absurd numerical output (although the source of the error may still be difficult to find). This makes good program debugging very important, even though it may not seem very rewarding immediately.

- *Uncertainty in physical data.* Most data from a physical experiment will contain error or uncertainty within it. This must affect the accuracy of any calculations based on the data, limiting the accuracy of the answers.
- *Machine errors.* This means the errors inherent in using the floating-point representation of numbers. Specifically, we mean the rounding/chopping errors and the underflow/overflow errors. The rounding/chopping errors are due to the finite length of the floating-point mantissa; and these errors occur with all of the computer arithmetic operations.
- *Mathematical truncation error.* This name refers to the error of approximation in numerically solving a mathematical problem, and it is the error generally associated with the subject of Numerical Analysis. It involves the approximation of infinite processes by finite ones, replacing noncomputable problems with computable ones, etc.

2.3 Propagation of Errors

We distinguish two types of problems:

Case 1. Given the errors in the approximations of input data, find the errors in the output.
We start with the case of a function of two variables $f : D \to \mathbb{R}$, $D = \{(x,y)|x,y \in \mathbb{R}\}$. Let (x^*, y^*) be the approximating values of (x,y). Their absolute errors are then

$$\Delta x = x - x^* \Rightarrow x = x^* + \Delta x,$$
$$\Delta y = y - y^* \Rightarrow y = y^* + \Delta y.$$

We want to compute the absolute error

$$\Delta f = f(x,y) - f(x^*, y^*) \tag{2.7}$$

and the relative error δf.

We use Taylor's formula in two variables (1.5), at x^* and y^*. We have:

$$\begin{aligned} f(x^*+\Delta x, y^*+\Delta y) &= f(x^*,y^*) + \Delta x f'_x(x^*,y^*) + \Delta y f'_y(x^*,y^*) \\ &+ \frac{(\Delta x)^2}{2!} f''_{x^2}(x^*,y^*) + 2\frac{\Delta x \Delta y}{2!} f''_{xy}(x^*,y^*) \\ &+ \frac{(\Delta y)^2}{2!} f''_{y^2}(x^*,y^*) + \ldots, \end{aligned} \tag{2.8}$$

or

$$\begin{aligned} f(x^*+\Delta x, y^*+\Delta y) - f(x^*,y^*) &= \Delta x f'_x(x^*,y^*) + \Delta y f'_y(x^*,y^*) \\ &+ \frac{1}{2!}\left[(\Delta x)^2 f''_{x^2}(x^*,y^*) + 2\Delta x \Delta y f''_{xy}(x^*,y^*)\right. \\ &+ \left.(\Delta y)^2 f''_{y^2}(x^*,y^*)\right] + \ldots \end{aligned} \tag{2.9}$$

From here, we get

$$\begin{aligned} \Delta f &= \Delta x f'_x(x^*,y^*) + \Delta y f'_y(x^*,y^*) \\ &+ \frac{1}{2!}\left[(\Delta x)^2 f''_{x^2}(x^*,y^*) + 2\Delta x \Delta y f''_{xy}(x^*,y^*)\right. \\ &+ \left.(\Delta y)^2 f''_{y^2}(x^*,y^*)\right] + \ldots \end{aligned}$$

If Δx and Δy are small, then $(\Delta x)^2$, $\Delta x \Delta y$ and $(\Delta y)^2$ can be neglected.

We find the **absolute error of function** f or the **maximum absolute error**:

$$\Delta f \simeq \Delta x f'_x(x^*,y^*) + \Delta y f'_y(x^*,y^*). \tag{2.10}$$

In general, for a function of n variables, we have

$$\begin{aligned} \Delta f &\simeq \Delta x_1 f'_{x_1}() + \Delta x_2 f'_{x_2}() + \cdots + \Delta x_n f'_{x_n}() = \\ &= \sum_{i=1}^{n} \Delta x_i f'_{x_i}() \end{aligned} \tag{2.11}$$

This is the **propagated error**.

Then the relative error δf is

$$\begin{aligned}\delta f = \frac{\Delta f}{f} &\simeq \sum_{i=1}^{n} \Delta x_i \frac{f'_{x_i}()}{f} = \sum_{i=1}^{n} \Delta x_i \frac{d}{dx_i} \ln f() = \\ &= \sum_{i=1}^{n} x_i \delta x_i \frac{d}{dx_i} \ln f() = \sum_{i=1}^{n} x_i \frac{d}{dx_i} \ln f() \delta x_i. \end{aligned} \tag{2.12}$$

Case 2. The inverse problem is to determine the precision needed in the input data that will guarantee a (given, preset) accuracy in the output data.
To do this, we use the so-called *principle of equal effects*. This assumes that all terms $f'_{x_i}\Delta x_i$ in (2.11) have the same effect, i. e.

$$f'_{x_1}\Delta x_1 = f'_{x_2}\Delta x_2 = \cdots = f'_{x_n}\Delta x_n.$$

Then (2.11) becomes

$$\Delta f \simeq n \Delta x_i f'_{x_i}(),$$

from which it follows that

$$\Delta x_i \simeq \frac{\Delta f}{n f'_{x_i}}, \tag{2.13}$$

or

$$|\Delta x_i| \simeq \frac{|\Delta f|}{n|f'_{x_i}|}. \tag{2.14}$$

Thus, by (2.12), the *relative error* is given by

$$\delta x_i \simeq \frac{|\delta f|}{n \left| x_i \frac{d}{dx_i} \ln f \right|} \tag{2.15}$$

3 Floating-Point Representation of Numbers. Significant Digits

Definition 3.1. *A number r written in base b (an even number) has the* ***floating-point representation*** *as*

$$r = \pm\, r_0 r_1 \ldots r_p \cdot r_{p+1} \ldots r_k \times b^e,$$

where $r_0, r_1, \ldots, r_k$ form the ***mantissa (significand)****, e is the* ***exponent*** *and b is the* ***base***.

In order to have uniqueness of the representation, the floating-point numbers are *normalized*, that is, we change the representation, not the value, such that $r_0 \neq 0$. The term *floating-point* *number* will be used to mean a real number that can be exactly represented in this format.

Definition 3.2. *The* ***significant digits*** *of a floating-point number written in base b are any of the digits $1, 2, \ldots, b - 1$ in its representation that are non-zero, or located between non-zero digits, or preceded by at least one non-zero digit.*

So 0 can be a significant digit if it does more than just indicate the floating decimal point or fills the places of unknown or omitted digits.
The leftmost significant digit is called the *most significant* digit.

For example,

- the number 7063 has all significant digits;
- the number 0.02340 – the last 4 digits are significant (the zeros in front of 2 merely indicate the decimal point, and are, therefore, *not* significant);
- 1.230×10^3 = 1230 – zero *is* a significant digit, as it is preceded by a nonzero digit.

Consider the number $r > 0$ with the following representation in base 10:

$$r = r_0 10^k + r_1 10^{k-1} + \cdots + r_{n-1} 10^{k-n+1} + r_n 10^{k-n} + \ldots.$$

Definition 3.3. *The number*

$$r^* = r_0^* 10^k + r_1^* 10^{k-1} + \cdots + r_{n-1}^* 10^{k-n+1}$$

approximates r **correctly with** n **significant digits** *if*

$$|\Delta r^*| \leq \frac{1}{2} \times 10^{k-n+1} \tag{3.1}$$

Example 3.4. Find the number of significant digits with which $e^* = 2.718282$ approximates correctly the number $e = 2.71828182\ldots$.

Solution Let e^* = 2.718282. This can be written as

$$e^* = 2 \cdot 10^0 + 7 \cdot 10^{-1} + 1 \cdot 10^{-2} + \cdots$$

So $k = 0$. Then

$$\Delta e^* = e^* - e = 0.00000018 = 0.18 \times 10^{-6} \leq \frac{1}{2} \times 10^{-6}.$$

Comparing it to (3.1), we get

$$10^{k-n+1} = 10^{-6} \Rightarrow 0 - n + 1 = -6 \Rightarrow n = 7.$$

Thus, the number $e^* = 2.718282$ approximates correctly $e = 2.71828182\ldots$ with 7 significant digits.

■

The following relation exists between the number de significant digits and the relative error:

$$\delta r^* \quad \leq \quad \frac{1}{r_0^* \times 10^{n-1}}, \tag{3.2}$$

where r^* is the correct approximation with n significant digits, with the normalized representation

$$r^* = r_0^* 10^k + r_1^* 10^{k-1} + \cdots + r_{n-1}^* 10^{k-n+1}. \tag{3.3}$$

This can be considered the *maximum relative error*.

Example 3.5. Let r^* = 0.0875 be a correct approximation with 3 significant digits of a number r. Find the maximum relative error of this approximation.

Solution The number r^* can be written as

$$r^* = 8 \cdot 10^{-2} + 7 \cdot 10^{-3} + 5 \cdot 10^{-4} = 8.75 \cdot 10^{-2} = 0.0875 \cdot 10^{-4},$$

from which we have $r_0^* = 8$ and $n = 3$.

The maximum relative error is

$$\frac{1}{r_0^* 10^{n-1}} = \frac{1}{8 \cdot 10^2} = \frac{1}{8} \cdot 10^{-2} = 0.125 \cdot 10^{-2} = 0.00125$$

Thus, the number r can be approximated correctly by the number $r^* = 0.0875$, with 3 significant digits , with a maximum relative error of $0.125\ \%$.

■