Chapter 3

Tutorial For the first two classes

3.0.1 Horner's method

Notice that the notation in the tutorial differs from that in the lecture. For example, the index of the coefficients of q starts from 1 in the tutorial, but from 0 in the lecture.

Another use of the Horner's method is for division of polynomials of degree $n \geq 1$ by first order polynomials in the form $(x - x_0)$. This application is based on the next result.

Theorem 3.0.1 (Polynomial remainder theorem)**.** *Let $p(x)$ be polynomial of degree $n \geq 1$ and let $x_0 \in \mathbb{R}$. Then the remainder of the division of $p(x)$ by $(x - x_0)$ is $p(x_0)$.*

The combination of this theorem with the Horner's method theorem gives that $q(x)$ is the quotient of the division of $p(x)$ by $(x - x_0)$.
Exercise Divide $x^3 - 6x^2 + 11x - 6$ by $(x - 2)$.
Solution We need to compute the quotient $q(x)$ and the remainder $p(2)$.

$$
\begin{array}{c|cccc}
 & 1 & -6 & 11 & -6 \\
 & & & & \\
2 & & 2 & -8 & 6 \\
\hline
 & 1 & -4 & 3 & 0
\end{array}
$$

Therefore, $q(x) = x^2 - 4x + 3$ and $p(2) = 0$.

3.0.2 Decimal machine numbers Floating-point numbers

A k-digit decimal machine number is a number of the form

$$\pm(0.d_1 d_2 d_3 \cdots d_k) \times 10^n \tag{3.1}$$

where d_i are decimal digits ($1 \leq d_1 \leq 9$ and $0 \leq d_i \leq 9$ for $i \geq 2$) and n is an integer. Any positive number admits the so-called normalized representation in this form

$$(0.d_1 d_2 d_3 \cdots d_k d_{k+1} d_{k+2} \cdots) \times 10^n, \quad 1 \leq d_1 \leq 9, \quad 0 \leq d_i \leq 9, i \geq 2, \quad n \in \mathbb{Z} \tag{3.2}$$

The k-digit floating-point representation of the number y is denoted by $fl(y)$ and it is obtained by terminating the representation of y at k digits. There are two common ways of doing this.

(a) By chopping: we chop off the digits $d_{k+1}, d_{k+2}, \ldots$. Then $fl(y) = (0.d_1 d_2 \cdots d_k) \times 10^n$.

(b) By rounding: we add $5 \times 10^{n-k-1}$ to y and then chop off the digits $d_{k+1}, d_{k+2}, \ldots$ to obtain the form $fl(y) = (0.\delta_1 \delta_2 \cdots \delta_k) \times 10^n$.

Notice that for rounding when $d_{k+1} \geq 5$, we add 1 to d_k and obtain $fl(y)$ and when $d_{k+1} < 5$, we have $\delta_i = d_i$ for $i = 1, 2, \ldots, k$.
Exercise Determine the five-digits (a) chopping and (b) rounding values of the number π.
Solution First, we write π in a normalized decimal form as $\pi = (0.314159265 \cdots) \times 10^1$. Here, $n = 1$ and $k = 5$.

(a) By chopping, we have $fl(\pi) = (0.31415) \times 10^1 = 3.1415$.

(b) By rounding: First, we compute $\pi + 5 \times 10^{1-5-1} = \pi + 0.00005 = 3.14159265 \cdots + 0.00005 = 3.14164 \cdots = (0.314164 \cdots) \times 10^1$. Then, by chopping at d_6, we have $fl(\pi) = (0.31416) \times 10^1 = 3.1416$.

3.0.3 Operations with floating point numbers

One common error-producing calculations involves the cancellation of significant digits due to the substraction of nearly equal number. Let x and y be two nearly equal numbers given by

$$x = 0.d_1 d_2 d_3 \cdots d_p \alpha_{p+1} \alpha_{p+2} \cdots \times 10^n \tag{3.3}$$

$$y = 0.d_1 d_2 d_3 \cdots d_p \beta_{p+1} \beta_{p+2} \cdots \times 10^n \tag{3.4}$$

Let $k > p$. Then the k-digits representation for x and y, for chopping for example, are

$$fl(x) = 0.d_1 d_2 d_3 \cdots d_p \alpha_{p+1} \alpha_{p+2} \cdots \alpha_k \times 10^n \tag{3.5}$$

$$fl(y) = 0.d_1 d_2 d_3 \cdots d_p \beta_{p+1} \beta_{p+2} \cdots \beta_k \times 10^n \tag{3.6}$$

then $fl(x) - fl(y) = 0.\delta_{p+1}\delta_{p+2} \cdots \delta_k \times 10^{n-p}$ where $\delta_{p+1}\delta_{p+2} \cdots \delta_k = \alpha_{p+1}\alpha_{p+2} \cdots \alpha_k - \beta_{p+1}\beta_{p+2} \cdots \beta_k$.

Notice that $fl(x) - fl(y)$ has at most $k - p$ significant digits. So maybe we are loosing information in the substraction operation.

Exercise Compute the solutions to $x^2 + 62.10x + 1 = 0$.

Solution We solve the floating point solution by the quadratic formula:

$$fl(x_1) = \frac{-62.10 + \sqrt{(62.10)^2 - 4.000 \times 1.000 \times 1.000}}{2.000 \times 1.000} = \frac{-62.10 + \sqrt{3852}}{2.000} = \frac{-62.10 + 62.06}{2.000} = -0.0200 \tag{3.7}$$

By using exact arithmetic, we get $x_1 = -0.01610723$. Similarly, we have that $f(x_2) = -62.10$ and $x_2 = -62.08390$. Notice that the relative errors are

$$e_1 = \frac{|-0.0200 + 0.01610723|}{|-0.01610723|} \approx 0.241678 \approx 2 \times 10^{-1} \tag{3.8}$$

$$e_2 = \frac{|x_2 - fl(x_2)|}{|x_2|} \approx 0.000259 \approx 2 \times 10^{-4} \tag{3.9}$$

We can improve the approximation of x_1 by simply doing this:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \cdot \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} = \frac{-2c}{b + \sqrt{b^2 - 4ac}} \tag{3.10}$$

Please, compute x_1 from the above formula and examine the relative error.