

Physics 234: Exercise 2

1. We know that the addition overflows because the most significant carry bit is on.

$$\begin{array}{|c|c|c|} \hline \color{red}{1} & \color{red}{1} & \\ \hline & \text{A} & \text{F} \\ \hline + & 7 & 2 \\ \hline & 2 & 1 \\ \hline \end{array} \iff \begin{array}{|c|c|c|c|c|c|c|c|} \hline \color{red}{1} & \color{red}{1} & \color{red}{1} & \color{red}{1} & \color{red}{1} & \color{red}{1} & \color{red}{1} & & \\ \hline & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ \hline + & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ \hline & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ \hline \end{array}$$

2. We begin by expanding in powers of δx .

$$\begin{aligned}
 a^p &= x_{n+1}^q = (x_n + \delta x)^q \\
 &= x_n^q \left(1 + \frac{\delta x}{x_n} \right)^q \\
 &= x_n^q \left(1 + \frac{q}{x_n} \delta x + \frac{q(q-1)}{2x_n^2} \delta x^2 + \dots \right)
 \end{aligned}$$

Truncating at linear order and solving for δx yields

$$\delta x = (x_n/q)(a^p/x_n^q - 1).$$

Hence,

$$x_{n+1} = x_n + \delta x = x_n + \frac{a^p}{q} x_n^{1-q} - \frac{1}{q} x_n = \left(1 - \frac{1}{q} \right) x_n + \frac{a^p}{q} x_n^{1-q}.$$

3. $t(x)$ is the Taylor series expansion of $\tanh x$ truncated at 6th order. Here, it's meant to serve as an approximating function to $\tanh x$. $r(x)$ is a rational polynomial which can be made to match $t(x)$ near $x = 0$ by suitable choice of the parameter b .

$$\begin{aligned}
 r(x) &= \frac{15x + x^3}{15 + bx^2} \\
 &= x \left(1 + \frac{1}{15}x^2\right) \left(1 + \frac{b}{15}x^2\right)^{-1} \\
 &= x \left(1 + \frac{1}{15}x^2\right) \left(1 - \frac{b}{15}x^2 + \frac{b^2}{15^2}x^4 - \dots\right) \\
 &= x + \frac{1-b}{15}x^3 + \frac{b(b-1)}{15^2}x^5 + O(x^7)
 \end{aligned}$$

And order-by-order comparison with $t(x) = x - \frac{1}{3}x^3 + \frac{2}{15}x^5$ shows that we need $b = 6$. By construction the two functions are nearly identical near the origin, but as x gets large, the Taylor series expansion begins to blow up. The rational polynomial, on the other hand, always stays bounded and is still accurate to about two or three decimal places all the way up to $x = 2$.

