

# Precision in the Computation of Polynomials

David W.H. Swenson

August 2004

Suppose we have a high-precision number  $x_d$ , which we approximate by its lower-precision counterpart,  $x_s$ . In this analysis, we will assume that  $x_d$  and  $x_s$  are positive. We define the precisions of these numbers such that  $x_d$  is represented by  $d$  digits in base  $b$ , and  $x_s$  is represented by  $s$  digits in that same base.

First, we note that  $x_d$  and  $x_s$  are separated by a constant,  $a$ :

$$a = x_s - x_d \quad (1)$$

Since the value of  $x_s$  is truncated at  $s$  digits in its base  $b$  representation, we have an upper borne on the value of  $a$ :

$$|a| \leq x_d b^{-s} \quad (2)$$

Using equation (1) and the binomial theorem, we expand  $x_s^n$  as:

$$(x_d + a)^n = \sum_{k=0}^n C_k^n x_d^k a^{n-k} = \left( \sum_{k=0}^{n-1} C_k^n x_d^k a^{n-k} \right) + x_d^n \quad (3)$$

We define  $\epsilon_n$  as the absolute error for a given exponent  $n$ , that is, the difference between  $x_s^n$  and  $x_d^n$ :

$$\epsilon_n = x_s^n - x_d^n = \sum_{k=0}^{n-1} C_k^n x_d^k a^{n-k} \quad (4)$$

Using the Cauchy-Schwartz inequality, we remark that

$$|\epsilon_n| = \left| \sum_{k=0}^{n-1} C_k^n x_d^k a^{n-k} \right| \leq \sum_{k=0}^{n-1} C_k^n x_d^k |a|^{n-k} = \sum_{k=0}^{n-1} C_k^n x_d^k |a|^{n-k} \quad (5)$$

Combining inequalities (2) and (5), we have

$$|\epsilon_n| \leq \sum_{k=0}^{n-1} C_k^n x_d^k |a|^{n-k} \leq \sum_{k=0}^{n-1} C_k^n x_d^k (x_d b^{-s})^{n-k} = \sum_{k=0}^{n-1} C_k^n x_d^n (b^{-s})^{n-k} \quad (6)$$

Since  $C_k^n$  is always a positive integer, we can safely say that  $C_k^n \gg (b^{-s})^i$  for  $s$  and  $i$  greater than 1. Therefore, we approximate the sum by the term when  $k$  is greatest, giving the lowest power of  $b^{-s}$ , and thus the largest number:

$$|\epsilon_n| \leq \sum_{k=0}^{n-1} C_k^n x_d^n (b^{-s})^{n-k} \approx C_{n-1}^n x_d^n (b^{-s})^{n-(n-1)} = n x_d^n (b^{-s}) \quad (7)$$

It may seem that, since this absolute error grows as  $x^n$ , the error in the lower-precision approximation will be significant at large values of  $n$ . However, a much better measure of the significance of the error is the relative error, defined as

$$RE_n = \frac{|x_s^n - x_d^n|}{x_d^n} = \frac{|\epsilon_n|}{x_d^n} \lesssim \frac{n x_d^n (b^{-s})}{x_d^n} = n(b^{-s}) \quad (8)$$