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:

$$a = x_s - x_d \tag{1}$$

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:

$$|a| \leq x_d b^{-s} \tag{2}$$

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

$$(x_d + a)^n = \sum_{k=0}^{n} \mathrm{C}_k^n x_d^k a^{n-k} = (\sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^k a^{n-k}) + x_d^n \tag{3}$$

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

$$\epsilon_n = x_s^n - x_d^n = \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^k a^{n-k} \tag{4}$$

Using the Cauchy-Schwartz inequality, we remark that

$$|\epsilon_n| = \left| \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^k a^{n-k} \right| \leq \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^k \left| a^{n-k} \right| = \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^k \left| a \right|^{n-k} \tag{5}$$

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

$$|\epsilon_n| \leq \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^k \left| a \right|^{n-k} \leq \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^k \left(x_d b^{-s} \right)^{n-k} = \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^n \left(b^{-s} \right)^{n-k} \tag{6}$$

Since C_k^n is always a positive integer, we can safely say that $\mathrm{C}_k^n \gg (b^{-s})^i$ 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:

$$|\epsilon_n| \leq \sum_{k=0}^{n-1} \mathrm{C}_k^n x_d^n \left(b^{-s} \right)^{n-k} \approx \mathrm{C}_{n-1}^n x_d^n (b^{-s})^{n-(n-1)} = n x_d^n (b^{-s}) \tag{7}$$

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

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