Accurate Floating Point Product

Stef Graillat

Laboratoire LIP6, Département Calcul Scientifique
Université Pierre et Marie Curie (Paris 6)
4 place Jussieu, F-75252, Paris cedex 05, France
email: stef.graillat@lip6.fr

Abstract

Several different techniques and softwares intend to improve the accuracy of results computed in a fixed finite precision [2]. Here we focus on a method to improve the accuracy of the product of floating point numbers. We show that the computed result is as accurate as if computed in twice the working precision. The algorithm (called `CompProd`) is simple since it only requires addition, subtraction and multiplication of floating point numbers in the same working precision as the given data.

More precisely, we assume to work with a floating point arithmetic adhering to IEEE 754 floating point standard. The set of floating point numbers is denoted by $\mathbb{F}$, the relative rounding error by `eps`. Let $a_i \in \mathbb{F}$, $1 \leq i \leq n$ be some floating point numbers, and set $p = \prod_{i=1}^{n} a_i$. If `res` is the result of our new algorithm `CompProd`, then

$$\frac{|\texttt{res} - p|}{|p|} \leq \texttt{eps} + \alpha n^2 \texttt{eps}^2,$$

where α is a moderate constant. We also give a validated computable error bound for our new algorithm. Additionally, using results from [3], we give sufficient conditions on the number of floating point numbers so as to get a faithfully rounded result (that is to say one of the two adjacent floating point numbers of the exact result). For example, in IEEE 754 double precision arithmetic ($\texttt{eps} = 2^{-53}$), we proved that if $n \leq 2^{25}$ then the result of `CompProd` is a faithful rounding of p. We also provide an *a posteriori* bound that makes it possible to check whether the result of our algorithm is faithfully rounded.

Such an algorithm can be useful for example to compute the determinant of a triangular matrix and to evaluate a polynomial when represented by the root product form. It can also be applied to compute the power of a floating point number [1].

References

[1] P. Kornerup, V. Lefevre, and J.-M Muller, Computing integer powers in floating-point arithmetic. arXiv:0705.4369v1 [cs.NA].
[2] T. Ogita, S. M. Rump, and S. Oishi, Accurate sum and dot product. *SIAM J. Sci. Comp.*, 26(6):1955-1988, 2005.
[3] S. M. Rump, T. Ogita, and S. Oishi, Accurate floating-point summation. Technical Report 05.12, Faculty for Information and Communication Sciences, Hamburg University of Technology, nov 2005.