Accurate Summation and Dot Products

S. M. Rump[1]

Abstract: We present two new algorithms `FastAccSum` and `FastPrecSum`, the first to compute a faithful rounding of the sum of floating-point numbers, and the second for a result as if computed in K-fold precision. Similar algorithms derive for the dot product of two vectors.

Faithful rounding means the result is one of the immediate floating-point neighbors of the exact result, or the exact result if it is a floating-point number.

The algorithms are based on so-called error-free transformation, that is a pair of floating-point numbers (a, b) is transformed without error into another pair (x, y) so that, for example, $a \cdot b = x + y$ where x is the nearest floating-point number to $a \cdot b$.

We improve the previously fastest algorithms `AccSum` and `PrecSum` by up to 25%. The first algorithm adapts to the condition number of the sum, i.e. the computing time is proportional to the logarithm of the condition number. The second algorithm does not need extra memory, and the computing time depends only on the dimension and K.

Both algorithms are the fastest known in terms of flops. They allow good instruction-level parallelism so that they are also fast in terms of measured computing time. The algorithms neither require special operations such as access to mantissa or exponent, they contain no branch in the inner loop, nor do they require some extra precision: Only standard floating-point addition, subtraction and multiplication in one working precision, for example double precision, is used.

The algorithms are even faster than XBLAS algorithms, although the result of the latter may be of much less quality.

[1] Institute for Reliable Computing, Hamburg University of Technology
Schwarzenbergstrasse 95, 21073 Hamburg, Germany
and
Waseda University, Faculty of Science and Engineering, 3–4–1 Okubo
Shinjuku-ku, Tokyo 169–8555, Japan
rump@tu-harburg.de