ANALYSIS OF SOME KNOWN METHODS
OF IMPROVING
THE ACCURACY OF FLOATING-POINT SUMS

SEppo LINNAINMAA

Abstract

Some well-known methods for calculating the round-off error in floating-point addition are analyzed in this paper. The methods have been introduced by Møller [16], Kahan [11] and Knuth [12]. The necessary and sufficient conditions under which these methods produce the value of the round-off error, for rounding, truncating and parity arithmetic, are given. The computer-oriented parity arithmetic is not commonly known, but it has some desirable properties, as this paper will demonstrate. Some experimental results are also reported.

Introduction

In 1951, Gill [5] examined the fourth order Runge-Kutta method for solving differential equations and noticed that the round-off error in summation could be estimated by subtracting one of the addition terms from the sum. Gill worked with fixed-point arithmetic and his ideas are not directly applicable to floating-point arithmetic. That is why, as Møller [16] points out, the expected reduction of roundoff errors is not reached in the algorithm for the Runge-Kutta method given by Romanelii [19] and corrected by Thompson [21]. These algorithms are discussed together with the empirical results in this paper.

In 1965 Møller [16], [17] applied Gill's idea for summation to floating-point arithmetic. He examined all possible cases in a binary arithmetic where one operand is truncated before addition, if its exponent is smaller than the other operand's. In the same year Kahan [11] gave a method for summation based on the same idea. These methods have been studied further or applied by Gram [7], Naur [18] and Dekker [4], as well as Babuška [2] and Gregory [8], who compare them with some other methods for improving the accuracy of floating-point sums. One such method consists of first summing groups of terms and then adding the partial sums [14]. Another method consists of using a group of summing registers for values of different magnitudes [22], [15].

Received Dec. 24, 1973.
In the 1960's, some writers stated that the problem itself was insignificant and would be solved by using double precision arithmetic, which became available at that time [10]. However, there are reasons to use those methods, in some circumstances, as we can deduce from the following.

In this paper, some general properties of floating-point arithmetic are introduced first. After that, the methods given by Møller, Kahan and Knuth are analyzed by proving a set of theorems giving the necessary and sufficient conditions under which these methods are valid. The method of proof used here is a slightly modified version of the one used by Knuth. As Knuth states, the methods to be analyzed are based on some regularity properties of floating-point arithmetic.

An analysis of the methods of Møller and Kahan, based on the universally accepted rules for rounding and truncating arithmetic, is discussed. Lemma 5 and Theorem 6 of this paper which have been proved by Knuth in rounding arithmetic, are generalized for truncating arithmetic.

The so-called parity arithmetic is also introduced for odd base numbers. An analysis is made for parity arithmetic as for the other arithmetics. This analysis gives further evidence of the usability of parity arithmetic.

At the end of this paper some empirical results are reported. The methods considered have been applied to summation of series, numerical integration, and numerical solution of differential equations.

Floating-point addition

An arbitrary nonzero real number $u$ can always be represented to the base $b$ as a floating-point number with signed-magnitude representation

$\begin{equation}
    u = \pm 0 \cdot u_1 u_2 u_3 \ldots \times b^{e_u} = f_u \times b^{e_u}, \\
    u_1 \neq 0,
\end{equation}$

where $b$, $e_u$ and $u_i$ ($i = 1, 2, 3, \ldots$) are integers, $b \geq 2$ and $0 \leq u_i < b$. The expression $0 \cdot u_1 u_2 u_3 \ldots$ here denotes the number $u_1 b^{-1} + u_2 b^{-2} + \ldots$. The numbers $u_1, u_2, u_3, \ldots$ are digits in the base $b$, $f_u$ is the fraction part and $e_u$ is the exponent of $u$. If $u_m = 0$ and $u_i = 0$, when $i > m$, then we say that $u_m$ is the least significant nonzero digit of $u$ in representation (1). In this case the value of $u$ is

$\begin{equation}
    u = \pm (u_1 b^{e_u-1} + u_2 b^{e_u-2} + \ldots + u_m b^{e_u-m}).
\end{equation}$

The exponent $e_u - i$ corresponding to the digit $u_i$ ($i = 1, \ldots, m$) indicates the unique position of this digit when $u$ is represented as a fixed-point