DATA PROCESSING SYSTEMS BULLETIN

MACHINE-OPTIMAL APPROXIMATIONS

The information contained in this bulletin was contributed by K. Spielberg, IBM Applied Programming.

Many methods are available to provide near-optimal polynomial or rational approximations for a given function f(x). Practical experience in applying such approximations when writing computer programs indicates that the efficiency of such programs depends on the programmer's ability to make the utmost use of machine characteristics. This bulletin gives an example of the proper adaptation of a well-known existing approximation to binary machines. The improvement described has been incorporated in part in the 709 subroutine IB SQ2, SHARE Distribution #721.

CALCULATION OF THE SQUARE ROOT

As shown in reference (1), the calculation of the square root depends essentially on the accuracy of the first "guess," which is improved with the aid of Heron's method. If the relative error of the first approximation is R_0, the relative error after two iterations, R_2, is given by $R_2 \approx R_0^4/8$. For eight-digit accuracy, we have $R_2 < 1/2 \times 10^{-8}$ and $R_0 < \sqrt{2} \times 10^{-2}$.

Linear approximations of the form $s_0 = Af + B$ can be found by the method discussed in reference (1). They give sufficient accuracy in the important intervals $1/4 \le f \le 1/2$ and $1/2 \le f \le 1$. The constant A will generally be such as to require a time-consuming multiplication. Therefore it is desirable to make A arbitrarily a convenient binary number that allows the multiplication to be replaced by additions and a shift. For instance, [2,3]

$$s_0 = 7/8f + 0.302734 \qquad (1/4 \le f \le 1/2) \qquad (1a)$$
$$s_0 = 37/64f + 0.421875 \qquad (1/2 \le f \le 1) \qquad (1b)$$
$$s_0 = 7/8f + 9/32 \qquad (1/4 \le f \le 1/2) \qquad (2a)$$
$$s_0 = 9/16f + 7/16 \qquad (1/2 \le f \le 1) \qquad (2b)$$

While equation (1) gives full accuracy, it is logically less favorable than equation (2), since $7/8f = f(1 - 1/8)$ and $9/16f = 1/2f(1 + 1/8)$. On the other hand, (2) is not sufficiently accurate. Therefore it is necessary to determine the constant B, for a given A, such as to give minimum relative error in the respective interval. This can be achieved by a modification of the procedure proposed in reference (1). Let $f = c(1 + t)$, and express the relative error of the approximation

$$\sqrt{f} \simeq c(1 - \epsilon + 1/2t) = \sqrt{c}\left\{1 - \epsilon + 1/2(f/c - 1)\right\} \qquad \text{as} \qquad (3)$$

$$R = (1 + t)^{-1/2}\left(\epsilon + \sqrt{1 + t} - 1 - 1/2t\right) \qquad (4)$$
$$\simeq (1 + t)^{-1/2}\left(\epsilon - t^2/8 + t^3/16 - t^4/32 + \ldots\right).$$

The method described in reference (1) corresponds to $\epsilon = 0$. It arrives at values of t_1 and t_2 (corresponding to the end points of the interval of f) by equalizing the relative errors at the ends: $|R(t_1)| = |R(t_2)|$. With this requirement the constant c and, consequently, both A and B are determined. In our case, on the other hand, c, t_1 and t_2 are already fixed by A; but we can still determine B so as to obtain an appropriate value (>0) for the parameter ϵ.

Within the region $t_1 \leq t \leq t_2$ the relative error R(t) has, evidently, only one relative maximum at $t = -2\epsilon$. ($R' = 0 \rightarrow t = -2\epsilon$). We can therefore determine ϵ by demanding that

$$R(-2\epsilon) = |R(t')|\,, \tag{5}$$

where t' is the argument (t_1 or t_2) for which $|R(t)|_{\epsilon = 0}$ is a maximum.

On the other hand, it will be found that $R(-2\epsilon)$ differs from R(0) only slightly:

$$\begin{aligned} R(-2\epsilon) &= 1 - \sqrt{1 - 2\epsilon} - \epsilon^2/2 + \epsilon^3/2 + \ldots \\ &= R(0) + \epsilon^2/2 + \epsilon^3/2 + \ldots \end{aligned} \tag{6}$$

Usually it will suffice to replace (5) by

$$R(0) = |R(t')| \tag{5'}$$

If so desired, the value of ϵ thus obtained can be taken more accurately as

$$\epsilon^* = \epsilon - \delta\epsilon\,, \quad \delta\epsilon = 1/2\left\{R(-2\epsilon) - R(0)\right\}. \tag{6'}$$

Use of the procedure described above will reduce the maximum relative error by approximately a factor of 2, because:

$$\begin{aligned} &|R(t')|_{\epsilon=0} - R(-2\epsilon) \approx |R(t')|_{\epsilon=0} - R(0) \approx \epsilon\,(1 + t')^{1/2} \\ &|R(t')|_{\epsilon=0} \approx \epsilon\left\{1 + 1/(1 + t')^{1/2}\right\} \approx 2R(0) \end{aligned} \tag{7}$$

Figure 1 represents schematically some of the quantities discussed.

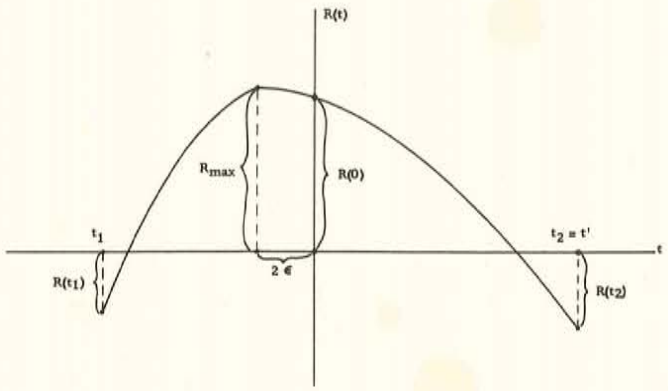

Figure 1

The procedure is illustrated by two examples presented below.

Example 1: (See equation (2a) and figure 2.)

$$s_0 = 7/8f + B \qquad (1/4 \leq f < 1/2)$$
$$A = \sqrt{c}(f/2c) = 7/8, \quad \sqrt{c} = 4/7$$
$$t_1 = -0.23438, \quad t_2 = 0.53125$$

Equation (5') gives

$$(1.53125)^{1/2}\,\epsilon = 0.02782 - \epsilon, \quad \epsilon \simeq 0.0124$$
$$|R(t')|_{\epsilon = 0} \simeq 0.0225$$
$$R(t_1) = 0.0071, \quad R(0) = |R(t_2)| \simeq 0.0124$$
$$R(-2\epsilon) = R_{max} = 0.01246$$

In this example, the relative error has been reduced by a factor close to 2.

$$s_0 \simeq 7/8f + 0.2786 \tag{8}$$

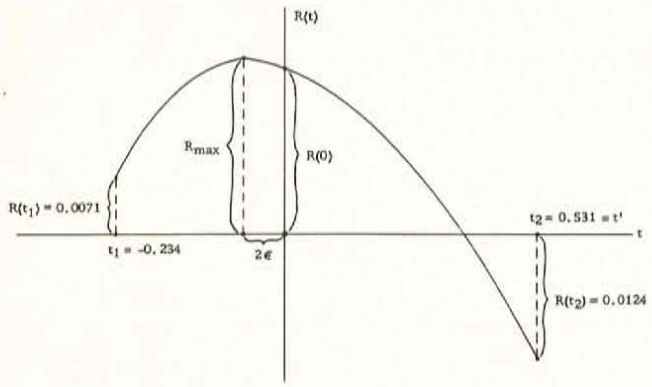

Figure 2

Example 2: (See equation (2b) and figure 3.)

$$s_0 = 9/16f + B \qquad (1/2 \leq f < 1)$$
$$A = \sqrt{c}(f/2c) = 9/16, \quad \sqrt{c} = 8/9$$
$$t_1 = -0.36719, \quad t_2 = 0.26563$$

ϵ is found from $R(0) = |R(t_1)|$ to $\epsilon = 0.011423$

$$|R(t')|_{\epsilon = 0} = 0.025789$$
$$|R(t_1)| = R(0) = \epsilon = 0.011423, \quad R(t_2) \simeq 0.0045$$
$$R(-2\epsilon) = R_{max} = 0.011488$$

In this example, the relative error has been reduced by a factor of 2.276. Hence the final relative error R_2 (after two Heron iterations) is reduced by as much as $(2.276)^4 \simeq 26.834$.

$$s_0 = 9/16f + 0.434 \tag{9}$$

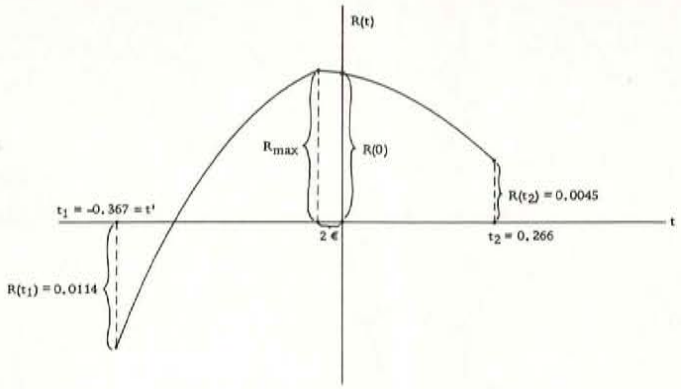

Figure 3

REFERENCES

1. E. G. Kogbetliantz, "Computation of Sin N, Cos N and $\sqrt[m]{N}$ Using an Electronic Computer," IBM Journal of Research and Development, Vol. 3, No. 2 (April, 1959).

2. S. Ross, IB SQRT, SHARE Distribution #507.

3. C. J. Swift, CS SQT1, SHARE Distribution #641.

IBM
International Business Machines Corporation
Data Processing Division, 112 East Post Road, White Plains, N. Y.

Printed in U. S. A. J28-8023 11/59