On New Dividing Method in Affine Arithmetic

Shinya Miyajima and Masahide Kashiwagi

Waseda University
61-404, Okubo 3-4-1, Shinjuku-ku, Tokyo 169-8555, Japan
Phone:+81-3-5286-3484, Fax:+81-3-5286-3362
Email: shin@kashi.info.waseda.ac.jp, kashi@kashi.info.waseda.ac.jp

1. Introduction

Numerical Computation with guaranteed accuracy means numerical methods which calculate not only numerical results but also mathematically rigorous error estimation of the results. Recently, a lot of algorithms with guaranteed accuracy for various numerical methods have been developed.

Those algorithms are realized by interval arithmetic. Interval arithmetic represents a real number by an interval [lower , upper] whose endpoints are machine representable numbers, and rounding errors of floating point arithmetic is able to be grasped by directed rounding technique.

On the other hand, interval arithmetic has an essential weak point that it causes unexpected overestimation of calculation results. The main reasons of the phenomenon are that correlation between variables is ignored. A great deal of effort has been put into finding a way of overcoming the problem.

In order to overcome this problem, Affine arithmetic [1][2](AA), an extension of standard interval arithmetic, was proposed in 1994. Because AA can express corelation between variables, it possesses advantages : to suppress explosive increases of interval width occured when operating ordinary interval arithmetic. AA estimate nonlinear images in the form of "(linear approxination)+(error term)". The defferences of the estimation is an important point because it is directly linked with performance in interval evaluation.

Because the dividing method in AA has not been defined in [1][2], Kashiwagi[3] defined it. Because [3] simplifies partly a complicated calculation to reduce calculating cost, Shirai[4] calculated strictly the part. However, in the case that correlation between variables exists, the most suitable evaluation is not be able to be obtained by operating [4]. Therefore, the methods that is able to supply more suitable evaluation than [4] exist in theory, but they are not known at present.

Therefore, the purpose of this paper is to suggest a new method that is able to supply more suitable evaluation than [4]. As a result, better evaluation than [4] is able to be obtained by operating the new method.

2. Affine arithmetic

Affine arithmetic (AA) [1] [2] is a variant of Interval arithmetic(IA). AA keeps correlation between quantities. So AA can overcome the error explosion problem often observed in standard IA, and it often produces much tighter bounds than a bound standard IA yields.

2.1. Affine form

In AA, a quantity x is represented as the following affine form:

$$x = x_0 + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n. \tag{1}$$

Here, coefficients x_i $(i = 0, 1, \cdots, n)$ are known real numbers, and ε_i $(i = 1, \cdots, n)$ are symbolic real variables whose values are unknown but assumed to lie in the interval $[-1, 1]$. Only coefficients x_i $(i = 0, 1, \cdots, n)$ are memoried in the computer. Remark that n, the number of the noise symbol ε_k, changes through the calculation.

2.2. Initialization

Given an interval $[\underline{x}, \overline{x}]$ representing some quantity x in standard IA, an equivalent affine form for the same quantity x is given by the following:

$$\frac{\underline{x} + \overline{x}}{2} + \frac{\overline{x} - \underline{x}}{2}\varepsilon, \quad (-1 \leq \varepsilon \leq 1). \tag{2}$$

ε represents the uncertainty in the value of x that is implicit in its standard interval representation. For a vector or a matrix, we must give a separate noise symbol ε_k for each component, because there are no relation between components in general.

2.3. Conversion from AA to IA

If a quantity x is described by the affine form like the formula (1), then it can be always converted to the following standard interval representation:

$$x \in [x_0 - \Delta_x, x_0 + \Delta_x], \quad \Delta_x = \sum_{i=1}^{n}(|x_i|) \qquad (3)$$

And the interval $[x_0 - \Delta_x, x_0 + \Delta_x]$ in the formula (3) is the smallest interval that contains all possible values of affine form x in the formula (1). But we had better not to convert quantities from affine form to standard interval representation, since we must lose any knowledge of correlations between quantities that was preserved in their affine forms.

2.4. Linear operations

For affine form x, y and a real number α, rules of linear operations for AA is defined as follows:

$$x \pm y = (x_0 \pm y_0) + \sum_{i=1}^{n}(x_i \pm y_i)\,\varepsilon_i \qquad (4)$$

$$x \pm \alpha = (x_0 \pm \alpha) + \sum_{i=1}^{n} x_i \varepsilon_i \qquad (5)$$

$$\alpha\, x = (\alpha x_0) + \sum_{i=1}^{n}(\alpha x_i)\,\varepsilon_i \qquad (6)$$

2.5. Nonlinear unary operations

Let f be defined as a nonlinear unary operation. Generary, $f(x)$ for some affine form x is not able to be expressed directly as an affine form. Therefore, linear approximation of f and maximum approximation error by introducing a new noise symbol ε_{n+1} is considered. First, X, domain of x is calculated as follows:

$$X = [x_0 - \Delta_x, x_0 + \Delta_x], \quad \Delta_x = \sum_{i=1}^{n}|x_i| \qquad (7)$$

Next, $ax + b$, linear approximation of $f(x)$ in X is calculated(see figure 1).

And maximum approximation error δ is calculated as follows:

$$\delta = \max_{x \in X}|f(x) - (ax + b)| \qquad (8)$$

By introducing a new noise symbol ε_{n+1}, the result of nonlinear unary operations is defined as follows:

$$\begin{aligned} f(x) &= ax + b + \delta\varepsilon_{n+1} \\ &= a(x_0 + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n) + b + \delta\varepsilon_{n+1} \quad (9) \end{aligned}$$

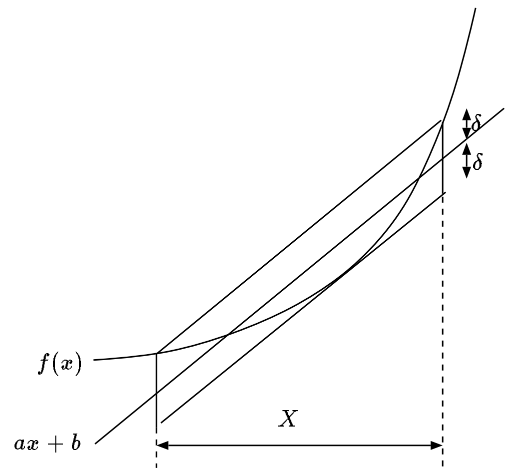

Figure 1: Linear approximation of nonlinear function.

2.6. Nonlinear binomial operations

In nonlinear binomial operation $f(x, y)$, the same method as nonlinear unary operation is refered to be possible. First, X, Y, domain of x, y is calculated. Next, $ax + by + c$, linear approximation of $f(x, y)$ in $X \times Y$ is calculated. And the result of nonlinear binomial operations is defined as follows:

$$\begin{aligned} f(x, y) &= ax + by + c + \delta\varepsilon_{n+1} \\ &= a(x_0 + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n) \\ &\quad + b(y_0 + y_1\varepsilon_1 + \cdots + y_n\varepsilon_n) \\ &\quad + c + \delta\varepsilon_{n+1} \qquad (10) \end{aligned}$$

2.7. The most suitable nonlinear operations

The most suitable nonlinear operations is defined as the nonlinear operations that minimize δ, coefficient of a new noise symbol ε_{n+1}. In nonlinear unary operations, the method described in section 2.5 provides the most suitable evaluation(formula(9) is defined as the most suitable result). However, in nonlinear binomial operation, the method described in section 2.6 does not always provide the most suitable evaluation(formula(10) is not always defined as the most suitable result). The reason of this matter is that in the case that correlation between variables exists(common ε is included between x and y), point(x, y) does not lie in all of the rectanglar domain $X \times Y$. Generally, in the case that $n = 1$, the domain that point(x, y) lies , which is called "joint range"forms a segment and in the case that $n \geq 2$, joint range forms a convex polygon , which is symmetrical with respect to point (x_0, y_0). For example, joint range for following x, y:

$$x = 5 + 2\varepsilon_1 - \varepsilon_2 - \varepsilon_3 \quad x \in [1, 9] \tag{11}$$

$$y = 9 - \varepsilon_1 + 4\varepsilon_2 + 2\varepsilon_3 \quad y \in [2, 16] \tag{12}$$

is defined as the slant line part of figure 2.7.

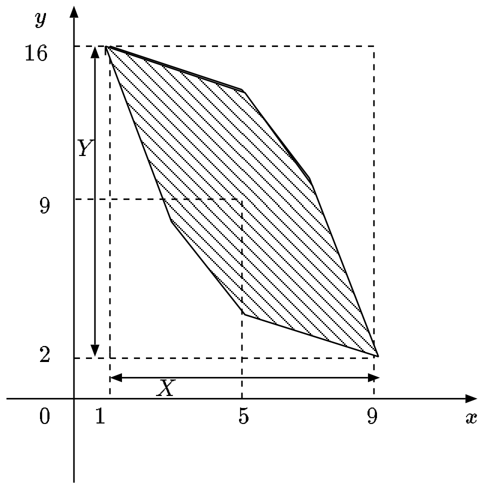

Figure 2: joint range

Therefore, in the most suitable nonlinear binomial operations, $ax+by+c$, linear approximation of $f(x, y)$ in joint range is calculated. And the result is defined as follows(formula(13) is defined as the most suitable result):

$$\begin{aligned} f(x, y) &= ax + by + c + \delta\varepsilon_{n+1} \\ &= a(x_0 + x_1\varepsilon_1 + \cdots + x_n\varepsilon_n) \\ &\quad +b(y_0 + y_1\varepsilon_1 + \cdots + y_n\varepsilon_n) \\ &\quad +c + \delta\varepsilon_{n+1} \end{aligned} \tag{13}$$

But the methods which realize these nonlinear binomial operations are refered to be difficult in general.

3. Dividing method at present

3.1. conventional method

The dividing method has not been defined in [1][2]. Thererore, up to now division has operated by $x/y = x \times 1/y$, combination of multiplication and reciprocal number. This method does not supply the most suitable evaluation because two εs add in one division by operating two nonlinear operations.

3.2. Kashiwagi method

Kashiwagi[3] defined the method which is able to operate division in one nonlinear operation. This method approximate directly x/y to $ax + by + c$ and operation result is defined as $x/y = ax + by + c + \delta\varepsilon_{n+1}$. In this method, to reduce calculating cost, b is approximated as follows:

$$b = -\frac{x_0}{\underline{y}\overline{y}} \tag{14}$$

For this b, p, q, r, s is calculated as follows:

$$p = \max_{\underline{y} \le y \le \overline{y}} \left(\frac{\underline{x}}{y} - by\right) \tag{15}$$

$$q = \min_{\underline{y} \le y \le \overline{y}} \left(\frac{\underline{x}}{y} - by\right) \tag{16}$$

$$r = \max_{\underline{y} \le y \le \overline{y}} \left(\frac{\overline{x}}{y} - by\right) \tag{17}$$

$$s = \min_{\underline{y} \le y \le \overline{y}} \left(\frac{\overline{x}}{y} - by\right) \tag{18}$$

And by following formula(19) and formula(20), a, c is calculated.

$$a\underline{x} + c = \frac{p+q}{2} \tag{19}$$

$$a\overline{x} + c = \frac{r+s}{2} \tag{20}$$

And δ is culculated as follows:

$$\delta = \max\left(\frac{p-q}{2}, \frac{r-s}{2}\right) \tag{21}$$

3.3. Shirai method

Shirai[4] calculated b strictly in the Kashiwagi method. The way of calculating a, c, δ is the same as the Kashiwagi method. In the case that corelation between variables does not exist, this method is able to provide the most suitable evaluation. However, in the case that corelation between variables exists, this method is not able to provide the most suitable evaluation.

4. New dividing method

In the Shirai metod, a, b which are the most suitable for the rectanglar domain $X \times Y$ are calculated, and by introducing this a, b , c, δ **which are the most suitable for the rectanglar domain** $X \times Y$ are calculated. Against this, in the new dividing method, a, b which are the most suitable for the rectanglar domain $X \times Y$ are calculated, and by introducing this a, b , c, δ **which are the most suitable for the joint range** are calculated. Concrete steps of the new dividing method is as follows:

step1 Calculate a, b by operating the Shirai method.

step2 Let D be defined as the joint range, and Let $d(x, y)$ be defined as follows:

$$d(x, y) = \frac{x}{y} - (ax + by) \tag{22}$$

And calculate following $d_{\max}$, $d_{\min}$.

$$d_{\max} = \max_{(x,\, y)\in D} d(x, y) \tag{23}$$

$$d_{\min} = \min_{(x,\, y)\in D} d(x, y) \tag{24}$$

step3 Calculate a, b by following formula(25) and formula(26).

$$c = \frac{d_{\max} + d_{\min}}{2} \tag{25}$$

$$\delta = \frac{d_{\max} - d_{\min}}{2} \tag{26}$$

At last, the result is defined as follows:

$$\frac{x}{y} = ax + by + c + \delta\varepsilon_{n+1} \tag{27}$$

□

Though the way of calculating a, b is not always defined as the most suitable, the way of calculating c,δ is always defined as the most suitable. The fact that this method provides more suitable evaluation than the Shirai method is expected.

5. Numerical example

As a numerical example, interval of expression(28) is evaluated.

$$\frac{-x + 2y}{2x + y} - 0.2y$$

$$x \in [0,\, 1],\, y \in [1,\, 3] \quad (x = 0.5 + 0.5\varepsilon_1,\, y = 2 + \varepsilon_2) \tag{28}$$

The method which provides operating result whose interval width is narrow guarantees high accuracy. Interval widths of the operating results when operating standard IA and AA(conventional method, Kashiwagi method, Shirai method and new dividing method) are shown in the following table 1.

By table1, the fact that the new dividing method is more accurate than the other methods is confirmed.

6. Conclusion

By section5, the fact that the new dividing method is more accurate than the other methods is confirmed. Therefore, in conclusion, the new method is recommended when dividing in AA.

Table 1: Comparison of interval width of the operating result.

methods	interval width
standard IA	6.200000
AA(conventional method)	5.869505
AA(Kashiwagi method)	5.113950
AA(Shirai method)	4.810998
AA(new dividing method)	3.524211

References

[1] Marcus Vinícius A. Andrade, João L. D. Comba and Jorge Stolfi: "Affine Arithmetic" INTERVAL'94, St. petersburg(Russia), March 5-10, 1994

[2] Luiz H. de Figueiredo and Jorge Stolfi: "Self-Validated Numerical Methods and Applications", Brazilian Mathematics Colloquium monograph, IMPA, Rio de Janeiro, Brazil; July 1997.

[3] Masahide Kashiwagi: "On the Division of the Affine Arithmetic", Technical Report of The Institute Electronics, Information and Communication Engineers, NLP 98-114, pp. 23-30 , 1999 (in Japanese)

[4] Kenichi Shirai, Takatomi Miyata, Masahide Kashiwagi: "On the Division of the Affine Arithmetic", proceedings of Engineering Sciences Society Conference in The Institute Electronics, Information and Communication Engineers , P. 45 , 2000 (in Japanese)