位宽优化中乘法运算的一种自动范围分析方法

孙瑞一，张 岩

（哈尔滨工业大学 深圳研究生院 网络环境智能计算重点实验室，518055 广东 深圳）

摘 要：乘法是硬件平台中最基本的非线性运算，而且在自动位宽优化过程中，目前的范围分析方法没有在精确的范围分析结果和计算复杂度之间做很好折衷.为了在较低的计算复杂度前提下更准确地分析乘法运算结果的范围，提出了改进的仿射近似法（NAA）.在改进的仿射近似法中，利用额外噪声项来表示近似产生的误差，并根据误差的特点把误差分成两部分，在不增加计算复杂度的前提下更准确地估计误差的范围.新方法的计算复杂度是 $O(M_1)$，其中 M_1 是乘法的两个操作数中非零噪声个数的和.实例分析表明，利用该方法得到的乘法结果范围的准确程度是用简单估计法得到的准确程度的 1.47 倍，和切比雪夫近似法的准确度接近.

关键词：位宽优化；范围分析；乘法；仿射算术；仿射近似法

中图分类号：TP391.72 **文献标志码**：A **文章编号**：0367-6234(2014)03-0043-06

A range analysis in automatic word length optimization for multiplication

SUN Ruiyi, ZHANG Yan

(Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, 518055 Shenzhen, Guangdong, China)

Abstract: To achieve more accurate result and lower computational complexity of range analysis for multiplication in automatic word length optimization, this paper presents a novel refined affine approximation method of multiplication for range analysis in automatic word length optimization, which is named novel affine arithmetic approximation (NAA). In NAA, a new noise term represents the error which is caused by approximation. This error is estimated more accurately without increasing the computational complexity. The computational complexity of NAA is $O(M_1)$, where M_1 denotes the total of the nonzero noise of the two multipliers. In experiments, the accuracy of the range using NAA is 1.47 times of that using trivial range estimation, and the same as that using Chebyshev approximation.

Keywords: word-length optimization; range analysis; multiplication; affine arithmetic; affine approximation method

在硬件系统的设计中，为了得到较高的数据精度及动态范围，算法模型都用浮点数表示.但是在硬件上实现浮点运算的代价很大，为了提高运行速度、降低功耗、节省面积，在硬件实现阶段数据一般都采用定点数表示.将浮点数转化为定点数的过程被称为位宽优化，也叫做定点化.位宽优化的目的是在满足系统规格要求的前提下寻找最佳的定点数位宽组合使系统的代价（面积、功耗、速度等）最小.位宽优化包括范围分析（对整数部分的优化）和精度分析（对小数部分的优化）两个步骤.

位宽优化是 NP-hard 问题[1]，常用的分类方法有动态法[2-6]和静态法[7-10].动态法通过对大量的测试向量进行反复的蒙特卡洛仿真来确定信号的位宽.它的结果质量较高，但优化时间长，甚至占到整个设计周期的 50%以上[11].而且，动态法不能保证测试向量没有覆盖到的情况也满足系统规格要求.静态法使用代码分析手段来推导出信号的位宽，它不需要测试向量，所以它的执行速度快、人为干扰因素小，而且简便易行，适于大规

收稿日期：2013-02-27.

基金项目：深圳市科技研发基础研究计划资助项目（JC201005260168A）.

作者简介：孙瑞一（1980—），女，博士研究生；
张 岩（1969—），男，教授，博士生导师.

通信作者：张 岩，ianzh@foxmail.com.

模系统的自动优化分析.

在硬件平台,如 Xilinx 公司的 FPGA 或全定制的 ASIC 芯片中,乘法运算是最基本也是最常用的非线性运算.其他的非线性运算,如除法、余弦、对数等超越函数一般都是通过加、减、乘这些基本运算计算的.但是目前乘法运算的范围分析的静态法得不到范围的精确解,所以对乘法运算结果的位宽优化的准确程度就关系到整个电路的位宽优化的准确程度.本文讨论基于静态法的对乘法结果的范围分析方法.

区间算术(interval arithmetic,IA)是 1960 年由 Moore[12] 提出的解决范围分析的方法. Cmar 等[13]首次采用 IA 对信号的变化范围进行分析.但是 IA 在估计信号范围时过于保守,甚至不切实际.

仿射算术(affine arithmetic,AA)保留了区间之间的相关性,这使得它比 IA 更精确.AA 非常适于对线性操作结果的范围分析.但是 AA 不能为乘法等非线性操作提供准确的仿射形式.为解决该问题,Stolfi 等[14]提出了乘法的仿射近似法,包括简单估计法和切比雪夫近似法.简单估计法计算效率高,但误差较大,分析范围最多可达到真实范围的 4 倍.这种误差沿着数据流累积到输出,会导致误差爆炸,这限制了它在大系统中的应用;切比雪夫近似法利用比简单估计法更精确的仿射形式,实现较好的范围分析结果,但是它的计算过程太复杂所以不适合在大系统中应用.

为了解决计算效率和计算复杂度之间的问题,Zhang 等[15]提出了一种叫 N - 级近似的,在分析精度和复杂度之间作折中的方法.其分析结果比用简单估计法得到的准确,比切比雪夫近似法简单.但是仍不能满足大系统对分析范围的精度和复杂度的要求.Pang 等[16]提出了一种新的范围分析方法,它混合了传统的 IA、AA 和算术变换(arithmetic transform,AT),能得到比 AA 更精确的范围结果,但是执行时间比 AA 长很多.所有这些对 AA 的改进方法,都以牺牲计算复杂度为代价来换取分析的精度.它们都是把自变量的仿射形式看成是一个整体,没有考虑仿射形式中的每一个噪声,这样就忽略了不同噪声之间的独立性和相同噪声在不同变量中的相关性.

为了更简便、更精确的实现对乘法运算结果的范围分析,本文提出一种名为改进的仿射近似法(novel affine arithmetic approximation,NAA)的新的仿射近似法.NAA 利用一个仿射形式近似的表示乘法运算的结果,并利用增加的额外噪声项来表示近似产生的误差.为了利用现有方法忽略的独立性和相关性,这个误差用各个噪声来表示,并且在不增加计算复杂度的前提下更准确的估计误差的范围,从而减少了新增噪声的系数.这种方法能比简单估计法得到更准确的分析范围,而且计算复杂度和简单估计法一样,比切比雪夫近似法更简单.

1 仿射算术

为了更清楚的分析仿射近似的过程和更方便的推导 NAA,首先介绍仿射算术的基础知识.

仿射算术用一次多项式的仿射形式 $\hat{x}$ 表示一个未知的信号 x 为

$$\hat{x} = x_0 + x_1\varepsilon_1 + x_2\varepsilon_2 + \cdots + x_n\varepsilon_n.$$

式中:$\varepsilon_i = [-1,1]$.对于未知信号 x,x_0 表示它的中心值;ε_i 为第 i 个噪声,代表信号 x 的一个独立的、未知的噪声源;x_i 为这个噪声的系数;$x_i\varepsilon_i$、$x_0 + x_i\varepsilon_i$ 分别为噪声项.

令 $x_0 = (x_{\max} + x_{\min})/2$,$x_1 = (x_{\max} - x_{\min})/2$,信号的范围区间 $\bar{x} = [x_{\min}, x_{\max}]$ 可以转化为等价的仿射形式为

$$\hat{x} = x_0 + x_1\varepsilon_1.$$

AA 通过与计算链中的其他信号共享噪声 ε_i 来保留信号之间的相关性.

乘法的简单估计法的仿射形式为

$$\hat{xy} = x_0y_0 + \sum_{i=1}^{n}(x_0y_i + y_0x_i)\varepsilon_i + \sum_{i=1}^{n}|x_i|\sum_{i=1}^{n}|y_i|\varepsilon_{n+1}.$$

这种方法的分析结果误差很大,但是其计算复杂度是 $O(M_1)$,其中 $M_1 = \max(n_1, n_2)$,n_1、n_2 分别为信号 x、y 中非零噪声的个数.

利用切比雪夫近似法可以得到更精确的范围分析结果.乘法的切比雪夫近似法的仿射形式为

$$\hat{xy} = x_0y_0 + \sum_{i=1}^{n}(x_0y_i + y_0x_i)\varepsilon_i + \frac{m+n}{2} + \frac{m-n}{2}\varepsilon_{n+1}.$$

式中:m、n 分别为 $(x - x_0)(y - y_0)$ 的最大值和最小值.

计算最值 m 和 n 的计算复杂度是 $O(M_2\log M_2)$,其中 $M_2 = n_1 + n_2$.

2 乘法的改进的仿射近似法

2.1 算法分析

自变量为仿射形式的乘法,可以表示为

$$z = \hat{x}\hat{y} = x_0y_0 + \sum_{i=1}^{n}x_0y_i\varepsilon_i + \sum_{i=1}^{n}y_0x_i\varepsilon_i + \sum_{i=1}^{n}x_i\varepsilon_i\sum_{j=1}^{n}y_j\varepsilon_j. \quad (1)$$

因为 z 不是仿射形式，选择一个仿射形式 f_z 来近似的表示它.式(1) 的前3项组成了一个仿射形式，为了简单，把这个仿射形式作为近似的仿射形式，即

$$f_z = x_0 y_0 + \sum_{i=1}^{n} x_0 y_i \varepsilon_i + \sum_{i=1}^{n} y_0 x_i \varepsilon_i. \tag{2}$$

式中：f_z 为 z 的近似表示，所以 f_z 与 z 之间存在误差，引入一个新的独立噪声项来表示这个误差 d_f，即

$$\hat{d} = z' + z_{n+1}\varepsilon_{n+1} = \frac{d_{f\max} - d_{f\min}}{2} + \frac{d_{f\max} + d_{f\min}}{2}\varepsilon_{n+1}.$$

计算 d_f 的真实最大值和最小值的计算量难以承受，所以 $d_{f\max}$、$d_{f\min}$ 分别为 d_f 的近似最大值和近似最小值.$d_{f\max}$、$d_{f\min}$ 与真实的最大值和最小值越接近，$\hat{z}$ 的形式越准确.

综上所述，z 的仿射形式可以记为

$$\hat{z} = f_z + \hat{d} = z_0 + z_1\varepsilon_1 + \cdots + z_n\varepsilon_n + z' + z_{n+1}\varepsilon_{n+1}.$$

2.2　仿射形式 $\hat{d}$

根据算法分析，仿射形式 $\hat{d}$ 是乘法运算结果的真实值与它的近似仿射形式之间的误差 d_f 的等价仿射形式.把 d_f 表示成各个噪声的函数，根据噪声项的特点，把它分成两部分，分别求解它们的仿射形式.这样可以更准确的估计各个部分的范围.这两个仿射形式的和就是 d_f 的仿射形式.利用这个仿射形式，d_f 的最大值和最小值的估计方法可以明显改善.

d_f 可以表示为

$$d_f = z - f_z = \left(x_0 + \sum_{i=1}^{n} x_i\varepsilon_i\right)\left(y_0 + \sum_{i=1}^{n} y_i\varepsilon_i\right) - x_0 y_0 - y_0\sum_{i=1}^{n} x_i\varepsilon_i - x_0\sum_{i=1}^{n} y_i\varepsilon_i = \sum_{i,j=1}^{n} x_i y_j \varepsilon_i\varepsilon_j. \tag{3}$$

式中：$\varepsilon_i, \varepsilon_j = [-1, 1]$.估计 d_f 的近似的最大值和最小值的最简单的方法就是把 $\varepsilon_i, \varepsilon_j$ 的最大值和最小值代入.根据式(2)，当 $i = j$ 时，有 $x_i y_j \varepsilon_i \varepsilon_j = x_i y_i \varepsilon_i^2$.把 $\varepsilon_i, \varepsilon_j = [-1, 1]$ 代入，$x_i y_i \varepsilon_i^2$ 估计的近似的最大值和最小值比 $x_i y_j \varepsilon_i \varepsilon_j$ 更精确，而且没有增加计算复杂度.因此把 d_f 分成两部分，即

$$d_1 = \sum_{i=1}^{n} x_i y_i \varepsilon_i^2, d_2 = \sum_{i,j=1, i\neq j}^{n} x_i y_j \varepsilon_i \varepsilon_j.$$

距离 d_f 可以记为

$$d_f = \sum_{i,j=1}^{n} x_i y_j \varepsilon_i \varepsilon_j = d_1 + d_2 = \sum_{i=1}^{n} x_i y_i \varepsilon_i^2 + \sum_{i,j=1, i\neq j}^{n} x_i y_j \varepsilon_i \varepsilon_j.$$

d_1 的最大值和最小值分别为

$$d_{1\max} = \begin{cases} \sum_{i=1}^{n} x_i y_i, & \sum_{i=1}^{n} x_i y_i \geqslant 0; \\ 0, & \text{其他.} \end{cases}$$

$$d_{1\min} = \begin{cases} 0, & \sum_{i=1}^{n} x_i y_i \geqslant 0; \\ \sum_{i=1}^{n} x_i y_i, & \text{其他.} \end{cases}$$

和它等价的仿射形式为

$$\hat{d}_1 = \frac{1}{2}\sum_{i=1}^{n} x_i y_i + \frac{1}{2}\sum_{i=1}^{n} x_i y_i \varepsilon_{n+1}.$$

d_2 的最大值和最小值分别为

$$d_{2\max} = \sum_{i,j=1, i\neq j}^{n} |x_i y_j|, d_{2\min} = -\sum_{i,j=1, i\neq j}^{n} |x_i y_j|.$$

和它等价的仿射形式为

$$\hat{d}_2 = \sum_{i,j=1, i\neq j}^{n} |x_i y_j| \varepsilon_{n+2}.$$

将 $\hat{d}_1$、$\hat{d}_2$ 相加，得 $\hat{d}$ 为

$$\hat{d} = \hat{d}_1 + \hat{d}_2 = \frac{1}{2}\sum_{i=1}^{n} x_i y_i + \frac{1}{2}\sum_{i=1}^{n} x_i y_i \varepsilon_{n+1} + \sum_{i,j=1, i\neq j}^{n} |x_i y_j| \varepsilon_{n+2}. \tag{4}$$

计算一次乘法要引入两个独立噪声会使接下来的计算噪声越来越多.所以需要化简 $\hat{d}$.根据式(4)，$\hat{d}$ 的范围为

$$\left[\frac{1}{2}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{2}\left|\sum_{i=1}^{n} x_i y_i\right| + \sum_{i,j=1, i\neq j}^{n} |x_i y_j|\right), \frac{1}{2}\sum_{i=1}^{n} x_i y_i + \left(\frac{1}{2}\left|\sum_{i=1}^{n} x_i y_i\right| + \sum_{i,j=1, i\neq j}^{n} |x_i y_j|\right)\right].$$

和它等价的仿射形式为

$$\hat{d} = \frac{1}{2}\sum_{i=1}^{n} x_i y_i + \left(\frac{1}{2}\left|\sum_{i=1}^{n} x_i y_i\right| + \sum_{i,j=1, i\neq j}^{n} |x_i y_j|\right)\varepsilon_{n+1}. \tag{5}$$

这样计算出的新增噪声项 $\hat{d}$ 的范围比简单估计法中新增噪声项的范围 $[-\sum_{i=1}^{n}|x_i|\sum_{i=1}^{n}|y_i|, \sum_{i=1}^{n}|x_i|\sum_{i=1}^{n}|y_i|]$ 要紧凑得多.$\hat{d}$ 将新增噪声的权重减少了 $\frac{1}{2}\sum_{i=1}^{n} x_i y_i$，也就减少了乘法运算近似产生的误差.因此，乘法的改进的仿射近似法比简单估计法更准确.当误差在数据链中的累计时，累计的误差也减少了.使 NAA 比简单估计法更精确.

式(4) 中的 ε_{n+1} 和 ε_{n+2} 之间有相关性，但是将 d_f 分为 d_1 和 d_2 并没有考虑它们之间的相关性，所以 d_f 的真实范围一定比 $\hat{d}$ 的范围小，这是过估

计.因此,NAA的准确程度比切比雪夫近似法的准确程度稍低.

2.3 乘法的改进的仿射近似法的表达式及计算复杂度分析

有了式(2)和新引入的式(5),乘法的改进的仿射近似法的表达式为

$$\hat{z}=\hat{x}\hat{y}=x_0y_0+x_0\sum_{i=1}^{n}y_i\varepsilon_i+y_0\sum_{i=1}^{n}x_i\varepsilon_i+\frac{1}{2}\sum_{i=1}^{n}x_iy_i+\left(\frac{1}{2}\left|\sum_{i=1}^{n}x_iy_i\right|+\sum_{i,j=1,i\neq j}^{n}|x_iy_j|\right)\varepsilon_{n+1}. \tag{6}$$

设 n_1、n_2 分别为 $\hat{x}$、$\hat{y}$ 中非零噪声的个数,$M_1=\max(n_1,n_2)$ 为 n_1 和 n_2 的最大值,$M_2=n_1+n_2$,则切比雪夫近似法的计算复杂度为$O(M_2\log M_2)$[14].简单估计法的计算复杂度为 $O(M_1)$.

为了计算NAA的计算复杂度,把式(6)改写为

$$\hat{z}=\hat{x}\hat{y}=x_0y_0+x_0\sum_{i=1}^{n}y_i\varepsilon_i+y_0\sum_{i=1}^{n}x_i\varepsilon_i+\frac{1}{2}\sum_{i=1}^{n}x_iy_i+\left(\frac{1}{2}\left|\sum_{i=1}^{n}x_iy_i\right|+\left(\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{j=1}^{n}y_j\right)-\sum_{i=1}^{n}|x_iy_i|\right)\right)\varepsilon_{n+1}. \tag{7}$$

式(7)的计算复杂度为

$$1+M_1+M_1+\frac{1}{2}M_1+\left(\frac{1}{2}M_1+(2M_1+M_1)\right)=O(M_1).$$

可以看出,NAA的计算复杂度和简单估计法的一样,都是$O(M_1)$.很明显$M_1<M_2$,所以有$O(M_1)<O(M_2\log M_2)$.根据计算复杂度的分析,NAA比切比雪夫近似法要简单.

3 实例分析

为了比较NAA的性能,本文以包含乘法的计算为例.第1个实例是多项式近似.第2个实例是*B*-splines.这两个实例均来自于文献[8].第3个实例是多变量多项式,来自于文献[17-19].将仿射算术用C++实现,乘法分别采用简单估计法、切比雪夫近似法和改进的仿射近似法.这些实例的计算平台是内存为2.99 G的Intel(R) Pentium(R) CPU G630 @2.7 GHz的32位双核PC.

3.1 实 例

实例1 多项式近似

对 $y=\ln(1+x)$,$x=[0,1]$ 的多项式近似进行信号范围分析,采用文献[20]的4级多项式表达为

$$y=(((-0.055\,0x+0.216\,8)x-0.464\,5)x+0.995\,6)x+0.000\,1.$$

其中,5个四舍五入到小数点后第4位的系数由多项式曲线拟合的方法获得.

实例2 *B*-splines

B-splines常用于图像卷绕,它的基本函数 B_0、B_1、B_2、B_3 分别定义如下,其中,输入 $u=[0,1]$.

$$B_0(u)=\frac{(1-u)^3}{6},B_1(u)=\frac{(3u^3-6u^2+4)}{6},$$

$$B_2(u)=\frac{-3u^3+3u^2+3u+1}{6},B_3(u)=\frac{-u^3}{6}.$$

实例3 变量多项式

考察表1中的8个多变量多项式.

表1 多变量多项式函数及输入范围

函数名称	函数	输入范围
Savitzky-Golay filter	$f_1(X)=7x_1^3-984x_2^3-76x_1^2x_2+92x_1x_2^2+7x_1^2-39x_1x_2-46x_2^2+7x_1-46x_2-75$	$X=[-2,2]^2$
Image rejection unit	$f_2(X)=16\,384(x_1^4+x_2^4)+64\,767(x_1^2-x_2^2)+x_1-x_2+57\,344x_1x_2(x_1-x_2)$	$X=[0,1]^2$
A random function	$f_3(X)=(x_1-1)(x_1+2)(x_2+1)(x_2-2)x_3^2$	$X=[-2,2]^3$
Mitchell function	$f_4(X)=4[x_1^4+(x_2^2+x_3^2)^2]+17x_1^2(x_2^2+x_3^2)-20(x_1^2+x_2^2+x_3^2)+17$	$X=[-2,2]^3$
Matyas function	$f_5(X)=0.26(x_1^2+x_2^2)-0.48x_1x_2$	$X=[-100,100]^2$
Three hump function	$f_6(X)=12x_1^2-6.3x_1^4+x_1^6+6x_2(x_2-x_1)$	$X=[-10,10]^2$
Goldstein Price function	$f_7(X)=[1+(x_1+x_2+1)^2(19-14x_1+3x_1^2-14x_2+6x_1x_2+3x_2^2)]\times[30+(2x_1-3x_2)^2(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2)]$	$X=[-2,2]^2$
Ratscheck function	$f_8(X)=4x_1^2-2.1x_1^4+\frac{1}{3}x_1^6+x_1x_2-4x_2^2+4x_2^4$	$X=[-100,100]^2$

3.2 3种方法分析范围的比较

利用仿射近似法计算得到的变量范围用分析范围来表示.用分析范围比率来表示分析范围的准确程度.分析范围比率是用3种方法得到的分析范围的区间间隔($y_{\max}-y_{\min}$)与真实范围的区间间隔的比值,表示了分析范围的区间间隔是真实范围的区间间隔的倍数.其中,$y_{\max}$、$y_{\min}$ 分别为变量的最大值和最小值.分析范围比率越接近1说明分析范围越接近真实范围.

表2给出了用3种方法计算得到的3个例子的分析范围和范围比率.它显示了利用NAA计算得到的分析范围全部覆盖了真实的输出范围.对

于这些例子，利用简单估计法计算得到的范围比率从 1.04～281.19；利用切比雪夫近似法计算得到的范围比率从 1.03～233.66；利用 NAA 计算得到的范围比率从 1.03～221.78.利用 NAA 得到的范围比率是利用简单估计法得到的范围比率的 0.18～0.99 倍，范围比率倍数的平均是0.68 倍；利用 NAA 计算得到的范围的比率是利用切比雪夫近似法计算得到的范围比率的 0.33～1.39 倍，范围比率倍数的平均是 0.99 倍.这说明，平均来看，利用 NAA 计算得到的分析范围的准确程度是用简单估计法计算得到的分析范围的准确程度的 1.47 倍，和切比雪夫近似法的分析范围的准确度接近.

表 2　3 种方法分析范围和范围比率的比较

实例	函数	真实范围	简单估计法		切比雪夫近似法		NAA	
			范围	比率	范围	比率	范围	比率
1	Y	[0,0.693 1]	[-0.054 1,0.865 0]	1.33	[-0.025 3,0.768 5]	1.15	[-0.027 0.0.780 7]	1.17
2	B_0	[0,0.17]	[-0.13,0.17]	1.76	[-0.05,0.17]	1.29	[-0.06,0.17]	1.37
	B_1	[0.17,0.67]	[-0.33,1.29]	3.24	[-0.05,0.98]	2.06	[-0.02,0.92]	1.88
	B_2	[0.17,0.67]	[-0.21,1.17]	2.76	[-0.02,0.89]	1.82	[0.04,0.85]	1.62
	B_3	[-0.17,0]	[-0.17,0.13]	1.76	[-0.17,0.05]	1.29	[-0.17,0.06]	1.37
3	f_1	[-945 3,930 3]	[-982 1,967 1]	1.04	[-979 3,948 7]	1.03	[-979 3,948 7]	1.03
	f_2	[-5.51e4,8.79e4]	[-1.75e5,1.79e5]	2.48	[-0.95e5,1.28e5]	1.56	[-1.21e5,1.41e5]	1.85
	f_3	[-36,64]	[-352,352]	7.04	[-192,192]	3.84	[-64,64]	1.28
	f_4	[-8,641]	[-1 087,1 121]	3.40	[-223,881]	1.70	[-367,641]	1.55
	f_5	[0,1e4]	[-1e4,1e4]	2.00	[-480 0,1e4]	1.48	[-480 0,1e4]	1.48
	f_6	[0,0.94e6]	[-1.07e6,1.07e6]	2.28	[-0.06e6,1.0e6]	1.13	[-3.6e5,9.4e5]	1.38
	f_7	[3,1.01e6]	[-1.42e8,1.42e8]	281.19	[-1.23e8,1.13e8]	233.66	[-1.12e8,1.12e8]	221.78
	f_8	[-1.03,0.33e12]	[-3.3e11,3.3e11]	2.00	[-2.1e8,3.3e11]	1.00	[-1.3e11,3.3e11]	1.39

对于实例 3 中的第 7 个函数，利用 3 种静态法得到的范围比率都在 200 以上.这是因为这个函数最终的乘法中的两个乘数都包含很多乘法，每一次乘法运算都会产生误差，而且这两个乘数都是相同自变量的函数，它们的相关性很强，这种情况下两个乘数的误差会相互放大，使最终得到的范围结果比真实范围大很多.所以静态法不适用于对这种情况的范围分析，动态法会得到更准确的分析范围.

从表2可以看出，这3种方法有一个共同的不足之处，即分析范围的下限误差较大.这是因为当变量 x 用仿射形式表示时，x 的范围是相对于 x_0 中心对称的.在计算过程中，x_0 是由计算链中变量决定的，很难保证它是在 x 范围的中心.当 $\hat{x}$ 要满足变量范围的上限或下限时，由于 x_0 不在 x 范围的中心，$\hat{x}$ 表示的 x 范围的下限或上限就有了较大的误差.在本文所举的例子中，都是上限大于下限的，所以表现为分析范围下限误差较大.

3.3　3 种方法整数位宽和执行时间的比较

变量的整数位宽可以由变量范围推导.表 3 列出了 3 种方法得到的输出变量位宽和 CPU 运行时间. B-splines 的 4 个基本函数是在同一个程序中完成的，所以只有一个 CPU 时间.NAA 最多比简单估计法节省 2 个 bit 位宽，比切比雪夫近似法最多节省 1 个 bit 位宽.与简单估计法相比，NAA 需要 0.95～1.39 倍的 CPU 时间，而切比雪夫近似法需要 54.01～57.97 倍的 CPU 时间.NAA 和简单估计法有着近似的执行时间，不到切比雪夫近似法执行时间的 1/50.

表 3　3 种方法得到的整数位宽和 CPU 时间的比较

实例	函数	整数位宽/bits				CPU 执行时间/ms		
		真实	简单	切比	NAA	简单	切比	NAA
1	Y	2	2	2	2	0.301	47.549	0.337
2	B_0	2	2	2	2	0.625	79.121	0.657
	B_1	2	2	2	2			
	B_2	2	2	2	2			
	B_3	2	2	2	2			
3	f_1	15	15	15	15	1.273	68.756	1.346
	f_2	18	19	18	18	0.904	104.489	0.859
	f_3	8	10	9	8	0.356	46.305	0.457
	f_4	11	12	11	11	0.883	58.611	1.107
	f_5	15	15	15	15	0.279	41.686	0.298
	f_6	21	22	21	21	0.498	49.845	0.545
	f_7	21	29	28	28	1.437	128.860	2.004
	f_8	40	40	40	40	0.873	85.112	0.835

4　结　论

1）利用一个仿射形式近似的表示乘法运算的结果，并利用额外噪声项来表示近似产生的误差.

2)把近似产生的误差分成两部分,分别用两个噪声来表示,能得到更精确的近似误差的范围.

3)与以往的范围分析的改进方法相比,新方法在不增加计算复杂度的前提下的,更准确的估计变量的范围.实例分析显示 NAA 的分析范围的准确程度和切比雪夫近似法的准确程度接近,比简单估计法更准确.而且,它的计算复杂度和简单估计法的一样,都是 $O(M_1)$,比切比雪夫近似法的小.

参考文献

[1] CONSTANTINIDES G, WOEGINGER G. The complxity of multiple wordlength assignment [J]. Applied Mathematics Letters, 2002, 15(2):137-140.

[2] KUM K I, SUNG W. Combined word-length optimization and highlevel synthesis of digital signal processing systems [J]. IEEE Trans on Computer-Aided Design Integrated Circuits, and Systems, 2001, 20(8): 921-930.

[3] CAFFARENA G, CARRERAS C, LOPEZ J A, et al. SQNR estimation of fixed-point DSP algorithms [J]. Eurasip Journal on Advances in Signal Processing, 2010, 2010(21):1-12.

[4] 朱珂, 华林, 周晓方,等. JPEG2000 中小波滤波器的定点分析及其 VLSI 实现[J].固体电子学研究与进展, 2004, 24(4): 466-471,504.

[5] 马志强, 季振洲, 胡铭曾. 基于超窄数据的低功耗数据 Cache 方案[J]. 计算机研究与发展, 2007, 44(5): 775-781.

[6] BANCIU A, CASSEAU E, MENARD D, et al. Stochastic modeling for floating-point to fixed-point conversion [C]//Proceedings of IEEE Workshop on Signal Processing Systems (SiPS). Beirut, Lebanon: IEEE Computer Society Press, 2011:180-185.

[7] SARBISHEI O, RADECKA K, ZILIC Z. Analytical optimization of bit-widths in fixed-point LTI systems[J]. IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems, 2012, 31(3): 343-355.

[8] LEE D-U, GAFFAR A A, CHEUNG R C, et al. Accuracy guaranteed bit-width optimisation [J]. IEEE Trans on Computer-Aided Design of Integrated Circuits and Systems, 2006, 25(10): 1990-2000.

[9] ROCHER R, MENARD D, SCALART P. Analytical approach for numerical accuracy estimation of fixed-point systems based on smooth operations[J]. IEEE Trans on Circuits and Systems I-Regular Papers, 2012, 59(10): 2326-2339.

[10] KINSMAN A B, NICOLICI N. Computational vector-magnitude-based range determination for scientific abstract data types [J]. IEEE Trans on Computers, 2011, 60(11): 1652-1663.

[11] KEDING H, WILLEMS M, COORS M, et al. FRIDGE: a fixed-point design and simulation environment [C]//Proceedings of Design, Automation and Test in Europe. Paris: DATE, 1998: 429-435.

[12] MOORE R E. Interval Arithmetic and Automatic Error Analysis in Digital Computing[D]. California: Stanford University, 1962.

[13] CMAR R, RIJNDERS L, SCHAUMONT P, et al. A methodology and design environment for DSP ASIC fixed point refinement [C]//Proceedings of Design, Automation and Test in Europe. Munich: ACM, 1999: 271-276.

[14] STOLFI J, de FIGUEIREDO L H. Self-validated Numerical Methods and Applications [M]. Rio De Janeiro: Brazilian Mathematics Colloquium monograph, IMPA, 1997.

[15] ZHANG Linsheng, ZHANG Yan, ZHOU Wenbiao. Tradeoff between Approximation accuracy and complexity for range analysis using affine arithmetic[J]. Journal of Signal Processing Systems, 2010, 61(3): 279-291.

[16] PANG Y, RADECKA K. An efficient algorithm of performing range analysis for fixed-point arithmetic circuits based on SAT checking [C]//Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS). Rio de Janeiro, BRAZIL: IEEE Computer Society Press, 2011: 1736-1739.

[17] SHEKHAR N, KALLA P, ENESCU F. Equivalence verification of arithmetic datapaths with multiple word-length operands [C]//Proceedings of Design, Automation and Test in Europe. Munich: DATE, 2006: 824-829.

[18] GOPALAKRISHNAN S, KALLA P, MEREDITH M B, et al. Finding linear building-blocks for RTL synthesis of polynomial datapaths with fixed-size bit-vectors [C]//Proceedings of International Conference on Computer-Aided Design. San Jose: IEEE Computer Society Press, 2007:143-148.

[19] SHOU Huahao, SONG Wenhao, SHEN Jie, et al. A recursive taylor method for ray casting algebraic surfaces [C]//Proceedings of International Conference on Computer Graphics and Virtual Reality, Las Vegas: CGVR, 2006:196-204.

[20] HORNER W G. A new method of solving numerical equations of all orders, by continuous approximation[J]. Philosophical Transactions of the Royal Society of London, 1819, 109:308-335.

(编辑 张 红)