文章编号: 1671-1114(2014)02-0016-03

同符号数相加"大数吃小数"的界限: 数值试验

曹 靖[1,2], 李建平[1]

(1. 中国科学院 a. 大气物理研究所, b. 研究生院, 北京 100029; 2. 天津理工大学 理学院, 天津 300384)

摘 要: 利用大量数值试验, 得出机器单双精度下关于两同符号数相加时"大数吃小数"界限的数据. 对试验数据进行分段线性拟合, 给出"大数吃小数"界限的近似公式, 与试验数据比较表明近似公式具有良好的估计效果, 可为单双精度下避免"大数吃小数"现象提供可靠的依据. 进一步, 将机器单双精度的结果推广至任意机器精度, 得到的结论可方便应用于实际的数值运算中.

关键词: "大数吃小数"; 分段线性拟合; 机器精度

中图分类号: O246 **文献标志码**: A

On bound of a large number annihilating a small number in addition operation of two numbers with same sign: numerical experiment

CAO Jing[1,2], LI Jianping[1]

(1a. Institute of Atmospheric Physics, 1b. Graduate University, Chinese Academy of Sciences, Beijing 100029, China;
2. College of Science, Tianjin University of Technology, Tianjin 300384, China)

Abstract: Through a lot of numerical experiments, the data of the bounds of a large number annihilating a small number in addition operation of two numbers with same sign under single and double machine precisions are obtained. By analyzing the experimental results, piecewise linear approximations of the bounds have been established. The approximate functions show good performances when comparing with the exact results in numerical experiments. Moreover, based on the results under single and double machine precisions, generalized approximate function for any given machine precisions has been deduced, which can be conveniently applied in practical numerical computations.

Keywords: a large number annihilating a small number; piecewise linear fitting; machine precision

"大数吃小数"现象[1]是数值运算中常见的一种影响计算精度的现象, 当以计算机计算一个实数 a 与实数 $b\neq 0$ 的代数和时, 如果$|b|$相对于$|a|$小到一定程度, 会出现 $a+b=a$ 的现象, 一般称作数 b 被数a"吃掉"了. "大数吃小数"现象一个典型的例子[2]就是计算 N 个实数的累加和 $\sum_{n=1}^{N} a_n$, 其中, $a_1 \gg a_i > 0$, $i=2, 3, \cdots, N$. 若以正常顺序累加, 则自 a_2 之后所有元素均被 a_1"吃掉", 累加结果为 $\sum_{n=1}^{N} a_n = a_1$, 其误差为 $\sum_{n=2}^{N} a_n$. 如果 N 取值很大, 那么 $\sum_{n=2}^{N} a_n$ 很可能是个较大的数, 这样数值计算就会有一个较大的误差. 另一个例子是在数值运算的网格剖分中, 如果剖分选取的分辨率(步长)过小, 那么计算过程中它有可能会被"吃掉", 出现网格的某些节点无法生成的情况, 从而进一步影响数值运算精度.

目前, 已存在一些避免"大数吃小数"现象的方法[2,3]. 做 N 个实数的累加和时, 可调换相加顺序, 将绝对值较小的数先进行累加, 以避免"大数吃小数"现象, 但仍存在2种风险: 1.前面"小数"的累加值相比于后面"大数"仍然足够小到被"吃掉", 调换顺序起不到作用; 2.前面若干"小数"的累加值远大于后面的"大数", 从而将其"吃掉", 使运算结果更加不可预料. 可

收稿日期: 2013-07-04
基金项目: 国家自然科学基金资助项目(41375110, 41175069), 中国科学院大气物理研究所大气科学和地球流体力学数值模拟国家重点实验室(LASG)2012年度开放课题
第一作者: 曹 靖(1981—), 女, 讲师, 博士, 主要从事计算数学方面的研究.
通信作者: 李建平(1969—), 男, 研究员, 主要从事气候动力学、数值模拟与计算等方面的研究.

见，仅靠目前的方法并不能完全避免危及精度的现象发生. 为真正避免此类现象的发生，找到准确判断其发生的界限，就显得十分迫切和必要. 本研究针对两同符号数相加的问题，通过数值试验与理论分析，给出“大数吃小数”现象发生的界限，从而为实际运算提供理论指导.

1　机器双精度下“大数吃小数”的界限

1.1　两正数相加的数值试验

设 $A>0$ 为一“大数”，$C_A>0$ 为可被 A“吃掉”的最大“小数”. 在计算机双精度下，分别应用 Matlab、Fortran 以及 C 等 3 种计算机语言进行大量数值试验，发现 3 种语言计算出的 A 与 C_A 的关系都呈现出相同的规律. 以 Matlab 为例，给出 $A\in[0.1, 10]$时的部分试验结果，如图 1 所示. 由图 1 可见，A 所处的区域可被分为若干区间段，在每个区间段内，虽然 A 的取值不同，但被其吃掉的最大“小数”C_A 却十分接近，从而图形呈现阶梯形状. 特别地，在每个区间段内，$\lg A$ 与 $\lg(C_A/A)$之间呈现线性关系，因此，下面考虑应用这种分段线性关系，推导出 C_A 的计算公式.

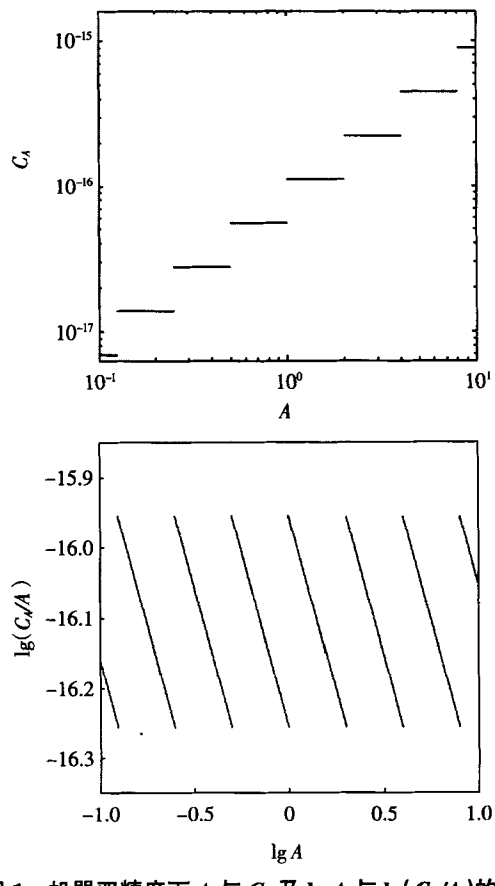

图 1　机器双精度下 A 与 C_A 及 $\lg A$ 与 $\lg(C_A/A)$的关系

Fig.1　A versus C_A and $\lg A$ versus $\lg(C_A/A)$, under double machine precision

1.2　两正数相加 $\lg(C_A/A)$关于 $\lg A$ 的分段线性拟合

为考察 $\lg(C_A/A)$关于 $\lg A$ 的分段线性关系，进行了大量数值试验. 首先，根据试验数据计算出 $\lg(C_A/A)$关于 $\lg A$ 斜率的变化规律，将 $\lg A$ 所属区间进行分段，即估计出试验所涉及的每段线性区间的左、右顶点坐标. 然后，分别计算出 $\lg A$ 每个分段区间的长度，以及区间内 $\lg(C_A/A)$关于 $\lg A$ 线性拟合的斜率. 部分试验结果见表 1.

表 1　机器双精度下，当 $A\in(10^{-2},10^{2})$时，$\lg(C_A/A)$与 $\lg A$ 线性关系

Tab.1　Linear relationship between $\lg(C_A/A)$ and $\lg A$, when $A\in(10^{-2},10^{2})$, under double machine precision

区间左顶点坐标 $(\lg A, \lg(C_A/A))$	区间右顶点坐标 $(\lg A, \lg(C_A/A))$	区间长度	区间内线性拟合斜率
(-1.806,-15.955)	(-1.505,-16.256)	0.301	-1.00
(-1.505,-15.955)	(-1.204,-16.256)	0.301	-1.00
(-1.204,-15.955)	(-0.903,-16.256)	0.301	-1.00
(-0.903,-15.955)	(-0.602,-16.256)	0.301	-1.00
(-0.602,-15.955)	(-0.301,-16.256)	0.301	-1.00
(-0.301,-15.955)	(0.000,-16.256)	0.301	-1.00
(0.000,-15.955)	(0.301,-16.256)	0.301	-1.00
(0.301,-15.955)	(0.602,-16.256)	0.301	-1.00
(0.602,-15.955)	(0.903,-16.256)	0.301	-1.00
(0.903,-15.955)	(1.204,-16.256)	0.301	-1.00
(1.204,-15.955)	(1.505,-16.256)	0.301	-1.00
(1.505,-15.955)	(1.806,-16.256)	0.301	-1.00

由表 1 可见，不同区间段之间满足十分相似的线性关系：$\lg A$ 分段区间长度都近似为 0.301，并且，每个区间段内，$\lg(C_A/A)$关于 $\lg A$ 线性拟合的斜率都接近于-1，且 $\lg(C_A/A)$均在[-16.256, -15.955]范围内单调递减. 由此便可估计 $\lg(C_A/A)$关于 $\lg A$ 的分段线性拟合函数.

首先，给出每个分段线性区间左顶点坐标. 由表 1 可见，$\lg A=0$ 为某一分段区间的左顶点，又因每个区间段长度均近似为 0.301，则可记“大数”A 属于第 k $(k\in\mathbf{Z})$段区间，其中 $k=\mathrm{floor}(\lg A/0.301)$，floor 表示向负无穷方向取整. 并且，此第 k 段区间的左顶点坐标应为$(0.301k, -15.955)$，$k\in\mathbf{Z}$. 然后，以 $\widetilde{C}_A$ 表示 C_A 的分段线性拟合近似值，若 A 属于第 k 段区间，结合上述试验中区间段内 $\lg(C_A/A)$关于 $\lg A$ 线性拟合的斜率以及左顶点坐标，得此区间段内 $\lg(\widetilde{C}_A/A)$关于 $\lg A$ 的线性拟合函数为

$$\lg(\widetilde{C}_A/A)=-(\lg A-0.301k)-15.955,\quad k=\mathrm{floor}(\lg A/0.301) \tag{1}$$

为验证式(1)的拟合效果，随机选取不同的“大数”A，将数值试验得到的 C_A 值与由式(1)计算出的拟合值 $\tilde{C}_A$ 进行比较，见表 2. 表 2 数据说明式(1)具有良好的拟合效果.

表 2 机器双精度下 $\tilde{C}_A$ 与 C_A 比较

Tab.2 Comparison between $\tilde{C}_A$ and C_A for random, under double machine precision

A	数值试验结果 C_A	拟合值 $\tilde{C}_A$
1.00	1.109×10^{-16}	1.109×10^{-16}
16.13	1.776×10^{-15}	1.774×10^{-15}
282.60	2.839×10^{-14}	2.838×10^{-14}
3 596.00	2.269×10^{-13}	2.270×10^{-13}
7.92×10^{6}	4.653×10^{-10}	4.645×10^{-10}
1.00×10^{-2}	8.670×10^{-19}	8.670×10^{-19}
3.45×10^{-3}	2.167×10^{-19}	2.168×10^{-19}
5.88×10^{-5}	3.383×10^{-21}	3.388×10^{-21}

1.3 两负数相加情况

设 $A' < 0$ 为一“大数”，应寻找 C'_A 为全体负数中可被 A' 吃掉的最小的“小数”，同时也是绝对值最大的“小数”. 大量数值试验结果表明，当 $|A'| = A$ 时，总有 $|C'_A| = C_A$. 因此，可将上述两正数相加情况的分段线性拟合结果直接推广至两负数相加情况. 令 $\tilde{C}'_A$ 为 C'_A 的拟合值，则结合式(1)可得

$$\lg|C'_A/A'| = -(\lg|A'| - 0.301k) - 15.955,$$
$$k = \mathrm{floor}(\lg|A'|/0.301) \tag{2}$$

1.4 双精度下“大数吃小数”界限的近似公式

结合式(1)与式(2)，可得机器双精度下两同号数相加时，“大数吃小数”现象发生界限的统一近似公式.

结论 1 在机器双精度下进行运算，对于任意符号相同的两实数 A 与 B，当且仅当 $|B| \leqslant C_A$ 时，机器运算结果为 $A + B = A$，其中

$$C_A \approx |A| \cdot 10^{-(\lg|A| - 0.301k) - 15.955},\ k = \mathrm{floor}(\lg|A|/0.301) \tag{3}$$

2 机器单精度下“大数吃小数”的界限

经与双精度类似的过程分析，可得机器单精度下“大数吃小数”界限的统一近似公式.

结论 2 在机器单精度下进行运算，对于任意符号相同的两实数 a 与 b，当且仅当 $|b| \leqslant c_a$ 时，机器运算结果为 $a + b = a$，其中

$$c_a \approx |a| \cdot 10^{-(\lg|a| - 0.301k) - 7.225},\ k = \mathrm{floor}(\lg|a|/0.301) \tag{4}$$

表 3 给出对于随机选出的不同“大数”a，可被其“吃掉”的最大“小数”c_a 的数值试验值与由式(4)计算出的近似值 $\tilde{c}_a$ 的比较结果，这些结果说明式(4)有良好的估计效果.

表 3 机器单精度下 c_a 试验值与式(4)近似值 $\tilde{c}_a$ 比较结果

Tab.3 Comparison between c_a and $\tilde{c}_a$ for random, under single machine precision

a	数值试验结果 c_a	拟合值 $\tilde{c}_a$
2.00	1.191×10^{-7}	1.191×10^{-7}
35.16	1.906×10^{-6}	1.905×10^{-6}
789.24	3.049×10^{-5}	3.048×10^{-5}
8 981.00	4.879×10^{-4}	4.875×10^{-4}
3.21×10^{5}	1.561×10^{-2}	1.560×10^{-2}
1.00×10^{-3}	5.808×10^{-11}	5.821×10^{-11}
9.28×10^{-4}	2.908×10^{-11}	2.911×10^{-11}
1.47×10^{-5}	4.542×10^{-13}	4.550×10^{-13}

3 任意机器精度“大数吃小数”界限的普适近似公式

公式(3)与公式(4)非常相似，只有指数上的一个常系数不同，单精度为 7.225，双精度为 15.955. 若记所选机器精度的二进制有效位数为 n，则机器单双精度下 n 分别为 24 和 53[4]. 注意到 $24\lg 2 \approx 7.224\ 7$，$53\lg 2 \approx 15.954\ 6$，也就是说，此常系数可由 n 近似推算出来，为 $n\lg 2$，由此推测 C_A 与 c_a 的近似表达式与 n 有关. 因此，在具有 n 位二进制有效数字的机器下进行运算，将被“大数”a“吃掉”的“小数”界限记为 $c_{a,n}$，结合结论 1 和结论 2，得任意机器精度下“大数吃小数”界限的普适近似公式.

结论 3 在具有 n 位二进制有效数字的机器精度下进行运算，对于任意符号相同的两实数 a 与 b，当且仅当 $|b| \leqslant c_{a,n}$ 时，机器运算结果为 $a + b = a$，其中

$$c_{a,n} = |a| \cdot 10^{-(\lg|a| - 0.301k) - n\lg 2},\ k = \mathrm{floor}(\lg|a|/0.301) \tag{5}$$

4 结论

通过数值试验，为数值运算中避免“大数吃小数”现象给出了界限，结论 3 适用于任意机器精度和任意计算机语言，可方便应用于实际的数值计算中. 下一步考虑利用二进制对位相加过程[5]，对于“大数吃小数”问题进行理论分析，并给出更为严格的理论界限.

另外，关于“大数吃小数”现象，本文只讨论了同号的情况，可应用类似的方法研究两异号数相加的情况，找到“大数吃小数”现象发生的统一界限公式.

参考文献：

[1] 刘目楼，汪卉琴. 数值分析[M]. 北京：冶金工业出版社，2005.
[2] 李庆扬，王能超，易大义. 数值分析[M]. 北京：清华大学出版社，2008.
[3] 徐士良. 计算机常用算法[M]. 北京：清华大学出版社，2005.
[4] 朱亚超. 基于 IEEE754 的浮点数存储格式分析研究[J]. 计算机与信息技术，2006(9)：50-52.
[5] GOLDBERG D. Computer Arithmetic[M]. Amsterdam: Elsevier Science, 2003.

(责任编校 马新光)

同符号数相加“大数吃小数”的界限:数值试验

作者： 曹靖， 李建平， CAO Jing， LI Jianping
作者单位： 曹靖,CAO Jing(中国科学院大气物理研究所,北京100029;中国科学院研究生院,北京100029;天津理工大学理学院,天津300384)， 李建平,LI Jianping(中国科学院大气物理研究所,北京100029;中国科学院研究生院,北京100029)

刊名： 天津师范大学学报（自然科学版） PKU
英文刊名： Journal of Tianjin Normal University(Natural Science Edition)
年，卷(期)： 2014,34(2)