第3卷 第2期 控制理论与应用 Vol. 3, No. 2
1986年6月 CONTROL THEORY AND APPLICATIONS June, 1986

计算矩阵指数函数的最佳计算参数选择

朱维彰 阴小军
（西安工业学院）

摘 要

收缩乘方法是计算矩阵指数函数的好方法．本文给出了这种方法新的相对误差上界和选择最佳计算参数的有效算法．

矩阵指数函数 e^A（$A-n\times n$ 实矩阵）的计算在系统数字仿真以及数值分析中均是重要的课题．〔1〕是这方面的重要综述．它把计算 e^A 的方法归纳为十九类，分别作了分析和评价．其中较好的方法之一是 Pade 收缩乘方法（Scaling and Squaring）和Taylor收缩乘方法．〔1〕给出了这两种方法的相对误差上界公式，原则上表明可以事先通过选择适当的计算参数来控制截断误差．

本文改进了〔1〕的相对误差上界公式，给出了便于编入计算机程序的最佳计算参数选择方法．

（一）相对误差上界公式的改进

〔1〕给出 Pade 收缩乘方法的相对误差上界公式为

如果
$$\|A\|/2^j\leq\frac{1}{2},\tag{1}$$

则
$$\frac{\|R_q(A/2^j)^{2^j}-e^A\|}{\|e^A\|}\leq\|E\|e^{\|E\|},\tag{2}$$

$$\frac{\|E\|}{\|A\|}\leq f(q,\ j),\tag{3}$$

其中
$$R_q(A)\triangleq D(-A)^{-1}D(A),$$

$$D(A)\triangleq\sum_{i=0}^{q}\frac{(2q-i)!\ q!}{(2q)!\ i!\ (q-i)!}A^i,$$

$$e^{A+E}\triangleq R_q(A/2^j)^{2^j},\tag{4}$$

本文于1984年6月25日收到．1984年12月5日收到修改稿．

$$f(q,\ j)\triangleq 8\left(\frac{\|A\|}{2^j}\right)^{2q}\frac{q!^2}{(2q)!\ (2q+1)!}. \tag{5}$$

定理 1　若 $\|E\|\leq 1$，则

$$\frac{\|R_q(A/2^j)^{2^j}-e^A\|}{\|e^A\|}\leq\|E\|(1+(e-2)\|E\|).$$

证　由（4）、$AE=EA$[1] 及条件 $\|E\|\leq 1$，可得

$$\frac{\|R_q(A/2^j)^{2^j}-e^A\|}{\|e^A\|}=\frac{\|e^A(e^E-I)\|}{\|e^A\|}\leq\|e^E-I\|$$

$$\leq\sum_{i=1}^{\infty}\frac{\|E\|^i}{i!}=\|E\|\left(1+\|E\|\left(\frac{1}{2!}+\frac{\|E\|}{2!}+\frac{\|E\|^2}{4!}+\cdots\right)\right)$$

$$\leq\|E\|\left(1+\|E\|\left(\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+\cdots\right)\right)=\|E\|(1+\|E\|(e-2)).$$ 证毕.

容易证明

$$\|E\|(1+\|E\|(e-2))<\|E\|e^{\|E\|},\qquad(\|E\|>0).$$

设 δ 为所希望的相对误差，求出满足不等式

$$x(1+(e-2)x)\leq\delta \tag{6}$$

的最大 x，记为 x_m，易得

$$x_m=(\sqrt{1+4(e-2)\delta}-1)/(2(e-2)). \tag{7}$$

若 δ 很小，则

$$x_m\cong\delta(1-\delta(e-2)). \tag{8}$$

选择适当的 $(q,\ j)$，使得

$$f(q,\ j)\leq x_m/\|A\|. \tag{9}$$

由（3）可得　$\|E\|\leq x_m$，即有

$$\|E\|(1+(e-2)\|E\|)\leq\delta.$$

故通过选择 $(q,\ j)$ 可以控制 Pade 方法中由于 q 取有限值所造成的截断误差.

若利用〔1〕的结果（(2)式）求 x_m，则需解超越方程 $x_m\exp(x_m)=\varepsilon$，所以定理 1，不但给出了更小的相对误差上界，而且还避免了解超越方程的困难. 定理 1 虽然引入了条件 $\|E\|\leq 1$，但只要 $\delta\leq 1$，由（6）可得 $x_m<1$，而选择计算参数过程中，又将使 $\|E\|\leq x_m$，所以实际上只要求 $\varepsilon\leq 1$，这在实际问题中总是满足的.

对于 Taylor 收缩乘方法，定理 1 结果同样适用，即

若 $\|A\|/2^j\leq\frac{1}{2}$，$\|E\|\leq 1$，则有

$$\frac{\|T_k(A/2^j)^{2^j}-e^A\|}{\|e^A\|}\leq\|E\|(1+(e-2)\|E\|)$$

$$\frac{\|E\|}{\|A\|}\leq 8\left(\frac{\|A\|}{2^j}\right)^k\frac{1}{(k+1)!},\qquad(10)$$

其中 $$T_k(A)\triangleq\sum_{i=0}^{k}A^i/i!\ ,$$

$$e^{A+E}\triangleq T_k(A/2^j)^{2^j}.$$

*〔1〕中此式有印刷错误，此处已更正。

（二）选择最佳计算参数（q，j）（或(k，j)）的方法

计算 $R_q(A/2^j)^{2^j}$ 的代价为 $(q+j+1/3)n^3$ “节拍”[1]，所以在满足(1)、(9)条件下，若 (q^*，j^*)使得($q+j$) 为最小，则 (q^*，j^*)是使计算代价最小的最佳参数.

由（5）可得

$$f(q-1,\ j+1)=N_j(q)f(q,\ j),\quad(q\geq 1)$$

其中

$$N_j(q)=\frac{16(2q-1)(2q+1)}{\|A\|^2}2^{2(j-q)}.$$

设 j_0 为使（1）成立的最小 j，$\varepsilon\triangleq x_m/\|A\|$，可以证明有下面结论：

结论 a　当$1\leq q\leq 6$时，若$f(q,j_0)>\varepsilon$，则有　$f(q-m,\ j_0+m)>\varepsilon$，$m=1,\ 2,\cdots,q$.

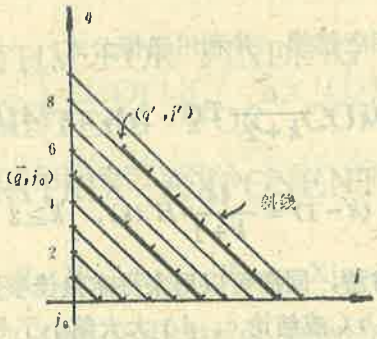
结论 b　当 $q\geq 1$，$j\geq j_0$ 时，若 $N_j(q)\geq 1$，$f(q,\ j)>\varepsilon$，则 $f(q-m,\ j+m)>\varepsilon$，$m=1,\ 2,\ \cdots,\ q$.

在 q—j 平面上，称斜率为 -1 的直线为斜线. 在 $q\geq 0$、$j\geq j_0$ 区域，记 $j=j_0$ 直线上的点为($\overline{q}$，j_0)，则过($\overline{q}$，j_0) 斜线上的点均有 $q+j=\overline{q}+j_0$. 结论 a 表明，当 $1\leq q\leq 6$，若某点($\overline{q}$，j_0) 使 $f(\overline{q},\ j_0)>\varepsilon$，则过该点斜线上的所有点都有 $f(q,\ j)>\varepsilon$. 结论 b 表明，若某点 (q'，j') 使 $f(q',\ j')>\varepsilon$，而且 $N_{j'}(q')\geq 1$，则过该点斜线上所有 $j>j'$ 的点都有 $f(q,\ j)>\varepsilon$. 如图所示.

据此，用下面步骤可以得到最佳计算参数 (q^*，j^*).

1）　对于给定的 A、δ 由（7）或（8）计算 x_m 及 $\varepsilon=x_m/\|A\|$；

2）　取 j_0 为大于 $\ln\|A\|/\ln 2+1$ 的最小正整数（包括 0），



图中粗线上的点都有 $f(q, j)>\varepsilon$

3） 依次取 $q=\overline{q}=0, 1, 2, \cdots \leq 6$, $j=j_0$，计算 $f(q, j)$，一旦有 $f(q, j)\leq\varepsilon$，执行 6，否则执行 4；

4） 将 $\overline{q}+1$ 赋给 $\overline{q}$；

5） 依次取 $m=0, 1, 2, \cdots, \overline{q}$，$q=\overline{q}-m$，$j=j_0+m$（即沿过$(\overline{q}, j_0)$斜线上取点），计算 $f(q, j)$ 及 $N_j(q)$，直到出现

（i） $f(q, j)\leq\varepsilon$，执行 6，

（ii） $f(q, j)>\varepsilon$，$N_j(q)\geq 1$，执行 4；

6） 最佳参数 $q^*=q$，$j^*=j$.

计算过程中，可应用下面递推公式

$$f(\overline{q}+1, j_0)=\frac{c}{(2\overline{q}+1)(2\overline{q}+3)}f(\overline{q}, j_0),$$

其中 $c=\|A\|^2/2^{2(j_0+1)}$

$$f(q-1, j+1)=N_j(q)f(q, j),\quad q\geq 1,$$

$$N_{j+1}(q-1)=16\frac{2q-3}{2q+1}N_j(q),\quad q\geq 2.$$

对于 Taylor 收缩乘方法，计算代价为 $(k+j-1)n^3$ “节拍”[1]。类似地，

令
$$T(k, j)\triangleq 8\left(\frac{\|A\|}{2^j}\right)^k\frac{1}{(k+1)!},$$

$$H_j(k)\triangleq 2^{(j-k+1)}\frac{k+1}{\|A\|},$$

则有 $T(k-1, j+1)=H_j(k)T(k, j)$，$k\geq 1$。可以证明有下面结果：

结论 c 当 $1\leq k\leq 4$ 时，若 $T(k, j_0)>\varepsilon$ 则 $T(k-m, j_0+m)>\varepsilon$，$m=1, 2, \cdots, k$。

结论 d 当 $k\geq 1$，$j\geq j_0$ 时，若 $T(k, j)>\varepsilon$，$H_j(k)\geq 1$，则 $T(k-m, j+m)>\varepsilon$，

$m=1,2,\cdots,k$。

只要将有关式子及量作相应替换，并利用递推公式

$$T(k+1,\ j_0)=\frac{c}{k+2}T(k,\ j_0),\ c=\|A\|/2^{j_0}$$

$$H_{j+1}(k-1)=\frac{4k}{k+1}H_j(k),\quad k\geq 2$$

上述寻求最佳 $(q,\ j)$ 的方法，同样可以用来寻求最佳参数$(k^*,\ j^*)$。

由于应用了结论 a、结论 b（或结论 c、d）大大缩小了寻优区域，各递推公式的使用，又使计算量减少，所以上述选择最佳计算参数的计算量并不大。用 Taylor、Pade 收缩乘方法计算 e^A 的计算代价与 A 的维数 n 的立方成正比，而且较小的$(q,\ j)$或$(k,\ j)$对于减少计算机舍入误差也有好处，所以花费一点代价来选择最佳计算参数，仍是合适的，对于较大的 n 更是如此。

对于给定的 $\|A\|$ 及 ε，用上面的方法可得下面最佳计算参数表，表中每格上面一行为$(q^*,\ j^*)$，下面一行为$(k^*,\ j^*)$。此表纠正了〔1〕中同样表中的某些错误。

$\|A\|$ \ ε	10^{-3}	10^{-6}	10^{-9}	10^{-12}	10^{-15}
10^{-2}	(1,0) (2,0)	(2,0) (3,0)	(2,0) (4,0)	(3,0) (6,0)	(3,0) (7,0)
10^{-1}	(2,0) (3,0)	(3,0) (5,0)	(3,0) (7,0)	(4,0) (8,0)	(5,0) (8,1)
10^{0}	(2,1) (5,1)	(4,1) (6,2)	(5,1) (8,2)	(5,1) (7,4)	(6,1) (10,3)
10^{1}	(2,5) (4,5)	(3,5) (7,5)	(4,5) (7,6)	(5,5) (9,6)	(6,5) (9,7)
10^{2}	(2,8) (5,8)	(3,8) (7,8)	(4,8) (9,8)	(5,8) (9,9)	(6,8) (8,11)
10^{3}		(4,11) (6,12)	(5,11) (8,12)	(5,11) (7,14)	(6,11) (10,13)

参考文献

〔1〕 Cleve Moler and Charles Van Loan, Ninteen Dubious Ways to Compute the Exponential of a Matrix, SIAM Review, **20**, 4, (1978).

AN ALGORITHM FOR FINDING THE OPTIMAL PARAMETERS FOR COMPUTING MATRIX EXPONENTIAL

Zu Weizhang, Yiang Xiaojun

(Xian Institute of Industry)

Abstract

The scaling and squaring method is very available for computing matrix exponential. New relative error upper-bound and an effective algorithm for finding the optimal parameters of the method is proposed in this paper.

系统与控制的数学理论和方法的发展讨论会于1986年5月在上海举行

系统与控制的数学理论和方法的发展讨论会于1986年5月9日至11日在上海复旦大学举行。会议主要议题是：

一、对系统与控制的数学理论与方法的国内外发展动向进行了交流。与会代表深切地感受到，随着科学技术的发展，一个以系统为研究对象，渗透于社会科学与自然科学之中，以定性研究与定量研究相结合的系统理论与控制理论，对于决策科学化和管理现代化等将产生重大的影响。历史和实践证明，系统理论与控制理论的发展必须得到现代数学的支持，而现代数学的发展也可以从它们得到模型与源泉。

二、探讨了系统与控制的数学理论的发展规划；

三、提出了关于人才培养规划和交流的建议；

四、讨论了有关教学、教材的建设问题。

为此会议建议：由国家教委会组织成立“系统科学与控制科学类的教材编审委员会”；在适当的时候召开一次系统科学与控制科学类专业的教学讨论会；建议国务院学位委员会在理学中成立平行于数学、物理等学科的“系统科学”学科组，以便在我国设立“系统科学”的硕士、博士学位点。

参加会议的十人代表，有高校中从事系统科学与控制理论研究的数学工作者：教授：李训经、陈祖浩、贺建勋、刘永清，副教授及正副系主任：王翼、胡启迪、吴占生、孙莱祥；中国科学院系统科学研究所所长助理陈翰馥研究员、航天工业部系统工程研究中心副主任、710所副所长于景元高级工程师。国家教委会高教一司徐沪生同志也列席了会议。

（永　清）