用 E 方法与高精度计算解线性方程组*

穆　默

（中国科学院计算中心）

SOLVING LINEAR SYSTEMS WITH E-METHOD AND HIGH ACCURACY ARITHMETIC

Mu Mo

(Computing Center, Academia Sinica)

Abstract

This paper deals with solving linear algebraic systems with the so-called E-method and high accuracy arithmetic. An algorithm with a new kind of stopping criteria is recommended and an error analysis is also contained.

误差分析一直是数值计算中的一个重要的基本问题. Wilkinson 提出的向后误差分析方法虽然能从理论上分析算法的数值稳定性，并给出误差的一些先验估计，但还不能解决实际计算解的误差估计问题. 六十年代发展起来的区间方法，基于用一个区间来表示有误差的数的想法. 理论上可以严格的得到解的上下界，并且实现了误差分析的自动化，但由于它是从严格的界出发，没有考虑大量计算中误差的随机相消的因素，常导致实际得到的误差界大得失去意义，因而阻碍了区间方法在数值计算中的应用和推广. 然而，区间方法在误差分析方面确有其显著的优点，近年来出现的 E 方法就是一类求解不动点方程 $x = f(x)$ 的区间方法[1-5]. E 方法的名称来源于 Existenz（存在性），Eindentigkeit（唯一性），Einschließung（包含性）三个德文词头. 它们刻划了该方法的特征，即在求出一个包含不动点在内的解区间的同时，还能数值上自动验证原问题解的存在与唯一性. E 方法与 [6] 提出的高精度计算技术结合起来，有效地克服了传统的区间方法在误差积累方面的缺点，取得了令人满意的数值效果. 例如，求 21×21 阶 Hilbert 矩阵的逆这样高度病态的问题（条件数约为 10^{30}），在 16 位（十进制计算机上的计算结果，各分量基本准确到第 15～16 位[5]. 因此，这一方法目前在国际上已开始受到重视. IBM 公司于 1983 年推出了相应的软件包 ACRITH[5].

本文就用 E 方法与高精度计算解线代数方程组，讨论迭代的收敛性，终止准则及误差估计的问题.

设区间 $X = [\underline{X}, \overline{X}]$. 定义 $m(X) = (\underline{X} + \overline{X})/2$, $d(X) = \overline{X} - \underline{X}$, $|X| = \max\{|\underline{X}|, |\overline{X}|\}$. 以 IR^n, $IR^{n\times n}$ 分别记全体 n 维区间向量及 $n \times n$ 阶区间矩阵. $\boxed{*}$, $* \in$

* 1986 年 6 月 23 日收到.

$\{+,-,\times,1,\cdot\}$ 表示相应的区间运算[6]. E 方法理论上基于下述引理.

引理 1[1] 设 $f: R^n\to R^n$ 为连续函数,映射 $F: IR^n\to IR^n$ 满足

$$f(x)\in F(X),\ \forall x\in X.$$

若存在 $\tilde{X}$, 使得

$$F(\tilde{X})\subseteq\tilde{X},$$

则 f 在 $F(\tilde{X})$ 中至少有一个不动点.

设求解线性方程组

$$Ax=b, \tag{1}$$

其中 $A\in R^{n\times n}$, $b\in R^n$. 若令

$$f(x)=x+R(b-Ax), \tag{2}$$

只要 $\det(R)\neq 0$, (1) 就与不动点方程 $x=f(x)$ 等价. 令

$$F(X)=B\boxdot X\boxplus C,\ \forall X\in \mathbb{R}^n, \tag{3}$$

其中 $B=I-RA$, $C=Rb$. f 和 F 显然满足引理 1 的条件.

引理 2[4] 设 $\mathbf{B}\in IR^{n\times n}$, $\mathbf{C}\in IR^n$, $X\in \mathbb{R}^n$. 如

$$\mathbf{B}\boxdot X\boxplus\mathbf{C}\subsetneqq X,$$

其中"$\subsetneqq$"按分量理解,则对谱半径,有

$$\rho(B)\leqslant\rho(|B|)\leqslant\rho(|\mathbf{B}|)<1,\ \forall B\in\mathbf{B}.$$

于是,构造区间迭代

$$X^{k+1}=F(X^k),\ k=0,1,2,\cdots,\ \forall X^0\in\mathbb{R}^n. \tag{4}$$

假如迭代到某步,成立

$$X^{k+1}\subsetneqq X^k, \tag{5}$$

则计算可终止. 由引理 1,2 断言: (1) 的解存在,唯一且属于 X^{k+1}.

对 (4) 的收敛性,我们证明了更一般的结论[7].

定理 1 设区间迭代

$$X^{k+1}=\mathbf{B}\boxdot X^k\boxplus\mathbf{C},\ k=0,1,2,\cdots,\ \forall X^0\in\mathbf{R}^n,$$

其中 $\mathbf{B}\in\mathbf{R}^{n\times n}$, $\mathbf{C}\in\mathbf{R}^n$, 则其收敛的充要条件为

$$\rho(|\mathbf{B}|)<1.$$

后来发现, [8] (1983) 也给出了同样的结果,并且证明的方法也类似,故这里不再证明.

注: 需要指出,在定理 1 中, 若区间迭代为特殊的 (4), 则极限区间 X^* 退化为普通的向量. 这是因为,由 $X^*=B\boxdot X^*\boxplus C$, 有 $d(X^*)=d(B\boxdot X^*)\leqslant|B|d(X^*)$. 再由 $\rho(|B|)<1$, 知 $d(X^*)=0$. 因此, X^* 就是 f 的不动点 x^*.

但是,容易举例说明, 即便 (4) 收敛, 甚至于已有 $x^*\in X^0$, (5) 也可能永远不会成立. 这样,仅以 (5) 作为迭代终止的判别条件就显得不够理想. 研究 (5) 是困难的[4]. 我们证明如下两个定理:

定理 2 若存在 K, 使得

$$X^K\subseteq\mathrm{Int}X^0, \tag{6}$$

则 $\rho(B)<1$, f 有唯一的不动点 x^*, 且 $x^*\in\mathrm{Int}X^0$, 其中 $\mathrm{Int}X^0$ 表示 X^0 的内部, 而, $x^*\in X^k$, $k=0,1,2,\cdots$.

证明　由 $X^K = F^K(X^0)$，有

$$F^K(X^0) \subseteq \mathrm{Int} X^0. \tag{7}$$

定义 g: $\mathbf{R}^n \to \mathbf{R}^n$ 为 $g(x) = f^K(x)$，则

$$g(x) = B^K x + \sum_{i=0}^{K-1} B^i C. \tag{8}$$

从而定义 G: $\mathbf{R}^n \to \mathbf{R}^n$ 为

$$G(X) = B^K \boxdot X \boxplus \left(\sum_{i=0}^{K-1} B^i C\right). \tag{9}$$

显然，g 与 G 满足引理 1 的条件，并且有

$$G(X) \subseteq F^K(X),\ \forall X \in \mathbf{R}^n, \tag{10}$$

从而

$$G(X^0) \subsetneqq X^0. \tag{11}$$

由引理 1，g 至少有一个不动点落在 X^0 中．再由引理 2，$\rho(B) = \rho(B^K)^{1/K} < 1$，从而 f 和 g 都有唯一的不动点．由于 f 的不动点就是 g 的不动点，故 $x^* \in X^0$．进一步，$x^* = f(x^*) \in F(X^0) \subseteq \mathrm{Int} X^0$．类似地可得 $x^* \in X^k$，$k = 0, 1, 2, \cdots$．证毕．

定理 3　设 $\rho(|B|) < 1$，$\forall X^0 \in \mathbf{R}^n$，下述三种情况有且仅有之一出现．

(i. a) $\exists K$，使 $X^K \subseteq \mathrm{Int} X^0$;

(ii. a) $\exists K$，使 $X^K \cap X^0 = \phi$;

(iii. a) $\lim\limits_{k\to\infty} m(X_i^k) = \partial X_i^0$，$i \in M$ (即 $\lim\limits_{k\to\infty} m(X^k) \in \partial X^0$)，

其中 ∂X^0 表示 X^0 的边界，M 是指标集 $\{1, 2, \cdots, n\}$ 的一个子集．这三种情况分别对应于

(i. b) $x^* \in \mathrm{Int} X^0$;

(ii. b) $x^* \bar{\in} X^0$;

(iii. b) $x^* \in \partial X^0$.

证明　$\forall X^0 \in \mathbf{R}^n$，(i. b)，(ii. b)，(iii. b) 有且仅有之一出现．由假设 $\rho(|B|) < 1$，利用定理 1，$\lim\limits_{k\to\infty} X^k = x^*$，于是，由 (i. b)，(ii. b)，(iii. b) 分别可推出 (i. a)，(ii. a)，(iii. a)．

反之，(i. a) ⇒ (i. b) 是定理 2 的一个结论．

(ii. a) ⇒ (ii. b)．反证法．若 $x^* \in X^0$，则 $x^* = f(x^*) \in F(X^0) = X^1$，依此类推，$x^* \in X^k$，$k = 0, 1, 2, \cdots$．从而 $x^* \in X^k \cap X^0$，$k = 0, 1, 2, \cdots$，与 (ii. a) 矛盾．

(iii. a) ⇒ (iii. b)．由假设，$\rho(|B|) < 1$ (事实上只要 $\rho(B) < 1$)，注意到 $m(X^{k+1}) = Bm(X^k) + C$，有 $\lim\limits_{k\to\infty} m(X^k) = x^*$，便得证．证毕．

由定理 2,3，我们建议下述更实际和更有效的终止准则与重新开始技巧．

算法

开始　① 适当选取初始 X^0;

$k := 0$;

② $X^{old} := X^0$;

③ $X^{new}:=F(X^{old})$

$k:=k+1$;

若 $X^{new}\subsetneqq X^{old}$, 则 $X^{old}:=X^{new}$; 转④;否则

若 $X^{new}\subseteq \mathrm{Int}X^0$, 则 $X^{old}:=X^{new}$; 转④;否则

若 $X^{new}\cap X^0=\phi$, 则 $X^0:=X^{new}$; 转②;否则

若 $m(X^{new})$ 很靠近 ∂X^0, 则 $X^0:=$ 适当放大的 X^0; 转②;否则

若 $k=K$ (控制常数),则打印失败;转⑤;否则

$X^{old}:=X^{new}$ 转③;

④ $X^{new}:=F(X^{old})$

若满足给定精度或已无改进,则打印 X^{new}; 转⑤;否则

$X^{old}:=X^{new}$; 转④.

⑤ 停机.

结束

由引理 1,2 及定理 2, 便知当计算过程成功终止时, $\rho(B)<1$, 相应的方程组的解存在,唯一且 $x^*\in X^{new}$.

一般说来,由于舍入误差的影响,在迭代过程中,实际计算的区间总大于准确区间.另一方面, $\lim\limits_{k\to\infty}m(X^k)=x^*$. 因而, x^* 必于某步落入某迭代区间. 故从实际计算的角度,由定理 3, 只要 $\rho(|B|)<1$, 即 (4) 理论上收敛. 计算过程一般都能成功地终止. B 在实际计算中通常是一个剩余量(即先用普通的浮点计算求出一个近似逆 R),故 $\rho(|B|)$ 一般比较小.

推论 若有 K,使 $X^{K+1}\subseteq X^K$, 则 $\rho(|B|)<1$; 或 $X^{K+1}=X^K$. 在后一情形, $x^*=m(X^K)$.

类似地,可以考虑迭代过程

$$X^{k+1}=F(X^k)\cap X^k,\quad k=0,1,2,\cdots,\quad \forall X^0\in \mathbf{R}^n. \tag{12}$$

定理 4 设 $\rho(|B|)<1$; 或 $\rho(|B|)=1$ 且 B 不可约, 则由 (12) 产生的序列具有下述性质:

a) $\exists K$, 使 $X^K=\phi$;

b) (i) $\lim\limits_{k\to\infty}X^k=X^*$;

(ii) $F(X^*)\supseteq X^*$;

(iii) $\lim\limits_{k\to\infty}m(X^k)=x^*$, 且 $x^*\in X^0$.

a) 出现当且仅当 $x^*\bar{\in} X^0$.

证明 若 $\rho(|B|)<1$, 由定理 3 立即可得. 以下设 $\rho(|B|)=1$ 且 B 不可约.

若 a) 不成立,由 (12) 产生一个单调的区间序列 $\{X^k\}$: $X^{k+1}\subseteq X^k$, $k=0,1,2,\cdots$, 于是 (3) 成立. 注意到 $X^{k+1}\subseteq F(X^k)$, 通过取极限便得 (ii). 由 (ii)

$$\begin{aligned} X^*\subseteq B\boxdot X^*\boxplus C &= B\boxdot(X^*\boxminus m(X^*))\boxplus B\boxdot m(X^*)\boxplus C\\ &=|B|\boxdot(X^*\boxminus m(X^*))\boxminus B\boxdot m(X^*)\boxplus C, \end{aligned} \tag{13}$$

若令 $z=\overline{X}^*-m(X^*)=m(X^*)-\underline{X}^*\geqslant 0$, 则由 (13) 可导出

$$\begin{aligned}\overline{X}^* &\leqslant C + Bm(X^*) + |B|z,\\ \underline{X}^* &\geqslant C + Bm(X^*) - |B|z,\end{aligned} \tag{14}$$

于是

$$|B|z - z \geqslant |m(X^*) - C - Bm(X^*)| \geqslant 0. \tag{15}$$

由于 $|B|$ 是非负不可约矩阵，$\rho(|B|) = 1$ 及 $|B|z \geqslant z$，利用[9]第二章第一部分的讨论,便知 $|B|z = z$. 再由 (15)，得 $m(X^*) = C + Bm(X^*)$. 若 $\det(I - B) \neq 0$，则

$$m(X^*) = x^* = (I - B)^{-1}C \in X^0.$$

从而 (iii) 得证. 类似于定理 2,3 的证明可得最后一个断言. 证毕.

下面估计舍入误差. 仍考虑 (4)，并假设 $\rho(|B|) < 1$ 及 $x^* \in X^0$，从而 $x^* \in X^k$，$k = 0, 1, 2, \cdots$. 利用最大精度的计算机算术[6],可以达到下述精度(其中 $X^{k+1} = \tilde{F}(X^k)$ 是实际计算结果，t 是尾数长度),即

$$X^{k+1} \subseteq [1 - 2^{1-t}, 1 + 2^{1-t}] \boxdot (B \boxdot X^k \boxplus C). \tag{16}$$

若记 $d^k = d(X^k)$，可验证

$$d^{k+1} \leqslant d(B \boxdot X^k \boxplus C) + 2 \cdot |B \boxdot X^k \boxplus| \cdot 2^{1-t}, \tag{17}$$

于是,

$$\begin{aligned}d^{k+1} &\leqslant d(B \boxdot X^k \boxplus C) + |B \boxdot X^k \boxplus C \boxminus x^*| \cdot 2^{2-t} + |x^*| \cdot 2^{2-t}\\ &\leqslant d(B \boxdot X^k \boxplus C)(1 + 2^{2-t}) + |x^*| \cdot 2^{2-t}\\ &= (1 + 2^{2-t})|B|d^k + |x^*| \cdot 2^{2-t},\end{aligned} \tag{18}$$

因此有

$$d^{k+1} \leqslant \{(1 + 2^{2-t})|B|\}^{k+1}d^0 + \sum_{i=0}^{k} \{(1 + 2^{2-t})|B|\}^i |x^*| 2^{2-t}. \tag{19}$$

若 $\rho(|B|) < \dfrac{1}{1 + 2^{2-t}}$, (19) 式的右端趋于 $\{I - (1 + 2^{2-t}) \cdot |B|\}^{-1}|x^*| \cdot 2^{2-t}$. 进一步,若 $\rho(|B|) \ll \dfrac{1}{1 + 2^{2-t}}$，则 $\{I - (1 + 2^{2-t})|B|\}^{-1} \approx I$.

定理 5　对迭代过程 (4)，若 $x^* \in X^0$，采用最大精度区间算术，则其宽度有估计式 (19). 若 $\rho(|B|) \ll \dfrac{1}{1 + 2^{2-t}}$，则对足够大的 k,

$$d_i^k \lessdot |x_i^*| \cdot 2^{2-t},\quad i = 1, 2, \cdots, n.$$

其中"$\lessdot$"表示近似地小于等于.

这与数值计算的结果很吻合. 因为实际计算中由于 $\rho(|B|)$ 一般很小,计算精度都非常高,通常可准确到计算机精度的最后 1～2 位.

本文是在黄鸿慈教授的精心指导下完成的,谨在此表示衷心感谢.

参　考　文　献

[1] S. M. Rump, E. Kaucher, Small Bounds for the Solution of Systems of Linear Equations. Computing, Suppl. 2, (1980), 115—164.

[2] E. Kaucher, S. M. Rump, Generalized Iteration Methods for Bounds of the Solution of Fixed Point Operator

Equations, Computing 24, (1980), 131—137.

[3] E. Kaucher, S. M. Rump, E-Methods for Fixed Point Equations $f(x)=x$. Computing 28(1982), 31—42.

[4] S. M. Rump, Solving Nonlinear Systems with Least Significant Bit Accuraey. Computing 29(1982), 183—200.

[5] High-Accuracy Arithmetic Subroutine Library General Information Manual. IBM. (1983).

[6] U. Kulish, W. Miranker, Computer Arithmetic in Theory and Practice. Academic Press (1981).

[7] 穆默，E 方法与高精度计算，硕士论文，中国科学院计算中心 (1984).

[8] G. Alefeld, J. Herzberger, Introduction to interval computations. Academic Press, (1983).

[9] R. S. Varga, Matrix Iterative Analysis. Prentice-Hall, Engle wood Cliffs. M. J. (1962).