

# 求解线性方程组的迭代法的统一

## ——二维迭代法\*



刘晓明 胡健栋

(北京邮电学院)

### 提 要

求解线性方程组  $Ax = b$  的迭代法有其独特的实用意义,但由于其收敛的问题而受到限制.本文导出了常用的雅各比法、高斯-塞德尔法和逐次超松弛法等统方法,称之为二维迭代法.并由此得到了从新的角度改进迭代法的收敛性和收敛速度的途径.理论分析和数值计算都表明该方法优于常用的迭代法.此方法在解大规模电路中,例如用于VLSI的模拟.

### 一、引 言

在电路和系统的分析和计算中,经常遇到线性方程组

$$Ax = b, \quad x, b \in R^n, \quad A \in R^{n \times n} \quad (1)$$

的求解问题.对于此类问题的求解有许多方法<sup>[1]</sup>,其中最常用的有 Gauss 消去法、LU 分解法等等.近年来,随着科学技术的发展,要处理的方程组规模越来越大,而其系数矩阵  $A$  却越来越稀疏,为了节省运算量和内存容量,通常都采用稀疏矩阵技术<sup>[2]</sup>.但这类算法需要进行方程的编序,且数据结构非常复杂,因之也要花费大量的机时.为解决这一矛盾,迭代法引起了人们的兴趣,因为迭代法具有如下几个突出的优点:(1)在整个迭代过程中, $A$  矩阵始终不变,故采用稀疏矩阵技术时,只用到静态存贮方式,无须编序,数据结构十分简单;(2)在每一步迭代中都得出变量的近似值,此性质对于某些实时处理系统特别有用.此外,在大系统的分裂法中,迭代法还有其独特的作用<sup>[3]</sup>.但迭代法也有如下几个主要缺点:一是收敛范围小,二是收敛速度慢,三是它不能直接求解在矩阵  $A$  的主对角线上有零元的方程组.由于这些问题一直未能得到较好的解决,使得迭代法的应用范围受到了极大的限制.多年来,人们设想了许多方法,试图克服迭代法的上述缺点,但多数工作都是围绕着后两个问题进行的,而对如何扩大迭代法的收敛范围则研究得较少,进展也不大.本文旨在在这方面作一些研究和探讨.文中主要做了以下几点工作:

- (1) 从参数拓展的概念<sup>[4]</sup>出发,导出了二维迭代法,并建立了几种具体的迭代格式.
- (2) 把二维迭代法的收敛性与所解系统的稳定性联系起来,证明了二维迭代法的收

\* 1984年11月10日收到,1985年2月25日修改定稿.

敛范围在理论上可达到迭代法所能应用的最大范围。

(3) 证明了常用的雅各比 (Jacobi) 法、高斯-塞德尔 (Gauss-Seidel) 法、逐步超松弛 (SOR) 法以及文献 [5] 中提出的一种迭代方法等都是二维迭代法的特例。同时, 阐述了二维迭代法与阻尼最小二乘法的关系。

(4) 定义了阻尼/步长因子, 讨论了调节该因子对迭代的收敛性的影响。

最后, 还给出了几个计算示例。

## 二、二维迭代的含义和迭代格式

二维迭代的概念是基于参数拓展法, 其主要思路是: 在方程组 (1) 中引入一参量  $t, t \in [t_0, t_M]$ , 并根据 (1) 式构造一个新的方程组

$$H(x, t) = 0, \quad (2)$$

其中  $H(x, t)$  满足如下要求: 在任意  $t_m \in [t_0, t_M]$  时,  $x(t_m)$  均可用迭代法求得; 而逐步变化  $t$  至  $t = t_M$  时, 有  $H(x, t_M) = Ax - b = 0$ , 并且在相邻  $t$  值下的  $x$  值之间具有一定的递推约束关系。显然, 只要满足以上条件, 我们就可以通过迭代法从 (2) 式得到 (1) 式的解, 即 (1) 式的求解化为一系列 (2) 式的求解。

现在的问题是如何构造出合乎要求的  $H(x, t)$  经过研究和提炼, 我们得出 (2) 式的一个可行方案为:

$$H(x, t) = C\dot{x} + Ax - b, \quad (3)$$

式中  $A$  是非奇异矩阵 (否则问题无意义), 且其主对角元均不小于零 (可等于零);  $t \in R^+$ ,  $t_M \rightarrow +\infty$ ;  $C = \text{diag}(C_1, C_2, \dots, C_n)$ , 且  $\forall i \in \{1, 2, \dots, n\}, C_i > 0$ 。

当我们用数值法求解 (3) 式时,  $C\dot{x}$  项的引入可以起到改变矩阵  $A$  的主对角元的作用。根据线性方程组的迭代求解法, 这样可以保证迭代的收敛性。与此同时,  $\dot{x}$  用多项式表示时, 多步积分公式决定了在相邻  $t$  值下的  $x$  值之间具有递推约束关系。

由于上述处理, 用数值法求解 (3) 式时, 整个过程可以分为两个交错的层次, 即在  $t$  方向上的迭代和在一定  $t$  值下的迭代。为了便于区别起见, 我们称前一迭代层次为递推迭代 (与  $t$  有关), 而后一迭代层次为递归迭代 (与  $t$  无关), 并把这一方法称为二维迭代法。

为证明这一方法的可行性, 先叙述两个定理:

**定理 1** 若矩阵  $C^{-1}A$  的所有特征值都在复平面的右半平面内, 则在 (3) 式中, 必有  $\lim_{t \rightarrow +\infty} \dot{x} = 0$ 。

**证明** (3) 式可化为如下形式:

$$\dot{x} = -C^{-1}Ax + C^{-1}b. \quad (4)$$

由微分方程理论可知, (4) 式的解析解为<sup>[6]</sup>:

$$x(t) = [x(t_0) - A^{-1}b]e^{[-C^{-1}A]t} + A^{-1}b. \quad (5)$$

由假定,  $C^{-1}A$  的所有特征值均在复平面的右半平面内, 此即表明:  $-C^{-1}A$  的所有特征值均具有负的实部, 故 (5) 式必然是渐近稳定的, 其稳定值为  $x = A^{-1}b$ , 于是  $\lim_{t \rightarrow \infty} \dot{x} = 0$

成立。(证毕)

**定理 2** 若矩阵  $\mathbf{A}$  的特征值都在复平面的右半平面内, 且  $\forall i \in \{1, 2, \dots, n\}$ ,  $C_i > 0$ , 则  $\mathbf{C}^{-1}\mathbf{A}$  的特征值也都在复平面的右半平面内, 反之亦然. 其中  $\mathbf{C} = \text{diag}(C_1, C_2, \dots, C_n)$ . 此定理可参照上面的方法予以证明, 这里从略.

由定理 1 和 2, 我们可以很自然地得到如下推论:

**推论 1** 若矩阵  $\mathbf{A}$  的特征值都在复平面的右半平面内, 则由 (3) 式所定义的方程组在  $t \rightarrow +\infty$  时与方程组 (1) 有相同的解.

显然, 上述推论完全概括了 (3) 式可行性的条件和结论.

至此, 我们已经把代数方程组 (1) 的求解问题化成了一个求微分方程组的稳态解的问题了. 此概念在文献 [7, 8] 中也曾应用过, 但本文的目的在于研究采用不同迭代方式求解方程 (3) 时所得结果的意义, 以及讨论如何改进迭代法的收敛性和收敛速度. 这是本文与文献 [7, 8] 的区别所在. 为简便起见, 下面所说的二维迭代法就是指求解 (3) 式所用的各种迭代方法, 并将推论 1 中的“矩阵  $\mathbf{A}$  的特征值都在复平面的右半平面内”这一条件称为“特征值条件”.

今假定 (3) 式中的矩阵  $\mathbf{A}$  已满足特征值条件. 以  $m$  代表递推推代码码,  $k$  代表递归迭代代码,  $i$  代表分量代码,  $\mathbf{x}_m$  代表  $\mathbf{x}(t_m)$ ,  $h$  代表  $t$  的步长, 则我们可得出几种迭代格式如下:

**格式 1** 用前向欧拉 (Euler) 公式代替 (3) 式中的微分项, 然后解所得代数方程组, 此法称为显 E 法, 其迭代式如下:

对于  $m = 1, 2, \dots; i = 1, 2, \dots, n$  计算

$$x_{m+1}^i = x_m^i + \tilde{d}_i^{-1} \left( b_i - \sum_{j=1}^n a_{ij} x_m^j \right) \quad (6)$$

式中  $\tilde{d}_i = C_i/h$ , 显然, 当取  $\tilde{d}_i = a_{ii}$  时, 显 E 法就是雅可比 (Jacobi) 法.

**格式 2** 顺序对 (3) 式中的各个方程用前向欧拉法求解, 并且把每个新求出的变量值立即用到求解下个变量的公式中去, 此法称为显 E-GS 法, 其迭代式如下:

对于  $m = 1, 2, \dots; k = 1, 2, \dots; i = 1, 2, \dots, n$  计算

$$x_{m+1, k+1}^i = x_m^i (1 - a_{ii}/\tilde{d}_i) + \frac{1}{\tilde{d}_i} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_{m+1, k+1}^j - \sum_{j=i+1}^n a_{ij} x_{m+1, k}^j \right), \quad (7)$$

式中  $\tilde{d}_i = C_i/h$ , 当  $\max_i |x_{m+1, k+1}^i - x_{m+1, k}^i| \leq \varepsilon_1$  时, 取  $x_{m+1} = x_{m+1, k+1}$ ,  $x_{m+2, 0} = x_{m+1}$ , 继续进行迭代, 直到  $\max_i |x_{m+1}^i - x_m^i| \leq \varepsilon_2$  时, 迭代终止. 在 (7) 式中, 取  $\tilde{d}_i = a_{ii}/\omega$  和  $k = 0$  (即在每个  $t_m$  点上只递归迭代一次) 时, (7) 式就是 SOR 法的迭代公式, 当  $\omega = 1$  时, (7) 式又变为高斯-塞德尔法.

**格式 3** 用后向欧拉公式代替 (3) 式中的微分项, 然后, 对所得之代数方程组用高斯-塞德尔法求解, 此法称为隐 E-GS 法, 其迭代式如下:

对于  $m = 1, 2, \dots; k = 1, 2, \dots; i = 1, 2, \dots, n$  计算

$$x_{m+1,k+1}^i = \frac{1}{a_{ii} + \tilde{d}_i} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_{m+1,k+1}^j - \sum_{j=i+1}^n x_{m+1,k}^j \right) + \frac{\tilde{d}_i}{a_{ii} + \tilde{d}_i} x_m^i \quad (8)$$

迭代终止条件同格式(2). 若取  $\tilde{d}_1 = \tilde{d}_2 = \cdots = \tilde{d}_n = \lambda$  和  $k \equiv 0$ , 则(8)式就是文献[5]中的公式(10).

**格式 4** 用二阶吉尔(Gear)公式  $\dot{x}_{m+1} = \frac{3}{2h} \left( x_{m+1} - \frac{4}{3} x_m + \frac{1}{3} x_{m-1} \right)$  替换(3)式中的微分项, 然后令  $\tilde{d}_i = 1.5C_i/h$ , 并用高斯-塞德尔法求解所得之代数方程组, 此种方法称为二阶 Gear-GS 法, 其迭代式如下:

对于  $m = 1, 2, \cdots; k = 1, 2, \cdots; i = 1, 2, \cdots, n$  计算

$$x_{m+1,k+1}^i = \frac{1}{a_{ii} + \tilde{d}_i} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_{m+1,k+1}^j - \sum_{j=i+1}^n a_{ij} x_{m+1,k}^j \right) + \frac{\tilde{d}_i}{a_{ii} + \tilde{d}_i} \left( \frac{4}{3} x_m - \frac{1}{3} x_{m-1} \right). \quad (9)$$

迭代终止的条件同格式 2 和 3.

仿照以上推导, 我们还可导出许多其它的二维迭代格式, 并可证明文献[5]中的公式(9)和(11)分别是隐 E-J 法和隐 E-SOR 法的特例. 限于篇幅, 在此不再详述.

最后, 我们想讨论一下二维迭代法与阻尼最小二乘法的关系.

线性问题的阻尼最小二乘法公式为:

$$x_{m+1} = x_m - (A^T A + W)^{-1} A^T (A x_m - b). \quad (10)$$

当  $W = A^T \tilde{D}$ ,  $A \in R^{n \times n}$  且  $A$  非奇异时, (10)式变为

$$x_{m+1} = [I - (A + \tilde{D})^{-1} A] x_m + (A + \tilde{D})^{-1} b,$$

应用谢尔曼-莫里森-伍德伯里(Sherman-Morrison-Woodbury)公式得

$$x_{m+1} = (I + \tilde{D}^{-1} A)^{-1} x_m + (A + \tilde{D})^{-1} b. \quad (11)$$

显然(11)式就是格式 3 中公式(8)的递推迭代的矢量形式. 由此可见, 对于一类特殊问题, 二维迭代与阻尼最小二乘法殊途同归. 因此, 我们把系数  $C_i$  称为阻尼系数. 又由于在各种迭代格式中,  $\tilde{d}_i$  均与  $C_i/h$  成比例, 而  $h$  是  $t$  的步长, 我们把  $\tilde{d}_i$  称为阻尼/步长因子.

### 三、 $\tilde{d}_i$ 的取值和迭代格式的收敛

上一节已经分析了常用的迭代法与二维迭代法的关系. 常用的迭代法可以看作二维迭代在  $t$  轴压缩成一点的情况.

常用迭代的收敛与迭代矩阵的谱半径有关. 以高斯-塞德尔法为例, 它的迭代矩阵的谱半径为  $\rho \leq \|(\mathbf{D} + \mathbf{L})^{-1} \mathbf{U}\|$ , 其中  $\mathbf{D}$  是矩阵  $\mathbf{A}$  的主对角阵,  $\mathbf{L}$  和  $\mathbf{U}$  分别是对角元为零的下和上三角阵. 迭代的收敛的充分条件是  $\rho < 1$ .

二维迭代的收敛决定于递归和递推迭代两个层次. 以格式 3 为例. 在递归迭代层

次, 相当于用高斯-塞德尔法求解方程组  $\mathbf{Ax} = \mathbf{b}$ , 其中  $\mathbf{A} = \mathbf{A} + \mathbf{D}$ . 由于  $C_i \in R^+$ ,  $h \in R^+$ , 故  $\tilde{d}_i \in R^+$ , 因此, 只要按照使矩阵  $\tilde{\mathbf{A}}$  成为强优对角阵的原则来选取  $\tilde{d}_i$  值 (亦即使  $a_{ii} + \tilde{d}_i > \sum_{j=1}^n |a_{ij}|, \forall i \in \{1, 2, \dots, n\}$ ), 则递归迭代的收敛是可以保证的. 在递推迭代层次, 方法的收敛决定于数值积分的方法和步长的选择. 在合理选择数值积分方法和步长, 收敛是可以保证的.

显然, 二维迭代法比常用的迭代法有较大的收敛范围. 只要按照使递归迭代收敛的准则选取  $\tilde{d}_i$ , 就可以保证二维迭代的收敛性.

其余格式的收敛性与  $\tilde{d}_i$  的关系可同样推导得出, 本文从略.

在各种迭代法中, 在迭代过程中变量  $x$  在空间  $R^n$  中的变化有如下几种情况:

- 情况 1 所有变量都单调地向一个方向变化, 且没有极限.
- 情况 2 所有变量都在精确解的邻域内摆动, 且摆幅越来越大.
- 情况 3 所有变量都在精确解的邻域内摆动, 且摆幅不变.
- 情况 4 所有变量都在精确解的邻域内摆动, 且摆幅越来越小.
- 情况 5 所有变量都单调地向精确解趋近.

在常用的迭代法中, 情况 4 和 5 是收敛的, 其余是发散的. 在二维迭代法中, 可以调节  $\tilde{d}_i$  使情况 2 和 3 变为收敛的. 因此, 二维迭代法的收敛范围比常用的迭代法大. 由推论 1 可知, 欲使 (3) 式的稳态解与 (1) 式的解相等, 唯一的条件是矩阵  $A$  必须满足特征值条件. 此条件对于任何稳定的系统来说是一定能够满足的.

虽然迭代法是一种无限步的计算方法, 其迭代的收敛速度与所解的具体问题有关, 无法作出一般性的比较, 我们可以对它们作出定性分析. 一般说, 常用的迭代法如高斯-塞德尔法的缺点是其阻尼/步长因子不能选择和改变, 而二维迭代法的  $\tilde{d}_i$  可任意选取以提高方法的收敛速度.

在实际应用中, 根据观察,  $\tilde{d}_i$  可按下列要求取值:

- (1) 对情况 2、3 和 4, 采用格式 3 或 4 取

$$\tilde{d}_i = \max \left[ \left( (1.1 - 1.8) \sum_{j=1}^n |a_{ij}| - a_{ii} \right), 0 \right].$$

- (2) 对情况 5, 采用格式 2, 取  $\tilde{d}_i = (0.55 - 0.95)a_{ii}$ .

- (3) 若矩阵  $A$  的主对角元中有零元, 比如  $a_{ii} = 0$ , 则无论采用何种格式, 均取

$$\tilde{d}_i = \sum_{j=1}^n |a_{ij}|.$$

#### 四、二维迭代法的电路解释

假定 (1) 式是一线性静态自主电网络  $\bar{N}$  的改进节点方程组, 则在  $\mathbf{x}$  中包含着节点电压和支路电流两类变量. 引入  $\mathbf{Cx}$  项后, 所得方程组 (3) 可视为一动态自主电网络  $\tilde{N}$  的状态方程组. 不难看出,  $\tilde{N}$  是在  $\bar{N}$  的每个节点到参考点之间并联一个电容, 而在每个选为电流变量的支路中串入一个电感构成. 这样,  $\tilde{N}$  中的状态变量就是电容电压和电感电

流, 恰好就是  $\bar{N}$  中的全部独立变量. 由于电容电压和电感电流的惰性, 它们的引入减缓了  $\bar{N}$  中各变量的变化速率. 因此, 从这个意义上说, 由  $\bar{N}$  变换到  $\tilde{N}$  相当于每个变量都增加了阻尼. 由此可见, 上面把  $C_i$  称为阻尼系数是有其物理背景的. 当动态网络在阶跃信号的激励下进入稳态后, 电容相当于开路, 电感相当于短路, 此时  $\tilde{N}$  的解就是  $\bar{N}$  的解, 亦即  $\tilde{N}$  的稳态解就是  $\bar{N}$  的解.

线性方程组的迭代法是在第  $i$  次迭代中求某一变量时, 其余变量采用当时已经确定的值, 从而使多变量的方程组的求解变为一列单变量方程的求解. 因此, 有把这种方法称为时间分裂法的<sup>[3]</sup>. 例如, 高斯-塞德尔法可表示为

$$\mathbf{x}_{m+1} = \mathbf{D}^{-1}(\mathbf{b} - \mathbf{L}\mathbf{x}_{m+1} - \mathbf{U}\mathbf{x}_m). \quad (12)$$

在求解  $\mathbf{x}$  的第  $i$  分量  $x_{m+1}^{(i)}$  时, 分量  $x^{(k)}$  当  $k < i$  时采用  $x_{m+1}^{(k)}$  值, 而当  $k > i$  时采用  $x_m^{(k)}$  值.  $x_m^{(k)}$  是在第  $m$  次迭代中确定的值, 而  $x_{m+1}^{(k)}$  是在第  $(m+1)$  次迭代中在解  $x_{m+1}^{(i)}$  时已确定的值. 因此方程 (12) 实际上变为一列单变量方程. 从这种意义上讲, 采用二维迭代法减缓了变量的变化, 从而改善了迭代的收敛.

## 五、计算示例

为了说明二维迭代法的实用意义, 我们在计算机上用此算法进行了实例计算, 效果是好的. 这里举几个简单的例子.

### 例 1 有方程组

$$\begin{bmatrix} 2 & -1 & 1 & 0 \\ -1 & 2 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -5 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

此方程组在用高斯-塞德尔法求解时, 各变量的变化如情况 2, 不收敛. 改用二阶 Gear-

GS 法取  $\tilde{d}_i = 1.1 \sum_{j=1}^n |a_{ij}| - a_{ii}$ , 若  $a_{ii} \geq \sum_{j=1}^n |a_{ij}|$ , 则取  $\tilde{d}_i = 0$ . 误差用最大残差

表示, 并取  $\varepsilon_1 = 10^{-2}$ ,  $\varepsilon_2 = 10^{-6}$ , 迭代 53 次即求得解为  $\mathbf{x}^T = [0.2, -0.4, 0.2, 1]$ .

### 例 2 有方程组

$$\begin{bmatrix} 4 & 2 & 1 & 1 \\ -1 & 3 & 1 & 1 \\ 0 & -2 & 3 & 2 \\ -1 & -2 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 4 \\ 3 \\ 1 \end{bmatrix},$$

此方程组在用高斯-塞德尔法求解时, 各变量的变化如情况 4, 有收敛趋势, 但由于计算机精度所限, 使它总是不能达到  $\varepsilon = 10^{-6}$  的要求, 采用格式 4, 取

$$\tilde{d}_i = \max \left[ \left( 1.6 \sum_{j=1}^n |a_{ij}| - a_{ii} \right), 0 \right], \quad \varepsilon_1 = 10^{-2}, \quad \varepsilon_2 = 10^{-6}$$

迭代 43 次可求得解为  $\mathbf{x}^T = [1.0, 1.0, 1.0, 1.0]$ .

## 六、结 束 语

二维迭代法是一类迭代方法的集合,它的导出使现有的几种常用的迭代方法得到了统一,并且更加丰富了迭代法的内容。另一方面,运用二维迭代的观念可从另一角度观察迭代法的性质。

理论分析和实际计算结果都表明,二维迭代法在收敛范围和收敛速度等方面都优于常用的迭代法。迭代法在第三代电路的计算机模拟中得到广泛的应用<sup>[3]</sup>,二维迭代法在电路模拟中的应用也是有希望的。

对于二维迭代法中尚未解决的问题,例如如何选择最佳的阻尼/步长因子以及如何合理控制递归迭代的次数或精度等,还需要作进一步的理论和实践研究。

### 参 考 文 献

- [1] Д. К. 法捷耶夫, B. H. 法捷耶娃著, 刘克武等译, 线性代数计算方法, 上海科技出版社, 1965年, 第231—291页。
- [2] I. S. Duff, *Proc. IEEE*, 65(1977), 500—35.
- [3] G. D. Hachtel, A. L. Sangiovanni-Vincentilli, *ibid.*, 69(1981), 1264—80.
- [4] S. L. Richter and R. A. Decarlo, *IEEE Trans. on CAS*, CAS-30 (1983), 347—52.
- [5] 谷荻隘嗣, 通信学会论文志, J65 A (1982), 802.
- [6] 张学铭等, 微分方程稳定性理论讲义, 山东人民出版社, 1958年, 第54—113页。
- [7] 韩天敏, 应用数学学报 1977年, 第3期, 第28页。
- [8] L. W. Nagel, Spice-II, A Computer Program to Simulate Semiconductor Circuits, Memorandum No. ERL, M 520, College of Engineering, University of California Berkeley, 1975.

## A UNIFIED APPROACH FOR SOLVING LINEAR EQUATIONS —TWO-DIMENSIONAL ITERATIVE METHOD

Liu Xiaoming, Hu Jiandong

(Beijing Institute of Posts and Telecommunications)

In this paper, a unified approach of iterative methods, such as Jacobi method, Gauss-Seidel method, SOR method, etc., for solving linear equations is discussed and studied. For the reason stated in this paper, this approach is called 2-dimensional iterative method. The convergence and the rate of convergence of iteration process are improved by using this new approach. The theoretical analysis and the computing results demonstrate that this approach has many advantages over the generally used iterative methods. It is useful in solving the large scale electric circuits, such as VLSI.