偏微分方程式の数値計算

辻 裕太

2017/3/26

1　差分法と連立一次方程式

$$-\frac{d^2u}{dx^2} = 1 \quad (0 < x < 1), \quad u(0) = u(1) = 0$$

という微分方程式を考える。これを差分法で離散化 (後に説明する) すると

$$-u_{i-1} + 2u_i - u_{i+1} = \frac{1}{n^2}, \quad i = 1, \ldots, n-1$$

という連立一次方程式を解くことになる。(ただし n は自然数で、n が大きいほど精度が上がる) $u_0 = u_n = 0$ を考慮してこれを行列で表示してみると、$x = (u_1, \ldots, u_{n-1})$ の係数行列 A は

$$\begin{pmatrix} 2 & -1 & \ldots & \ldots & 0 \\ -1 & 2 & -1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ldots & -1 & 2 & -1 \\ 0 & \ldots & \ldots & -1 & 2 \end{pmatrix}$$

という $(n-1) \times (n-1)$ 行列になる。このような行列は大半の成分が 0 だから疎行列という。連立一次方程式をどのように解くかについて述べる。

2　線形反復法

$$A \in M_n(\mathbb{R}), \quad x, f \in \mathbb{R}^n$$

に対して (A,f は既知、x は未知)

$$Ax = f$$

という形の連立一次方程式を解いて x を求めたい。特に A が正則なら

$$x = A^{-1}f$$

として解が求まるが、A^{-1} を求める、あるいはそれに近いことをして厳密に x を求める事は A のサイズが大きいときは手計算では不可能に近い。実際の数値計算においては A のサイズはきわめて大きい (1 億以上) ため、厳密に x を求めるよりは反復法と呼ばれる手法を用いることが多い。まず線形反復法と呼ばれる手法について述べた後、CG 法と呼ばれる手法について解説する。

2.1 一般論

$$A = M + N, \quad M \in GL_n(\mathbb{R})$$

として A を分解すると、元の方程式は

$$(M+N)x = f \quad \Leftrightarrow \quad x = M^{-1}(f - Nx)$$

と書き直せる。これを元に、

1. $x^{(0)} \in \mathbb{R}^n$: 出発ベクトル
2. $x^{(k)} = M^{-1}(f - Nx^{(k-1)}) \quad k = 1, 2, \ldots$

としてベクトルの列を作り、これが元の方程式の解に収束することを期待する。収束のためには A に条件が必要である。まず分解の仕方の具体例を挙げ、A に課した条件の下で収束することを示す。

2.2 具体例

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$$

に対して、行列 D, L, U を

$$D = \begin{pmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn} \end{pmatrix}, L = \begin{pmatrix} 0 & & & 0 \\ a_{21} & 0 & & \\ \vdots & & \ddots & \\ a_{n1} & \ldots & a_{nn-1} & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & a_{12} & \ldots & a_{1n} \\ & 0 & & \vdots \\ & & \ddots & a_{n-1n} \\ 0 & & & 0 \end{pmatrix}$$

で定める。

2.2.1 Jacobi 法

$$M = D, \quad N = L + U$$

とする。M は A の全対角成分がゼロでなければ正則である。この時、反復の式 $x^{(k)} = M^{-1}(f - Nx^{(k-1)})$ を成分で書くと

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(f_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k-1)} \right) \quad i = 1, \ldots, n$$

となる。

2.2.2 Gauss-Seidel 法

$$M = D + L, \quad N = U$$

とする。M は A の全対角成分がゼロでなければ正則である。この時の反復の式を成分で書くと

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(f_i - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} \right) \quad i = 1, \ldots, n$$

となる。右辺の x に添え字 k が入っているので、$i = 1, 2, \ldots, n$ の順に計算しないといけないことに注意。

2.2.3 SOR(Successive Over-Relaxation method) 法

$\omega \neq 0$ をパラメータとして、

$$M = \frac{1}{\omega} D + L, \quad N = \left(1 - \frac{1}{\omega}\right) D + U$$

とする。特に $\omega = 1$ なら Gauss-Seidel 法に一致する。反復の式 $Mx^{(k)} = f - Nx^{(k-1)}$ に代入してみると

$$(D + \omega L)x^{(k)} = (1 - \omega)D - \omega U x^{(k-1)} + \omega f$$

となる。これを成分で書き換えると

$$a_{ii} x_i^{(k)} = -\omega \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \omega \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} + (1 - \omega) a_{ii} x_i^{(k-1)} + \omega f_i$$

となる。そこでやはり A の対角成分がゼロでなければ、

$$x_i^{(k)} = \frac{\omega}{a_{ii}} \left(f_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k-1)} \right) + (1 - \omega) x_i^{(k-1)}$$

となる。ここで右辺第一項は G-S 法の右辺の ω 倍だから、SOR 法は $x_i^{(k)}$ として G-S 法と $x_i^{(k-1)}$ の加重平均を取っているという見方ができる。実際に、上手く ω を取ると SOR 法は G-S 法よりも少ない反復回数で収束する。

2.3　数値例

実際に上記の三つの方法で方程式を解いてみる。

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad f = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

とし、$Ax = f$ を解く。厳密解は $x_1 = 3/4, x_2 = 1/2, x_3 = 1/4$ である。$\varepsilon = 10^{-7}$ として

$$\max_{1 \leq i \leq 3} \left| \frac{x_i^{(k)} - x_i^{(k-1)}}{x_i^{(k)}} \right| < \varepsilon$$

となった時に反復を打ち切り、その時の反復回数と、SOR 法のパラメータ ω を出力した。Jacobi 法は 47 回、G-S 法は 23 回で収束した。SOR 法は ω の取り方によって回数が大きく異なったが、ω を 0.1 と 1.9 の間で 0.1 刻みに動かしたところ、$\omega = 1.2$ の時が最も少なく、12 回で収束した。ちなみに $\omega = 0.1$ の時は 426 回である。

2.4　狭義対角優位行列に対する収束の証明

前述した計算法は A が狭義対角優位という行列ならば収束することを見る。まず行列のノルムを定義する。

Definition 2.1. $n \times n$ 実行列全体を $M_n(\mathbb{R})$ とする。$M_n(\mathbb{R})$ は通常の行列の和やスカラー倍について実ベクトル空間となり、その次元は n^2 である。$A \in M_n(\mathbb{R})$ に対して $\|A\|$ を、$\mathbb{R}^n$ のノルム $\|\cdot\|$ を用いて

$$\|A\| := \sup_{\substack{x \neq 0 \\ x \in \mathbb{R}^n}} \frac{\|Ax\|}{\|x\|}$$

として定める。

これがノルムになる事の証明は省略する。このノルムをユークリッド空間のノルムに誘導されるノルムという。有限次元であるから収束の議論をする際には証明しやすいものを取れば良いのだが、どういうノルムを取っているのかは非常に重要である。例えば実対称行列全体 (これも有限次元ベクトル空間になる) に通常のユークリッドノルムを用いてノルムを定義すると、 $\|A\|$ は A の固有値の最大の絶対値になるが、場合によっては固有値の計算の方が面倒になる事もある。後に収束を示す際に次の定理を用いる。

Theorem 2.2. $\mathbb{R}^n$ のノルムとして、$x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ に対し

$$\|x\| := \max_{1 \leq i \leq n} |x_i|$$

というものを取る。これを用いて定めた $M_n(\mathbb{R})$ のノルムについて、A の (i, j) 成分を a_{ij} とすると次が成立する。

$$\|A\| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

Proof.

$$\begin{aligned}\|Ax\| &= \max_{1\le i\le n}\left|\sum_{j=1}^{n} a_{ij}x_j\right| \\ &\le \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}||x_j| \\ &\le \left(\max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|\right)\|x\|\end{aligned}$$

だから、$x\neq 0$ に対して

$$\frac{\|Ax\|}{\|x\|}\le \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

となり、左辺の sup を取って

$$\|A\|\le \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

を得る。

$$\max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

を与える i の一つを I とする。$x\in\mathbb{R}^n$ を $a_{Ij}\ge 0$ の時 $x_j=1$、$a_{Ij}<0$ の時 $x_j=-1$ として定めると、$\|x\|=1$ であり、

$$\|Ax\|\ge \|x\|\sum_{j=1}^{n}|a_{Ij}|=\|x\|\max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

となる。よってこの x に対して

$$\frac{\|Ax\|}{\|x\|}\ge \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

となり、左辺の sup を取って

$$\|A\|\ge \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

を得る。よって

$$\|A\|= \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

となる。 □

収束を示す際に基礎となるのが次の定理である。

Theorem 2.3.

$$x_0, b \in \mathbb{R}^n, \quad P \in M_n(\mathbb{R})$$

$$x_{k+1} = Px_k + b, \quad k \geq 0$$

とする。$\mathbb{R}^n$ のノルムを $\|\cdot\|$ とし、このノルムに誘導される $M_n(\mathbb{R})$ のノルムについて

$$\|P\| < 1$$

ならば x_k は $\|\cdot\|$ について収束し、それを x とすると、x は

$$y = Py + b$$

のただ一つの解となる。

Proof. $\{x_k\}$ が $\mathbb{R}^n$ のコーシー列であることを示す。$k \geq 1$ に対して

$$\begin{aligned}\|x_{k+1} - x_k\| &= \|P(x_k - x_{k-1})\| \\ &\leq \|P\|\|x_k - x_{k-1}\|\end{aligned}$$

となるから、これを繰り返し用いて

$$\|x_{k+1} - x_k\| \leq \|P\|^k \|x_1 - x_0\| \quad (k \geq 1)$$

を得る。$\|x_1 - x_0\| = 0$ ならば $x_k = x_0$ となってコーシー列である。そこで以降は $\|x_1 - x_0\| \neq 0$ とする。$m, n \in \mathbb{N}, m > n > 1$ に対して

$$\begin{aligned}\|x_m - x_n\| &\leq \|x_m - x_{m-1}\| + \ldots + \|x_{n+1} - xn\| \\ &= \|x_1 - x_0\| \sum_{i=n+1}^{m} \|P\|^i\end{aligned}$$

となる。$S_k := \sum_{i=0}^{j} \|P\|^i$ とすると $\|P\| < 1$ より S_k は $\mathbb{R}$ のコーシー列である。よって

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad m, n > N$$

$\Longrightarrow$

$$|S_m - S_n| = \sum_{i=n+1}^{m} \|P\|^i < \frac{\varepsilon}{\|x_1 - x_0\|}$$

となる。この時

$$\|x_m - x_n\| < \varepsilon$$

となるから $\{x_k\}$ は $\mathbb{R}^n$ のコーシー列であり、

$$\lim_{k \to \infty} x_k = x, \quad x \in \mathbb{R}^n$$

となる。$x_{k+1} = Px_k + b$ において $k \to \infty$ とすれば

$$x = Px + b$$

となる。今、もう一つの解 z があったとすると

$$\|x-z\| = \|P(x-z)\| \leq \|P\|\|x-z\|$$

となるが、$\|P\| < 1$ だから $\|x-z\| = 0$、つまり $x = z$ となる。以上より示された。 □

反復法を上の定理の形に即して書き直すと

$$x_{k+1} = -M^{-1}Nx_k + M^{-1}f$$

となり、P として $-M^{-1}N$ を取っているので、次が分かる。

Theorem 2.4. $\|-M^{-1}N\| < 1$ ならば、*Jacobi* 法、*G-S* 法、*SOR* 法は共に任意の初期値に対して収束する。

従って、分解前の行列 A に条件を課して $-M^{-1}N$ のノルムを 1 より小さくすれば収束する。特に次の条件での収束を示す。

Definition 2.5. $A \in M_n(\mathbb{R})$ が狭義対角優位であるとは、任意の i に対して次が成立すること。

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

次の定理が目標である。

Theorem 2.6. A が狭義対角優位であれば、*Jacobi* 法は収束する。

Proof. Theorem 2.2 のノルムについて $\|-M^{-1}N\| = \|M^{-1}N\| < 1$ となる事を示せばよい。A の定義から対角成分は全て零ではなく、M^{-1} が存在する。$M^{-1}N$ の (i,j) 成分は、$i = j$ の時 0、$i \neq j$ の時 a_{ij}/a_{ii} となる。よって Theorem 2.2 から、

$$\|M^{-1}N\| = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|$$

となる。狭義対角優位の定義から全ての i に対して

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1$$

となり、したがって $\|M^{-1}N\| < 1$ となる。 □

2.5 既約優対角行列に対する Jacobi 法の収束

狭義対角優位行列に対する収束を示したわけだが、ポアソン方程式の差分法などに現れる係数行列は狭義対角優位ではない。そこで、既約優対角行列と呼ばれる行列に対して Jacobi 法の収束を示す。まずは行列のスペクトル半径を定義する。

Theorem 2.7.

$$A \in M_n(\mathbb{R})$$

に対し、A のスペクトル半径 $\rho(A)$ を

$$\rho = \max_{\lambda \in Sp(A)} |\lambda|$$

で定義する。ただし $Sp(A)$ は A の固有値全体の集合である。

次の定理が重要である。

Theorem 2.8. Theorem 2.3 は仮定 $\|P\| < 1$ を $\rho(P) < 1$ で置き換えても成立する。

これが成立すれば $\rho(-M^{-1}N) < 1$ なら Jacobi 法、G-S 法、SOR 法は収束する。Theorem 2.8 を示すためにいくつか補題を準備する。

Lemma 2.9. $A \in M_n(\mathbb{R})$ とする。$M_n(\mathbb{R})$ のノルムを $\|\cdot\|$ とすると次が成立する。

$$\rho(A) < 1 \Rightarrow \lim_{m \to \infty} \|A^m\| = 0$$

Proof. $M_n(\mathbb{R})$ は有限次元だから、ノルムとして $\|A\| \coloneqq \max_{1 \leq i,j \leq n} |a_{ij}|$ を取り、これについて示せばよい。A の Jordan 標準形は

$$P^{-1}AP = \begin{pmatrix} J_1 & 0 & \dots & \dots & 0 \\ 0 & J_2 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & 0 & j_{r-1} & 0 \\ 0 & \dots & \dots & 0 & J_r \end{pmatrix}$$

と書ける。ただし $J_i (i = 1, \dots, r)$ は Jordan 細胞で、$\lambda_i \in Sp(A)$ として

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}$$

または

$$J_i = (\lambda_i)$$

と書ける。よって、任意の自然数 m に対して

$$P^{-1}A^mP = \begin{pmatrix} J_1^m & 0 & \dots & \dots & 0 \\ 0 & J_2^m & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & 0 & j_{r-1}^m & 0 \\ 0 & \dots & \dots & 0 & J_r^m \end{pmatrix}$$

となる。

$$J_i^m = \begin{pmatrix} \lambda_i^m & m\lambda_i^{m-1} & \dots & & {}_mC_{p-1}\lambda^{m-p+1} \\ & \lambda_i^m & m\lambda_i^{m-1} & & \vdots \\ & & \ddots & \ddots & \\ & & & \lambda_i^m & m\lambda_i^{m-1} \\ & & & & \lambda_i^m \end{pmatrix}$$

だから、$|\lambda_i| \le \rho(A) < 1$ より、$\lim_{m\to\infty} \|J_i^m\| = 0$ である。したがって $\lim_{m\to\infty} \|A^m\| = 0$ となる。 □

Lemma 2.10. $A \in M_n(\mathbb{R})$ とする。

$$\rho(A) < 1 \Rightarrow I - A \in GL_n(\mathbb{R}).$$

Proof.

$$0 \notin Sp(I - A)$$

を示せばよい。$0 \in Sp(I - A)$ とすると、$1 \in Sp(A)$ となり、$\rho(A) < 1$ に矛盾する。よって示された。 □

Remark 1. この補題の仮定の下で

$$(I - A)^{-1} = \sum_{m=0}^{\infty} A^m$$

が示せるので、以前これを示した時の仮定 $\|A\| < 1$ は $\rho(A) < 1$ に精密化できる事が分かる。($\lim_{m\to\infty} A^m = 0$ を用いればできる)

これら二つの補題を用いて Theorem 2.8 を示す。

Proof. $\rho(P) < 1$ だから、Lemma 2.10 より

$$I - P \in GL_n(\mathbb{R})$$

である。よって

$$y = Py + b$$

には解が一意に存在する。それを x とすると、任意の自然数 k に対し

$$\begin{aligned} \|x_k - x\| &= \|P(x_{k-1} - x)\| \\ &= \|P^k(x_0 - x)\| \\ &\le \|P^k\|\|x_0 - x\| \end{aligned}$$

が成立する。Lemma 2.9 より $\lim_{k\to\infty} \|P^k\| = 0$ である。よって $\lim_{k\to\infty} x_k = x$ となる。 □

Remark 2. この定理を示すのに用いた補題二つの仮定 $\rho(A) < 1$ は実は必要十分条件であり、補題自体は仮定を $\|A\| < 1$ に変えても成立する。従って Thorem 2.3 を上と同様に示すことも出来る。次の定理が示す通り、$\|A\| < 1 \Rightarrow \rho(A) < 1$ が成り立つ。

Theorem 2.11. $\mathbb{C}^n$ のノルムから誘導された $M_n(\mathbb{R})$ のノルムに対し、

$$\rho(A) \leq \|A\|$$

が成立する。

Proof. $|\lambda| = \rho(A)$ となる $\lambda \in Sp(A)$ を取り、対応する固有ベクトルを x とする。

$$\|Ax\| = \|\lambda x\| = |\lambda|\|x\|$$

固有ベクトルの定義から $x \neq 0$ であり、

$$\frac{\|Ax\|}{\|x\|} = |\lambda|$$

となるから、左辺の sup を取って

$$\rho(A) \leq \|A\|$$

となる。 □

Theorem 2.12. $A \in M_n(\mathbb{R})$ の (i, j) 成分を a_{ij} とする。

$$\Lambda_i \coloneqq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

$$D_A \coloneqq \cup_{i=1}^{n} \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \Lambda_i\}$$

と定義すると、次が成立する。

$$Sp(A) \subset D_A$$

Proof. $\lambda \in Sp(A)$ を任意に取り、対応する固有ベクトル x を適当に定数倍して

$$|x_i| \leq |x_r| = 1 \quad (i = 1, \ldots, n)$$

となるように取る。$Ax = \lambda x$ の第 r 成分を書くと

$$\sum_{j=1}^{n} a_{rj} x_j = \lambda x_r$$

となる。よって移項して

$$(\lambda - a_{rr}) x_r = \sum_{\substack{j=1 \\ j \neq r}}^{n} a_{rj} x_j$$

だから、両辺の絶対値をとって

$$\begin{aligned} |\lambda - a_{rr}| &= \left| \sum_{\substack{j=1 \\ j \neq r}}^{n} a_{rj} x_j \right| \\ &\leq \sum_{\substack{j=1 \\ j \neq r}}^{n} |a_{rj}||x_j| \\ &\leq \Lambda_r \end{aligned}$$

となり、$\lambda \in D_A$ となる。よって、$Sp(A) \subset D_A$ $\square$

Definition 2.13. $\lambda \in Sp(A)$ とする。対応する固有ベクトル x について、$|x_r| = \|x\|_\infty$(sup ノルム) となる全ての r に対して (Theorem 2.12 の証明中の r) $|\lambda - a_{rr}| = \Lambda_r$ となるとき、λ は D_A の境界にあるという。

後に Theorem 2.12 を精密化する際にこの定義を用いる。

Remark 3. A が狭義対角優位ならば正則であるということが、$0 \in Sp(A)$ と仮定して固有ベクトルを上と同様にとれば r 番目の成分が狭義対角優位に反するという方法で示せる。

行列 A が既約であることの定義をする。

Definition 2.14. $A \in M_n(\mathbb{R})$ が可約であるとは、A の (i,j) 成分を a_{ij} とすると、ある置換

$$\begin{pmatrix} 1 & \dots & r & r+1 & \dots & n \\ \rho_1 & \dots & \rho_r & \sigma_1 & \dots & \sigma_{n-r} \end{pmatrix}$$

が存在して、

$$\forall \nu \in \{1, \dots, r\}, \quad \forall \mu \in \{1, \dots, n-r\}, \quad a_{\rho_\nu \sigma_\mu} = 0$$

が成立すること。A が可約でないとき、A は既約であるという。

Remark 4. 置換は互換の積で書けることを考えれば、A が可約であるとは、A を基本変形して

$$\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$$

ただし

$$A_{11} \in M_r(\mathbb{R}), \quad A_{22} \in M_{n-r}(\mathbb{R})$$

という形にできる、ということである。この時、元の方程式 $Ax = f$ は、$A_{11}x^1 = f^1, A_{22}x^2 = f^2$ というサイズの小さい二つの方程式に分解できる。

Definition 2.15.

$$M = \{P_1, \dots, P_n\}$$

を n 個の点の集合とする。

$$G \subset M \times M$$

を M のグラフという。G の元を G の辺という。$i, j \in \{1, \dots, n\}$ に対し P_i, P_j の間に道があるとは、ある $a_1, \dots, a_{r-1} \in \{1, \dots, n\}$ が存在して、$P_i = a_0, P_j = a_r$ とすると

$$(P_{a_k}, P_{a_{k+1}}) \in G, \quad (k = 0, \dots, r-1)$$

となることである。グラフ G が連結であるとは、任意の $i, j \in \{1, \dots, n\}$ に対し、P_i, P_j の間に道があることである。

Remark 5. この定義では n 個の点の結び方をグラフという。$(P_i, P_j) \neq (P_j, P_i)$ だから結び方には向きがついている。このように向きのついているグラフを特に有向グラフという。有向グラフが連結であるとは、M の任意の二点 P_i, P_j に対して、P_i から P_j へと順番にたどっていけるような辺の組が存在するというものである。

行列 A に対応するグラフを次で定義する。

Definition 2.16.

$$A \in M_n(\mathbb{R})$$

とし、A の (i,j) 成分を a_{ij} とする。

$$M = \{P_1, \ldots, P_n\}$$

に対し、M のグラフ G_A を、

$$(P_i, P_j) \in G_A \Leftrightarrow a_{ij} \neq 0 \quad (\forall i, j \in \{1, \ldots, n\})$$

として定める。

Remark 6. もし A が対称行列ならば $a_{ij} = a_{ji}$ だから $(P_i, P_j) \in G_A \Leftrightarrow (P_j, P_i) \in G_A$ となる。このような場合はグラフの向きを考える必要がないので、連結の定義なども変わってくるが、今は向きのついたものを考える。

Theorem 2.17.

$$A \in M_n(\mathbb{R}) \text{ が既約} \Leftrightarrow G_A \text{が連結}$$

Proof. 十分性. 対偶を示す。A が可約であるとする。Definition 2.14 の置換を取ると

$$\forall k \in \{1, \ldots, r\}, \quad \forall l \in \{1, \ldots, n-r\}, \quad a_{\rho_k \sigma_l} = 0$$

となるので、

$$\forall k \in \{1, \ldots, r\}, \quad \forall l \in \{1, \ldots, n-r\}, \quad (P_{\rho_k}, P_{\sigma_l}) \notin G_A$$

となる。そこで、ある $k_1 \in \{1, \ldots, r\}, l_1 \in \{1, \ldots, n-r\}$ において $P_{\rho_{k_1}}, P_{\sigma_{l_1}}$ の間に道が存在したとすると、ある $k_2 \in \{1, \ldots, r\}, l_2 \in \{1, \ldots, n-r\}$ が存在して $(P_{\rho_{k_2}}, P_{\sigma_{l_2}}) \in G_A$ となる。これは明らかに上記と矛盾する。よって G_A は連結ではない。必要性. 対偶を示す。G_A が連結でないとする。このとき、ある $i, j \in \{1, \ldots, n\}$ が存在して P_i, P_j の間に道がない。そこで

$$B \coloneqq \{P_\alpha \in M \mid P_i \text{と } P_\alpha \text{は道で結べる }\} \cup P_i$$

$$C \coloneqq M \setminus B$$

とし、適当に M の点の番号を取り直して $B = \{P_{\rho_1}, \ldots, P_{\rho_r}\}, C = \{P_{\sigma_1}, \ldots, P_{\sigma_{n-r}}\}$ とする。$P_j \in C$ より $C \neq \phi$ であり、したがって

$$\begin{pmatrix} 1 & \ldots & r & r+1 & \ldots & n \\ \rho_1 & \ldots & \rho_r & \sigma_1 & \ldots & \sigma_{n-r} \end{pmatrix}$$

として置換が定義できる。これが Definition 2.14 の条件を満たすことを示す。もし、ある $x \in \{1, \ldots, r\}, y \in \{1, \ldots, n-r\}$ があって $a_{\rho_x \sigma_y} \neq 0$ となったとすると、$(P_{\rho_x}, P_{\sigma_y}) \in G_A$ となる。この時、B の定義から P_i と P_{ρ_x} は道で結べるので、P_i と P_{σ_y} も道で結べる。これは $P_{\sigma_y} \in C$ に反する。よって

$$\forall \nu \in \{1, \ldots, r\}, \quad \forall \mu \in \{1, \ldots, n-r\}, \quad a_{\rho_\nu \sigma_\mu} = 0$$

が成立し、A は可約である。 □

次の定理は、A が既約であることと優対角性を結び付ける。Theorem 2.12 の精密化である。

Theorem 2.18. *$A \in M_n(\mathbb{R})$ は既約であるとする。$\lambda \in Sp(A)$ が D_A の境界上にあるならば、次が成立する。*

$$|\lambda - a_{ii}| = \Lambda_i (\forall i = 1, \ldots, n)$$

Proof. λ は D_A の境界上にあるので、対応する固有ベクトル x を

$$\|x\|_\infty = 1$$

となるように取ると、$|x_r| = \|x\|_\infty = 1$ となる全ての r で

$$\begin{aligned} |\lambda - a_{rr}| &= \left| \sum_{\substack{j=1 \\ j \neq r}}^{n} a_{rj} x_j \right| \\ &= \sum_{\substack{j=1 \\ j \neq r}}^{n} |a_{rj}||x_j| \\ &= \Lambda_r \end{aligned}$$

が成立する。したがってこのような r の一つを r_1 とすると、$j = 1, \ldots, n$ に対し

$$a_{r_1 j} \neq 0 \Rightarrow |x_j| = 1$$

が成立する。Theorem 2.17 から、G_A は連結である。したがって任意の $i \in \{1, \ldots, n\}$ に対して P_{r_1} と P_i の間には道がある、つまり

$$r_1 = \alpha_0, \alpha_1, \ldots, \alpha_{s-1}, \alpha_s = i \quad (\alpha_k \in \{1, \ldots, n\}, k = 1, \ldots, s)$$

が存在して

$$(P_{\alpha_k}, P_{\alpha_{k+1}}) \in G_A \quad (k = 0, \ldots, s-1)$$

を満たす。G_A の定義から

$$a_{\alpha_k \alpha_{k+1}} \neq 0 \quad (k = 0, \ldots, s-1)$$

となる。特に $k = 0$ の時、$a_{\alpha_0 \alpha_1} = a_{r_1 \alpha_1} \neq 0$ だから

$$|x_{\alpha_1}| = 1 = \|x\|_\infty$$

となり、r_1 の時の議論と同様にして、$j = 1, \ldots, n$ に対し

$$a_{\alpha_1 j} \neq 0 \Rightarrow |x_j| = 1$$

が成立する。この議論を $k = 1, \ldots, s-1$ で繰り返して

$$|x_{\alpha_s}| = |x_i| = 1 = \|x\|_\infty$$

を得られるので、

$$
\begin{aligned}
|\lambda - a_{ii}| &= \left| \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j \right| \\
&= \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}||x_j| \\
&= \Lambda_i
\end{aligned}
$$

となる。i は任意だったから、示された。 □

Definition 2.19. *$A \in M_n(\mathbb{R})$ が既約優対角であるとは、A が既約で、次が成り立つこと。*

1. *$|a_{ii}| \geq \Lambda_i, \quad \forall i \in \{1, \ldots, n\}$*
2. *$\exists i_0 \in \{1, \ldots, n\}, \quad |a_{i_0 i_0}| > \Lambda_{i_0}$*

Remark 7. たとえばポアソン方程式の差分法では境界条件によって狭義の優対角性が成り立つようになる。

Theorem 2.20. *A が既約優対角行列の時、Jacobi 法は任意の初期値に対して収束する。*

Proof. まず、既約優対角の時に $a_{ii} = 0$ となる i が存在すると、その行は全ての成分が 0 になる。基本変形によってこの行を一行目に移せば、A が既約であることに反する (Definition 2.14 の Remark 参照)。従って対角成分は全て 0 でなく、Jacobi 法は定義できる。後は $\rho(-M^{-1}N) < 1$ を示せばよい。まず A が既約の時 $-M^{-1}N$ も既約であることを示す。Jacobi 法において $-M^{-1}N$ の (i,j) 成分 p_{ij} は $i = j$ の時 0、$i \neq j$ の時 $-a_{ij}/a_{ii}$ だから、$i \neq j$ の時は $a_{ij} \neq 0 \leftrightarrow p_{ij} \neq 0$ となって $-M^{-1}N$ のグラフも連結であり、$-M^{-1}N$ も既約である。Theorem 2.2 は $\mathbb{C}^n$ のノルムとそれに誘導される $M_n(\mathbb{R})$ のノルムについても全く同様に成立する。$\mathbb{C}^n$ のノルムとして sup ノルム $\|\cdot\|_\infty$ を取り、それに誘導される $M_n(\mathbb{R})$ のノルム $\|\cdot\|$ を取る。$i = j$ の時 $p_{ij} = 0$、$i \neq j$ の時 $p_{ij} = a_{ij}/a_{ii}$ だから、$-M^{-1}N$ に対して

$$
\begin{aligned}
\Lambda_i &= \sum_{\substack{j=1 \\ j \neq i}}^{n} |p_{ij}| \\
&= \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|}{|a_{ii}|}
\end{aligned}
$$

であって、Theorem 2.2 と既約優対角の定義から、

$$
\rho(-M^{-1}N) \leq \| - M^{-1}N\| = \|M^{-1}N\| = \max_{1 \leq i \leq n} \Lambda_i \leq 1
$$

が成立する。もし $\rho(-M^{-1}N) = 1$ が成立したとすると、ある $\lambda \in Sp(-M^{-1}N)$ が存在して

$$
|\lambda| = \max_{1 \leq i \leq n} \Lambda_i = 1
$$

を満たす。この時、λ は $D_{-M^{-1}N}$ の境界にあることを示す。λ に対応する固有ベクトル x を $\|x\|_\infty = 1$ となるように取る。$|x_r| = \|x\|_\infty = 1$ を満たす r に対して $Ax = \lambda x$ の r 成分を書くと、Theorem 2.12 と同様にして

$$
|\lambda - p_{rr}| \leq \Lambda_r
$$

となる。すると、$p_{rr} = 0$ であること、$|\lambda| = \max_{1\leq i\leq n} \Lambda_i$ となることから、

$$|\lambda - p_{rr}| = \max_{1\leq i\leq n} \Lambda_i \leq \Lambda_r$$

となるが、max の定義から

$$\max_{1\leq i\leq n} \Lambda_i = \Lambda_r$$

となる。よって

$$|\lambda - p_{rr}| = \Lambda_r$$

が成立して、λ は $D_{-M^{-1}N}$ の境界にある。$-M^{-1}N$ は既約だから、Theorem 2.18 より、任意の $i \in \{1, \ldots, n\}$ に対して

$$|\lambda - p_{ii}| = \Lambda_i$$

が成立する。これは $p_{ii} = 0$、$|\lambda| = 1$、Λ_i の定義から

$$1 = \frac{\sum_{\substack{j=1\\j\neq i}}^{n} |a_{ij}|}{|a_{ii}|}$$

となり、さらに

$$|a_{ii}| = \sum_{\substack{j=1\\j\neq i}}^{n} |a_{ij}|$$

と書き換えられるが、これが任意の $i \in \{1, \ldots, n\}$ に対して成立することは、A が既約優対角であることに反する。よって $\rho(-M^{-1}N) < 1$ であり、示された。 $\square$