PHƯƠNG PHÁP SỐ TRONG ĐẠI SỐ TUYẾN TÍNH

PGS.TS. Nguyễn Thanh Bình
ThS. Trần Thị Mỹ Huỳnh

Tp. Hồ Chí Minh - 2023

Mục lục

Chương 1

Các khái niệm cơ bản

1.1 Ma trận và vector

1.1.1 Ma trận

Cho $\mathbb{R}$ là tập hợp các số thực. Khi đó, $\mathbb{R}^{m\times n}$ là không gian vector của các ma trận thực có m dòng và n cột

$$A \in \mathbb{R}^{m\times n} \Longleftrightarrow A = (a_{ij}) = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & amn \end{bmatrix},\ a_{ij} \in \mathbb{R}$$

Ngoài ra, chúng ta còn sử dụng $[A]_{ij}$ hay $A(i,j)$ để chỉ những phần tử của một ma trận.

1.1.2 Các phép toán trên ma trận

Các phép toán cơ bản trên ma trận gồm:

- Ma trận chuyển vị ($\mathbb{R}^{m\times n} \to \mathbb{R}^{n\times m}$),

$$C = A^T \Longrightarrow c_{ij} = a_{ji}$$

- Cộng hai ma trận ($\mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$),

$$C = A + B \Longrightarrow c_{ij} = a_{ij} + b_{ij}$$

- Nhân một số với ma trận ($\mathbb{R} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$),

$$C = \alpha A \Longrightarrow c_{ij} = \alpha a_{ij},$$

- Nhân hai ma trận ($\mathbb{R}^{m\times p} \times \mathbb{R}^{p\times n} \to \mathbb{R}^{m\times n}$),

$$C = AB \Longrightarrow c_{ij} = \sum_{k=1}^{r} a_{ik} b_{kj}.$$

- Nhân ma trận theo từng điểm ($\mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$),

$$C = A. * B \Longrightarrow c_{ij} = a_{ij} b_{ij}$$

- Phép chia theo từng điểm ($\mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$),

$$C = A./B \Longrightarrow c_{ij} = a_{ij}/b_{ij}.$$

1.1.3 Khái niệm vector

Cho $\mathbb{R}^n$ là không gian vector của các vector thực có n phần tử

$$x \in \mathbb{R}^n \Longleftrightarrow x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, x_i \in \mathbb{R}$$

trong đó, x_i là thành phần thứ i của vector x.
Chú ý, ta đồng nhất $\mathbb{R}^n$ với $\mathbb{R}^{n\times 1}$ nên mỗi phần tử của $\mathbb{R}^n$ là một vector *cột*. Mặc khác, những phần tử của $\mathbb{R}^{1\times n}$ là những vector *dòng*:

$$x \in \mathbb{R}^{1\times n} \Longleftrightarrow x = [x_1, \ldots, x_n].$$

Nếu x là một vector cột thì $y = x^T$ là một vector dòng.

1.1.4 Các phép toán trên vector

Cho $a \in \mathbb{R}, x \in \mathbb{R}^n$ và $y \in \mathbb{R}^n$. Khi đó, các phép toán cơ bản trên vector gồm:

- Nhân một số với một vector ,

$$z = ax \Longrightarrow z_i = ax_i,$$

- Cộng hai vector

$$z = x + y \Longrightarrow z_i = x_i + y_i,$$

- Tích vô hướng của hai vector (hay *tích trong*),

$$c = x^T y \Rightarrow c = \sum_{i=1}^{n} x_i y_i,$$

- Nhân vector theo từng điểm

$$z = x. * y \Rightarrow z_i = x_i y_i$$

- Chia vector theo từng điểm

$$z = x./y \Rightarrow z_i = x_i/y_i$$

1.1.5 Ma trận phức

Không gian vector của các ma trận phức có m dòng và n cột được ký hiệu bởi $\mathbb{C}^{m\times n}$. Phép nhân với vô hướng, phép cộng và phép nhân của các ma trận phức tương ứng như trong ma trận thực. Tuy nhiên, phép chuyển vị trở thành chuyển vị liên hợp

$$C = A^H \Rightarrow c_{ij} = \overline{a_{ji}}$$

Không gian vector của các vector phức n chiều được ký hiệu là $\mathbb{C}^n$. Tích vô hướng của hai vector x và y được cho bởi

$$s = x^H y = \sum_{i=1}^{n} \overline{x_i} y_i.$$

Cho $A = B + iC \in \mathbb{C}^{m\times n}$, phần thực và phần ảo của A tương ứng là $Re(A) = B$ và $Im(A) = C$. Liên hợp của A là ma trận $\overline{A} = (\overline{a}_{ij})$.

1.1.6 Ma trận đường chéo

Các ma trận với 0 băng thông dưới và 0 băng thông trên là ma trận đường chéo. Nếu $D \in \mathbb{R}^{m\times n}$ là ma trận đường chéo thì khi đó

$$D = diag(d_1, \ldots, d_q), q = min(m, n) \Longleftrightarrow d_i = d_{ii}.$$

Nếu $D = diag(d) \in \mathbb{R}^{n\times n}$ và $x \in \mathbb{R}^n$ thì $Dx = d. * x$. Nếu $A \in \mathbb{R}^{m\times n}$ thì phép nhân trái với $D = diag(d_1, \ldots, d_m) \in \mathbb{R}^{m\times m}$,

$$B = DA \Longleftrightarrow B(i,:) = d_i \cdot A(i,:), i = 1 : m$$

và phép nhân phải với $D = diag(d_1, \ldots, d_m) \in \mathbb{R}^{n\times n}$,

$$B = AD \Longleftrightarrow B(:,j) = d_j \cdot A(:,j), j = 1 : n.$$

1.1.7 Ma trận đối xứng

Ma trận $A \in \mathbb{R}^{n\times n}$ là đối xứng nếu $A^T = A$ và là phản đối xứng nếu $A^T = -A$. Tương tự, ma trận $A \in \mathbb{C}^{n\times n}$ là Hermit nếu $A^H = A$ và là phản Hermit nếu $A^H = -A$.
Ví dụ: Ma trận đối xứng:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix},$$

Ma trận phản đối xứng:

$$\begin{bmatrix} 0 & -2 & 3 \\ 2 & 0 & -5 \\ -3 & 5 & 0 \end{bmatrix},$$

Ma trận Hermit:

$$\begin{bmatrix} 1 & 2-3i & 4-5i \\ 2+3i & 6 & 7-8i \\ 4+5i & 7+8i & 9 \end{bmatrix},$$

Ma trận phản Hermit:

$$\begin{bmatrix} i & -2+3i & -4+5i \\ 2+3i & 6i & -7+8i \\ 4+5i & 7+8i & 9i \end{bmatrix}.$$

1.1.8 Ma trận hoán vị và ma trận đơn vị

Ta ký hiệu ma trận đơn vị $n \times n$ là I_n, ví du,

$$I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Ta sử dụng ký hiệu e_i để chỉ cột thứ i của I_n. Nếu các dòng của I_n được sắp xếp lại thì ma trận kết quả được biểu diễn như là một ma trận hoán vị. Ví dụ,

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}. \tag{1.1.1}$$

1.1.9 Ma trận khối

Ma trận khối là ma trận mà các phần tử cũng là các ma trận. Chẳng hạn, một ma trận 8×15 của các vô hướng có thể được xem như là ma trận khối 2×3 với các phần tử là các ma trận 4×5.

Cho $A \in \mathbb{R}^{m \times n}$, *phân tích dạng dòng* của A là một mảng các vector dòng:

$$\iff A = \begin{bmatrix} r_1^T \\ \vdots \\ r_m^T \end{bmatrix}, r_k \in \mathbb{R}^n. \tag{1.1.2}$$

Ví dụ 1.1.1. phân tích dạng dòng của ma trận $\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$, ta xem A như là một tập hợp của các vector dòng với

$$r_1^T = [1 \quad 2], \qquad r_2^T = [3 \quad 4], \qquad r_3^T = [5 \quad 6].$$

Tương tự, ta cũng có *phân tích dạng cột* của ma trận A là một tập hợp các vector cột:

$$A \in \mathbb{R}^{m\times n} \Longleftrightarrow A = [c_1|\dots|c_n], c_k \in \mathbb{R}^m. \tag{1.1.3}$$

Ở ví dụ trên, ta đặt c_1 và c_2 lần lượt là cột thứ nhất và cột thứ hai của A:

$$c_1 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \qquad c_2 = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}.$$

Phân tích dạng dòng và cột của một ma trận là các trường hợp đặc biệt của việc tạo khối ma trận. Tổng quát, ta có phân tích dạng dòng và cột của ma trận A có m dòng và n cột

$$A = \begin{matrix} \begin{bmatrix} A_{11} & \dots & A_{1r} \\ \vdots & & \vdots \\ A_{q1} & \cdots & A_{qr} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_q \end{matrix} \\ \begin{matrix} n_1 & & n_r \end{matrix} \end{matrix}$$

trong đó $m_1 + \ldots + m_q = m, n_1 + \ldots + n_r = n$, và $A_{\alpha\beta}$ ký hiệu khối (α, β) (ma trận con) có số chiều là $m_\alpha \times n_\beta$ và ta nói $A = (A_{\alpha\beta})$ là một ma trận khối $q \times r$.
Ta sử dụng các số hạng này để miêu tả các cấu trúc dải phổ biến cho các ma trận với các khối tương tự như các phần tử vô hướng. Do đó,

$$diag(A_{11}, A_{22}, A_{33}) = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A_{33} \end{bmatrix}$$

là đường chéo khối,

$$L = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix}, U = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix}, T = \begin{bmatrix} T_{11} & T_{12} & 0 \\ T_{21} & T_{22} & T_{23} \\ 0 & T_{32} & T_{33} \end{bmatrix},$$

lần lượt là *ma trận tam giác khối dưới, tam giác khối trên*, và *ba đường chéo khối.*

1.1.10 Các phép toán ma trận khối

Các ma trận khối có thể được nhân với vô hướng và chuyển vị:

$$\mu \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} = \begin{bmatrix} \mu A_{11} & \mu A_{12} \\ \mu A_{21} & \mu A_{22} \\ \mu A_{31} & \mu A_{32} \end{bmatrix},$$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix}^T = \begin{bmatrix} A_{11}^T & A_{21}^T & A_{31}^T \\ A_{12}^T & A_{22}^T & A_{32}^T \end{bmatrix}.$$

Chú ý, chuyển vị của khối (i,j) trở thành khối (j,i). Tương tự, ta có phép cộng hai ma trận khối

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \\ B_{31} & B_{32} \end{bmatrix} = \begin{bmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ A_{21}+B_{21} & A_{22}+B_{22} \\ A_{31}+B_{31} & A_{32}+B_{32} \end{bmatrix}.$$

Phép nhân hai ma trận khối cần nhiều điều kiện về số chiều. Chẳng hạn, nếu

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \\ A_{31}B_{11}+A_{32}B_{21} & A_{31}B_{12}+A_{32}B_{22} \end{bmatrix}$$

thì số cột của A_{11}, A_{21} và A_{31} phải bằng với số dòng của cả B_{11} và B_{12}. Tương tự, số cột của A_{12}, A_{22} và A_{32} phải bằng với số dòng của cả B_{21} và B_{22}.

Mỗi khi cộng hoặc nhân một ma trận khối thì số dòng và cột của các khối thỏa mãn tất cả các ràng buộc cần thiết. Trong trường hợp đó, ta nói các toán hạng được *phân tích đúng với* định lý theo sau.

Định lý 1.1.1 *Nếu*

$$A = \begin{bmatrix} A_{11} & \dots & A_{1s} \\ \vdots & & \vdots \\ A_{q1} & \cdots & A_{qs} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_q \end{matrix}, \qquad B = \begin{bmatrix} B_{11} & \dots & B_{1r} \\ \vdots & & \vdots \\ B_{s1} & \cdots & B_{sr} \end{bmatrix} \begin{matrix} p_1 \\ \\ p_s \end{matrix}$$
$$\begin{matrix} p_1 & & p_s \end{matrix} \qquad\qquad \begin{matrix} n_1 & & n_r \end{matrix}$$

và phân tích tích $C = AB$ như sau,

$$C = \begin{bmatrix} C_{11} & \dots & C_{1r} \\ \vdots & & \vdots \\ C_{q1} & \cdots & C_{qr} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_q \end{matrix}$$
$$\begin{matrix} n_1 & & n_r \end{matrix}$$

thì với $\alpha = 1:q$ và $\beta = 1:r$ *ta có* $C_{\alpha\beta} = \sum_{\gamma=1}^{s} A_{\alpha\gamma}B_{\gamma\beta}$.

Chứng minh. Giả sử $1 \leqslant \alpha \leqslant q$ và $1 \leqslant \beta \leqslant r$. Đặt $M = m_1 + \ldots + m_{\alpha-1}$ và $N = n_1 + \ldots + n_{\beta-1}$. Nếu $1 \leqslant i \leqslant m_\alpha$ và $1 \leqslant j \leqslant n_\beta$ thì

$$\begin{aligned}[C_{\alpha\beta}]_{ij} &= \sum_{k=1}^{p_1+\ldots+p_s} a_{M+i,k}b_{k,N+j} = \sum_{\gamma=1}^{s} \sum_{p_1+\ldots+p_{\gamma-1}+1}^{p_1+\ldots+p_\gamma} a_{M+i,k}b_{k,N+j} \\ &= \sum_{\gamma=1}^{s}\sum_{k=1}^{p_\gamma} [A_{\alpha\gamma}]_{ik}[B_{\gamma\beta}]_{kj} = \sum_{\gamma=1}^{s} [A_{\alpha\gamma}B_{\gamma\beta}]_{ij} = \left[\sum_{\gamma=1}^{s} A_{\alpha\gamma}B_{\gamma\beta}\right]_{ij}.\end{aligned}$$

Do đó, $C_{\alpha\beta} = A_{\alpha,1}B_{1,\beta} + \ldots + A_{\alpha,s}B_{s,\beta}$.

Nếu $A_{11}B_{11} + A_{12}B_{21} \neq B_{11}A_{11} + B_{21}A_{12}$ thì thao tác trên ma trận khối chính là thao tác trên ma trận ban đầu với a_{ij} và b_{ij} được viết như là A_{ij} và B_{ij}.

1.1.11 Các ma trận con

Cho $A \in \mathbb{R}^{m\times n}$. Nếu $\alpha = [\alpha_1, \ldots, \alpha_s]$ và $\beta = [\beta_1, \ldots, \beta_t]$ là các vector nguyên với các phần tử phân biệt thỏa mãn $1 \leqslant \alpha_i \leqslant m$ và $1 \leqslant \beta_i \leqslant n$ thì

$$A(\alpha, \beta) = \begin{bmatrix} a_{\alpha_1,\beta_1} & \cdots & a_{\alpha_1,\beta_t} \\ \vdots & \ddots & \vdots \\ a_{\alpha_s,\beta_1} & \cdots & a_{\alpha_s,\beta_t} \end{bmatrix}$$

là ma trận con của A có s dòng và t cột. Ví dụ, cho $A \in \mathbb{R}^{8\times 6}, \alpha = \begin{bmatrix} 2 & 4 & 6 & 8 \end{bmatrix}$, và $\beta = \begin{bmatrix} 4 & 5 & 6 \end{bmatrix}$,

$$A(\alpha, \beta) = \begin{bmatrix} a_{24} & a_{25} & a_{26} \\ a_{44} & a_{45} & a_{46} \\ a_{64} & a_{65} & a_{66} \\ a_{84} & a_{85} & a_{86} \end{bmatrix}.$$

Nếu $\alpha = \beta$ thì $A(\alpha, \beta)$ là *ma trận con chính*. Nếu $\alpha = \beta = 1 : k$ và $1 \leqslant k \leqslant min\{m, n\}$ thì $A(\alpha, \beta)$ là *ma trận con chính dẫn đầu*
Nếu $A \in \mathbb{R}^{m\times n}$ và

$$A = \begin{bmatrix} A_{11} & \ldots & A_{1s} \\ \vdots & & \vdots \\ A_{q1} & \cdots & A_{qs} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_q \end{matrix}$$
$$\begin{matrix} n_1 & & n_s \end{matrix}$$

thì ký hiệu dấu hai chấm có thể được sử dụng để xác định các khối riêng biệt. Đặc biệt,

$$A_{ij} = A(\tau + 1 : \tau + m, \mu + 1 : \mu + n_j)$$

trong đó $\tau = m_1 + \ldots + m_{i-1}$ và $\mu = n_1 + \ldots + n_{j-1}$.

1.2 Phép nhân ma trận với vector

1.2.1 Định nghĩa

Cho x là một vector cột n chiều và cho A là ma trận có $m \times n$ chiều. Khi đó, tích ma trận với vector $b = Ax$ là một vector cột m chiều xác định như sau:

$$b_i = \sum_{j=1}^{n} a_{ij}x_j, \quad i = 1, \ldots, m. \tag{1.2.1}$$

với b_i, a_{ij} và x_j tương ứng là các phần tử của b, A và x. Nếu các giá trị được cho này nằm trong $\mathbb{C}$ thì không gian của vector m chiều là $\mathbb{C}^m$ và không gian của các ma trận $m \times n$ là $\mathbb{C}^{m\times n}$.

Ánh xạ $x \mapsto Ax$ là *tuyến tính*, nghĩa là $x, y \in \mathbb{C}^n$ và $\alpha \in \mathbb{C}$ bất kỳ,

$$A(x + y) = Ax + Ay,$$
$$A(\alpha x) = \alpha Ax.$$

Ngược lại, mọi ánh xạ tuyến tính từ $\mathbb{C}^n$ vào $\mathbb{C}^m$ có thể được biểu diễn như phép nhân ma trận $m \times n$.

1.2.2 Nhân ma trận với vector

Cho a_j là cột thứ j của A và cũng là một vector m chiều. Khi đó (1.2.1) có thể được viết lại

$$b = Ax = \sum_{j=1}^{n} x_j a_j. \tag{1.2.2}$$

và b cũng có thể được viết dưới dạng tổ hợp tuyến tính của các cột trong ma trận A như sau:

$$\left[\begin{array}{c} b \end{array}\right] = \left[\begin{array}{c|c|c|c} a_1 & a_2 & \dots & a_n \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \left[a_1\right] + x_2 \left[a_2\right] + \ldots + x_n \left[a_n\right].$$

Ví dụ 1.2.1.(**ma trận Vandermonde**) Cố định một chuỗi các số $\{x_1, x_2, \ldots, x_m\}$. Nếu p, q là các đa thức bậc nhỏ hơn n và α là một vô hướng, thì $p + q$ và αp cũng là các đa thức bậc nhỏ hơn n. Hơn nữa, các giá trị của các đa thức này tại các điểm x_i thỏa mãn các tính chất tuyến tính sau:

$$(p + q)(x_i) = p(x_i) + q(x_i)$$
$$(\alpha p)(x_i) = \alpha(p(x_i)).$$

Do đó, ánh xạ từ các vector của các hệ số của các đa thức p có bậc nhỏ hơn n vào các vector $(p(x_1), p(x_2), \ldots, p(x_m))$ là tuyến tính. Ánh xạ tuyến tính bất kỳ có thể được biểu diễn như phép nhân ma trận. Thực vậy, nó biểu diễn bằng ma trận Vandermonde như sau

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{n-1} \end{bmatrix}.$$

Nếu c là vector cột của các hệ số của p,

$$c = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{bmatrix}, \quad p(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_{n-1} x^{n-1},$$

thì tích Ac

$$(Ac)_i = c_0 + c_1 x_i + c_2 x_i^2 + \ldots + c_{n-1} x_i^{n-1} = p(x_i), \quad \forall i = 1, \ldots, m. \tag{1.2.3}$$

Trong ví dụ này, tích Ac không cần thông qua m phép cộng vô hướng phân biệt mà mỗi phép cộng là một tổ hợp tuyến tính khác nhau của các phần tử của c (như (1.2.1)). Hơn nữa, A có thể được xem như là một ma trận của các cột mà mỗi cột cho các giá trị được lấy của cùng một đơn thức,

$$A = \left[1 \middle| x \middle| x^2 \middle| \ldots \middle| x^{n-1} \right], \tag{1.2.4}$$

và tích Ac là một tổ hợp tuyến tính của các đơn thức này,

$$Ac = c_0 + c_1 x + c_2 x^2 + \ldots + c_{n-1} x^{n-1} = p(x).$$

1.2.3 Nhân ma trận với ma trận

Cho A là ma trận $l \times m$ và C là ma trận $m \times n$, thì $B = AC$ là ma trận $l \times n$, với các phần tử xác định bởi

$$b_{ij} = \sum_{k=1}^{m} a_{ik} c_{kj} \tag{1.2.5}$$

với b_{ij}, a_{ik} và c_{kj} tương ứng là các phần tử của B, A và C. Viết dưới dạng cột, ta được tích như sau:

$$\left[b_1 \middle| b_2 \middle| \ldots \middle| b_n \right] = \left[a_1 \middle| a_2 \middle| \ldots \middle| a_m \right] \left[c_1 \middle| c_2 \middle| \ldots \middle| c_n \right],$$

và (1.2.5) trở thành

$$b_j = Ac_j = \sum_{k=1}^{m} c_{kj} a_k. \tag{1.2.6}$$

Do đó, b_j là tổ hợp tuyến tính của các cột a_k với các hệ số c_{kj}.

Ví dụ 1.2.2. (Tích ngoài). Tích của vector cột u có m chiều với vector dòng v có n chiều là một ma trận $m \times n$ hạng 1 như sau

$$\left[u \right] \left[v_1 \middle| v_2 \middle| \ldots \middle| v_n \right] = \left[v_1 u \middle| v_2 u \middle| \ldots \middle| v_n u \right] = \begin{bmatrix} v_1 u_1 & \ldots & v_n u_1 \\ \vdots & & \vdots \\ v_1 u_m & \ldots & v_n u_m \end{bmatrix}.$$

Các cột là bội của cùng vector u và tương tự, các dòng là bội của cùng vector v.

Ví dụ 1.2.3. Xét $B = AR$, với R là ma trận tam giác trên $n \times n$ mà $r_{ij} = 1$ với $i \leqslant j$ và $r_{ij} = 0$ với $i > j$. Khi đó, B được viết như sau

$$\left[\begin{array}{c|c|c} b_1 & \dots & b_n \end{array}\right] = \left[\begin{array}{c|c|c} a_1 & \dots & a_n \end{array}\right] \begin{bmatrix} 1 & \dots & 1 \\ & \ddots & \vdots \\ & & 1 \end{bmatrix}.$$

Công thức (1.2.6) cho

$$b_j = Ar_j = \sum_{k=1}^{j} a_k. \tag{1.2.7}$$

hay cột thứ j của B là tổng của j cột đầu tiên của A.

1.2.4 Range và không gian đầy đủ

Range của ma trận A (được ký hiệu là $range(A)$) là tập hợp các vector có dạng Ax với x bất kỳ. Công thức (1.2.2) cho ta một đặc trưng của $range(A)$.

Định lý 1.2.1 *$range(A)$ là không gian được sinh bởi các cột của A.*

Chứng minh. Do (1.2.2), Ax bất kỳ là một tổ hợp tuyến tính các cột của A. Ngược lại, vector y bất kỳ trong không gian sinh bởi các cột của A có thể được viết như là một tổ hợp tuyến tính của các cột, $y = \sum_{j=1}^{n} x_j a_j$. Do đó, y nằm trong $range(A)$.

Trong Định lý 1.2.1, range của ma trận A cũng được gọi là *không gian cột của A*.

Không gian đầy đủ của $A \in \mathbb{C}^{m\times n}$ (được ký hiệu là $null(A)$) là tập hợp các vector x thỏa mãn $Ax = 0$, với 0 là vector không trong $\mathbb{C}^m$. Các phần tử của mỗi vector $x \in null(A)$ cho khai triển các hệ số của 0 như là một tổ hợp tuyết tính các cột của ma trận A: $0 = x_1a_1 + x_2a_2 + \ldots + x_na_n$.

1.2.5 Hạng

Hạng cột của một ma trận là số chiều không gian cột của nó. Tương tự, *hạng dòng* của một ma trận là số chiều của không gian sinh bởi các dòng của nó. Hạng dòng thường bằng với hạng cột nên để đơn gian ta gọi là *hạng* của một ma trận.

Ma trận *hạng đầy đủ* $m \times n$ là ma trận có hạng có thể lớn nhất (nhỏ hơn m và n). Nghĩa là một ma trận hạng đầy đủ với $m \geqslant n$ phải có n cột độc lập tuyến tính. Ma trận như vậy cũng được xác định bởi tính chất mà ánh xạ xác định nó là đơn ánh.

Định lý 1.2.2 *Ma trận $A \in \mathbb{C}^{m\times n}$ với $m \geqslant n$ có hạng đầy đủ nếu và chỉ nếu nó ánh xạ 2 vector không phân biệt thành cùng một vector.*

Chứng minh. ($\Longrightarrow$) Nếu A là một ma trận hạng đầy đủ thì các cột của nó là độc lập tuyến tính, nên chúng hình thành một cơ sở cho $range(A)$. Nghĩa là mọi $b \in range(A)$ có duy nhất mở rộng tuyến tính các cột của A. Do đó, theo (1.2.2), mọi $b \in range(A)$ có duy nhất x sao cho $b = Ax$.
($\Longleftarrow$) Ngược lại, nếu A không có hạng đầy đủ thì các cột a_j của nó là phụ thuộc tuyến tính, và $\sum_{j=1}^{n} c_j a_j = 0$ là một tổ hợp tuyến tính không tầm thường. Vector c khác 0 hình thành từ các hệ số c_j thỏa $Ac = 0$. Nhưng khi đó A ánh xạ các vector phân biệt thành cùng một vector vì với x bất kỳ, $Ax = A(x + c)$.

1.2.6 Nghịch đảo

Ma trận *khả nghịch* hoặc *không suy biến* là ma trận vuông có hạng đầy đủ. Vì m cột của một ma trận không suy biến $m \times m$ tạo thành một cơ sở cho toàn bộ không gian $\mathbb{C}^m$ nên một vector bất kỳ được biểu diễn duy nhất dưới dạng là một tổ hợp tuyến tính của chúng. Đặc biệt, vector đơn vị chính tắc e_j với 1 ở vị trí thứ j và 0 ở những vị trí còn lại

$$e_j = \sum_{i=1}^{m} z_{ij} a_i. \tag{1.2.8}$$

Cho Z là ma trận với các phần tử là z_{ij}, và cho z_j là cột thứ j của Z. Khi đó, (1.2.8) có thể được viết $e_j = Az_j$. Phương trình này có dạng của (1.2.6) và được viết lại như sau

$$\left[e_1 \middle| \dots \middle| e_m \right] = I = AZ,$$

với I là ma trận *đơn vị* $m \times m$. Ma trận Z là *ma trận nghịch đảo* của A. Ma trận không suy biến vuông A bất kỳ có duy nhất một nghịch đảo, được ký hiệu bởi A^{-1}, thỏa $AA^{-1} = A^{-1}A = I$.
Định lý sau cho số điều kiện tương đương khi ma trận vuông A không suy biến.

Định lý 1.2.3 *Cho $A \in \mathbb{C}^{m \times m}$, các điều kiện sau là tương đương:*

(a) A có nghịch đảo A^{-1},

(b) $rank(A) = m$,

(c) $range(A) = \mathbb{C}^m$,

(d) $null(A) = \{0\}$,

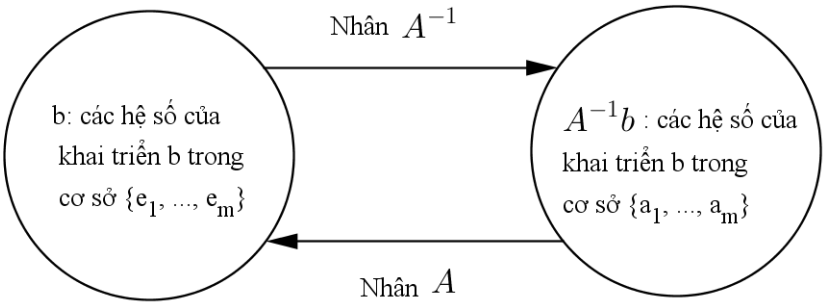
(e) 0 không là trị riêng của A,

(f) 0 không là giá trị suy biến của A,

(g) $det(A) \neq 0$.

1.2.7 Nhân ma trận nghịch đảo với một vector

Theo (1.2.6), tích $x = A^{-1}b$ là vector khai triển tuyến tính duy nhất các hệ số của b trong cơ sở các cột của A. Nhân với A^{-1} là một phép toán *chuyển cơ sở*



1.3 Vector và ma trận trực giao

1.3.1 Phụ hợp

Liên hợp phức của một vô hướng z, được ký hiệu bởi $\overline{z}$ hoặc z^*, có được bằng việc phủ định phần ảo của nó. Cho z là số thực, $\overline{z} = z$.
Liên hợp Hermit hay *phụ hợp* của một ma trận A có kích thước $m \times n$, được ký hiệu bởi A^*, là ma trận $n \times m$ mà phần tử i, j của nó là liên hợp phức của phần tử j, i của A. Ví dụ,

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \Longrightarrow A^* = \begin{bmatrix} \overline{a_{11}} & \overline{a_{21}} & \overline{a_{31}} \\ \overline{a_{12}} & \overline{a_{22}} & \overline{a_{32}} \end{bmatrix}$$

Nếu $A = A^*$ thì A là ma trận *hermit*. Theo định nghĩa, một ma trận hermit phải là ma trận vuông. Cho A là ma trận thực, ma trận phụ hợp chỉ đơn giản là hoán đổi các dòng và các cột của A. Trong trường hợp này, ma trận phụ hợp cũng là *ma trận chuyển vị* A^T. Do đó, nếu một ma trận thực là hermit, nghĩa là $A = A^T$, thì nó cũng là *ma trận đối xứng*.

1.3.2 Tích trong

Tích trong của hai vector cột $x, y \in \mathbb{C}^m$ là tích phụ hợp của x với y:

$$x^* y = \sum_{i=1}^{m} \overline{x_i} y_i. \tag{1.3.1}$$

Độ dài Euclidean của x được ký hiệu là $\|x\|$ (chuẩn vector) và được xác định như căn bậc hai của tích trong của x với chính nó:

$$\|x\| = \sqrt{x^*x} = (\sum_{i=1}^{m} |x_i|^2)^{1/2}. \tag{1.3.2}$$

Cos của góc α giữa x và y được biểu diễn trong các số hạng của tích trong:

$$\cos\alpha = \frac{x^*y}{\|x\|\|y\|}. \tag{1.3.3}$$

Tích trong là *song tuyến tính*, nghĩa là nó tuyến tính theo từng vector riêng biệt

$$\begin{aligned}(x_1 + x_2)^*y &= x_1^*y + x_2^*y,\\ x^*(y_1 + y_2) &= x^*y_1 + x^*y_2,\\ (\alpha x)^*(\beta y) &= \overline{\alpha}\beta x^*y.\end{aligned}$$

Ta cũng sẽ thường xuyên sử dụng tính chất này cho các ma trận hay các vector bất kỳ A và B có các chiều tương thích,

$$(AB)^* = B^*A^*. \tag{1.3.4}$$

Tương tự cho tích của các ma trận vuông khả nghịch,

$$(AB)^{-1} = B^{-1}A^{-1}. \tag{1.3.5}$$

Ký hiệu A^{-*} là một tốc ký của $(A^*)^{-1}$ hay $(A^{-1})^*$; hai ký hiệu này là tương đương, được kiểm tra bằng việc áp dụng (1.3.4) với $B = A^{-1}$.

1.3.3 Các vector trực giao

Hai vector x và y được gọi là *trực giao* nếu $x^*y = 0$. Nếu x và y là các vector thực thì điều này có nghĩa là chúng nằm vuông góc với nhau trong $\mathbb{R}^m$. Tập hợp các vector X và tập hợp các vector Y là trực giao (hay X trực giao Y) nếu với mọi $x \in X$ trực giao với mọi $y \in Y$.

Tập hợp các vector S khác không là *trực giao* nếu các phần tử của nó là trực giao từng đôi một, nghĩa là, nếu $x, y \in S, x \neq y \Longrightarrow x^*y = 0$. Tập hợp các vector là *trực chuẩn* nếu nó trực giao và với mọi $x \in S$, $\|x\| = 1$.

Định lý 1.3.1 *Các vector trong một tập hợp trực giao S là độc lập tuyến tính.*

Chứng minh. Nếu các vector trong S không độc lập tuyến tính thì tồn tại $v_k \in S$ bất kỳ được biểu diễn dưới dạng tổ hợp tuyến tính của $v_1, \ldots, v_n \in S$,

$$v_k = \sum_{\substack{i=1\\ i\neq k}}^{n} c_i v_i.$$

Vì $v_k \neq 0, v_k^* v_k = \|v_k\|^2 > 0$. Sử dụng tính song tuyến tính của tích trong và tính trực giao của S, ta có

$$v_k^* v_k = \sum_{\substack{i=1 \\ i \neq k}}^{n} c_i v_k^* v_i = 0,$$

mâu thuẫn với giả thuyết các vector trong S là khác không.

Như là một hệ quả của Định lý 1.3.1, nếu một tập trực giao $S \subseteq \mathbb{C}^m$ chứa m vector thì nó là một cơ sở của $\mathbb{C}^m$.

1.3.4 Các thành phần của một vector

Ý tưởng quan trọng nhất từ các khái niệm của tích trong và trực giao là các tích trong có thể được sử dụng để phân tích các vector tùy ý thành các thành phần trực giao.

Ví dụ 1.3.1. Giả sử $\{q_1, q_2, \ldots, q_n\}$ là một tập trực giao, v là một vector bất kỳ. Con số $q_j^* v$ là một vô hướng. Khi đó, ta có vector

$$r = v - (q_1^* v)q_1 - (q_2^* v)q_2 - \ldots - (q_n^* v)q_n \tag{1.3.6}$$

là trực giao với $\{q_1, q_2, \ldots, q_n\}$. Điều này có thể được kiểm tra bằng việc tính $q_i^* v$

$$q_i^* v = q_i^* v - (q_1^* v)(q_i^* q_1) - \ldots - (q_n^* v)(q_i^* q_n).$$

Vì $q_i^* q_j = 0$ với $i \neq j$ nên tổng này được rút gọn như sau

$$q_i^* v = q_i^* v - (q_i^* v)(q_i^* q_i) = 0.$$

Do đó ta thấy rằng v được phân tích thành $n+1$ thành phần trực giao:

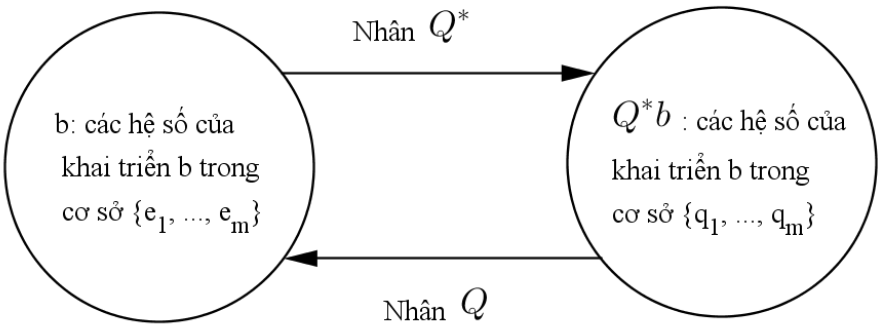
$$v = r + \sum_{i=1}^{n} (q_i^* v)q_i = r + \sum_{i=1}^{n} (q_i q_i^*)v. \tag{1.3.7}$$

Trong phân tích này, r là phần của v trực giao với tập các vector $\{q_1, q_2, \ldots, q_n\}$, hay không gian con sinh bởi tập các vector này và $(q_i^* v)q_i$ là phần của v trong phương của q_i.

Nếu $\{q_i\}$ là cơ sở của $\mathbb{C}^m$ thì n phải bằng m và r phải là vector không, nên v được phân tích đầy đủ thành m thành phần trực giao trong các phương của q_i:

$$v = \sum_{i=1}^{m} (q_i^* v)q_i = \sum_{i=1}^{m} (q_i q_i^*)v. \tag{1.3.8}$$

Cả hai công thức (1.3.7) và (1.3.8), ta đã viết công thức trong hai cách khác nhau, một với $(q_i^* v)q_i$ và một với $(q_i q_i^*)v$, là tương đương nhau nhưng chúng có các giải thích khác nhau. Trong trường hợp đầu tiên, ta xem v như một tổng của các hệ số $q_i^* v$ nhân với các vector q_i. Trong trường hợp hai, ta xem v như một tổng của các phép chiếu trực giao của v vào các phương khác nhau q_i. Phép chiếu thứ i được thực hiện bởi ma trận hạng một rất đặc biệt $q_i q_i^*$.



1.3.5 Các ma trận Unita

Ma trận vuông $Q \in \mathbb{C}^{m\times m}$ là *unita* (trong trường hợp thực, *trực giao*) nếu $Q^* = Q^{-1}$, nghĩa là, nếu $Q^*Q = I$. Trong các dạng cột của Q, tích này được viết như sau

$$\begin{bmatrix} q_1^* \\ q_2^* \\ \vdots \\ q_m^* \end{bmatrix} \quad \begin{bmatrix} q_1 \Big| q_2 \Big| \cdots \Big| q_m \end{bmatrix} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & \ddots \\ & & & 1 \end{bmatrix}.$$

Mặt khác, $q_i^* q_j = \delta_{ij}$, và các cột của một ma trận unita Q tạo thành một cơ sở trực giao của $\mathbb{C}^m$. Ký hiệu δ_{ij} là *Kronecker delta*, $\delta_{ij} = 1$ nếu $i = j$ và $\delta_{ij} = 0$ nếu $i \neq j$.

1.3.6 Nhân với ma trận Unita

Nếu A là một ma trận unita Q thì Ax và $A^{-1}b$ trở thành Qx và Q^*b. Như trong những mục trước, Qx là tổ hợp tuyến tính của các cột của Q với các hệ số x. Ngược lại,

Q^*b là vector các hệ số của khai triển b trong cơ sở các cột của Q.

Các quá trình của phép nhân này với một ma trận Unita hoặc phụ hợp của nó bảo toàn cấu trúc hình học trong ý nghĩa Euclidean bởi vì các tích trong được bảo toàn. Đó là, với ma trận unita Q,

$$(Qx)^*(Qy) = x^*y, \tag{1.3.9}$$

được kiểm tra bởi (1.3.4). Tính bất biến của tích trong có nghĩa là các góc giữa các vector được bảo toàn, và chiều dài của chúng là

$$\|Qx\| = \|x\|. \tag{1.3.10}$$

Trong trường hợp thực, phép nhân với ma trận trực giao Q tương ứng với phép quay cố định (nếu $detQ = 1$) hoặc phép đối xứng (nếu $detQ = -1$) của không gian vector.

1.4 Trực chuẩn

1.4.1 Các chuẩn vector

Một *chuẩn* là một hàm $\|\cdot\| : \mathbb{C}^m \to \mathbb{R}$ thỏa mãn 3 điều kiện theo sau: Với mọi vector x và y và với mọi vô hướng $\alpha \in \mathbb{C}$,

$$\begin{aligned} &(1) \quad \|x\| \geqslant 0, \text{ và } \|x\| = 0 \text{ nếu } x = 0, \\ &(2) \quad \|x + y\| \leqslant \|x\| + \|y\|, \\ &(3) \quad \|\alpha x\| = |\alpha|\|x\|. \end{aligned} \tag{1.4.1}$$

Trong đó, các điều kiện này yêu cầu rằng (1) chuẩn của một vector khác không là dương, (2) chuẩn của một tổng vector là không vượt quá tổng các chuẩn của các phần của nó - *bất đẳng thức tam giác*, và (3) co giãn toàn bộ các phần tử của một vector bằng một hằng số α thì chuẩn của nó cũng co giãn theo giá trị tuyệt đối của hằng số đó.
Quả cầu đơn vị đóng $\{x \in \mathbb{C}^m : \|x\| \leqslant 1\}$ tương ứng với mỗi chuẩn được minh họa ở hình bên dưới cho trường hợp $m = 2$.

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{m} |x_i|, \\ \|x\|_2 &= \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2} = \sqrt{x^* x}, \\ \|x\|_\infty &= \max_{1 \leqslant i \leqslant m} |x_i|, \\ \|x\|_p &= \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p} \qquad (1 \leqslant p < \infty). \end{aligned} \tag{1.4.2}$$

Tổng quát, cho chuẩn $\|\cdot\|$ bất kỳ, một chuẩn có trọng số có thể được viết như

$$\|x\|_W = \|Wx\|. \tag{1.4.3}$$

W ở đây là ma trận đường chéo mà phần tử đường chéo ở vị trí thứ i là trọng số $w_i \neq 0$. Ví dụ, chuẩn 2 có trọng số $\|\cdot\|_W$ trong $\mathbb{C}^m$ được thiết lập như sau:

$$\|x\|_W = \left(\sum_{i=1}^{m} |w_i x_i|^2 \right)^{1/2}. \tag{1.4.4}$$

Ta cũng có thể tổng quát hóa ý tưởng các chuẩn có trọng số bằng việc cho phép W là một ma trận không suy biến tùy ý, không cần thiết là đường chéo.

1.4.2 Các chuẩn ma trận bao gồm các chuẩn vector

Ma trận $m \times n$ có thể được xem như một vector trong không gian mn chiều mà mỗi phần tử mn của ma trận là một tọa độ độc lập. Chuẩn mn chiều bất kỳ có thể được sử dụng cho việc đo "kích thước" của một ma trận như vậy.

Cho các chuẩn vector $\|\cdot\|_{(n)}$ và $\|\cdot\|_{(m)}$ trong miền xác định và range của $A \in \mathbb{C}^{m\times n}$ tương ứng, chuẩn ma trận được bao gồm $\|A\|_{(m,n)}$ là số nhỏ nhất C sao cho bất đẳng thức sau đúng với mọi $x \in \mathbb{C}^n$:

$$\|Ax\|_{(m)} \leqslant C\|x\|_{(n)}. \tag{1.4.5}$$

Mặt khác, $\|A\|_{(m,n)}$ là cận trên của tỉ số $\|Ax\|_{(m)}/\|x\|_{(n)}$ với mọi vector $x \in \mathbb{C}^n$ - thừa số lớn nhất mà A có thể "giãn" một vector x. Ta nói rằng $\|\cdot\|_{(m,n)}$ là chuẩn ma trận được bao gồm bởi $\|\cdot\|_{(n)}$ và $\|\cdot\|_{(m)}$.

Bởi vì điều kiện (3) của (1.4.1), tác động của A được xác định bởi tác động của nó trong các vector đơn vị. Do đó, chuẩn ma trận có thể được xác định một cách tương đương với các ảnh của các vector đơn vị dưới A:

$$\|Ax\|_{(m,n)} = \sup_{\substack{x\in\mathbb{C}^n \\ x\neq 0}} \frac{||\mathrm{Ax}||_{(m)}}{||x||_{(n)}} = \sup_{\substack{x\in\mathbb{C}^n \\ ||x||_{(n)}=1}} ||\mathrm{Ax}||_{(m)}. \tag{1.4.6}$$

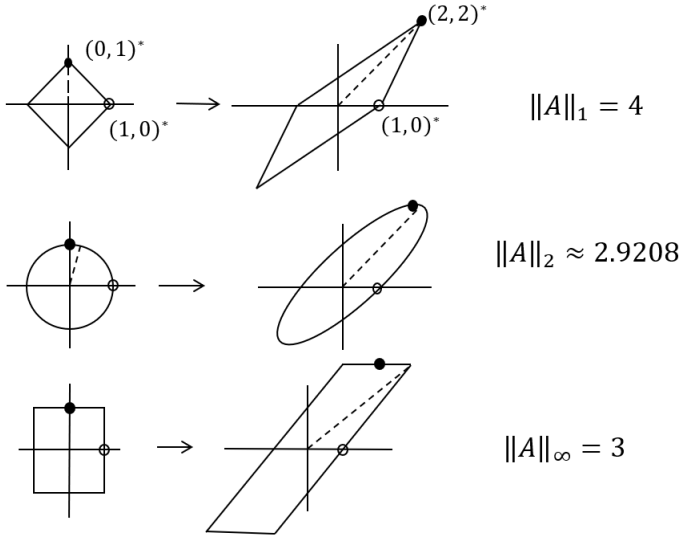
1.4.3 Các ví dụ

Ví dụ 1.4.1. Ma trận

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix} \tag{1.4.7}$$

ánh xạ từ $\mathbb{C}^2$ vào $\mathbb{C}^2$. Nó cũng ánh xạ từ $\mathbb{R}^2$ vào $\mathbb{R}^2$.

Hình 1.1 miêu tả tác động của A vào các quả cầu đơn vị của $\mathbb{R}^2$ xác định bởi chuẩn 1,



Hình 1.1: Các quả cầu đơn vị ứng với các chuẩn 1, 2 và ∞

chuẩn 2 và chuẩn ∞. Không quan tâm tới chuẩn, A ánh xạ $e_1 = (1,0)^*$ thành cột đầu tiên của A, cụ thể e_1 thành chính nó, và $e_2 = (0,1)^*$ thành cột thứ 2 của A, cụ thể là $(2,2)^*$. Trong chuẩn 1, vector đơn vị x được khuếch đại hầu hết bởi A là $(0,1)^*$ (hoặc phủ định của nó), và thừa số khuếch đại là 4. Trong chuẩn ∞, vector đơn vị x được

khuếch đại hầu hết bởi A là $(1,1)^*$ (hoặc phủ định của nó), và thừa số khuếch đại là 3. Trong chuẩn 2, vector đơn vị được khuếch đại hầu hết bởi A là vector được bao gồm bởi đường đứt nét trong hình (hoặc phủ định của nó), và thừa số khuếch đại là xấp xỉ 2.9208. (Chú ý rằng nó phải ít nhất là $\sqrt{8} \approx 2.8284$, vì $(0,1)^*$ ánh xạ thành $(2,2)^*$.)

Ví dụ 1.4.2.(Chuẩn p của ma trận đường chéo). Cho D là một ma trận đường chéo

$$D = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m \end{bmatrix}$$

Khi đó, trong dòng thứ hai của Hình 1.1, ảnh của hình cầu đơn vị chuẩn 2 của D là một ellip m chiều mà các chiều dài bán trục của nó được cho bởi các số $|d_i|$. Các vector đơn vị khuếch đại lớn nhất bởi D được ánh xa tới bán trục dài nhất của ellip, của chiều dài $max_i\{|d_i|\}$. Do đó, ta có $\|D\|_2 = max_{1 \leqslant i \leqslant m}\{|d_i|\}$. Tổng quá hóa kết quả này cho chuẩn p bất kỳ: nếu D là đường chéo thì $\|D\|_p = max_{1 \leqslant i \leqslant m}|d_i|$.

Ví dụ 1.4.3. (Chuẩn 1 của một ma trận). Nếu A là một ma trận $m \times n$ bất kỳ thì $\|A\|_1$ là bằng với "tổng cột lớn nhất" của A. Ta giải thích kết quả này như sau. Viết A dưới dạng các cột của nó

$$A = \left[a_1 \middle| \dots \middle| a_n \right], \tag{1.4.8}$$

với mỗi a_j là vector m chiều. Xét quả cầu đơn vị chuẩn 1 có hình dạng giống kim cương trong $\mathbb{C}^n$, được minh hoa như trong (1.4.2). Đó là tập hợp $\{x \in \mathbb{C}^n : \sum_{j=1}^n |x_j| \leqslant 1\}$. Vector Ax bất kỳ trong hình này thỏa mãn

$$\|Ax\|_1 = \|\sum_{j=1}^n x_j a_j\|_1 \leqslant \sum_{j=1}^n |x_j| \|a_j\|_1 \leqslant max_{1 \leqslant j \leqslant n} \|a_j\|_1.$$

Do đó, chuẩn 1 của ma trận được bao gồm thỏa mãn $\|A\|_1 \leqslant max_{1 \leqslant j \leqslant n} \|a_j\|_1$. Bằng việc chọn $x = e_j$ với j cực đại hóa $\|a_j\|_1$, ta thu được chuẩn ma trận như sau

$$\|A\|_1 = max_{1 \leqslant j \leqslant n} \|a_j\|_1. \tag{1.4.9}$$

Ví dụ 1.4.4. (Chuẩn ∞ của một ma trận). Chuẩn ∞ của một ma trận $m \times n$ là tương đương với "tổng dòng lớn nhất",

$$\|A\|_\infty = \max_{1 \leqslant i \leqslant m} \|a_i^*\|_1, \tag{1.4.10}$$

với a_i^* là dòng thứ i của A.

1.4.4 Bất đẳng thức Cauchy - Schwarz và Holder

Việc tính toán chuẩn p của ma trận với $p \neq 1, \infty$ là khó khăn hơn, và để xấp xỉ bài toán này, ta chú ý rằng các tích trong có thể được chặn bằng việc sử dụng chuẩn p. Cho p và q thỏa mãn $\frac{1}{p} + \frac{1}{q} = 1$, với $1 \leqslant p, q \leqslant \infty$. Khi đó, *bất đẳng thức Holder* phát biểu như sau: với các vector x và y bất kỳ,

$$|x^* y| \leqslant \|x\|_p \|y\|_q. \tag{1.4.11}$$

Bất đẳng thức Cauchy - Schwarz là trường hợp đặc biệt $p = q = 2$:

$$|x^* y| \leqslant \|x\|_2 \|y\|_2. \tag{1.4.12}$$

Các kết quả này có thể được tìm thấy trong các sách Đại số tuyến tính.

Ví dụ 1.4.5. (Chuẩn 2 của một vector dòng). Xét một ma trận A chứa một dòng đơn. Ma trận này có thể được viết như $A = a^*$, với a là một vector cột. Theo bất đẳng thức Cauchy - Schwarz, với x bất kỳ, ta có $\|Ax\|_2 = |a^* x| \leqslant \|a\|_2 \|x\|_2$ mà $\|Aa\|_2 = \|a\|_2^2$. Do đó, ta có

$$\|A\|_2 = \sup_{x \neq 0} \{\|Ax\|_2 / \|x\|_2\} = \|a\|_2.$$

Ví dụ 1.4.6. Chuẩn 2 của tích ngoài. Tổng quát, xét tích ngoài hạng một $A = uv^*$, với u là một vector m chiều và v là một vector n chiều. Cho vector n chiều x bất kỳ, ta có thể chặn $\|Ax\|_2$ như sau

$$\|Ax\|_2 = \|uv^* x\|_2 = \|u\|_2 |v^* x| \leqslant \|u\|_2 \|v\|_2 \|x\|_2. \tag{1.4.13}$$

Do đó $\|A\|_2 \leqslant \|u\|_2 \|v\|_2$. Dấu "=" xảy ra khi $x = v$.

1.4.5 Chặn của $\|AB\|$ trong chuẩn ma trận được bao gồm

Cho $\|\cdot\|_{(l)}, \|\cdot\|_{(m)}$ và $\|\cdot\|_{(n)}$ là các chuẩn tương ứng trong $\mathbb{C}^l, \mathbb{C}^m$, và $\mathbb{C}^n$, và cho A là một ma trận $l \times m$ và B là ma trận $m \times n$. Với $x \in \mathbb{C}^n$ bất kỳ, ta có

$$\|ABx\|_{(l)} \leqslant \|A\|_{(l,m)} \|Bx\|_{(m)} \leqslant \|A\|_{(l,m)} \|B\|_{(m,n)} \|x\|_{(n)}.$$

Do đó, chuẩn được bao gồm của AB phải thỏa mãn

$$\|AB\|_{(l,n)} \leqslant \|A\|_{(l,m)} \|B\|_{(m,n)}. \tag{1.4.14}$$

Tổng quát, bất đẳng thức này là không bằng nhau. Ví dụ, bất đẳng thức $\|A^n\| \leqslant \|A\|^n$ đúng với ma trận vuông bất kỳ trong chuẩn ma trận bất kỳ được bao gồm bởi một chuẩn vector, nhưng $\|A^n\| = \|A\|^n$ không đúng trong trường hợp tổng quát với $n \geqslant 2$.

1.4.6 Các chuẩn ma trận tổng quát

Tổng quát, một chuẩn ma trận phải thỏa mãn 3 điều kiện chuẩn vector (1.4.1) áp dụng trong không gian vector mn chiều của các ma trận:

$$\begin{aligned} &(1) \quad \|A\| \geqslant 0, \text{ và } \|A\| = 0 \text{ chỉ nếu } A = 0, \\ &(2) \quad \|A + B\| \leqslant \|A\| + \|B\|, \\ &(3) \quad \|\alpha A\| = |\alpha|\|A\|. \end{aligned} \tag{1.4.15}$$

Chuẩn ma trận phổ biến nhất không được bao gồm bởi một chuẩn vector là *chuẩn Hilbert - Schmidt* hay *chuẩn Frobenius*, xác định bởi

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}. \tag{1.4.16}$$

Công thức của chuẩn Frobenius cũng có thể được viết dưới dạng các cột hoặc các dòng riêng biệt. Ví dụ, nếu a_j là cột thứ j của A, ta có

$$\|A\|_F = \left(\sum_{j=1}^{n} \|a_j\|_2^2 \right)^{1/2}. \tag{1.4.17}$$

Kết quả tương tự cho các dòng được biểu diễn bởi phương trình sau:

$$\|A\|_F = \sqrt{tr(A^*A)} = \sqrt{tr(AA^*)}, \tag{1.4.18}$$

với $tr(B)$ là *vết* của B, tổng các phần tử trên đường chéo của nó.
Giống như chuẩn ma trận được bao gồm, chuẩn Frobenius có thể được sử dụng để chặn các tích của các ma trận. Cho $C = AB$ với các phần tử c_{ik}, và cho a_i^* là dòng thứ i của A và b_j là cột thứ j của B. Khi đó, $c_{ij} = a_i^* b_j$ nên theo bất đẳng thức Cauchy - Schwarz, ta có $|c_{ij}| \leqslant \|a_i\|_2 \|b_j\|_2$. Bình phương cả hai vế và tính tổng trên tất cả chỉ số i, j, ta được

$$\begin{aligned} \|AB\|_F^2 &= \sum_{i=1}^{n} \sum_{j=1}^{m} |c_{ij}|^2 \\ &\leqslant \sum_{i=1}^{n} \sum_{j=1}^{m} (\|a_i\|_2 \|b_j\|_2)^2 \\ &= \sum_{i=1}^{n} (\|a_i\|_2)^2 \sum_{j=1}^{m} (\|b_j\|_2)^2 = \|A\|_F^2 \|B\|_F^2. \end{aligned}$$

1.4.7 Bất biến dưới phép nhân Unita

Một trong số những tính chất đặc biệt của chuẩn 2 ma trận là tính bất biến dưới phép nhân các ma trận Unita. Tính chất này cũng đúng cho chuẩn Frobenius.

Định lý 1.4.1 *Cho $A \in \mathbb{C}^{m\times n}$ bất kỳ và ma trận Unita $Q \in \mathbb{C}^{m\times m}$, ta có*

$$\|QA\|_2 = \|A\|_2,\ \|QA\|_F = \|A\|_F.$$

Chứng minh. Vì $\|Qx\|_2 = \|x\|_2$ với mọi x (theo (1.3.10)) nên tính bất biến trong chuẩn 2 theo sau từ (1.4.6). Cho chuẩn Frobenius, ta có thể sử dụng (1.4.18).

Theo Định lý 1.4.1, nếu Q được tổng quát hóa thành ma trận hình chữ nhật với các cột trực giao, nghĩa là $Q \in \mathbb{C}^{p\times m}$ với $p > m$. Tương tự tính đồng nhất cũng đúng cho phép nhân các ma trận Unita trong vế phải, hoặc tổng quát hơn, nhân các ma trận hình chữ nhật với các dòng trực giao.

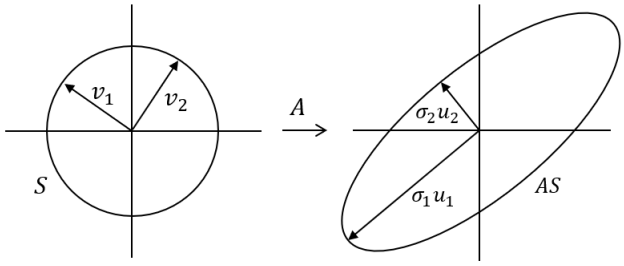
1.5 Phân tích giá trị suy biến

1.5.1 Quan sát hình học

Phân tích các giá trị suy biến (Singular Value Decomposition - SVD) được thúc đẩy bởi lập luận hình học sau:

Ảnh của một quả cầu đơn vị dưới ma trận $m \times n$ bất kỳ là một siêu ellip.

SVD có thể áp dụng được cho cả ma trận thực và ma trận phức. Tuy nhiên, trong mô tả hình học, ma trận sử dụng là ma trận thực.
Thuật ngữ "siêu ellip" có thể là xa lạ, nhưng đó là sự tổng quát hóa m chiều của một ellip. Ta có thể định nghĩa một siêu ellip trong $\mathbb{R}^m$ như là một mặt thu được bằng việc kéo căng quả cầu đơn vị trong $\mathbb{R}^m$ bởi các thừa số $\sigma_1, \ldots, \sigma_m$ (có thể là 0) trong các phương trực giao bất kỳ $u_1, \ldots, u_m \in \mathbb{R}^m$. Cho thuật lợi, ta lấy u_i là các vector đơn vị, nghĩa là $\|u_i\|_2 = 1$. Các vector $\{\sigma_i u_i\}$ là *các bán trục chính* của siêu ellip, với các độ dài $\sigma_1, \ldots, \sigma_m$. Nếu A có hạng r thì một cách chính xác r của các độ dài σ_i sẽ trả ra giá trị khác 0, và đặc biệt, nếu $m \geqslant n$ thì tối đa n trong số chúng sẽ là khác 0.
Ảnh của quả cầu đơn vị mà ta muốn nói là quả cầu Euclidean thông thường trong không gian n chiều, nghĩa là quả cầu đơn vị trong chuẩn 2; ta ký hiệu nó là S. Khi đó AS, ảnh của S dưới ánh xạ A, là một siêu ellip như vừa được xác định.



Hình 1.2: SVD của ma trận 2×2

Cho S là quả cầu đơn vị trong $\mathbb{R}^n$ và $A \in \mathbb{R}^{m\times n}$ bất kỳ với $m \geqslant n$. Để đơn giản, giả sử A có hạng đầy đủ là n. Ảnh AS là một siêu ellip trong $\mathbb{R}^m$. Bây giờ ta định nghĩa một vài tính chất của A liên quan tới hình dạng của AS. Ý tưởng chính được miêu tả trong Hình 1.2.

Đầu tiên, ta định nghĩa n *giá trị suy biến* của A. Các giá trị này là các độ dài của n bán trục chính của AS, được viết là $\sigma_1, \sigma_2, \ldots, \sigma_n$. Theo qui ước, giả sử rằng các giá trị suy biến được đánh số trong thứ tự giảm dần, $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n > 0$.

Tiếp theo, ta định nghĩa n *vector suy biến trái* của A. Các vector này là các vector đơn vị $\{u_1, u_2, \ldots, u_n\}$ trực giao với phương của các bán trục chính của AS, tương ứng với các giá trị suy biến. Do đó vector $\sigma_i u_i$ là bán trục chính lớn nhất thứ i của AS.

Cuối cùng, ta định nghĩa n *vector suy biến phải* của A. Các vector này là các vector đơn vị của $\{v_1, v_2, \ldots, v_n\} \in S$ mà chúng là ảnh ngược của các bán trục chính của AS, được ký hiệu bởi $Av_j = \sigma_j u_j$.

1.5.2 SVD được giảm

Như đề cập ở trên, ta có phương trình liên hệ giữa các vector suy biến phải $\{v_j\}$ với các vector suy biến trái $\{u_j\}$

$$Av_j = \sigma_j u_j,\ 1 \leqslant j \leqslant n. \tag{1.5.1}$$

Tập hợp các phương trình vector này có thể được biểu diễn như một phương trình ma trận, hay $AV = \widehat{U}\widehat{\Sigma}$. Trong đó, $\widehat{\Sigma}$ là ma trận đường chéo $n \times n$ với các phần tử thực dương (vì A được giả sử có hạng đầy đủ là n), $\widehat{U}$ là ma trận $m \times n$ với các cột trực giao, và V là ma trận $n \times n$ với các cột trực giao.

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix}$$

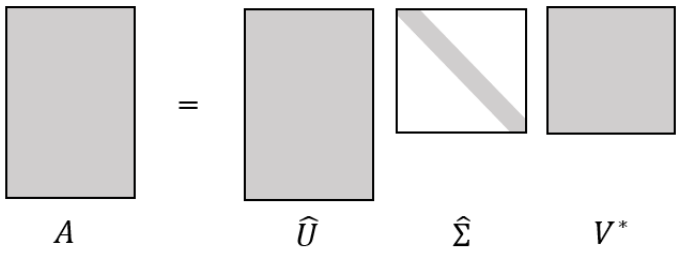
Do đó, V là ma trận Unita, và ta có thể nhân V^* bên phải nó để được

$$A = \widehat{U}\widehat{\Sigma}V^*. \tag{1.5.2}$$

Phân tích này của A được gọi là *phân tích giá trị suy biến được giảm* hay *SVD được giảm* của A. Dưới dạng biểu đồ, nó trông giống điều này

1.5.3 SVD đầy đủ

Các cột của $\widehat{U}$ là n vetor trực giao trong không gian m chiều $\mathbb{C}^m$. Trừ khi $m = n$, chúng không tạo thành một cơ sở của $\mathbb{C}^m$, $\widehat{U}$ cũng không là ma trận Unita. Tuy nhiên, bằng



Hình 1.3: SVD được giảm ($m \geqslant n$)

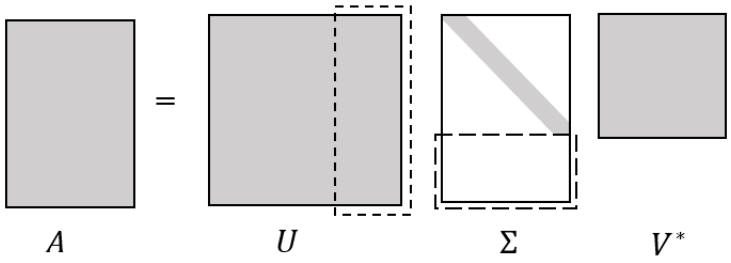
việc thêm vào $m-n$ cột trực giao, $\widehat{U}$ có thể được mở rộng thành một ma trận Unita.

Nếu $\widehat{U}$ được thay thế bởi U trong (1.5.2) thì $\widehat{\Sigma}$ sẽ phải thay đổi như vậy. Cho tích giữ nguyên không thay đổi, $m-n$ cột cuối của U sẽ được nhân với 0. Do đó, cho Σ là ma trận $m \times n$ gồm có $\widehat{\Sigma}$ trong $n \times n$ khối bên trên với $m-n$ dòng 0 bên dưới. Phân tích giá trị suy biến đầy đủ hay SVD đầy đủ của A

$$A = U\Sigma V^*. \tag{1.5.3}$$

với U là ma trân Unita $m \times m$, V là ma trận Unita $n \times n$, và Σ là ma trận đường chéo $m \times n$ với các phần tử thực dương. Dưới dạng biểu đồ

Nếu A có hạng không đầy đủ thì phân tích (1.5.3) vẫn là thích hợp. Tất cả các thay



Hình 1.4: SVD đầy đủ ($m \geqslant n$)

đổi đó không phải là n mà là r vector suy biến trái của A được xác định bởi hình học của siêu ellip. Để xây dựng ma trận Unita U, ta đưa vào $m-r$ thay cho $m-n$ cột trực giao tùy ý. Ma trận V cũng sẽ cần $n-r$ cột trực giao tùy ý để mở rộng thành r cột xác định bởi hình học. Ma trận Σ sẽ có r phần tử đường chéo dương, với $n-r$ phần tử còn lại bằng 0.

Vì vậy, SVD được giảm (1.5.2) cũng làm số chiều các ma trận A nhỏ hơn hạng đầy đủ. Ta có thể lấy ma trận $\widehat{U}$ là $m \times n$, với các số chiều của Σ là $n \times n$ với một vài số 0 trên đường chéo, hoặc xa hơn nén $\widehat{U}$ thành $m \times r$ và $\widehat{\Sigma}$ thành $r \times r$ và dương ngặt trên đường chéo.

1.5.4 Định nghĩa

Cho m và n tùy ý và cho $A \in \mathbb{C}^{m\times n}$, *phân tích giá trị suy biến* (SVD) của A là một phân tích

$$A = U\Sigma V^* \tag{1.5.4}$$

với

$$U \in \mathbb{C}^{m\times m} \text{ là Unita,}$$
$$V \in \mathbb{C}^{n\times n} \text{ là Unita,}$$
$$\Sigma \in \mathbb{C}^{m\times n} \text{ là đường chéo.}$$

Hơn nữa, giả sử các phần tử trên đường chéo σ_j của Σ là không âm và sắp thứ tự không tăng, nghĩa là, $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_p \geqslant 0$, với $p = \min(m, n)$.
Chú ý, ma trận đường chéo Σ có hình dạng giống như A khi A không là ma trận vuông, nhưng U và V thường là các ma trận Unita vuông.
Rõ ràng ảnh của quả cầu đơn vị trong $\mathbb{R}^n$ dưới ánh xạ $A = U\Sigma V^*$ phải là một siêu ellip trong $\mathbb{R}^m$. Ánh xạ Unita V^* bảo toàn quả cầu, ma trận đường chéo Σ kéo quả cầu thành một siêu ellip được căn lề với cơ sở chính tắc, và ánh xạ Unita cuối cùng U quay hoặc phản xạ siêu ellip không thay đổi hình dạng của nó. Do đó, nếu ta có thể chứng minh mọi ma trận có một SVD thì ta sẽ chứng minh ảnh của quả cầu đơn vị dưới ánh xạ tuyến tính bất kỳ là một siêu ellip.

1.5.5 Sự tồn tại và tính duy nhất

Định lý 1.5.1 *Mọi ma trận $A \in \mathbb{C}^{m\times n}$ có một phân tích giá trị suy biến (1.5.4). Hơn nữa, các giá trị suy biến $\{\sigma_j\}$ được xác định duy nhất và, nếu A là ma trận vuông và σ_j là phân biệt thì các vector suy biến trái $\{u_j\}$ và phải $\{v_j\}$ được xác định duy nhất thành các ký hiệu phức (nghĩa là, các thừa số vô hướng phức của giá trị tuyệt đối 1).*

Chứng minh. Để chứng minh sự tồn tại của SVD, ta tách phương tác động lớn nhất của A, và khi đó tiếp tục phương pháp quy nạp theo số chiều của A.
Đặt $\sigma_1 = \|A\|_2$. Theo đối số tính compact, phải có các vector $v_1 \in \mathbb{C}^n$ và $u_1 \in \mathbb{C}^m$ với $\|v_1\|_2 = \|u_1\|_2 = 1$ và $Av_1 = \sigma_1 u_1$. Xét các khai triển bất kỳ của v_1 thành một cơ sở trực giao $\{v_j\}$ của $\mathbb{C}^n$ và khai triển của u_1 thành cơ sở trực giao $\{u_j\}$ của $\mathbb{C}^m$, và cho U_1 và V_1 là các ma trận Unita với u_j và v_j cột tương ứng là. Khi đó, ta có

$$U_1^* A V_1 = S = \begin{bmatrix} \sigma_1 & w^* \\ 0 & B \end{bmatrix}, \tag{1.5.5}$$

với 0 là vector cột $m-1$ chiều, w^* là vector dòng $n-1$ chiều, và B có $(m-1)\times(n-1)$ chiều. Hơn nữa,

$$\left\| \begin{bmatrix} \sigma_1 & w^* \\ 0 & B \end{bmatrix} \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2 \geqslant \sigma_1^2 + w^*w = (\sigma_1^2 + w^*w)^{1/2} \left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2,$$

kéo theo $\|S\|_2 \geqslant (\sigma_1^2 + w^*w)^{1/2}$. Vì U_1, V_1 là ma trận Unita và $\|S\|_2 = \|A\|_2 = \sigma_1$ nên điều này kéo theo $w = 0$.

Nếu $n = 1$ hoặc $m = 1$ thì ta đã hoàn thành. Mặt khác, ma trận con B miêu tả tác động của A vào không gian con trực giao với v_1. Theo giả thiết quy nạp, B có một SVD $B = U_2\Sigma_2V_2^*$. Bây giờ ta dễ dàng kiểm tra

$$A = U_1 \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sum_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_2 \end{bmatrix}^* V^*$$

là một SVD của A, hoàn thành chứng minh sự tồn tại.

Cho tính duy nhất, chứng minh hình học là không phức tạp: nếu độ dài các bán trục của một siêu ellip là phân biệt, khi đó các bán trục của chúng được xác định bởi hình học, lên tới các ký hiệu. Về phương diện đại số, ta có thể chứng minh như sau. Đầu tiên, từ (1.5.4) ta chú ý rằng σ_1 là được xác định duy nhất bởi điều kiện mà nó bằng $\|A\|_2$. Giả sử có một vector w độc lập tuyến tính khác với $\|w\|_2 = 1$ và $\|Aw\|_2 = \sigma_1$. Xác định một vector đơn vị v_2 mà nó trực giao với v_1 là một tổ hợp tuyến tính của v_1 và w

$$v_2 = \frac{w - (v_1^*w)v_1}{\|w - (v_1^*w)v_1\|_2}.$$

Vì $\|A\|_2 = \sigma_1, \|Av_2\|_2 \leqslant \sigma_1$ nhưng điều này phải bằng nhau cho trường hợp khác. Vì $w = v_1c + v_2s$ với các hằng số c và s bất kỳ thỏa $|c|^2 + |s|^2 = 1$, ta sẽ có $\|Aw\|_2 \leqslant \sigma_1$. Vector v_2 này là vector suy biến phải thứ hai của A ứng với giá trị suy biến σ_1 nên tồn tại một vector y (bằng với $n-1$ thành phần cuối của $V_1^*v_2$) thỏa $\|y\|_2 = 1$ và $\|By\|_2 = \sigma_1$. Nếu vector suy biến v_1 là không duy nhất thì giá trị suy biến σ_1 tương ứng là không đơn giản. Để hoàn thành chứng minh tính duy nhất ta chú ý, như được cho ở trên, σ_1, v_1 và u_1 được xác định, phần còn lại của SVD được xác định bởi tác động của A vào không gian trực giao với v_1. Vì v_1 là duy nhất nên không gian trực giao này được xác định duy nhất, và tính duy nhất của các giá trị và vector suy biến còn lại theo sau phương pháp quy nạp.

1.5.6 Sự thay đổi của các cơ sở

Cho $b \in \mathbb{C}^m$ bất kỳ có thể được khai triển trong cơ sở của các vector suy biến trái của A (các cột của U), và $x \in \mathbb{C}^n$ bất kỳ có thể được khai triển trong cơ sở của các vector

suy biến phải của A (các cột của V). Các vector tọa độ cho các khai triển này là

$$b^{'} = U^{*}b,\ x^{'} = V^{*}x.$$

Theo (1.5.3), $b = Ax$ có thể được biểu diễn theo $b^{'}$ và $x^{'}$

$$b = Ax \Longleftrightarrow U^{*}b = U^{*}Ax = U^{*}U\Sigma V^{*}x \Longleftrightarrow b^{'} = \Sigma x^{'}.$$

Khi $b = Ax$, ta có $b^{'} = \Sigma x^{'}$. Do đó, A rút gọn thành ma trận đường chéo Σ khi range được biểu diễn trong cơ sở các cột của U và miền xác định được biểu diễn trong cơ sở các cột của V.

1.5.7 SVD so với phân tích trị riêng

Một ma trận vuông đầy đủ không quan trọng A có thể được biểu diễn như là một ma trận đường chéo của các trị riêng Λ, nếu range và miền xác định được biểu diễn trong một cơ sở của các vector riêng.
Nếu các cột của ma trận $X \in \mathbb{C}^{m\times m}$ chứa các vector riêng độc lập tuyến tính của $A \in \mathbb{C}^{m\times m}$, *phân tích trị riêng* của A là

$$A = X\Lambda X^{-1}, \tag{1.5.6}$$

với Λ là ma trận đường chéo $m \times m$ mà các phần tử của nó là các trị riêng của A. Cho $b,\ x \in \mathbb{C}^{m}$ thỏa mãn $b = Ax$, ta định nghĩa

$$b^{'} = X^{-1}b,\ x^{'} = X^{-1}x,$$

thì các vector được khai triển mới $b^{'}$ và $x^{'}$ thỏa mãn $b^{'} = \Lambda x^{'}$.
Có sự khác nhau cơ bản giữa SVD và phân tích trị riêng. Một là SVD sử dụng hai cơ sở khác nhau (các tập hợp của các vector suy biến trái và phải), trong khi đó phân tích trị riêng sử dụng đúng một cơ sở (các vector riêng). Thứ hai là SVD sử dụng cơ sở trực giao, trong khi đó phân tích trị riêng sử dụng một cơ sở nói chung không phải là trực giao. Thứ ba là không phải tất cả các ma trận (ngay cả ma trận vuông) đều có phân tích trị riêng, nhưng tất cả các ma trận (ngay cả ma trận hình chữ nhật) có phân tích giá trị suy biến, như được thiết lập trong Định lý 1.5.1. Trong các ứng dụng, các trị riêng hướng về các bài toán có liên quan tới các dạng được lặp lại của A, như ma trận lũy thừa A^{k} hay các hàm mũ e^{tA}, trong khi các vector suy biến hướng về các bài toán có liên quan tới xử lý của chính A hoặc nghịch đảo của nó.

1.5.8 Các tính chất ma trận thông qua SVD

Giả sử A có $m \times n$ chiều. Cho p là số nhỏ nhất của m và n, $r \leqslant p$ là số các giá trị suy biến khác 0 của A, và cho $< x, y, \ldots, z >$ là không gian sinh bởi các vector $x, y, \ldots, z$. Khi đó, ta có các định lý sau

Định lý 1.5.2 *$rank(A) = r$, số các giá trị suy biến khác 0.*

Chứng minh. Hạng của một ma trận đường chéo là bằng số các phần tử khác 0 của nó, và trong phân tích $A = U\Sigma V^*$, U và V là hạng đầy đủ. Do đó, $rank(A) = rank(\Sigma) = r$.

Định lý 1.5.3 *$range(A) =< u_1, \ldots, u_r >$ và $null(A) =< v_{r+1}, \ldots, v_n >$.*

Chứng minh. Đây là chuỗi mà $range(\Sigma) =< e_1, \ldots, e_r > \subseteq \mathbb{C}^m$ và
$null(\Sigma) =< e_{r+1}, \ldots, e_n > \subseteq \mathbb{C}^n$.

Định lý 1.5.4 $\|A\|_2 = \sigma_1$ *và* $\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_r^2}$.

Chứng minh. Kết quả đầu tiên đã được thiết lập trong chứng minh của Định lý 1.5.1 vì $A = U\Sigma V^*$ với ma trận Unita U và V, $\|A\|_2 = \|\Sigma\|_2 = max\{|\sigma_j|\} = \sigma_1$ (do Định lý 1.4.1). Kết quả thứ hai, do Định lý 1.4.1 và nhận xét theo sau, chuẩn Frobenius là bất biến dưới phép nhân Unita nên $\|A\|_F = \|\Sigma\|_F$, và do (1.4.16) nên ta có công thức như trên.

Định lý 1.5.5 *Các giá trị suy biến khác không của A là các căn bậc hai của các trị riêng khác không của A^*A hay AA^*. (Các ma trận này có cùng các trị riêng khác không.)*

Chứng minh. Từ kết quả tính toán

$$A^*A = (U\Sigma V^*)^*(U\Sigma V^*) = V\Sigma^*U^*U\Sigma V^* = V(\Sigma^*\Sigma)V^*,$$

ta thấy A^*A tương tự với $\Sigma^*\Sigma$ và do đó có cùng n trị riêng. Các trị riêng của ma trân đường chéo $\Sigma^*\Sigma$ là $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$, với $n - p$ trị riêng 0 thêm vào nếu $n > p$. Tính toán tương tự với m trị riêng của AA^*.

Định lý 1.5.6 *Nếu $A = A^*$ thì các giá trị suy biến của A là giá trị tuyệt đối của các trị riêng của A.*

Chứng minh. Một ma trận Hermit là một tập đầy đủ các vector riêng trực giao và tất cả các trị riêng này là thực. Một phát biểu tương đương là (1.5.6) đúng với X bằng với một ma trận Unita Q bất kỳ và Λ là ma trận đường chéo thực. Khi đó, ta có thể viết

$$A = Q\Lambda Q^* = Q|\Lambda|sign(\Lambda)Q^*, \tag{1.5.7}$$

với $|\Lambda|$ và $sign(\Lambda)$ là các ma trận đường chéo mà các phần tử của nó tương ứng là $|\lambda_j|$ và $sign(\lambda_j)$. (Ta có thể đặt $sign(\Lambda)$ bên trái Λ thay vì bên phải.) Vì $sign(\Lambda)Q^*$ là Unita khi Q là Unita, (1.5.7) là SVD của A, với các giá trị suy biến bằng với các phần tử trên đường chéo của $|\Lambda|, |\lambda_j|$. Nếu được như mong muốn thì các số này có thể được sắp xếp thành thứ tự không tăng bằng việc thêm các ma trận hoán vị phù hợp như là các thừa số trong vế trái ma trận Unita của (1.5.7), Q và vế phải ma trận Unita, $sign(\Lambda)Q^*$.

Định lý 1.5.7 *Cho* $A \in \mathbb{C}^{m\times m}, |det(A)| = \prod_{i=1}^{m} \sigma_i$.

Chứng minh. Định thức của tích các ma trận vuông là tích các định thức của các thừa số. Hơn nữa, do công thức $U^*U = I$ và tính chất $det(U^*) = (det(U))^*$ nên trị tuyệt đối của định thức của một ma trận Unita thường là 1. Do đó,

$$|det(A)| = |det(U\Sigma V^*)| = |det(U)||det(\Sigma)||det(V^*)| = |det(\Sigma)| = \prod_{i=1}^{m} \sigma_i.$$

1.5.9 Xấp xỉ ma trận hạng thấp

Định lý 1.5.8 *A là tổng của r ma trận hạng 1:*

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^*. \tag{1.5.8}$$

Chứng minh. Nếu ta viết Σ như là tổng của r ma trận Σ_j, với
$\Sigma_j = diag(0, \ldots, 0, \sigma_j, 0, \ldots, 0)$, thì (1.5.8) theo sau từ (1.5.3).

Có nhiều cách để khai triển một ma trận A có $m \times n$ chiều như là tổng của các ma trận hạng 1. Ví dụ, A có thể được viết như tổng của m dòng của nó, hoặc tổng của n cột của nó, hoặc tổng của mn phần tử của nó. Cho ví dụ khác, khử Gauss giảm A thành tổng của ma trận hạng 1 đầy đủ, một ma trận hạng 1 mà 2 dòng và cột đầu tiên của nó là 0, ...
Tuy nhiên, công thức (1.5.8) biểu diễn phân tích thành các ma trận hạng 1 với tích chất sâu hơn: *tổng riêng phần thứ ν thu giữ nhiều năng lượng của A ngay khi có thể thực hiện được.* Phát biểu này đúng với "năng lượng" xác định bởi hoặc là chuẩn 2 hoặc là chuẩn Frobenius. Ta có thể làm nó tỉ mỉ bằng việc đưa ra công thức một bài toán xấp xỉ tốt nhất của một ma trận A bằng các ma trận có hạng nhỏ hơn.

Định lý 1.5.9 *Cho ν bất kỳ với $0 \leqslant \nu \leqslant r$, định nghĩa*

$$A_\nu = \sum_{j=1}^{\nu} \sigma_j u_j v_j^*; \tag{1.5.9}$$

Nếu $\nu = p = \min\{m, n\}$ thì định nghĩa $\sigma_{\nu+1} = 0$. Khi đó,

$$\|A - A_\nu\|_2 = \inf_{\substack{B\in\mathbb{C}^{m\times n} \\ rank(B)\leqslant\nu}} \|A - B\|_2 = \sigma_{\nu+1}.$$

Chứng minh. Giả sử có B bất kỳ với $rank(B) \leqslant \nu$ sao cho $\|A-B\|_2 < \|A-A_\nu\|_2 = \sigma_{\nu+1}$. Khi đó, có một không gian con $W \subseteq \mathbb{C}^n$ có $(n-\nu)$ chiều sao cho $w \in W \Rightarrow Bw = 0$. Do đó, với $w \in W$ bất kỳ, ta có $Aw = (A-B)w$ và

$$\|Aw\|_2 = \|(A-B)w\|_2 \leqslant \|A-B\|_2\|w\|_2 < \sigma_{\nu+1}\|w\|_2.$$

Do đó, W là không gian con $n-\nu$ chiều với $\|Aw\| < \sigma_{\nu+1}\|w\|$. Nhưng có không gian con $(\nu+1)$ chiều với $\|Aw\| \geqslant \sigma_{\nu+1}\|w\|$, cụ thể là không gian sinh bởi $\nu+1$ vector suy biến trái đầu tiên của A. Vì tổng các chiều của các không gian này vượt quá n, nên phải có một vector khác 0 nằm trong cả hai, mâu thuẫn.

Ta phát biểu kết quả tương tự cho chuẩn Frobenius mà không chứng minh.

Định lý 1.5.10 *Cho ν bất kỳ với $0 \leqslant \nu \leqslant r$, ma trận A_ν của (1.5.8) cũng thỏa mãn*

$$\|A - A_\nu\|_F = \inf_{\substack{B\in\mathbb{C}^{m\times n}\\ rank(B)\leqslant\nu}} \|A-B\|_F = \sqrt{\sigma_{\nu+1}^2 + \ldots + \sigma_r^2}.$$

1.5.10 Ví dụ:

Ví dụ 1.5.1. (SVD đầy đủ) Cho ma trận

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

Đầu tiên, ta tính các giá trị suy biến σ_i bằng việc tìm các trị riêng của AA^T. Ma trận chuyển vị của A là

$$A^T = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}$$

nên

$$AA^T = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$$

Đa thức đặc trưng là $det(AA^T - \lambda I) = \begin{vmatrix} 11-\lambda & 1 \\ 1 & 11-\lambda \end{vmatrix} = (11-\lambda)(11-\lambda)-1 = (\lambda-12)(\lambda-10)$ nên các giá trị suy biến là $\sigma_1 = \sqrt{12}$ và $\sigma_2 = \sqrt{10}$.
Tiếp theo, ta tìm các vector suy biến phải bằng việc tìm một tập các vector riêng trực giao của A^TA nên ta có

$$A^TA = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}$$

Tìm các vector riêng và các trị riêng tương ứng của A^TA. Các vector riêng xác định bởi phương trình $Ax = \lambda x$ nên áp dụng cho A^TA ta được

$$\begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Đa thức đặc trưng là

$$\begin{aligned} det(A^TA - \lambda I) &= \begin{vmatrix} 10-\lambda & 0 & 2 \\ 0 & 10-\lambda & 4 \\ 2 & 4 & 2-\lambda \end{vmatrix} \\ &= (10-\lambda)\begin{vmatrix} 10-\lambda & 4 \\ 4 & 2-\lambda \end{vmatrix} + 2\begin{vmatrix} 0 & 10-\lambda \\ 2 & 4 \end{vmatrix} \\ &= (10-\lambda)[(10-\lambda)(2-\lambda) - 16] + 2[0 - (20 - 2\lambda)] \\ &= \lambda(\lambda - 10)(\lambda - 12) \end{aligned}$$

nên $\lambda = 0, \lambda = 10, \lambda = 12$ là các trị riêng của A^TA. Do A^TA là ma trận đối xứng nên các vector riêng tương ứng sẽ trực giao.
Với $\lambda = 12$, ta có

$$A^TA - \lambda I = \begin{bmatrix} -2 & 0 & 2 \\ 0 & -2 & 4 \\ 2 & 4 & -10 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

hay $x_1 - x_3 = 0$ và $x_2 - 2x_3 = 0$ nên ta chọn $x_1 = 1, x_2 = 2, x_3 = 1$. Do đó, vector riêng tương ứng với trị riêng $\lambda = 12$ là $v_1 = (1, 2, 1)$.
Với $\lambda = 10$, ta có

$$A^TA - \lambda I = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 4 \\ 2 & 4 & -8 \end{bmatrix} \sim \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

hay $x_1 + 2x_2 = 0$ và $x_3 = 0$ nên ta chọn $x_1 = 1, x_2 = -2, x_3 = 0$. Do đó, vector riêng tương ứng với trị riêng $\lambda = 10$ là $v_2 = (1, -2, 0)$.
Tương tự với $\lambda = 0$, ta có

$$A^TA - \lambda I = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix} \sim \begin{bmatrix} 5 & 0 & 1 \\ 0 & 5 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

hay $5x_1 + x_3 = 0$ và $5x_2 + 2x_3 = 0$ nên ta chọn $x_1 = 1, x_2 = 2, x_3 = -5$. Do đó, vector riêng tương ứng với trị riêng $\lambda = 0$ là $v_3 = (1, 2, -5)$. Khi đó, ta có

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{bmatrix}$$

Sử dụng trực giao hóa Gram-Schmidt để chuyển ma trận ở trên thành ma trận trực giao

$$\begin{aligned} u_1 &= \frac{v_1}{\|v_1\|} = \left(\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right) \\ w_2 &= v_2 - u_1 v_2^* u_1 = (2, 1, 0) \\ u_2 &= \frac{w_2}{\|w_2\|} = \left(\frac{2}{\sqrt{5}}, \frac{-1}{\sqrt{5}}, 0\right) \\ w_3 &= v_3 - u_1 v_3^* u_1 - u_2 v_3^* u_2 = \left(\frac{-2}{3}, \frac{-4}{3}, \frac{10}{3}\right) \\ u_3 &= \frac{w_3}{\|w_3\|} = \left(\frac{1}{\sqrt{30}}, \frac{2}{\sqrt{30}}, \frac{-5}{\sqrt{30}}\right) \end{aligned}$$

Khi đó, ta được

$$V = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

và

$$V^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

Cuối cùng, ta tính U bằng công thức $\sigma_i u_i = A v_i$ với $i = 1, 2$ hay $u_i = \frac{1}{\sigma_i} A v_i$

$$U = [u_1 | u_2] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

Do đó, ta được

$$\begin{aligned}
A &= U\Sigma V^* = U\Sigma V^T \\
&= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{6} & \sqrt{5} & 0 \\ \sqrt{6} & -\sqrt{5} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix} \\
&= \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}.
\end{aligned}$$

Ví dụ 1.5.2. (SVD được giảm) Cho ma trận

$$\begin{bmatrix} 2 & 0 & 8 & 6 & 0 \\ 1 & 6 & 0 & 1 & 7 \\ 5 & 0 & 7 & 4 & 0 \\ 7 & 0 & 8 & 5 & 0 \\ 0 & 10 & 0 & 0 & 7 \end{bmatrix}$$

Tương tự, ta tính các giá trị suy biến σ_i bằng việc tìm các trị riêng của AA^T

$$AA^T = \begin{bmatrix} 2 & 0 & 8 & 6 & 0 \\ 1 & 6 & 0 & 1 & 7 \\ 5 & 0 & 7 & 4 & 0 \\ 7 & 0 & 8 & 5 & 0 \\ 0 & 10 & 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} 2 & 1 & 5 & 7 & 0 \\ 0 & 6 & 0 & 0 & 10 \\ 8 & 0 & 7 & 8 & 0 \\ 6 & 1 & 4 & 5 & 0 \\ 0 & 7 & 0 & 0 & 7 \end{bmatrix} = \begin{bmatrix} 104 & 8 & 90 & 108 & 0 \\ 8 & 87 & 9 & 12 & 109 \\ 90 & 9 & 90 & 111 & 0 \\ 108 & 12 & 111 & 138 & 0 \\ 0 & 109 & 0 & 0 & 149 \end{bmatrix}$$

Đa thức đặc trưng là

$$det(AA^T - \lambda I) = \begin{vmatrix} 104-\lambda & 8 & 90 & 108 & 0 \\ 8 & 87-\lambda & 9 & 12 & 109 \\ 90 & 9 & 90-\lambda & 111 & 0 \\ 108 & 12 & 111 & 138-\lambda & 0 \\ 0 & 109 & 0 & 0 & 149-\lambda \end{vmatrix}$$

Các trị riêng của AA^T là

$$\lambda = 321.07, \lambda = 230.17, \lambda = 12.70, \lambda = 3.94, \lambda = 0.12$$

Tiếp theo, ta tìm các vector suy biến phải bằng việc tìm một tập các vector riêng trực giao của A^TA nên ta có

$$A^TA = \begin{bmatrix} 79 & 6 & 107 & 68 & 7 \\ 6 & 136 & 0 & 6 & 112 \\ 107 & 0 & 177 & 116 & 0 \\ 68 & 6 & 116 & 78 & 7 \\ 7 & 112 & 0 & 7 & 98 \end{bmatrix}$$

Tương tự, ta được

$$V^T = \begin{bmatrix} -0.46 & 0.02 & -0.87 & -0.00 & 0.17 \\ -0.07 & -0.76 & 0.06 & 0.60 & 0.23 \\ -0.74 & 0.10 & 0.28 & 0.22 & -0.56 \\ -0.48 & 0.03 & 0.40 & -0.33 & 0.70 \\ -0.07 & -0.64 & -0.04 & -0.69 & -0.32 \end{bmatrix}$$

Ta tính U bằng công thức $\sigma_i u_i = Av_i$ với $i = 1, 2$ hay $u_i = \frac{1}{\sigma_i} Av_i$

$$U = \begin{bmatrix} -0.54 & 0.07 & 0.82 & -0.11 & 0.12 \\ -0.10 & -0.59 & -0.11 & -0.79 & -0.06 \\ -0.53 & 0.06 & -0.21 & 0.12 & -0.81 \\ -0.65 & 0.07 & -0.51 & 0.06 & 0.56 \\ -0.06 & -0.80 & 0.09 & 0.59 & 0.04 \end{bmatrix}$$

Để minh họa tác động của việc giảm kích thước trên tập dữ liệu này, ta sẽ giới hạn Σ ở ba giá trị đầu tiên

$$\Sigma = \begin{bmatrix} 17.92 & 0 & 0 \\ 0 & 15.17 & 0 \\ 0 & 0 & 3.56 \end{bmatrix}$$

Một xấp xỉ của A sử dụng 3 chiều thay vì 5 chiều b

$$\hat{A} = \begin{bmatrix} -0.54 & 0.07 & 0.82 \\ -0.10 & -0.59 & -0.11 \\ -0.53 & 0.06 & -0.21 \\ -0.65 & 0.07 & -0.51 \\ -0.06 & -0.80 & 0.09 \end{bmatrix} \begin{bmatrix} 17.92 & 0 & 0 \\ 0 & 15.17 & 0 \\ 0 & 0 & 3.56 \end{bmatrix} \begin{bmatrix} -0.46 & 0.02 & -0.87 & -0.00 & 0.17 \\ -0.07 & -0.76 & 0.06 & 0.60 & 0.23 \\ -0.74 & 0.10 & 0.28 & 0.22 & -0.56 \end{bmatrix}$$

$$= \begin{bmatrix} 2.29 & -0.66 & 9.33 & 1.25 & -3.09 \\ 1.77 & 6.76 & 0.90 & -5.50 & -2.13 \\ 4.86 & -0.96 & 8.01 & 0.38 & -0.97 \\ 6.62 & -1.23 & 9.58 & 0.24 & -0.71 \\ 1.14 & 9.19 & 0.33 & -7.19 & -3.13 \end{bmatrix}$$

Bài tập

1. Chứng minh Định lý 1.2.3

2. Cho B là ma trận 4×4, ta áp dụng các phép toán theo sau:

   a) 2 lần cột 1,

   b) 1/2 dòng 3,

   c) Dòng 3 + dòng 1,

   d) Hoán đổi cột 1 và cột 4,

   e) Trừ dòng 2 từ mỗi dòng khác,

   f) Thay thế cột 4 bởi cột 3,

   g) Xóa cột 1 (sao cho số chiều của cột được giảm xuống 1)

   Viết kết quả tích của 8 ma trận và viết lại kết quả với tích của ABC (giống B) của 3 ma trận.

3. Chứng minh rằng nếu một ma trận A vừa là ma trận tam giác vừa là ma trận Unita thì A là ma trận đường chéo.

4. Định lý Pythagorean khẳng định rằng với n vector trực giao $\{x_i\}$,

$$\|\sum_{i=1}^{n} x_i\|^2 = \sum_{i=1}^{n} \|x_i\|^2$$

   a) Chứng minh điều này trong trường hợp $n = 2$ bằng một tính toán rõ ràng của $\|x_1 + x_2\|^2$.

b) Cho thấy rằng tính toán này cũng thiết lặp trong trường hợp tổng quát (bằng quy nạp).

5. Cho $A \in \mathbb{C}^{m\times m}$ là ma trận hermit. Một trị riêng của A là một vector khác 0 $x \in \mathbb{C}^m$ sao cho $Ax = \lambda x$ với $\lambda \in \mathbb{C}$ là trị riêng tương ứng.

   a) Chứng minh tất cả các trị riêng của A là số thực.

   b) Chứng minh rằng nếu x và y là 2 vector riêng tương ứng với 2 trị riêng riêng phân biệt thì x và y trực giao.

6. Chứng minh Ví dụ 1.4.3.

7. Chứng minh Định lý 1.4.1.

8. Chứng minh rằng nếu W là ma trận không suy biến (kỳ dị) bất kỳ thì hàm $\|\cdot\|_W$ được xác định bởi 1.4.3 là một chuẩn vector.

9. Cho $\|\cdot\|$ là chuẩn bất kỳ trong $\mathbb{C}^m$ và cũng là chuẩn ma trận được bao gồm trong $\mathbb{C}^{m\times m}$. Chứng minh $\rho(A) \leqslant \|A\|$, với $\rho(A)$ là *bán kính phổ* của A, nghĩa là trị tuyệt đối lớn nhất $|\lambda|$ của trị riêng λ của A.

10. Chứng minh Ví dụ 1.4.4.

11. Ví dụ 1.4.4 cho thấy rằng nếu E là một tích ngoài $E = uv^*$ thì $\|E\|_2 = \|u\|_2\|v\|_2$. Kết quả này vẫn đúng cho chuẩn Frobenius, nghĩa là $\|E\|_F = \|u\|_F\|v\|_F$ không? Chứng minh nếu nó vẫn còn đúng ngược lại cho một phản ví dụ.

12. Chứng minh Định lý 1.5.1.

13. Xác định SVD của các ma trận sau (tính tay)

    a) $\begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$, b) $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$, c) $\begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, d) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$, e) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

14. Hai ma trận $A, B \in \mathbb{C}^{m\times m}$ là *tương đương unita* nếu $A = QBQ^*$ với $Q \in \mathbb{C}^{m\times m}$ là ma trận unita nào đó. A và B là tương đương unita khi và chỉ khi chúng có cùng các giá trị suy biến, điều này đúng hay sai?

15. Định lý 1.5.1 khẳng định rằng mọi ma trận $A \in \mathbb{C}^{m\times n}$ có phân tích SVD $A = U\Sigma V^*$. Chứng minh rằng nếu A là ma trận thực thì A có phân tích SVD thực $(U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n})$.

16. Xét ma trận

$$\begin{bmatrix} -2 & 11 \\ -10 & 5 \end{bmatrix}$$

a) Xác định trên giấy phân tích SVD thực của A dưới dạng $A = U\Sigma V^T$. Phân tích SVD là không duy nhất nên tìm một phân tích có số dấu "-" ít nhất trong U và V.

b) Liệt kê các giá trị suy biến, các vector suy biến trái và các vector suy biến phải của A. Vẽ ảnh được gán nhãn của quả cầu đơn vị trong $\mathbb{R}^2$, ảnh của nó dưới tác động của A cùng với các vector suy biến với các tọa độ của các đỉnh được đánh dấu.

c) Chuẩn 1, 2, ∞ và Frobenius của A?

d) Tìm A^{-1} thông qua SVD.

e) Tìm các trị riêng λ_1, λ_2 của A.

f) Kiểm tra $detA = \lambda_1\lambda_2$ và $|detA| = \sigma_1\sigma_2$.

g) Diện tích của ellip mà A ánh xạ quả cầu đơn vị của $\mathbb{R}^2$ lên trên?

17. Cho $A \in \mathbb{C}^{m\times m}$ có phân tích SVD $A = U\Sigma V^*$. Tìm phân tích trị riêng 1.5.6 của ma trận hermitian $2m \times 2m$

$$\begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}.$$

18. Chứng minh rằng nếu $A \in \mathbb{C}^{m\times n}$ có hạng là r thì $\|A(A^TA)^{-1}A^T\|_2 = 1$.

19. Chứng minh rằng nếu thêm một dòng khác 0 vào ma trận A thì giá trị suy biến lớn nhất và nhỏ nhất của A đều tăng.

20. Viết chương trình tạo ngẫu nhiên ma trận đối xứng cấp N.

21. Viết chương trình tạo ngẫu nhiên ma trận trực chuẩn cấp N.

22. Cho ma trận

$$A = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$$

a) Tính SVD của ma trận A.

b) Viết chương trình vẽ các vector suy biến phải của ma trận V trong phân tích SVD như Hình 1.2.

c) Viết chương trình vẽ các vector suy biến trái của ma trận U trong phân tích SVD như Hình 1.2.

Chương 2

Phân tích QR và bình phương tối tiểu

2.1 Phép chiếu

2.1.1 Phép chiếu

Một *phép chiếu* là một ma trận vuông P thỏa mãn

$$P^2 = P. \tag{2.1.1}$$

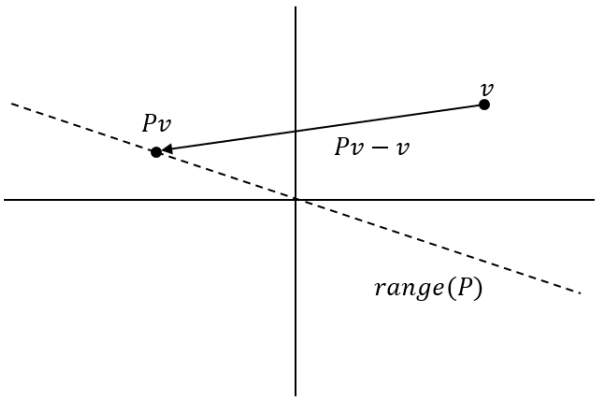
(Ma trận như vậy cũng được nói là ma trận *lũy đẳng*.) Định nghĩa này bao gồm cả phép chiếu trực giao và không trực giao. Để tránh lộn xộn ta sử dụng thuật ngữ *phép chiếu nghiêng* trong trường hợp không trực giao.

Thuật ngữ phép chiếu có được thông qua như việc xuất hiện từ ký hiệu rằng nếu người ta chiếu ánh sáng vào không gian con $range(P)$ chỉ từ phương thẳng thì Pv sẽ là bóng được chiếu bởi vector v.

Quan sát thấy rằng nếu $v \in range(P)$ thì nó nằm một cách chính xác trong cái bóng của nó và việc áp dụng các kết quả của phép chiếu trong chính v. Theo toán học, ta có $v = Px$ với x bất kì và

$$Pv = P^2x = Px = v.$$

Ánh sáng chiếu vào phương như thế nào khi $v \neq Pv$? Tổng quát, câu trả lời phụ thuộc



Hình 2.1: Phép chiếu nghiêng

vào v nhưng với v đặc biệt bất kì, nó dễ dàng được suy ra bằng việc vẽ đường từ v tới $Pv, Pv - v$ (Hình 2.1). Việc áp dụng phép chiếu tới vector này cho một kết quả

$$P(Pv - v) = P^2v - Pv = 0.$$

nghĩa là $Pv - v \in null(P)$. Phương của ánh sáng có thể là khác nhau cho v khác nhau nhưng nó thường được miêu tả bởi một vector trong $null(P)$.

2.1.2 Phép chiếu bù

Nếu P là phép chiếu thì $I - P$ cũng là một phép chiếu, cũng là một lũy đẳng:

$$(I - P)^2 = I - 2P + P^2 = I - P.$$

Ma trận $I - P$ được gọi là *phép chiếu bù* tới P.
Phép chiếu $I - P$ vào không gian đầy đủ của P. Ta biết rằng $range(I - P) \supseteq null(P)$ bởi vì nếu $Pv = 0$, ta có $(I - P)v = v$. Ngược lại, $range(I - P) \subseteq null(P)$ vì với v bất kì, ta có $(I - P)v = v - Pv \in null(P)$. Do đó, với phép chiếu P bất kì,

$$range(I - P) = null(P). \tag{2.1.2}$$

Bằng việc viết $P = I - (I - P)$ ta suy ra phần bù

$$null(I - P) = range(P). \tag{2.1.3}$$

Ta cũng thấy rằng $null(I - P) \cap null(P) = \{0\}$: vector v bất kì trong cả 2 tập thỏa mãn $v = v - Pv = (I - P)v = 0$. Mặt khác,

$$range(P) \cap null(P) = \{0\}. \tag{2.1.4}$$

Các tính toán này cho thấy *một phép chiếu tách $\mathbb{C}^m$ thành 2 không gian*. Ngược lại, cho S_1 và S_2 là hai không gian con của $\mathbb{C}^m$ sao cho $S_1 \cap S_2 = \{0\}$ và $S_1 + S_2 = \mathbb{C}^m$, với $S_1 + S_2$ là không gian sinh của S_1 và S_2, nghĩa là $S_1 + S_2$ là tập các vector $s_1 + s_2$ với $s_1 \in S_1$ và $s_2 \in S_2$. (Một cặp như vậy được nói là *các không gian con bù*). Khi đó, có một phép chiếu P thỏa mãn $range(P) = S_1$ và $null(P) = S_2$. Ta nói rằng P là phép chiếu *vào* S_1 *dọc theo* S_2. Phép chiếu này và bù của nó có thể được xem như là lời giải duy nhất của bài toán theo sau:

Cho v, tìm các vector $v_1 \in S_1$ và $v_2 \in S_2$ sao cho $v_1 + v_2 = v$.

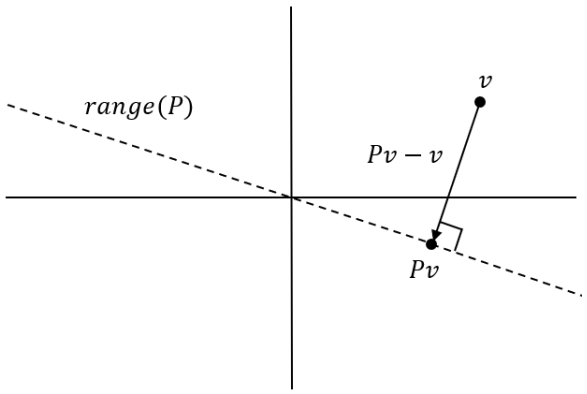
Phép chiếu Pv cho v_1, và phép chiếu bù $(I - P)v$ cho v_2. Các vector này là duy nhất bởi vì tất cả các lời giải phải có dạng

$$(Pv + v_3) + ((I - P)v - v_3) = v,$$

rõ ràng v_3 phải nằm trong cả S_1 và S_2, nghĩa là, $v_3 = 0$.

Giả sử ma trận A (kích thước $m \times m$) có một tập đầy đủ các vector riêng $\{v_j\}$, như trong (1.5.6), nghĩa là $\{v_j\}$ là một cơ sở của $\mathbb{C}^m$. Chúng thường được liên quan tới các bài toán kết hợp với các khai triển của các vector trong cơ sở này. Cho $x \in \mathbb{C}^m$, ví dụ, thành phần của x trong phương của một vector riêng đặc biệt v là gì? Câu trả lời là Px, với P là phép chiếu hạng 1 nào đó.

2.1.3 Phép chiếu trực giao



Hình 2.2: Phép chiếu trực giao

Một *phép chiếu trực giao* (Hình 2.2) là một phép chiếu lên một không gian con S_1 dọc theo không gian S_2, với S_1 và S_2 là trực giao.
Định nghĩa về mặt đại số: một phép chiếu trực giao là phép chiếu bất kì mà nó là Hermit, thỏa mãn $P^* = P$ như trong (2.1.1).

Định lý 2.1.1 *Phép chiếu P là trực giao nếu và chỉ nếu $P = P^*$.*

Chứng minh. ($\Rightarrow$) Nếu $P = P^*$ thì tích trong của vector $Px \in S_1$ và vector $(I-P)y \in S_2$ là 0:

$$x^* P^* (I - P)y = x^* (P - P^2)y = 0.$$

Khi đó, phép chiếu là trực giao.
($\Rightarrow$) Giả sử P chiếu lên S_1 dọc theo S_2, với $S_1 \perp S_2$ và S_1 có số chiều là n. Khi đó, SVD của P có thể được xây dựng như sau. Cho $\{q_1, q_2, \ldots, q_m\}$ là một cơ sở trực giao của $\mathbb{C}^m$, với $\{q_1, \ldots, q_n\}$ là 1 cơ sở của S_1 và $\{q_{n+1}, \ldots, q_m\}$ là một cơ sở của S_2. Cho $j \leqslant n$, ta có $Pq_j = q_j$, và cho $j > n$, ta có $Pq_j = 0$. Cho Q là ma trận Unita mà cột thứ j là q_j. Khi đó, ta có

$$PQ = \left[q_1 \middle| \ldots \middle| q_n \middle| 0 \middle| \ldots \right],$$

để cho

$$Q^*PQ = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix} = \Sigma,$$

một ma trận đường chéo với 1 nằm ở n phần tử đầu tiên và 0 nằm ở những nơi khác. Khi đó, ta đã xây dựng một phân tích giá trị suy biến của P:

$$P = Q\Sigma Q^*. \tag{2.1.5}$$

(Chú ý đây cũng là một phân tích trị riêng (1.5.6). Từ đây ta thấy P là Hermit, vì $P^* = (Q\Sigma Q^*)^* = Q\Sigma^* Q^* = Q\Sigma Q^* = P$.)

2.1.4 Phép chiếu với cơ sở trực giao

Vì phép chiếu trực giao có một vài giá trị suy biến bằng 0 (ngoại trừ trường hợp tầm thường $P = I$) nên các cột của Q trong (2.1.5) và sử dụng SVD được giảm, ta thu được biểu thức đơn giản

$$P = \widehat{Q}\widehat{Q}^*, \tag{2.1.6}$$

với các cột của $\widehat{Q}$ là trực giao.

Trong (2.1.6), ma trận $\widehat{Q}$ không cần thiết đến từ SVD. Cho $\{q_1, \ldots, q_n\}$ là một tập bất kì của n vecotor trực giao trong $\mathbb{C}^m$, và cho $\widehat{Q}$ là ma trận $m \times n$ tương ứng. Từ (1.3.7), ta biết rằng

$$v = r + \sum_{i=1}^{n} (q_i q_i^*)v$$

biểu diễn phân tích của vector $v \in \mathbb{C}^m$ thành một phân tích trong không gian cột của $\widehat{Q}$ cộng với một phân tích trong không gian trực giao. Do đó, ánh xạ

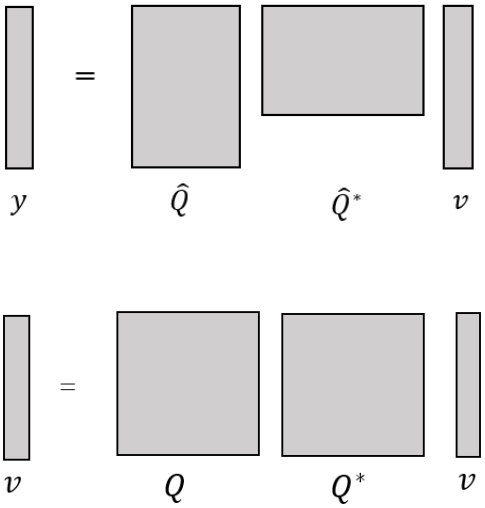
$$v \mapsto \sum_{i=1}^{n} (q_i q_i^*)v \tag{2.1.7}$$

là một phép chiếu trực giao lên $range(\widehat{Q})$, và trong dạng ma trận, nó có thể được viết $y = \widehat{Q}\widehat{Q}^* v$

Do đó, tích $\widehat{Q}\widehat{Q}^*$ bất kì thường là một phép chiếu vào không gian cột của $\widehat{Q}$, bất chấp thu được $\widehat{Q}$ như thế nào miễn là các cột của nó là trực giao. Có thể $\widehat{Q}$ được thu được bằng việc giảm một vài cột và dòng từ phân tích đầy đủ $v = \widehat{Q}\widehat{Q}^* v$

và có thể là không.

Phần bù của một phép chiếu trực giao cũng là một phép chiếu trực giao (chứng minh: $I - \widehat{Q}\widehat{Q}^*$ là hermit). Các phép chiếu bù lên không gian trực giao tới $range(\widehat{Q})$.



Một trường hợp đặc biệt quan trọng của các phép chiếu trực giao là phép chiếu trực giao hạng 1 tách thành phần trong một phương q

$$P_q = qq^*. \tag{2.1.8}$$

Các phần bù của chúng là các phép chiếu trực giao giao hạng $m-1$ mà chúng ước lượng thành phần trong phương của q:

$$P_{\perp q} = I - qq^*. \tag{2.1.9}$$

Phương trình (2.1.8) và (2.1.9) giả sử q là một vector đơn vị. Cho một vector a khác không tùy ý, các công thức tương tự là

$$P_a = \frac{aa^*}{a^*a}, \tag{2.1.10}$$

$$P_{\perp a} = I - \frac{aa^*}{a^*a}. \tag{2.1.11}$$

2.1.5 Phép chiếu với cơ sở tùy ý

Một phép chiếu trực giao lên một không gian con của $\mathbb{C}^m$ cũng có thể được xây dựng với một cơ sở tùy ý, không cần thiết là trực giao. Giả sử không gian con được sinh bởi các vector độc lập tuyến tính $\{a_1, \ldots, a_n\}$, và cho A là ma trận $m \times n$ mà cột thứ j của nó là a_j.

Ngẫu nhiên từ v tới phép chiếu trực giao $y \in range(A)$ của nó, $y - v$ phải trực giao với $range(A)$. Điều này tương đương với phát biểu y phải thỏa mãn $a_j^*(y - v) = 0$ với mọi j. Vì $y \in range(A)$ nên ta có thể đặt $y = Ax$ và viết điều kiện này như $a_j^*(Ax - v) = 0$ với mọi j, hay $A^*(Ax - v) = 0$ hoặc $A^*Ax = A^*v$. Dễ dàng thấy vì A có hạng đầy đủ nên A^*A là không suy biến. Do đó,

$$x = (A^*A)^{-1}A^*v. \tag{2.1.12}$$

Cuối cùng, phép chiếu của $v, y = Ax$, là $y = A(A^*A)^{-1}A^*v$. Do đó, phép chiếu trực giao lên $range(A)$ có thể được biểu diễn bởi công thức

$$P = A(A^*A)^{-1}A^*A. \tag{2.1.13}$$

2.2 Phân tích QR

2.2.1 Phân tích QR được giảm

Các không gian *liên tiếp* được sinh bởi các cột $a_1, a_2, \ldots$ của A:

$$< a_1 > \subseteq < a_1, a_2 > \subseteq < a_1, a_2, a_3 > \subseteq \ldots$$

Do đó, $< a_1 >$ là không gian 1 chiều sinh bởi a_1, $< a_1, a_2 >$ là không gian 2 chiều sinh bởi a_1 và a_2, ... Ý tưởng của phân tích QR là xây dựng một chuỗi các vector trực giao $q_1, q_2, \ldots$ mà chúng sinh ra các không gian liên tiếp này.
Giả sử $A \in \mathbb{C}^{m\times n}(m \geqslant n)$ có hạng đầy đủ là n. Ta muốn chuỗi $q_1, q_2, \ldots$ có tính chất

$$< q_1, q_2, \ldots, q_j > = < a_1, a_2, \ldots, a_j >,\ j = 1, \ldots, n. \tag{2.2.1}$$

Từ mục 1.2, ta có điều kiện

$$\left[\begin{array}{c|c|c|c} a_1 & a_2 & \ldots & a_n \end{array}\right] = \left[\begin{array}{c|c|c|c} q_1 & q_2 & \ldots & q_n \end{array}\right] \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ & r_{22} & & \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}, \tag{2.2.2}$$

với các phần tử đường chéo r_{kk} khác 0 - nếu (2.2.2) đúng thì $a_1, \ldots, a_k$ có thể được biểu diễn như là tổ hợp tuyến tính của $q_1, \ldots, q_k$, và nghịch đảo của khối $k \times k$ ở trên bên trái của ma trận tam giác. Do đó, $q_1, \ldots, q_k$ có thể được biểu diễn như tổ hợp tuyến tính của $a_1, \ldots, a_k$. Các phương trình này có dạng

$$\begin{aligned} a_1 &= r_{11}q_1, \\ a_2 &= r_{12}q_1 + r_{22}q_2, \\ a_3 &= r_{13}q_1 + r_{23}q_2 + r_{33}q_3, \\ &\vdots \\ a_n &= r_{1n}q_1 + r_{2n}q_2 + \ldots + r_{nn}q_n. \end{aligned} \tag{2.2.3}$$

Khi đó, ta có

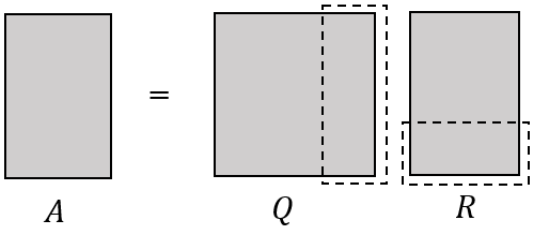
$$A = \widehat{Q}\widehat{R}, \tag{2.2.4}$$

với $\widehat{Q}$ là ma trận $m \times n$ với các cột trực giao và $\widehat{R}$ là ma trận tam giác trên $n \times n$. Phân tích như vậy được gọi là *phân tích QR được giảm của A*.

2.2.2 Phân tích QR đầy đủ

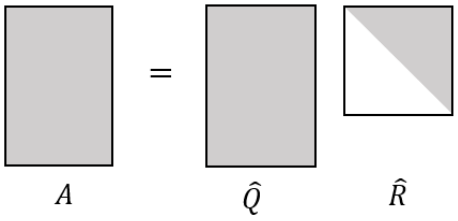
Tương tự SVD được giảm thành SVD đầy đủ ở trong mục trước. *Phân tích QR đầy đủ* của $A \in \mathbb{C}^{m\times n}(m \geqslant n)$ có được bằng việc thêm $m-n$ cột trực giao vào $\widehat{Q}$ sao cho nó trở thành ma trận Unita Q có kích thước $m \times m$. Các dòng 0 được thêm vào $\widehat{R}$ để nó trở thành ma trận R có $m \times n$, vẫn là ma trận tam giác trên. Phân tích QR đầy đủ và được giảm có mối quan hệ như sau

Trong phân tích QR đầy đủ, Q là ma trận $m \times m$, R là ma trận $m \times n$, và $m-n$ cột



Hình 2.3: Phân tích QR đầy đủ ($m \geqslant n$)

cuối cùng của Q được nhân với 0 trong R (bao bọc bởi các đường đứt nét). Trong phân tích QR được giảm, các cột và các dòng không được nói đến bị loại bỏ. Ma trận $\widehat{Q}$ là ma trận $m \times n$, $\widehat{R}$ là ma trận $n \times n$, và $\widehat{R}$ không có dòng nào là 0.



Hình 2.4: Phân tích QR được giảm

Chú ý trong phân tích QR đầy đủ, các cột q_j với $j > n$ là trực giao với $range(A)$. Giả sử A là ma trận có hạng đầy đủ là n thì chúng tạo thành một cơ sở trực giao cho $range(A)^{\perp}$ (không gian trực giao với $range(A)$), hoặc tương đương cho $null(A^*)$.

2.2.3 Trực giao hóa Gram - Schmidt

Phương trình (2.2.3) đưa ra một phương pháp cho việc tính phân tích QR được giảm. Cho $a_1, a_2, \ldots,$ ta có thể xây dựng các vector $q_1, q_2, \ldots$ và các phần tử r_{ij} bằng một qúa trình trực giao hóa liên tiếp, được biết như *trực giao hóa Gram - Schmidt.*

Tại bước thứ j, ta mong tìm một vector đơn vị $q_j \in < a_1, \ldots, a_j >$ trực giao với $q_1, \ldots, q_{j-1}$. Khi điều này xảy ra, ta đã xét kỹ thuật trực giao hóa cần thiết trong (1.3.6). Từ phương

trình đó, ta thấy rằng

$$v_j = a_j - (q_1^* a_j)q_1 - (q_2^*)q_2 - \ldots - (q_{j-1}^* a_j)q_{j-1} \tag{2.2.5}$$

là một loại vector được yêu cầu, ngoại trừ nó không được trực chuẩn hóa. Nếu ta chia cho $\|v_j\|_2$ thì kết quả là một vector phù hợp q_j.
Ta viết lại (2.2.3) thành dạng

$$\begin{aligned} q_1 &= \frac{a_1}{r_{11}}, \\ q_2 &= \frac{a_2 - r_{12}q_1}{r_{22}}, \\ q_3 &= \frac{a_3 - r_{13}q_1 - r_{23}q_2}{r_{33}}, \\ &\vdots \\ q_n &= \frac{a_n - \sum_{i=1}^{n-1} r_{in}q_i}{r_{nn}}. \end{aligned} \tag{2.2.6}$$

Từ (2.2.5), một định nghĩa xấp xỉ cho các hệ số r_{ij} trong các tử số của (2.2.6) là

$$r_{ij} = q_i^* a_j \ (i \neq j). \tag{2.2.7}$$

Các hệ số r_{ij} trong các mẫu số được chọn cho sự trực chuẩn hóa:

$$|r_{ij}| = \|a_j - \sum_{i=1}^{j-1} r_{ij}q_i\|_2. \tag{2.2.8}$$

Chú ý dấu của r_{ij} không được xác định nên ta có thể chọn $r_{ij} > 0$, trong trường hợp mà ta sẽ hoàn thành phân tích $A = \widehat{Q}\widehat{R}$ mà $\widehat{R}$ có các phần tử dương trên đường chéo.
Thuật toán được thể hiện trong (2.2.6) - (2.2.8) là bước lặp Gram - Schmidt. Theo toán học, nó đưa ra một dây truyền đơn giản để hiểu và chứng minh các tính chất khác nhau của các phân tích QR. Theo số học, nó trả ra kết quả là không ổn định bởi vì việc làm tròn các sai số trong máy tính. Để nhấn mạnh tính không ổn định, các nhà phân tích số xem điều này như *bước lặp Gram - Schmidt cổ điển*, đối lập với *bước lặp Gram - Schmidt được giảm.*

Thuật toán 2.1 Gram - Schmidt cổ điển (không ổn định)

```
for j = 1 to n do
    v_j = a_j
    for i = 1 to j − 1 do
        r_ij = q_i^* a_j
        v_j = v_j − r_ij q_i
    end for
    r_ij = ‖v_j‖_2
    q_j = v_j / r_jj
end for
```

2.2.4 Sự tồn tại và tính duy nhất

Tất cả các ma trận có các phân tích QR, và dưới các hạn chế phù hợp, chúng là duy nhất. Ta bắt đầu kết quả tồn tại đầu tiên.

Định lý 2.2.1 *Mọi* $A \in \mathbb{C}^{m\times n}$ $(m \geqslant n)$ *có một phân tích QR đầy đủ, do đó cũng có một phân tích QR được giảm.*

Chứng minh. Giả sử A có hạng đầy đủ và ta chỉ muốn phân tích QR được giảm. Trong trường hợp này, chứng minh tồn tại được cung cấp bởi chính thuật toán Gram - Schmidt. Quá trình này sinh ra các cột trực giao của $\widehat{Q}$ và các phần tử của $\widehat{R}$ sao cho (2.2.4) đúng. Thất bại có thể xảy ra khi tại một vài bước bất kì, v_j là 0 và do đó nó không thể được trực chuẩn hóa để đưa ra q_j.
Tuy nhiên, điều này sẽ kéo theo $a_j \in < q_1, \ldots, q_{j-1} > = < a_1, \ldots, a_{j-1} >$, mâu thuẫn với giả thuyết A có hạng đầy đủ.
Giả sử A không có hạng đầy đủ. Khi đó tại nhiều hơn một bước j, (2.2.5) cho $v_j = 0$. Bây giờ, ta chọn một cách đơn giản q_j tùy ý để là vector được chuẩn hóa bất kì trực giao với $< q_1, \ldots, q_{j-1} >$, và khi đó tiếp tục quá trình Gram - Schmidt.
Cuối cùng, phân tích QR đầy đủ (hay xa hơn QR được giảm) của một ma trận $m \times n$ với $m > n$ có thể được xây dựng bằng việc đưa ra các vector trực giao tùy ý trong cùng mô hình. Quá trình Gram - Schmidt qua bước n, khi đó tiếp tục thêm vào $m - n$ bước, đưa ra các vector q_j tại mỗi bước.

Bây giờ ta chuyển sang tính duy nhất. Giả sử $A = \widehat{Q}\widehat{R}$ là một phân tích QR được giảm. Nếu cột thứ i của $\widehat{Q}$ được nhân với z và dòng thứ i của $\widehat{R}$ được nhân với z^{-1} với vô hướng z bất kì thỏa $|z| = 1$, ta được phân tích QR khác của A.

Định lý 2.2.2 *Mỗi* $A \in \mathbb{C}^{m\times n}$ $(m \geqslant n)$ *hạng đầy đủ có duy nhất một phân tích QR được giảm* $A = \widehat{Q}\widehat{R}$ *với* $r_{ij} > 0$.

Chứng minh. Nhắc lại, chứng minh được cung cấp bởi bước lặp Gram - Schmidt. Từ (2.2.4), tính trực giao các cột của $\widehat{Q}$, và tính chất tam giác trên của $\widehat{R}$, phân tích QR được giảm bất kì của A phải thỏa (2.2.6) - (2.2.8). Theo giả thuyết hạng đầy đủ, các mẫu số (2.2.8) của (2.2.6) là khác 0, và do đó tại mỗi bước j liên tiếp, các công thức này xác định một cách đầy đủ r_{ij} và q_j, ngoài trừ dấu của r_{ij} chưa được chỉ định trong (2.2.8). Điều này được cố định bằng điều kiện $r_{ij} > 0$ như trong Thuật toán 2.1, phân tích được xác định một cách đầy đủ.

2.2.5 Khi các vector trở thành các hàm liên tục

Giả sử ta thay thế $\mathbb{C}^m$ bằng $L^2[-1, 1]$, không gian vector của các hàm có giá trị phức trong [-1,1]. Ta sẽ không đưa ra các tính chất của không gian này. Tích trong của f và

g có dạng

$$(f,g) = \int_{-1}^{1} \overline{f(x)} g(x) dx. \tag{2.2.9}$$

Ví dụ, xét "ma trận" theo sau mà "các cột" của nó là các đơn thức x^j:

$$A = \left[1 \middle| x \middle| x^2 \middle| \dots \middle| x^{n-1} \right], \tag{2.2.10}$$

Mỗi cột là một hàm trong $L^2[-1,1]$. Do đó, trong khi A là rời rạc như trong phương nằm ngang thông thường, nó liên tục trong phương thẳng đứng. Nó là mô hình liên tục của các ma trận Vandermonde (4.3.4) của mục Ví dụ (1.2.1).

"Phân tích QR liên tục" của A có dạng

$$A = QR = \left[q_0(x) \middle| q_1(x) \middle| \dots \middle| q_{n-1}(x) \right] \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & & \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix},$$

với các cột của Q là các hàm của x, trực giao đối với tích trong (2.2.9)

$$\int_{-1}^{1} \overline{q_i(x)} q_j(x) dx = \delta_{ij} = \begin{cases} 1 & \text{nếu } i = j, \\ 0 & \text{nếu } i \neq j. \end{cases}$$

Từ việc xây dựng Gram - Schmidt ta có thể thấy rằng q_j là một đa thức bậc j. Các đa thức này là các bội vô hướng của *các đa thức Legendre*, P_j, mà chúng được trực chuẩn để $P_j(1) = 1$. Một vài P_j đầu tiên là

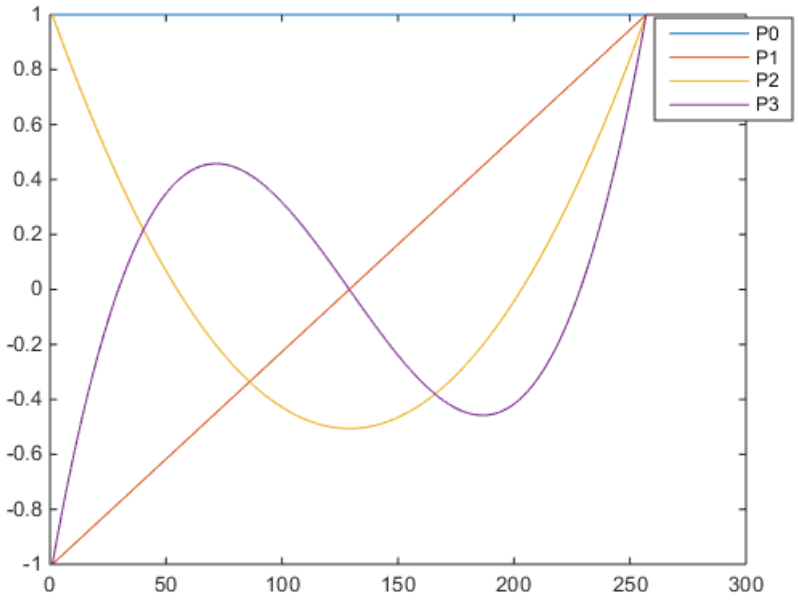
$$P_0(x) = 1,\ P_1(x) = x,\ P_2(x) = \frac{3}{2}x^2 - \frac{1}{2},\ P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x; \tag{2.2.11}$$

thấy trong Hình 2.5. Giống như các đơn thức $1, x, x^2, \dots$, chuỗi các đa thức này sinh ra các không gian các đa thức bậc cao hơn liên tiếp nhau. Tuy nhiên, $P_0(x), P_1(x), P_2(x), \dots$ là trực giao nhau. Thật vậy, tính toán với các đa thức như vậy tạo thành cơ sở trực giao của *các phương pháp phổ*, một trong những kỹ thuật mạnh nhất cho lời giải số của các phương trình đạo hàm riêng.

"Phép chiếu ma trận" $\widehat{Q}\widehat{Q}^*$ (2.1.6) kết hợp với $\widehat{Q}$ là một "ma trận $[-1,1]\times[-1,1]$", nghĩa là một toán tử tích phân

$$f(\cdot) \mapsto \sum_{j=0}^{n-1} q_j(\cdot) \int_{-1}^{1} \overline{q_j(x)} f(x) dx \tag{2.2.12}$$

ánh xạ các hàm trong $L^2[-1,1]$ vào các hàm trong $L^2[-1,1]$.



Hình 2.5: Bốn đa thức Legendre đầu tiên trong (2.2.11) ($[1, x, x^2, x^3]$)

2.2.6 Giải phương trình Ax = b bằng phân tích QR

Giả sử ta muốn giải phương trình $Ax = b$ cho biến x, với $A \in \mathbb{C}^{m\times m}$ là không suy biến. Nếu $A = QR$ là một phân tích QR thì ta có thể viết $QRx = b$, hoặc

$$Rx = Q^*b. \tag{2.2.13}$$

Vế bên phải của phương trình này tính dễ dàng nếu biết Q và hệ phương trình tuyến tính ẩn trong vế bên trái cũng giải dễ dàng bởi vì nó là tam giác. Phương pháp giải phương trình $Ax = b$ được đề xuất như sau:

1. Tính phân tích QR $A = QR$.
2. Tính $y = Q^*b$.
3. Giải $Rx = y$ cho x.

Kết hợp từ 1-3 là một phương pháp thông minh cho việc giải hệ phương trình tuyến tính. Tuy nhiên, nó không là phương pháp cho các bài toán như vậy. Khử Gauss là một thuật toán tổng quát được sử dụng trong thực hành vì nó chỉ yêu cầu phân nửa phép toán số học.

2.3 Trực giao hóa Gram - Schmit

2.3.1 Phép chiếu Gram - Schmidt

Cho $A \in \mathbb{C}^{m\times n}(m \geqslant n)$ là ma trận có hạng đầy đủ với các cột $\{a_j\}$. Trước đó, ta biểu diễn bước lặp Gram - Schmidt bằng các công thức (2.2.6) - (2.2.8). Xét chuỗi các công

thức

$$q_1 = \frac{P_1 a_1}{\|P_1 a_1\|}, \quad q_2 = \frac{P_2 a_2}{\|P_2 a_2\|}, \ldots, q_n = \frac{P_n a_n}{\|P_n a_n\|}. \tag{2.3.1}$$

Trong các công thức này, mỗi P_j là một phép chiếu trực giao. Đặc biệt, P_j là ma trận $m \times m$ có hạng $m-(j-1)$ mà nó chiếu trực giao $\mathbb{C}^m$ lên không gian trực giao với $< q_1, \ldots, q_{j-1} >$. (Trong trường hợp $j = 1$, $P_1 = I$). Ta thấy q_j được xác định như trong (2.3.1) là trực giao với $q_1, \ldots, q_{j-1}$, nằm trong không gian $< a_1, \ldots, a_j >$, và có chuẩn bằng 1. Khi đó, (2.3.1) tương đương với (2.2.6) - (2.2.8) và do đó tương đương với Thuật toán 2.1.

Cho $\widehat{Q}_{j-1}$ là ma trận $m \times (j-1)$ chứa $j-1$ cột đầu tiên của $\widehat{Q}$,

$$\widehat{Q}_{j-1} = \left[q_1 \middle| q_2 \middle| \ldots \middle| q_{j-1} \right]. \tag{2.3.2}$$

Khi đó, P_j được cho bởi

$$P_j = I - \widehat{Q}_{j-1} \widehat{Q}^*_{j-1}. \tag{2.3.3}$$

2.3.2 Thuật toán Gram - Schmidt được sửa đổi

Với mỗi giá trị j, Thuật toán 2.1 tính phép chiếu trực giao đơn có hạng $m-(j-1)$,

$$v_j = P_j a_j. \tag{2.3.4}$$

Ngược lại, thuật toán Gram - Schmidt được sửa đổi tính kết quả giống nhau bằng một chuỗi $j-1$ phép chiếu có hạng $m-1$. Nhắc lại từ (2.1.9), $P_{\perp q}$ là phép chiếu trực giao có hạng $m-1$ lên không gian trực giao với một vector $q \in \mathbb{C}^m$ khác 0. Theo định nghĩa của P_j, ta có

$$P_j = P_{\perp q_{j-1}} \ldots P_{\perp q_2} P_{\perp q_1}, \tag{2.3.5}$$

nhắc lại $P_1 = I$. Do đó một phát biểu tương đương với (2.3.4) là

$$v_j = P_{\perp q_{j-1}} \ldots P_{\perp q_2} P_{\perp q_1} a_j. \tag{2.3.6}$$

Thuật toán Gram - Schmidt được sửa đổi sử dụng (2.3.6) thay vì (2.3.4).

Theo toán học, (2.3.4) và (2.3.6) tương đương nhau. Tuy nhiên, các chuỗi phép toán số học bao hàm bởi công thức này là khác nhau. Thuật toán được sửa đổi tính v_j bằng việc đánh giá các công thức sau

$$\begin{aligned}
v_j^{(1)} &= a_j, \\
v_j^{(2)} &= P_{\perp q_1} v_j^{(1)} = v_j^{(1)} - q_1 q_1^* v_j^{(1)}, \\
v_j^{(3)} &= P_{\perp q_1} v_j^{(2)} = v_j^{(2)} - q_2 q_2^* v_j^{(2)}, \\
&\vdots \qquad\qquad \vdots \\
v_j = v_j^{(j)} &= P_{\perp q_{j-1}} v_j^{(j-1)} = v_j^{(j-1)} - q_{j-1} q_{j-1}^* v_j^{(j-1)}.
\end{aligned} \tag{2.3.7}$$

Trong số học tính toán độ chính xác hữu hạn, ta sẽ thấy rằng (2.3.7) đưa ra các sai số nhỏ hơn (2.3.4).

Khi thuật toán được thực thi, phép chiếu $P_{\perp q_i}$ có thể được ứng dụng một cách thuận lợi cho $v_j^{(i)}$ với $j > i$ ngay sau khi q_i được biết.

Thuật toán 2.2 Gram - Schmidt được sửa đổi

```
for i = 1 to n do
    v_i = a_i
end for
for i = 1 to n do
    r_ii = ||v_i||
    q_i = v_i / r_ii
    for j = i + 1 to n do
        r_ij = q_i^* v_j
        v_j = v_j − r_ij q_i
    end for
end for
```

2.3.3 Đếm số phép toán

Mỗi phép cộng, phép trừ, phép nhân, phép chia hoặc căn bậc hai đếm như là một phép toán dấu chấm động (floating point operations - flops). Ta không phân biệt giữa số học thực và phức, mặc dù trong thực hành trong hầu hết các máy tính có sự khác nhau khá lớn.

Thật vậy, có nhiều hơn cho chi phí của thuật toán hơn là đếm số phép toán. Trong máy tính xử lý đơn, thời gian thực thi bị ảnh hưởng bởi sự di chuyển dữ liệu giữa các phần tử của hệ thống cấp bậc trong bộ nhớ và việc cạnh tranh các công việc đang chạy trong cùng một xử lý. Trong các máy hệ thống đa xử lý việc này trở nên phức tạp hơn, sự giao tiếp giữa các xử lý thỉnh thoảng đưa thông tin quan trọng lớn hơn nhiều của các "tính toán" hiện nay.

Định lý 2.3.1 *Thuật toán 2.1 và 2.2 cần* $\sim 2mn^2$ *phép toán dấu chấm động để tính phân tích QR của một ma trận A có* $m \times n$.

Chú ý rằng thuật toán chỉ biểu diễn số hạng dẫn đầu của số phép toán dấu chấm động. Ký hiệu "$\sim$" có nghĩa tiệm cận thông thường của nó:

$$\lim_{m,n\to\infty} \frac{\text{số phép toán dấu chấm động}}{2mn^2} = 1.$$

Định lý 2.3.1 có thể được thiết lặp như sau. Để xác định, xét thuật toán Gram - Schmidt được sửa đổi (thuật toán 2.2). Khi m và n lớn, các phép toán trong vòng lặp ở trong cùng:

$$r_{ij} = q_i^* v_j,$$

$$v_j = v_j - r_{ij} q_i.$$

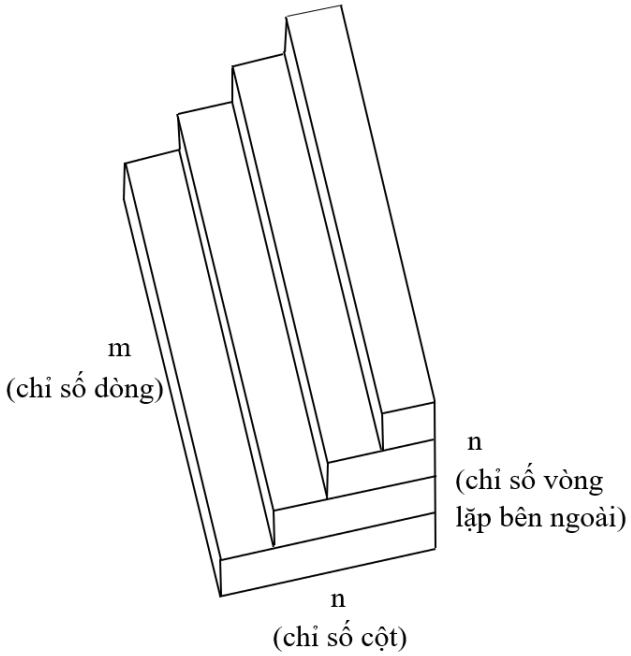
Dòng đầu tiên tính tích trong $q_i^* v_j$ cần m phép nhân và $m-1$ phép cộng. Dòng thứ hai tính $v_j - r_{ij} q_i$ cần m phép nhân và m phép trừ. Tổng số việc được bao gồm môt bước lặp đơn bên trong là $\sim 4m$ phép toán dấu chấm động, hay 4 phép toán dấu chấm động trên phần tử vector cột. Do đó, số phép toán dấu chấm động cần cho thuật toán là tiệm cận

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} 4m \sim \sum_{i=1}^{n} (i) 4m \sim 2mn^2. \tag{2.3.8}$$

2.3.4 Đếm số phép toán theo hình học

Tại bước đầu tiên của vòng lặp bên ngoài, Thuật toán 2.2 chạy trong toàn bộ ma trận, việc trừ một bôi của cột 1 từ các cột khác. Tại bước thứ hai, nó tính toán trong một ma trận con, việc trừ một bội của cột 2 từ cột $3, \ldots, n$. Tiếp tục cách này, tại mỗi bước số chiều cột rút lại bởi 1 cho tới bước cuối cùng, chỉ cột n được sửa đổi. Thủ tục này có thể được biểu diễn bằng biểu đồ theo sau:

Hình chữ nhật $m \times n$ tại đáy tương ứng bước đầu tiên qua vòng lặp bên ngoài, hình



chữ nhật $m \times (n-1)$ ở trên nó tương ứng bước thứ hai, ...

Khi $m, n \to \infty$, số phép toán cho trực giao hoá Gram - Schmidt tỉ lệ với thể tích của hình ở trên. Hai bước của vòng lặp bên trong tương ứng với 4 phép toán tại vị trí mỗi ma trận là 4 flop . Khi $m, n \to \infty$, hình hội tụ tới lăng trụ tam giác vuông với thể tích $mn^2/2$. Nhân với 4 flop trên 1 đơn vị thể tích

$$\text{Trực giao hóa Gram - Schmidt: } \sim 2mn^2 \text{ flop.} \tag{2.3.9}$$

2.3.5 Gram - Schmidt như trực giao hóa tam giác

Mỗi bước bên ngoài của thuật toán Gram - Schmidt được sửa đổi có thể được làm sáng tỏ như phép nhân phải với một ma trận tam giác trên vuông. Ví dụ, bắt đầu với A, bước lặp đầu tiên nhân cột đầu tiên a_1 với $1/r_{11}$ và khi đó trừ r_{1j} lần kết quả này với mỗi cột còn lại a_j. Điều này tương đương với phép nhân phải với ma trận R_1:

$$\left[\begin{array}{c|c|c|c} v_1 & v_2 & \dots & v_n \end{array}\right] \begin{bmatrix} \frac{1}{r_{11}} & \frac{-r_{12}}{r_{11}} & \frac{-r_{13}}{r_{11}} & \dots \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{bmatrix} = \left[\begin{array}{c|c|c|c} q_1 & v_2^{(2)} & \dots & v_n^{(2)} \end{array}\right].$$

Tổng quát, bước thứ i của Thuật toán 2.2 trừ r_{ij}/r_{ii} lần cột i của A với các cột $j > i$ và thay thế cột i bằng $1/r_{ii}$ lần cột i. Điều này tương ứng với phép nhân ma trận với ma trận tam giác trên R_i:

$$R_2 = \begin{bmatrix} 1 & & & \\ & \frac{1}{r_{22}} & \frac{-r_{23}}{r_{22}} & \dots \\ & & 1 & \\ & & & \ddots \end{bmatrix}, \; R_3 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \frac{1}{r_{33}} & \\ & & & \ddots \end{bmatrix}, \dots$$

Tại bước lặp cuối cùng, ta có

$$A \underbrace{R_1 R_2 \dots R_n}_{\widehat{R}^{-1}} = \widehat{Q}. \tag{2.3.10}$$

Công thức này chứng minh rằng thuật toán Gram - Schmidt là một phương pháp của *trực giao hóa tam giác*. Nó áp dụng các phép toán tam giác vào bên phải của một ma trận để giảm nó thành một ma trận với các cột trực giao. Dĩ nhiên, trong thực hành, ta không làm thành các ma trận R_i và nhân chúng với nhau rõ ràng.

2.4 Tam giác hóa Householder

Thuật toán Householder là một quá trình của "tam giác hóa trực giao", làm một ma trận tam giác bằng một chuỗi các phép toán ma trận Unita.

2.4.1 Householder và Gram - Schmidt

Như ta thấy trong mục (2.3), bước lặp Gram - Schmidt áp dụng liên tiếp các ma trận tam giác cơ bản R_k vào bên phải của A, để được ma trận kết quả

$$A \underbrace{R_1 R_2 \dots R_n}_{\widehat{R}^{-1}} = \widehat{Q}$$

có các cột trực giao. Tích $\widehat{R} = R_n^{-1} \dots R_2^{-1} R_1^{-1}$ cũng là ma trận tam giác trên, và do đó $A = \widehat{Q}\widehat{R}$ là một phân tích QR được giảm của A.

Ngược lại, phương pháp Householder áp dụng liên tiếp các ma trận Unita Q_k cơ bản vào bên trái của A nên ma trận kết quả

$$\underbrace{Q_n \dots Q_2 Q_1}_{Q^*} A = R$$

là ma trận tam giác trên. Tích $Q = Q_1^* Q_2^* \dots Q_n^*$ cũng là ma trận Unita, và do đó $A = QR$ là một phân tích QR đầy đủ của A.

Do đó, hai phương pháp có thể được tóm tắt như sau:

Gram - Schmidt: trực giao hóa tam giác,
Householder: tam giác hóa trực giao.

2.4.2 Tam giác hóa bằng việc đưa vào các số 0

Phương pháp Householder được đưa ra đầu tiên bởi Alston Householder trong năm 1958. Đây là một cách khéo léo của việc thiết kế các ma trận Unita Q_k sao cho $Q_n \dots Q_2 Q_1 A$ là ma trận tam giác trên.

Ma trận Q_k được chọn để đưa ra các số 0 bên dưới đường chéo trong cột thứ k trong khi nó bảo toàn các số 0 được đưa ra trước đó. Ví dụ, trong trường hợp 5×3, 3 phép toán Q_k được áp dụng. Trong các ma trận này, ký hiệu $\times$ biểu diễn một phần tử khác 0, và kiểu chữ đậm cho biết một phần tử vừa được thay đổi. Các phần tử để trống là 0.

$$\underset{A}{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}} \xrightarrow{Q_1} \underset{Q_1 A}{\begin{bmatrix} \mathbf{\times} & \mathbf{\times} & \mathbf{\times} \\ \mathbf{0} & \mathbf{\times} & \mathbf{\times} \\ \mathbf{0} & \mathbf{\times} & \mathbf{\times} \\ \mathbf{0} & \mathbf{\times} & \mathbf{\times} \\ \mathbf{0} & \mathbf{\times} & \mathbf{\times} \end{bmatrix}} \xrightarrow{Q_2} \underset{Q_2 Q_1 A}{\begin{bmatrix} \times & \times & \times \\ & \mathbf{\times} & \mathbf{\times} \\ & \mathbf{0} & \mathbf{\times} \\ & \mathbf{0} & \mathbf{\times} \\ & \mathbf{0} & \mathbf{\times} \end{bmatrix}} \xrightarrow{Q_3} \underset{Q_3 Q_2 Q_1 A}{\begin{bmatrix} \times & \times & \times \\ & \times & \times \\ & & \mathbf{\times} \\ & & \mathbf{0} \\ & & \mathbf{0} \end{bmatrix}} \tag{2.4.1}$$

Đầu tiên, Q_1 tính toán trong các dòng 1, ..., 5, việc đưa ra các số 0 nằm ở các vị trí (2,1), (3,1), (4,1) và (5,1). Tiếp theo, Q_2 tính toán trong các dòng 2, ..., 5, đưa ra các số 0 nằm ở các vị trí (3,2), (4,2) và (5,2) nhưng không triệt tiêu các số 0 được đưa ra bởi Q_1. Cuối cùng, Q_k tính toán trong các dòng 3, ..., 5, đưa ra các số 0 ở các vị trí (4,3), (5,3) không triệt tiêu bất kì số 0 nào được đưa ra trước đó.

Tổng quát, Q_k tính toán trong các dòng $k, \dots, m$. Bắt đầu của bước k, có 1 khối các số 0 trong $k-1$ cột đầu tiên của các dòng này. Áp dụng của Q_k hình thành các tổ hợp tuyến tính của các dòng này, và các tổ hợp tuyến tính của các phần tử 0 còn lại là 0.

Sau n bước, tất cả các phần tử nằm bên dưới đường chéo đã được khử và $Q_n \dots Q_2 Q_1 A$ là ma trận tam giác trên.

2.4.3 Phản xạ Householder

Mỗi Q_k được chọn để là một ma trận Unita dạng

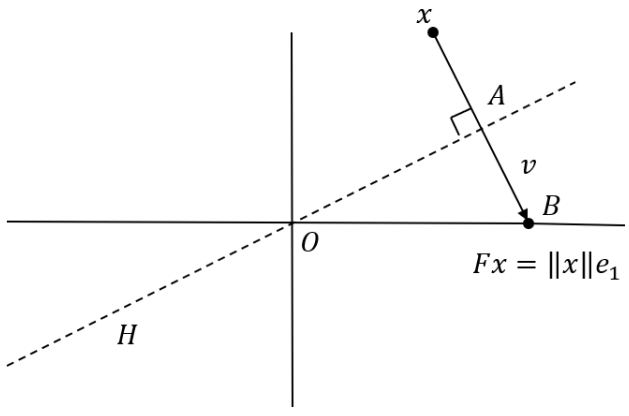
$$Q_k = \begin{bmatrix} I & 0 \\ 0 & F \end{bmatrix}, \tag{2.4.2}$$

với I là ma trận đơn vị $(k-1) \times (k-1)$ và F là ma trận Unita $(m-k+1) \times (m-k+1)$. Phép nhân với F phải đưa vào các số 0 vào cột thứ k. Thuật toán Householder chọn F là một ma trận đặc biệt được gọi là *phản xạ Householder*.

Giả sử, bắt đầu bước k, các phần tử $k, \dots, m$ của cột thứ k được cho bởi vector $x \in \mathbb{C}^{m-k+1}$. Để đưa chính xác các số 0 vào cột thứ k, phản xạ Householder F nên tác động ánh xạ theo sau

$$x = \begin{bmatrix} \times \\ \times \\ \times \\ \vdots \\ \times \end{bmatrix} \xrightarrow{F} Fx = \begin{bmatrix} \|x\| \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \|x\| e_1. \tag{2.4.3}$$

Ý tưởng cho việc thực hiện này được cho biết trong Hình 2.6. Phản xạ F sẽ phản xạ không gian $\mathbb{C}^{m-k+1}$ qua một siêu phẳng H trực giao với $v = \|x\| e_1 - x$. Một *siêu phẳng* là sự tổng quát hóa số chiều cao hơn của mặt phẳng 2 chiều trong không gian 3 chiều - một không gian con 3 chiều của một không gian 4 chiều, một không gian con 4 chiều của một không gian 5 chiều, ... Tổng quát, một siêu phẳng có thể được đặc trưng như tập hợp các điểm trực giao với một vector khác 0 được cố định. Trong Hình 2.6, vector đó là $v = \|x\| e_1 - x$, và đường nét gạch như là một miêu tả của H được xem là "bờ".



Hình 2.6: Phản xạ Householder

Khi phản xạ được áp dụng, mọi điểm trong bờ của siêu phẳng H được ánh xạ thành ảnh phản xạ của nó trong bờ khác. Đặc biệt, x được ánh xạ thành $\|x\|e_1$. Trong (2.1.11), với $y \in \mathbb{C}^m$ bất kì, vector

$$Py = \left(I - \frac{vv^*}{v^*v}\right) y = y - v\left(\frac{v^*y}{v^*v}\right) \tag{2.4.4}$$

là một phép chiếu trực giao của y vào không gian H. Để lấy phản xạ y qua H, ta phải lấy 2 lần hơn là trong cùng một phương. Do đó, phép chiếu Fy sẽ là

$$Fy = \left(I - 2\frac{vv^*}{v^*v}\right) y = y - 2v\left(\frac{v^*y}{v^*v}\right).$$

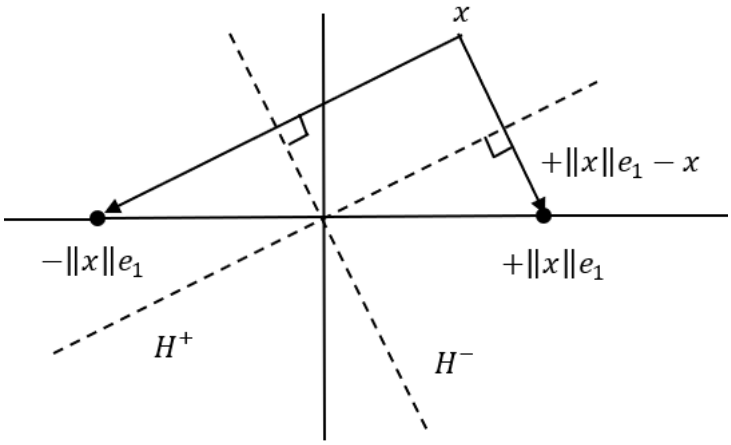
Do đó, ma trận F là

$$F = I - 2\frac{vv^*}{v^*v}. \tag{2.4.5}$$

Chú ý rằng phép chiếu P (hạng $m-1$) và phản xạ F (hạng đầy đủ, Unita) chỉ khác nhau trong biểu biễn một thừa số của 2.

2.4.4 Ưu thế của 2 phản xạ

Trong (2.4.3) và trong Hình 2.6 ta có các vấn đề đã được đơn giản, thật vậy, có nhiều phản xạ Householder mà chúng sẽ đưa ra các số 0 cần thiết. Vector x có thể được lấy phản xạ thành $z\|x\|e_1$, với z là vô hướng bất kì thỏa $|z| = 1$. Trong trường hợp số phức, có một vòng tròn của các phản xạ có thể thực hiện được, và ngay trong trường hợp số thực, có 2 lựa chọn, được miêu tả bởi các phản xạ qua hai siêu phẳng khác nhau, H^+ và H^-, như được miêu tả trong Hình 2.7.



Hình 2.7: Hai phản xạ có thể thực hiện được

Theo toán học, mỗi sự lựa chọn dấu là thỏa mãn. Tuy nhiên, đây là trường hợp mà mục tiêu của tính ổn định số đưa ra một lựa chọn sẽ được lấy hơn một cái khác. Cho tính ổn định số, đáng mong ước để lấy đối xứng x thành vector $z\|x\|e_1$ mà không quá gần x.

Để đạt được điều này, ta có thể chọn $z = -sign(x_1)$, với x_1 là thành phần đầu tiên của x, để mà vector phản xạ trở thành $v = -sign(x_1)\|x\|e_1 - x$, hay theo các thừa số -1

$$v = sign(x_1)\|x\|e_1 + x. \tag{2.4.6}$$

Để làm điều này đầy đủ, ta có thể đặt vào tùy ý quy ước $sign(x_1) = 1$ nếu $x_1 = 0$.
Không quá khó để thấy vì sao sự lựa chọn dấu làm một sự khác nhau cho tính ổn định. Giả sử trong Hình 2.7, góc giữa H^+ và trục e_1 là rất nhỏ. Khi đó vector $v = \|x\|e_1 - x$ là nhỏ hơn x hoặc $\|x\|e_1$ nhiều. Do đó tính toán v biểu diễn một phép trừ các lượng gần và sẽ hướng tới để cho các sai số triệt tiêu nhau. Nếu ta lựa chọn dấu như trong (2.4.6), ta đảm bảo rằng $\|v\|$ là không bao giờ nhỏ hơn $\|x\|$.

2.4.5 Thuật toán

Bây giờ ta đưa ra thuật toán Householder đầy đủ. Nếu A là một ma trận, ta xác định $a_{i:i',j:j'}$ là ma trận con $(i' - i + 1) \times (j' - j + 1)$ của A với góc trái trên là a_{ij} và góc phải dưới là $a_{i',j'}$. Trong trường hợp đặc biệt mà ma trận con giảm xuống thành một vector con của một dòng hay cột, ta viết tương ứng là $A_{i,j:j'}$ hoặc $A_{i:i',j}$.
Thuật toán theo sau tính thừa số R của phân tích QR của một ma trận A có $m \times n$ với $m \geqslant n$. Theo cách này, n vector phản xạ $v_1, \ldots, v_n$ được lưu trữ sử dụng sau.

Thuật toán 2.3 Phân tích QR Householder

1: **for** $k = 1$ to n **do**
2: $\quad x = A_{k:m,k}$
3: $\quad v_k = sign(x_1)\|x\|_2 e_1 + x$
4: $\quad v_k = v_k / \|v_k\|_2$
5: $\quad A_{k:m,k:n} = A_{k:m,k:n} - 2v_k(v_k^* A_{k:m,k:n})$
6: **end for**

2.4.6 Áp dụng hoặc tạo thành Q

Theo Thuật toán 2.3, A đã được giảm xuống thành dạng tam giác trên. Đó là ma trận R trong phân tích QR $A = QR$. Tuy nhiên, ma trận Unita Q đã không được xây dựng, không có ma trận con $\widehat{Q}$ n cột của nó tương ứng thành phân tích QR được giảm. Việc xây dựng Q hoặc $\widehat{Q}$ đưa thêm vào, và trong nhiều ứng dụng, ta có thể tránh điều này bằng cách làm một cách trực tiếp với công thức

$$Q^* = Q_n \ldots Q_2 Q_1 \tag{2.4.7}$$

hoặc liên hợp của nó

$$Q = Q_1 Q_2 \ldots Q_n. \tag{2.4.8}$$

Ví dụ, trong Mục 2.2 ta thấy rằng một hệ thống vuông của các phương trình $Ax = b$ có thể được giải thông qua phân tích QR của A. Cách mà trong đó Q được sử dụng trong

quá trình này là nằm trong tính toán của tích Q^*b. Theo (2.4.7) ta có thể tính Q^*b bằng một chuỗi n phép toán được áp dụng cho b, các phép toán tương tự nhau được áp dụng cho A để làm nó thành ma trận tam giác. Thuật toán như sau.

Thuật toán 2.4 Tính toán ngầm của tích Q^*b

1: **for** $k = 1$ to n **do**
2: $\quad b_{k:m} = b_{k:m} - 2v_k(v_k^* b_{k:m})$
3: **end for**

Tương tự, tính toán tích Qx có thể được thực hiện bằng quá trình tương tự được thực thi trong thứ tự ngược lại.
Các thuật toán này có độ phức tạp là $O(mn)$ mà không phải là $O(mn^2)$ như trong Thuật toán 2.3.

Thuật toán 2.5 Tính toán ngầm của tích Q^*b

1: **for** $k = n$ down 1 **do**
2: $\quad x_{k:m} = x_{k:m} - 2v_k(v_k^* x_{k:m})$
3: **end for**

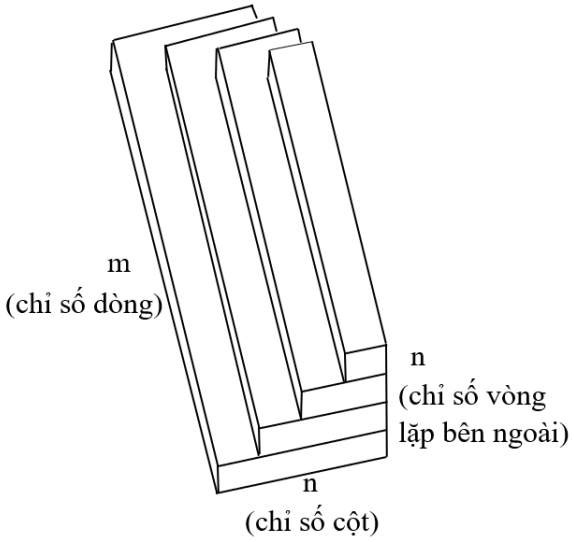
Ta có thể xây dựng QI thông qua Thuật toán 2.5 bằng việc tính các cột $Qe_1, Qe_2, \ldots, Qe_m$ của nó. Ngoài ra, ta có thể xây dựng Q^*I thông qua Thuật toán 2.4 và do đó kết quả là liên hợp. Một biến thể của ý tưởng này là xây dựng IQ bằng việc tính các dòng $e_1^*Q, e_2^*Q, \ldots, e_m^*Q$ như được đề nghị bởi (2.4.8). Ý tưởng tốt nhất là ý tưởng đầu tiên, dựa vào Thuật toán 2.5. Lý do là vì nó bắt đầu với các phép toán bao gồm Q_n, Q_{n-1}, và chỉ sửa đổi một phần nhỏ vector được áp dụng.
Nếu $\widehat{Q}$ cần thiết hơn là Q thì nó đủ để tính toán các cột $Qe_1, Qe_2, \ldots, Qe_n$.

2.4.7 Đếm số phép toán

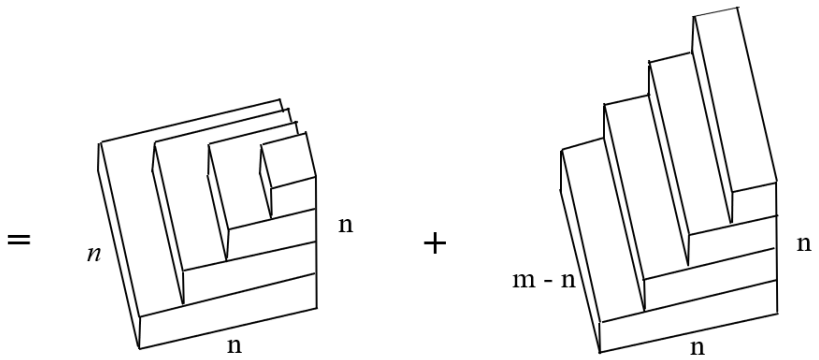
Thuật toán 2.3 được chi phối bởi vòng lặp ở trong cùng,

$$A_{k:m,j} - 2v_k(v_k^* A_{k:m,j}). \tag{2.4.9}$$

Nếu chiều dài vector là $l = m - k + 1$ thì tính toán này cần $4l - 1 \sim 4l$ phép toán vô hướng: l cho phép trừ, l cho phép nhân vô hướng, và $2l - 1$ cho tích vô hướng. Đó là ~ 4 phép toán dấu chấm động cho mỗi phần tử được thực hiện.
Ta có thể lấy tổng 4 phép toán dấu chấm động này trên một phần tử bằng lý do hình học như trong Mục 2.3. Mỗi bước liên tiếp của vòng lặp bên ngoài tính toán trong vài dòng bởi vì trong suốt bước k, các dòng $1, \ldots, k-1$ không được thay đổi. Hơn nữa, mỗi bước tính toán trong một vài cột bởi vì các cột của $1, \ldots, k-1$ của các dòng được tính toán là 0 và được bỏ qua. Do đó, công việc được hoàn thành bởi một bước bên ngoài có thể được biểu diễn bằng một lớp của hình ba chiều theo sau:



Tổng số phép toán tương ứng với 4 lần thể tích của hình ba chiều. Để xác định thể tích bằng hình ảnh ta có thể chia hình ba chiều thành 2 mảnh:



Hình ba chiều bên trái có hình dạng của kim tự tháp và hội tụ tới một hình chóp khi $n \to \infty$, với thể tích là $\frac{1}{3}n^3$. Hình ba chiều bên phải có hình dạng của một cầu thang và hội tụ tới hình lăng trụ khi $m, n \to \infty$, với thể tích là $\frac{1}{2}(m-n)n^2$. Kết hợp lại, thể tích $\sim \frac{1}{2}mn^2 - \frac{1}{6}n^3$. Việc nhân 4 phép toán dấu chấm động trên đơn vị thể tích, ta thấy

$$\text{Trực giao hóa Householder: } \sim 2mn^2 - \frac{2}{3}n^3 \text{ phép toán dấu chấm động.} \tag{2.4.10}$$

2.5 Các bài toán bình phương nhỏ nhất

2.5.1 Bài toán

Xét một hệ thống tuyến tính các phương trình có m phương trình, n ẩn ($m > n$). Một cách hình thức, ta mong ước tìm một vector $x \in \mathbb{C}^n$ sao cho $Ax = b$, với $A \in \mathbb{C}^{m \times n}$ và $b \in \mathbb{C}^m$. Tổng quát, một bài toán như vậy không có lời giải. Một vector x phù hợp chỉ

tồn tại nếu b nằm trong $range(A)$, và vì b là một vector m chiều, trong khi $range(A)$ có nhiều nhất n chiều, nên điều này chỉ đúng cho các lựa chọn ngoại lệ của b. Ta nói rằng một hệ thống hình chữ nhật của các phương trình với $m > n$ là *được xác định hầu hết*. Vector được biết như là *thặng dư*,

$$r = b - Ax \in \mathbb{C}^m, \tag{2.5.1}$$

có thể được làm khá nhỏ bằng một lựa chọn phù hợp của x nhưng tổng quát nó không thể được làm bằng 0.

Vì thặng dư r không thể bằng 0 nên làm nó nhỏ như có thể thực hiện được. Việc đo sự nhỏ nhất của r dẫn đến việc chọn một chuẩn. Nếu ta chọn chuẩn 2, bài toán đưa về dạng như sau:

$$\begin{gathered} \text{Cho } A \in \mathbb{C}^{m\times n}, m \geqslant n, b \in \mathbb{C}^m, \\ \text{tìm } x \in \mathbb{C}^n \text{ sao cho } \|b - Ax\|_2 \text{ được cực tiểu hóa.} \end{gathered} \tag{2.5.2}$$

Đó là công thức của *bài toán bình phương nhỏ nhất* tổng quát (tuyến tính). Việc chọn chuẩn 2 có thể được bảo vệ bởi các đối số hình học và thống kê khác nhau để đưa ra các thuật toán đơn giản bởi vì đạo hàm của một hàm bậc hai mà nó được đặt là 0 cho sự cực tiểu hóa là tuyến tính.

Chuẩn 2 tương ứng với khoảng cách Euclidean, nên có một sự giải thích hình học đơn giản của (2.5.2). Ta tìm một vector $x \in \mathbb{C}^n$ sao cho vector $Ax \in \mathbb{C}^m$ là một điểm gần b nhất trong $range(A)$.

2.5.2 Ví dụ: việc điều chỉnh dữ liệu đa thức

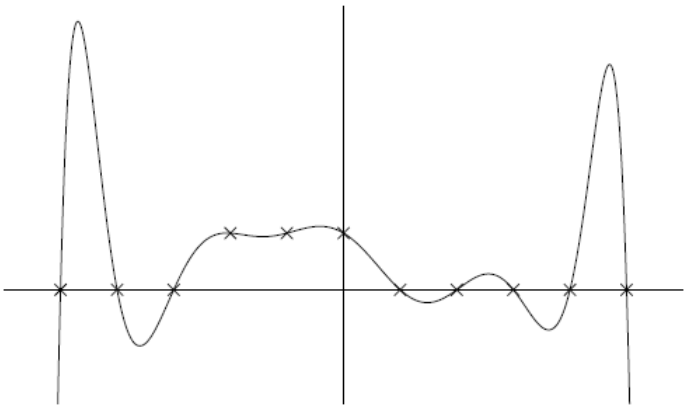
Ví dụ 2.5.1. (Nội suy đa thức) Giả sử ta được cho m điểm phân biệt $x_1, \ldots, x_m \in \mathbb{C}$ và dữ liệu $y_1, \ldots, y_m \in \mathbb{C}$ tại các điểm này. Khi đó, tồn tại duy nhất một *nội suy đa thức* bậc lớn nhất là $m-1$ tới các dữ liệu này trong các điểm này

$$p(x) = c_0 + c_1 x + \ldots + c_{m-1}x^{m-1}, \tag{2.5.3}$$

với tính chất mà tại mỗi điểm x_i, $p(x_i) = y_i$. Quan hệ giữa $\{x_i\}, \{y_i\}$ với các hệ số $\{c_i\}$ có thể được biểu diễn bởi hệ thống Vandermonde vuông được thấy trong Ví dụ 1.2.1:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{m-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{m-1} \\ 1 & x_3 & x_3^2 & \ldots & x_3^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \ldots & x_m^{m-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{m-1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix} \tag{2.5.4}$$

Để xác định các hệ số $\{c_i\}$ cho một tập dữ liệu được cho, ta có thể giải hệ thống các phương trình này mà nó được đảm bảo là không suy biến miễn là các điểm $\{x_i\}$ là phân biệt.



Hình 2.8: Nội suy đa thức bậc 10 của 11 điểm dữ liệu

Hình 2.8 đưa ra một ví dụ của quá trình nội suy đa thức. Ta có 11 điểm dữ liệu trong dạng sóng vuông rời rạc được biểu diễn bởi các dấu chữ thập và đường cong $p(x)$ đi qua chúng. Tuy nhiên, sự điều chỉnh là không dễ chịu chút nào. Gần cuối khoảng, $p(x)$ đưa ra sự giao động lớn mà chúng rõ ràng là một thành phần lạ của quá trình nội suy, không phản ánh dữ liệu hợp lý.

Xử lý không thỏa mãn này là đặc trưng của nội suy đa thức. Các điều chỉnh mà nó đưa ra kết quả thường là xấu và chúng đề cập đến trường hợp xấu nhất hơn là trường hợp tốt hơn nếu nhiều dữ liệu được sử dụng. Ngay cả khi sự điều chỉnh là tốt, quá trình nội suy có thể là điều kiện xấu, nghĩa là bị ảnh hưởng bởi các nhiễu của dữ liệu. Để tránh các bài toán này, ta có thể sử dụng một tập không đồng nhất các điểm nội suy như là các điểm Chebyshev trong khoảng [-1,1]. Tuy nhiên, trong nhiều ứng dụng, nó sẽ không thường xuyên có thể thực hiện được để chọn các điểm nội suy.

Ví dụ 2.5.2. (Điều chỉnh bình phương nhỏ nhất đa thức) Cho $x_1, \ldots, x_m$ và $y_1, \ldots, y_m$ như trong Ví dụ 2.6.1, xét một đa thức bậc $n-1$

$$p(x) = c_0 + c_1 x + \ldots + c_{n-1} x^{n-1} \tag{2.5.5}$$

với $n < m$. Một đa thức như vậy là điều chỉnh bình phương nhỏ nhất tới dữ liệu nếu nó cực tiểu hóa tổng bình phương của độ lệch từ dữ liệu,

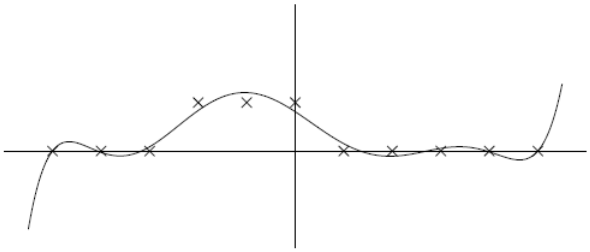
$$\sum_{i=1}^{m} |p(x_i) - y_i|^2. \tag{2.5.6}$$

Tổng bình phương này là tương đương với bình phương của chuẩn thặng dư $\|r\|_2^2$ cho hệ

thống Vandermonde hình chữ nhật

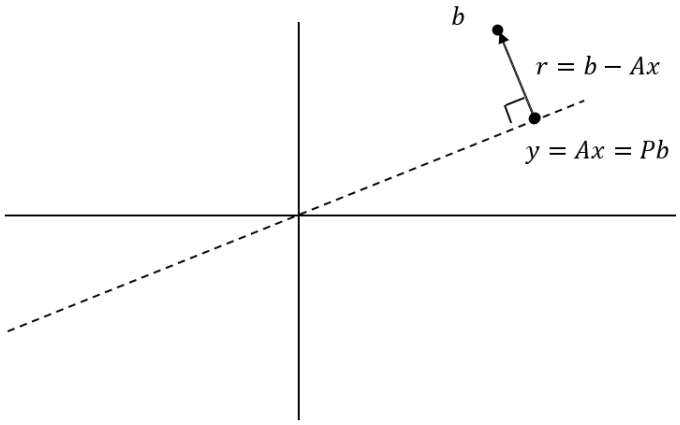
$$\begin{bmatrix} 1 & x_1 & & x_1^{n-1} \\ 1 & x_2 & \dots & x_2^{n-1} \\ 1 & x_3 & \dots & x_3^{n-1} \\ \vdots & & & \vdots \\ 1 & x_m & \dots & x_m^{n-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} \approx \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix} \tag{2.5.7}$$

Hình 2.8 minh họa điều mà ta thu được nếu ta điều chỉnh 11 điểm dữ liệu giống nhau từ ví dụ cuối cùng với đa thức bậc 7. Một đa thức mới không nội suy dữ liệu nhưng nó thu nạp tất cả tập tính của chúng nhiều hơn là đa thức trong Ví dụ 2.6.1. Mặc dù ta không thể thấy điều này trong hình, nó cũng chỉ bị ảnh hưởng ít hơn tới các nhiễu.



Hình 2.9: Đa thức bậc 7 điều chỉnh bình phương nhỏ nhất của 11 điểm dữ liệu giống nhau

2.5.3 Phép chiếu trực giao và các phương trình chuẩn tắc



Hình 2.10: Công thức của bài toán bình phương nhỏ nhất liên quan tới phép chiếu trực giao

Ý tưởng được minh họa trong Hình 2.10. Mục tiêu của chúng ta là tìm một điểm Ax gần với b nhất trong $range(A)$ để cho chuẩn của thặng dư $r = b - Ax$ được cực tiểu hóa. Rõ ràng theo hình học, điều này sẽ đưa ra $Ax = Pb$, trong đó $P \in \mathbb{C}^{m\times m}$ là một phép chiếu trực giao mà nó ánh xạ $\mathbb{C}^m$ vào $range(A)$. Mặt khác, *thặng dư* $r = b - Ax$ *phải trực giao với* $range(A)$.

Định lý 2.5.1 *Cho $A \in \mathbb{C}^{m\times n}(m \geqslant n)$ và $b \in \mathbb{C}^m$ được cho. Một vector $x \in \mathbb{C}^m$ cực tiểu hóa chuẩn thặng dư $\|r\|_2 = \|b - Ax\|_2$, do đó việc giải bài toán bình phương nhỏ nhất (2.5.2) nếu và chỉ nếu $r \perp range(A)$, đó là,*

$$A^*r = 0, \tag{2.5.8}$$

tương đương

$$A^*Ax = A^*b, \tag{2.5.9}$$

hoặc tương đương

$$Pb = Ax, \tag{2.5.10}$$

với $P \in \mathbb{C}^{m\times m}$ là phép chiếu trực giao vào $range(A)$. Hệ thống $n \times n$ phương trình (2.5.9) được biết như là các phương trình chuẩn tắc là không suy biến nếu và chỉ nếu A có hạng đầy đủ. Do đó lời giải x là duy nhất nếu và chỉ nếu A có hạng đầy đủ.

Chứng minh. Sự tương đương của (2.5.8) và (2.5.10) theo sau từ các tính chất của các phép chiếu trực giao được thảo luận trong Mục 2.1, và sự tương đương của (2.5.8) và (2.5.9) theo sau từ định nghĩa của r. Để chứng minh $y = Pb$ là điểm duy nhất trong $range(A)$ mà nó cực tiểu hóa $\|b - y\|_2$, giả sử $z \neq y$ là một điểm khác trong $range(A)$. Vì $z - y$ trực giao với $b - y$, định lý Pythagorean cho $\|b - z\|_2^2 = \|b - y\|_2^2 + \|y - z\|_2^2 > \|b - y\|_2^2$ như được yêu cầu. Cuối cùng, ta chú ý rằng nếu A^*A là suy biến thì $A^*Ax = 0$ với x khác 0 bất kì mà điều này kéo theo $x^*A^*Ax = 0$. Do đó $Ax = 0$ mà nó kéo theo A có hạng không đầy đủ. Ngược lại, nếu A có hạng không đầy đủ thì $Ax = 0$ với x khác 0, cũng kéo theo $A^*Ax = 0$, nên A^*A là suy biến. Do (2.5.9), đặc trưng của các ma trận không suy biến A^*A này đưa ra phát biểu về tính duy nhất của x.

2.5.4 Giả nghịch đảo

Nếu A có hạng đầy đủ thì lời giải x cho bài toán bình phương nhỏ nhất (2.5.2) là duy nhất và được cho bởi $x = (A^*A)^{-1}A^*b$. Ma trận $(A^*A)^{-1}A^*$ được biết như là *giả nghịch đảo* của A, được ký hiệu là A^+:

$$A^+ = (A^*A)^{-1}A^* \in \mathbb{C}^{n,m}. \tag{2.5.11}$$

Ma trận này ánh xạ các vector $b \in \mathbb{C}^m$ thành các vector $x \in \mathbb{C}^n$.
Bài toán bình phương nhỏ nhất tuyến tính hạng đầy đủ (2.5.2) được tóm tắt như sau. Bài toán tính một hoặc cả hai vector

$$x = A^+b, \quad y = Pb, \tag{2.5.12}$$

với A^+ là giả nghịch đảo của A và P là phép chiếu trực giao vào $range(A)$.

2.5.5 Các phương trình chuẩn tắc

Cách cổ điển để giải các bài toán bình phương nhỏ nhất là để giải các phương trình chuẩn tắc (2.5.9). Nếu A có hạng đầy đủ thì đây là hệ thống xác định dương Hermit, vuông của các phương trình n chiều. Phương pháp tiêu chuẩn của việc giải một hệ thống này là *phân tích Cholesky*, được thảo luận ở mục sau. Phương pháp này xây dựng một phân tích $A^*A = R^*R$, với R là ma trận tam giác trên, giảm (2.5.9) thành các phương trình

$$R^*Rx = A^*b. \tag{2.5.13}$$

Dưới đây là thuật toán.

Thuật toán 2.6 Bình phương nhỏ nhất qua các phương trình chuẩn tắc

1: Thiết lặp ma trận A^*A và vector A^*b.
2: Tính phân tích Cholesky $A^*A = R^*R$.
3: Giải hệ thống tam giác dưới $R^*w = A^*b$ cho biến w.
4: Giải hệ thống tam giác trên $Rx = w$ cho biến x.

Các bước mà nó chi phối công việc cho tính toán này là 2 bước đầu tiên (cho bước 3 và 4). Bởi vì tính đối xứng, A^*A chỉ cần mn^2 phép toán dấu chấm động, phân nửa chi phí nếu A và A^* là các ma trận tùy ý có cùng số chiều. Phân tích Cholesky yêu cầu $n^3/3$ phép toán dấu chấm động. Kết hợp lại, việc giải bài toán bình phương nhỏ nhất bằng các phương trình chuẩn tắc bao gồm tổng số phép toán theo sau:

$$\text{Thuật toán 2.6: } \sim mn^2 + \frac{1}{3}n^3 \text{ phép toán dấu chấm động.} \tag{2.5.14}$$

2.5.6 Phân tích QR

Phương pháp "mô hình cổ điển" cho việc giải các bài toán bình phương nhỏ nhất phổ biến từ những năm 1960 được dựa vào phân tích QR được giảm. Theo trực giao hóa Gram - Schmidt hoặc thường xuyên hơn là tam giác hóa Householder, ta xây dựng phân tích $A = \widehat{Q}\widehat{R}$. Khi đó phép chiếu trực giao P có thể được viết $P = \widehat{Q}\widehat{Q}^*$ (2.1.6), nên ta có

$$y = Pb = \widehat{Q}\widehat{Q}^*b. \tag{2.5.15}$$

Vì $y \in range(A)$ nên hệ thống $Ax = y$ có một lời giải chính xác. Kết hợp phân tích QR và (2.5.15) cho

$$\widehat{Q}\widehat{R}x = \widehat{Q}\widehat{Q}^*b, \tag{2.5.16}$$

và nhân trái với $\widehat{Q}^*$

$$\widehat{R}x = \widehat{Q}^*b. \tag{2.5.17}$$

(Việc nhân với $\widehat{R}^{-1}$ cho công thức $A^+ = \widehat{R}^{-1}\widehat{Q}^*$ cho giả nghịch đảo.) Phương trình (2.5.17) là hệ thống tam giác trên, không suy biến nếu A có hạng đầy đủ, và nó được giải bằng phép thế ngược.

Thuật toán 2.7 Bình phương nhỏ nhất thông qua phân tích QR

1: Tính phân tích QR được giảm $A = \widehat{Q}\widehat{R}$.
2: Tính vector $\widehat{Q}^*b$.
3: Giải hệ thống tam giác trên $\widehat{R}x = \widehat{Q}^*b$ cho biến x.

Chú ý (2.5.17) cũng có thể được suy ra từ các phương trình chuẩn tắc. Nếu $A^*Ax = A^*b$, thì $\widehat{R}^*\widehat{Q}^*\widehat{Q}\widehat{R}x = \widehat{R}^*\widehat{Q}^*b$, kéo theo $\widehat{R}x = \widehat{Q}^*b$.

Thuật toán 2.7 bị ảnh hưởng bởi chi phí của phân tích QR. Nếu các phản xạ Householder được sử dụng ở bước này thì từ (2.4.10) ta có

$$\text{Thuật toán 2.7: } \sim 2mn^2 - \frac{2}{3}n^3 \text{ phép toán dấu chấm động.} \tag{2.5.18}$$

2.5.7 SVD

Một phương pháp khác cho việc giải bài toán bình phương nhỏ nhất là phân tích giá trị suy biến được giảm $A = \widehat{U}\widehat{\Sigma}V^*$. P được biểu diễn trong dạng $P = \widehat{U}\widehat{U}^*$, cho

$$y = Pb = \widehat{U}\widehat{U}^*b, \tag{2.5.19}$$

và tương tự (2.5.16) và (2.5.17) là

$$\widehat{U}\widehat{\Sigma}V^*x = \widehat{U}\widehat{U}^*b \tag{2.5.20}$$

và

$$\widehat{\Sigma}V^*x = \widehat{U}^*b. \tag{2.5.21}$$

(Việc nhân với $V\widehat{\Sigma}^{-1}$ cho $A^+ = V\widehat{\Sigma}^{-1}\widehat{U}^*$.)

Chú ý trong khi phân tích QR giảm bài toán bình phương nhỏ nhất thành một hệ

Thuật toán 2.8 Bình phương nhỏ nhất qua SVD

1: Tính SVD được sửa đổi $A = \widehat{U}\widehat{\Sigma}V^*$.
2: Tính vector $\widehat{U}^*b$.
3: Giải hệ thống đường chéo $\widehat{\Sigma}w = \widehat{U}^*b$ cho biến w.
4: Đặt $x = Vw$.

thống tam giác của các phương trình, SVD giảm nó thành một hệ thống đường chéo của các phương trình. Nếu A có hạng đầy đủ thì hệ thống đường chéo là không suy biến. Như trước đó, (2.5.21) có thể được suy ra từ các phương trình chuẩn tắc. Nếu $A^*Ax = A^*b$ thì $V\widehat{\Sigma}^*\widehat{U}^*\widehat{U}\widehat{\Sigma}V^*x = V\widehat{\Sigma}^*\widehat{U}^*b$, kéo theo $\widehat{\Sigma}V^*x = \widehat{U}^*b$.

Đếm số phép toán cho Thuật toán 2.8 bị ảnh hưởng bởi sự tính toán của SVD. Cho $m \gg n$ chi phí này được xấp xỉ giống như phân tích QR nhưng với $m \approx n$ SVD là tốn kém hơn nhiều. Một ước lượng tiêu chuẩn là

$$\text{Thuật toán 2.8: } \sim 2mn^2 + 11n^3 \text{ phép toán dấu chấm động.} \tag{2.5.22}$$

2.5.8 Ví dụ

Ví dụ 2.5.3. (Phân tích QR sử dụng phản xạ Householder) Cho ma trận

$$A = \begin{bmatrix} 0.8147 & 0.0975 & 0.1576 \\ 0.9058 & 0.2785 & 0.9706 \\ 0.1270 & 0.5469 & 0.9572 \\ 0.9134 & 0.9575 & 0.4854 \\ 0.6324 & 0.9649 & 0.8003 \end{bmatrix}$$

Ta làm việc với cột đầu tiên của A

$$x_1 = \begin{bmatrix} 0.8147 \\ 0.9058 \\ 0.1270 \\ 0.9134 \\ 0.6324 \end{bmatrix}, \quad \|x_1\|_2 = 1.6536.$$

Vector Householder tương ứng là

$$\tilde{v_1} = x_1 + \|x_1\|_2 e_1 = \begin{bmatrix} 0.8147 \\ 0.9058 \\ 0.1270 \\ 0.9134 \\ 0.6324 \end{bmatrix} + 1.6536 \begin{bmatrix} 1.0000 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2.4684 \\ 0.9058 \\ 0.1270 \\ 0.9134 \\ 0.6324 \end{bmatrix}$$

Từ vector này, ta xây dựng phản xạ Householder

$$c = \frac{2}{\tilde{v_1}^T \tilde{v_1}} = 0.2450, \quad \tilde{H_1} = I - c\tilde{v_1}\tilde{v_1}^T$$

Áp dụng phản xạ này cho A, ta được

$$\tilde{H_1} A^{(1)}_{1:5,1:3} = \begin{bmatrix} -1.6536 & -1.1405 & -1.2569 \\ 0 & -0.1758 & 0.4515 \\ 0 & 0.4832 & 0.8844 \\ 0 & 0.4994 & -0.0381 \\ 0 & 0.6477 & 0.4379 \end{bmatrix}, \quad A^{(2)} = \begin{bmatrix} -1.6536 & -1.1405 & -1.2569 \\ 0 & -0.1758 & 0.4515 \\ 0 & 0.4832 & 0.8844 \\ 0 & 0.4994 & -0.0381 \\ 0 & 0.6477 & 0.4379 \end{bmatrix}$$

Tiếp theo, ta lấy phần thú vị của cột 2 của ma trận được cập nhật $A^{(2)}$ từ dòng 2 tới 5

$$x_2 = a^{(2)}_{2:5,2} = \begin{bmatrix} -0.1758 \\ 0.4832 \\ 0.4994 \\ 0.6477 \end{bmatrix}, \quad \|x_2\|_2 = 0.9661.$$

Vector Householder tương ứng là

$$\tilde{v_2} = x_2 - \|x_2\|_2 e_1 = \begin{bmatrix} -0.1758 \\ 0.4832 \\ 0.4994 \\ 0.6477 \end{bmatrix} - 0.9661 \begin{bmatrix} 1.0000 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1.1419 \\ 0.4832 \\ 0.4994 \\ 0.6477 \end{bmatrix}$$

Từ vector này, ta xây dựng phản xạ Householder

$$c = \frac{2}{\tilde{v_2}^T \tilde{v_2}} = 0.9065, \quad \tilde{H_2} = I - c\tilde{v_2}\tilde{v_2}^T$$

Áp dụng phản xạ này cho $A^{(2)}$, ta được

$$\tilde{H_2} A^{(1)}_{2:5,2:3} = \begin{bmatrix} 0.9661 & 0.6341 \\ 0 & 0.8071 \\ 0 & -0.1179 \\ 0 & 0.3343 \end{bmatrix}, \quad A^{(3)} = \begin{bmatrix} -1.6536 & -1.1405 & -1.2569 \\ 0 & 0.9661 & 0.6341 \\ 0 & 0 & 0.8071 \\ 0 & 0 & -0.1179 \\ 0 & 0 & 0.3343 \end{bmatrix}$$

Tiếp theo, ta lấy phần thú vị của cột 3 của ma trận được cập nhật $A^{(3)}$ từ dòng 3 tới 5

$$x_3 = a^{(3)}_{3:5,3} = \begin{bmatrix} 0.8071 \\ -0.1179 \\ 0.3343 \end{bmatrix}, \quad \|x_3\|_2 = 0.8816.$$

Vector Householder tương ứng là

$$\tilde{v_3} = x_3 + \|x_3\|_3 e_1 = \begin{bmatrix} 0.8071 \\ -0.1179 \\ 0.3343 \end{bmatrix} + 0.8816 \begin{bmatrix} 1.0000 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.6887 \\ -0.1179 \\ 0.3343 \end{bmatrix}$$

Từ vector này, ta xây dựng phản xạ Householder

$$c = \frac{2}{\tilde{v_3}^T \tilde{v_3}} = 0.6717, \quad \tilde{H_3} = I - c\tilde{v_2}\tilde{v_2}^T$$

Áp dụng phản xạ này cho $A^{(3)}$, ta được

$$\tilde{H_3} A^{(1)}_{3:5,2:3} = \begin{bmatrix} -0.8816 \\ 0 \\ 0 \end{bmatrix}, \quad R = A^{(4)} = \begin{bmatrix} -1.6536 & -1.1405 & -1.2569 \\ 0 & 0.9661 & 0.6341 \\ 0 & 0 & -0.8816 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Áp dụng các Householder này vào bên phải của ma trận đơn vị, ta được ma trận trực giao

$$Q = H_1H_2H_3 = \begin{bmatrix} -0.4927 & -0.4806 & 0.1780 & -0.6015 & -0.3644 \\ -0.5478 & -0.3583 & -0.5777 & 0.3760 & 0.3104 \\ -0.0768 & 0.4754 & -0.6343 & -0.1497 & -0.5859 \\ -0.5523 & 0.3391 & 0.4808 & 0.5071 & -0.3026 \\ -0.3824 & 0.5473 & 0.0311 & -0.4661 & 0.5796 \end{bmatrix}$$

sao cho

$$R = Q^TA = H_3H_2H_1A = \begin{bmatrix} -1.6536 & -1.1405 & -1.2569 \\ 0 & 0.9661 & 0.6341 \\ 0 & 0 & -0.8816 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

là tam giác trên với

$$H_1 = \tilde{H}_1, H_2 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{H}_2 \end{bmatrix}, H_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \tilde{H}_3 \end{bmatrix}.$$

Chú ý rằng, với $j = 1, 2, 3$

$$H_j = I - 2v_jv_j^T, v_j = \begin{bmatrix} 0 \\ \tilde{v_j} \end{bmatrix}, \|v_j\|_2 = \|\tilde{v_j}\|_2 = 1$$

với $j-1$ thành phần đầu tiên của v_j là bằng 0.

Ví dụ 2.5.4. (Phân tích QR sử dụng trực giao hóa Gram-Schmidt) Tìm phân tích QR cho ma trận sau

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Đầu tiên, $v_1 = a_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ và $r_{11} = \|v_1\| = \sqrt{2}$ ta được

$$q_1 = \frac{v_1}{\|v_1\|} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Tiếp theo,

$$v_2 = a_2 - \underbrace{(q_1^* a_2)}_{r_{12}} q_1$$
$$= \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} - \frac{\sqrt{2}}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

Tính toán này cần $r_{12} = q_1^* a_2 = \dfrac{2}{\sqrt{2}} = \sqrt{2}$. Hơn nữa, $r_{22} = \|v_2\| = \sqrt{3}$ và

$$q_2 = \frac{v_2}{\|v_2\|} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

Trong bước lặp thứ 3, ta có

$$v_3 = a_3 - \underbrace{(q_1^* a_3)}_{r_{13}} q_1 - \underbrace{(q_2^* a_3)}_{r_{23}} q_2$$
$$= \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} - 0 = \frac{1}{2} \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$$

với $r_{13} = \dfrac{1}{\sqrt{2}}$ và $r_{23} = 0$.

Cuối cùng, $r_{33} = \|v_3\| = \dfrac{\sqrt{6}}{2}$ và

$$q_3 = \frac{v_3}{\|v_3\|} = \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}.$$

Khi đó, ta được

$$Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \end{bmatrix} \text{ và } R = \begin{bmatrix} \sqrt{2} & \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & \frac{\sqrt{6}}{2} \end{bmatrix}.$$

Bài tập

1. Nếu P là một phép chiếu trực giao thì $I - 2P$ là unita. Chứng minh điều này theo phương diện đại số và cho 1 sự giải thích hình học.

2. Cho E là ma trận $m \times m$ mà nó trích ra (extract) "phần chẵn" của m vector: $Ex = (x + Fx)/2$, với F là ma trận $m \times m$ mà nó lật (flip) $(x_1, \ldots, x_m)^*$ thành $(x_m, \ldots, x_1)^*$. E là một phép chiếu trực giao, phép chiếu xiên (oblique), hoặc hoàn toàn không phải là 1 phép chiếu?Các phần tử của nó là gì?

3. Cho $A \in \mathbb{R}^{m\times n}$ với $m > n$. Chứng minh rằng nếu AA^* không suy biến nếu và chỉ nếu A có hạng đầy đủ.

4. Giả sử $P \in \mathbb{R}^{m\times m}$ thỏa $\|P^TP - I_m\|_2 = \epsilon < 1$. Chứng minh rằng tất cả các giá trị suy biến của P nằm trong khoảng $[1 - \epsilon, 1 + \epsilon]$ và $\|P - UV^T\|_2 \leqslant \epsilon$, với $P = U\sum V^T$ là SVD của P.

5. Cho ma trận
$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$
   a) Xác định phép chiếu trưc giao P lên $range(A)$ và ảnh dưới P của vector $(1, 2, 3)^*$?
   b) Làm tương tự cho B.

6. Cho $P \in \mathbb{C}^{m\times m}$ là một phép chiếu khác 0. Chứng minh rằng $\|P\|_2 \geqslant 1$, với dấu "=" xảy ra khi và chỉ khi P là 1 phép chiếu trực giao.

7. Cho A là ma trận có các cột $1, 3, 5, 7, \ldots$ trực giao với các cột $2, 4, 6, 8, \ldots$. Trong phân tích QR được sửa đổi $A = \widehat{Q}\widehat{R}$, $\widehat{R}$ có cấu trúc đặc biệt gì? Giả sử A có hạng đầy đủ.

8. Cho A là ma trận $m \times n$ $(m \geqslant n)$ và $A = \widehat{Q}\widehat{R}$ là một phân tích QR được sửa đổi
   a) Chứng minh rằng A có hạng n nếu và chỉ nếu tất cả các phần tử nằm trên đường chéo của $\widehat{R}$ khác 0.
   b) Giả sử $\widehat{R}$ có k phần tử nằm trên đường chéo khác 0 với $0 \leqslant k < n$. Điều này suy ra hạng của A là gì? một cách chính xác k? ít nhất k? tối đa k? Cho câu trả lời chính xác và chứng minh nó.

9. Cho A là ma trận $m \times n$. Xác định chính xác số phép cộng, phép trừ, phép nhân và phép chia dấu chấm động được bao gồm trong việc tính phân tích $A = \widehat{Q}\widehat{R}$ bởi Thuật toán 2.2.

10. Xét các ma trận trực giao 2×2
$$F = \begin{bmatrix} -c & s \\ s & c \end{bmatrix}, \quad J = \begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

với $s = \sin\theta$ và $c = \cos\theta$, θ bất kì. Ma trận đầu tiên có $detF = -1$ và là 1 phép đối xứng - trường hợp đặc biệt của phép đối xứng Householder trong chiều 2. Ma trận thứ 2 có $detJ = 1$ và tác động một phéo quay thay vì phép đối xứng. Một ma trận như vậy được gọi là *phép quay Givens*.

a) Miêu tả một cách chính xác hình học tác động của các phép nhân trái bởi F và J có trong mặt phẳng $\mathbb{R}^2$. (J quay mặt phẳng bằng 1 góc θ theo chiều kim đồng hồ hay ngược chiều kim đồng hồ?)

b) Miêu tả 1 thuật toán cho phân tích QR mà nó tương tự với Thuật toán 2.3 nhưng dựa vào các phép quay Givens thay cho các phép đối xứng Householder.

c) Chứng minh rằng thuật toán ở câu b) bao gồm 6 phép toán dấu chấm động trên phần tử được tính toán hơn là 4 phép toán, sao cho đếm số phép toán tiệm cận là 50% lớn hơn (2.4.10).

11. Cài đặt thuật toán trực giao hóa Gram – Schmidt của họ $\{u_1, u_2, \ldots, u_n\}$ độc lập tuyến tính.

12. Cho n vector $a_1, a_2, \ldots, a_n$ trong không gian vector $\mathbb{R}^m$.

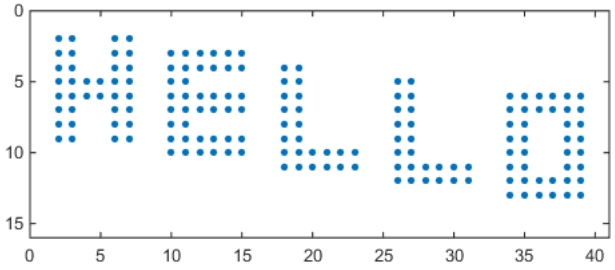
a) Hãy viết thuật toán Gram-schmidt cổ điển với n vector $a_1, a_2, \ldots, a_n$.

b) Viết chương trình đếm số phép gán và số phép so sánh của thuật toán vừa viết theo m và n.

c) Ước lượng độ phức tạp thuật toán theo tổng số phép toán gán và so sánh.

13. Cho A là ma trận $m \times n$ ($m \geqslant n$) và cho $A = \widehat{Q}\widehat{R}$ là SVD được sửa đổi của A. Chứng minh rằng A có hạng đầy đủ nếu và chỉ nếu các phần tử trên đường chéo của $\widehat{R}$ đều khác 0.

14. a) Viết một chương trình Matlab cài đặt một ma trận 15×40 với các phần tử 0 khắp nơi ngoại trừ 1 ở các vị trí được cho biết trong hình bên dưới. Số 1 ở vị trí nhất ở trên là nằm ở vị trí (2,2), và số 1 ở vị trí phải nhất ở phía dưới là nằm ở vị trí (13,39). Hình này được đưa ra với lệnh $spy(A)$.



b) Gọi SVD để tính các giá trị suy biến của A, và in các kết quả. Vẽ các số này sử dụng cả *plot* và *semilogy*. Hạng chính xác của A? Điều này cho thấy các giá trị suy biến được tính toán như thế nào?

c) Với mỗi i từ 1 tới $rank(A)$, xây dựng các ma trận hạng i B mà nó là xấp xỉ tốt nhất cho A trong chuẩn 2. Sử dụng các lệnh $pcolor(B)$ với $colormap(gray)$ để khởi tạo các ảnh của các xấp xỉ khác nhau này.

15. Cho x và y là các vector khác không của $\mathbb{R}^m$. Cho một thuật toán xác định ma trận Householder P sao cho Px là bội của y.

16. a) Viết một hàm $[W, R] = house(A)$ trong Matlab tính biểu diễn ẩn của phân tích QR đầy đủ $A = QR$ của một ma trận A $m \times n$ với $m \geqslant n$ sử dụng các phản xạ Householder. Các biến đầu ra là một ma trận tam giác dưới $W \in \mathbb{C}^{m\times n}$ mà các cột của nó là các vector v_k xác định các phản xạ Householder liên tiếp, và một ma trận tam giác $R \in \mathbb{C}^{m\times n}$.

b) Viết một hàm $Q = formQ(W)$ mà nó lấy ma trận W đưa ra bởi *house* như đầu vào và sinh ra một ma trận Q trực giao $m \times m$ tương ứng.

Chương 3

Điều kiện và tính ổn định

3.1 Điều kiện và các số điều kiện

3.1.1 Điều kiện của một bài toán

Về mặt lý thuyết, ta có thể xem một *bài toán* như là một hàm $f : X \to Y$ từ một không gian vector định chuẩn X của dữ liệu vào một không gian vector định chuẩn Y của các lời giải. Hàm f này thường là không tuyến tính (ngay trong đại số tuyến tính) nhưng ít nhất nó cũng là hàm liên tục.
Một bài toán *điều kiện tốt* là một bài toán với tính chất mà tất cả các nhiễu nhỏ của x chỉ dẫn đến các thay đổi nhỏ trong $f(x)$. Một bài toán *điều kiện xấu* là một bài toán với tính chất mà một nhiễu nhỏ bất kì của x dẫn tới một thay đổi lớn trong $f(x)$.

3.1.2 Số điều kiện tuyệt đối

Cho δx là một nhiễu nhỏ của x, và viết $\delta f = f(x + \delta x) - f(x)$. *Số điều kiện tuyệt đối* $\widehat{\kappa} = \widehat{\kappa}(x)$ của bài toán f tại x được xác định như sau

$$\widehat{\kappa} = \lim_{\delta \to 0} \sup_{\|\delta x\| \leqslant \delta} \frac{\|\delta f\|}{\|\delta x\|}. \tag{3.1.1}$$

Cho hầu hết các bài toán, giới hạn của cân trên đúng trong công thức này có thể được làm sáng tỏ như một cân trên đúng trên tất cả các nhiễu nhỏ vô cùng δx. Thông thường ta sẽ viết công thức một cách đơn giản như sau

$$\widehat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}, \tag{3.1.2}$$

với δx và δf là nhỏ vô cùng.
Nếu f là khả vi thì ta có thể ước lượng số điều kiện bằng các trung bình đạo hàm của f. Cho $J(x)$ là một ma trận mà phần tử i, j của nó là đạo hàm riêng phần $\partial f_i / \partial x_j$ được ước lượng tại x, được biết như là *Jacobian* của f tại x. Định nghĩa của đạo hàm cho $\delta f \approx J(x)\delta x$ với đẳng thức trong giới hạn $\|\delta x\| \to 0$. Số điều kiện tuyệt đối trở thành

$$\widehat{\kappa} = \|J(x)\|, \tag{3.1.3}$$

với $\|J(x)\|$ là chuẩn của $J(x)$ được bao gồm bởi các chuẩn trong X và Y.

3.1.3 Số điều kiện tương đối

Số điều kiện tương đối $\kappa = \kappa(x)$ được xác định bởi

$$\kappa = \lim_{\delta \to 0} \sup_{\|\delta x\| \leqslant \delta} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right), \tag{3.1.4}$$

hoặc giả sử δx và δf là nhỏ vô cùng

$$\kappa = \sup_{\delta x} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right). \tag{3.1.5}$$

Nếu f là khả vi thì ta có thể biểu diễn con số này trong các số hạng của Jacobian:

$$\kappa = \frac{\|J(x)\|}{\|f(x)\|/\|x\|}. \tag{3.1.6}$$

Một bài toán *điều kiện tốt* nếu κ là nhỏ (ví dụ, $1, 10, 10^2$), và *điều kiện xấu* nếu κ là lớn (ví dụ, $10^6, 10^{16}$).

3.1.4 Ví dụ

Ví dụ 3.1.1. Xét bài toán $f : x \mapsto x/2$, với $x \in \mathbb{C}$. Jacobian của hàm f phải là đạo hàm $J = f' = 1/2$, nên theo (3.1.6),

$$\kappa = \frac{\|J\|}{\|f(x)\|/\|x\|} = \frac{1/2}{(x/2)/x} = 1.$$

Bài toán này là có điều kiện tốt với chuẩn bất kỳ.

Ví dụ 3.1.2. Xét bài toán $f : x \mapsto \sqrt{x}$ với $x > 0$. Jacobian của f là đạo hàm $J = f' = 1/(2\sqrt{x})$, nên ta có

$$\kappa = \frac{\|J\|}{\|f(x)\|/\|x\|} = \frac{1/(2\sqrt{x})}{\sqrt{x}/x} = \frac{1}{2}.$$

Đây cũng là bài toán có điều kiện tốt.

Ví dụ 3.1.3. Xét bài toán $f(x) = x_1 - x_2$ từ vector $x = (x_1, x_2)^* \in \mathbb{C}^2$. Cho đơn giản, ta sử dụng chuẩn ∞ trong không gian dữ liệu $\mathbb{C}^2$. Jacobian của f là

$$J = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \dfrac{\partial f}{\partial x_2} \end{bmatrix} = [1 \ -1],$$

với $\|J\|_\infty = 2$. Khi đó, số điều kiện là

$$\kappa = \frac{\|J\|_\infty}{\|f(x)\|/\|x\|} = \frac{2}{|x_1 - x_2|/\max\{|x_1|, |x_2|\}}.$$

Con số này là lớn nếu $|x_1 - x_2| \approx 0$ nên bài toán này là điều kiện xấu khi $x_1 \approx x_2$.

Ví dụ 3.1.4. Xác định các nghiệm của một đa thức với các hệ số được cho trước là một ví dụ cổ điển của bài toán điều kiện xấu. Xét $x^2 - 2x + 1 = (x-1)^2$, với nghiệm bội

$x = 1$. Một nhiễu nhỏ trong các hệ số có thể dẫn đến một thay đổi lớn hơn trong các nghiệm. Ví dụ, $x^2 - 2x + 0.9999 = (x - 0.99)(x - 1.01)$. Thât vậy, các nghiệm có thể thay đổi tương ứng với căn bậc hai của thay đổi trong các hệ số nên trong trường hợp này Jacobian là vô hạn (bài toán là không khả vi), và $\kappa = \infty$.

Việc tìm nghiệm đa thức là đặc thù của bài toán điều kiện xấu ngay trong các trường hợp mà chúng không bao gồm các nghiệm bội. Nếu hệ số a_i của đa thức $p(x)$ được làm nhiễu bằng một con số nhỏ vô cùng δa_i thì nhiễu của nghiệm thứ j (x_j) là $\delta x_j = -\dfrac{(\delta a_i)x_j^i}{p'(x_j)}$, với p' là đạo hàm của p. Do đó, số điều kiện của x_j tương ứng với các nhiễu của hệ số đơn a_i là

$$\kappa = \frac{|\delta x_j|}{|x_j|} \Big/ \frac{|\delta a_i|}{|a_i|} = \frac{|a_i x_j^{i-1}|}{|p'(x_j)|}. \tag{3.1.7}$$

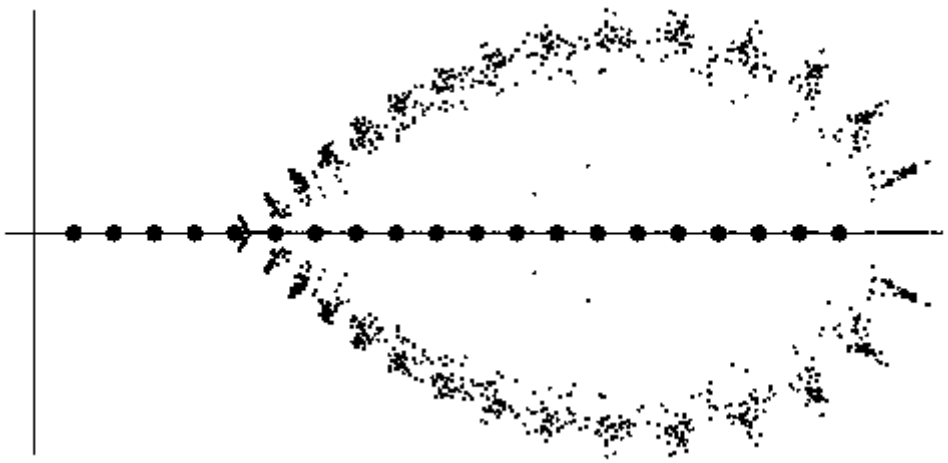
Số này thường là rất lớn. Xét "đa thức Wilkinson"

$$p(x) = \prod_{i=1}^{20}(x - i) = a_0 + a_1 x + \ldots + a_{19}x^{19} + x^{20}. \tag{3.1.8}$$

Nghiệm dễ bị ảnh hưởng nhất của đa thức này là $x = 15$ và dễ bị ảnh hưởng nhất để thay đổi hệ số $a_{15} \approx 1.67 \times 10^9$. Số điều kiện là

$$\kappa \approx \frac{1.67 \times 10^9 \cdot 15^{14}}{5!14!} \approx 5.1 \times 10^{13}.$$

Hình 3.1 miêu tả bài toán điều kiện xấu.



Hình 3.1: Ví dụ điều kiện xấu kinh điển của Wilkinson. Các dấu chấm lớn là các nghiệm của đa thức (3.1.8) chưa bị nhiễu. Các dấu chấm nhỏ là các nghiệm được thêm vào trong mặt phẳng phức của 100 đa thức bị nhiễu ngẫu nhiên với các hệ số được xác định bởi $\tilde{a}_k = a_k(1 + 10^{-10}r_k)$, với r_k là một số từ phân phối chuẩn của trung bình 0 và phương sai 1

Ví dụ 3.1.5. Bài toán tính trị riêng của một ma trận không đối xứng cũng thường là

bài toán điều kiện xấu. So sánh hai ma trận

$$\begin{bmatrix} 1 & 1000 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1000 \\ 0.001 & 1 \end{bmatrix},$$

mà các trị riêng của chúng tương ứng là $\{1, 1\}$ và $\{0, 2\}$. Mặt khác, nếu một ma trận A là đối xứng (tổng quát hơn, nếu nó là chuẩn tắc) thì các trị riêng của nó là điều kiện tốt. Nếu λ và $\lambda + \delta\lambda$ tương ứng với các trị riêng của A và $A + \delta A$ thì $|\delta\lambda| \leqslant \|\delta A\|_2$, dấu bằng xảy ra nếu δA là một bội của ma trận đơn vị. Do đó, số điều kiện tuyệt đối của bài toán trị riêng đối xứng là $\widehat{\kappa} = 1$, nếu các nhiễu được đo trong chuẩn 2 thì số điều kiện tương đối là $\kappa = \dfrac{\|A\|_2}{|\lambda|}$.

3.1.5 Điều kiện của phép nhân ma trận với vector

Cố định $A \in \mathbb{C}^{m\times n}$ và xét bài toán tính Ax từ số liệu đầu vào x, nghĩa là ta sẽ xác định số điều kiện tương ứng với các nhiễu của x mà không phải là của A. Từ định nghĩa của κ, với $\|\cdot\|$ là chuẩn vector tùy ý và chuẩn ma trận được bao gồm tương ứng, ta thấy

$$\kappa = \sup_{\delta x} \left(\frac{\|A(x+\delta x) - Ax\|}{\|Ax\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right) = \sup_{\delta x} \frac{\|A\delta x\|}{\|\delta x\|} \Big/ \frac{\|Ax\|}{\|x\|}$$

Do đó,

$$\kappa = \|A\| \frac{\|x\|}{\|Ax\|} \tag{3.1.9}$$

(trường hợp đặc biệt của (3.1.6)). Đây là một công thức chính xác cho κ, phụ thuộc vào cả A và x.

Giả sử trong tính toán ở trên A là ma trận vuông và không suy biến. Khi đó, ta có thể sử dụng $\dfrac{\|x\|}{\|Ax\|} \leqslant \|A^{-1}\|$ để nới lõng (3.1.9) thành một chặn độc lập với x:

$$\kappa \leqslant \|A\|\|A^{-1}\|. \tag{3.1.10}$$

Hoặc

$$\kappa = \alpha\|A\|\|A^{-1}\| \tag{3.1.11}$$

với

$$\alpha = \frac{\|x\|}{\|Ax\|} \Big/ \|A^{-1}\|. \tag{3.1.12}$$

Cho các sự lựa chọn nào đó của x, ta có $\alpha = 1$, và do đó $\kappa = \|A\|\|A^{-1}\|$. Nếu $\|\cdot\| = \|\cdot\|_2$ thì điều này sẽ xuất hiện bất kỳ lúc nào x là một bội của vector suy biến phải cực tiểu của A.

Thật vậy, A không cần phải là ma trận vuông. Nếu $A \in \mathbb{C}^{m\times n}$ với $m \geqslant n$ có hạng đầy đủ thì các phương trình (3.1.10) - (3.1.12) đúng với A^{-1} được thay thế bằng giả nghịch đảo A^{+} được xác định trong (2.5.11).

Định lý 3.1.1 *Cho $A \in \mathbb{C}^{m\times m}$ là không suy biến và xét phương trình $Ax = b$. Bài toán tính b với x được cho trước, có số điều kiện*

$$\kappa = \|A\|\frac{\|x\|}{\|b\|} \leqslant \|A\|\|A^{-1}\| \tag{3.1.13}$$

tương ứng với các nhiễu của x. Bài toán tính x với b được cho trước, có số điều kiện

$$\kappa = \|A^{-1}\|\frac{\|b\|}{\|x\|} \leqslant \|A\|\|A^{-1}\| \tag{3.1.14}$$

tương ứng với các nhiễu của b. Nếu $\|\cdot\| = \|\cdot\|_2$ thì dấu bằng trong (3.1.13) xảy ra nếu x là một bội của vector suy biến phải của A tương ứng với giá trị suy biến nhỏ nhất σ_m, và dấu bằng trong (3.1.14) xảy ra nếu b là bội của một vector suy biến trái của A tương ứng với giá trị suy biến lớn nhất σ_1.

3.1.6 Số điều kiện của một ma trận

Tích $\|A\|\|A^{-1}\|$ là *số điều kiện* của A (liên quan tới chuẩn $\|\cdot\|$), được ký hiệu bởi $\kappa(A)$:

$$\kappa(A) = \|A\|\|A^{-1}\|. \tag{3.1.15}$$

Nếu $\kappa(A)$ nhỏ thì A được gọi là *điều kiện tốt*. Nếu $\kappa(A)$ lớn thì A là *điều kiện xấu*. Nếu A suy biến thì $\kappa(A) = \infty$.
Nếu $\|\cdot\| = \|\cdot\|_2$ thì $\|A\| = \sigma_1$ và $\|A^{-1}\| = \dfrac{1}{\sigma_m}$. Do đó,

$$\kappa(A) = \frac{\sigma_1}{\sigma_m} \tag{3.1.16}$$

trong chuẩn 2, và công thức này được sử dụng cho việc tính số điều kiện chuẩn 2 của các ma trận. Tỉ số $\dfrac{\sigma_1}{\sigma_m}$ được giải thích như là độ lệch tâm của siêu ellip mà nó là ảnh của quả cầu đơn vị của $\mathbb{C}^m$ dưới A (Hình 1.2).
Cho một ma trận hình chữ nhật $A \in \mathbb{C}^{m\times n}$ có hạng đầy đủ ($m \geqslant n$), số điều kiện được xác định: $\kappa(A) = \|A\|\|A^{+}\|$. Vì A^{+} được thúc đẩy bởi các bài toán bình phương nhỏ nhất nên định nghĩa này là hữu ích trong trong trường hợp $\|\cdot\| = \|\cdot\|_2$. Ta có

$$\kappa(A) = \frac{\sigma_1}{\sigma_n} \tag{3.1.17}$$

3.1.7 Điều kiện của một hệ thống các phương trình

Trong Định lý 3.1.1, ta cho A được cố định và x hoặc b được làm nhiễu. Điều gì sẽ xảy ra nếu ta làm nhiễu A? Đặc biệt, ta cho b cố định và xét bài toán $A \mapsto x = A^{-1}b$ khi A được làm nhiễu bằng δA nhỏ vô cùng. Khi đó, x phải thay đổi bằng δx nhỏ vô cùng, với

$$(A + \delta A)(x + \delta x) = b.$$

Sử dụng phương trình $Ax = b$ và việc giảm số hạng nhỏ vô cùng $(\delta A)(\delta x)$ xuống, ta được $(\delta A)x + A(\delta x) = 0$. Do đó, $\delta x = -A^{-1}(\delta A)x$. Phương trình này kéo theo $\|\delta x\| \leqslant \|A^{-1}\|\|\delta A\|\|x\|$, hoặc tương đương,

$$\frac{\|\delta x\|}{\|x\|} \Big/ \frac{\|\delta A\|}{\|A\|} \leqslant \|A\|\|A^{-1}\| = \kappa(A).$$

Đẳng thức xảy ra khi δA thỏa

$$\|A^{-1}(\delta A)x\| = \|A^{-1}\|\|\delta A\|\|x\|,$$

và nó có thể được chứng minh bằng sử dụng các chuẩn đối ngẫu mà với A, b, và chuẩn $\|\cdot\|$ bất kỳ, các nhiễu δA như vậy tồn tại.

Định lý 3.1.2 *Cho b cố định và xét bài toán $x = A^{-1}b$, với A là ma trận vuông và không suy biến. Số điều kiện của bài toán này với các nhiễu tương ứng trong A là*

$$\kappa = \|A\|\|A^{-1}\| = \kappa(A). \tag{3.1.18}$$

Các Định lý 3.1.1 và 3.1.2 là quan trọng cơ bản trong phương pháp số trong đại số tuyến tính vì chúng xác định một cách chính xác các hệ thống của các phương trình được giải như thế nào. Nếu một bài toán $Ax = b$ chứa một ma trận điều kiện xấu A thì thường ta phải hy vọng "mất $log_{10}\kappa(A)$ chữ số" trong việc tính toán tìm lời giải, ngoại trừ dưới các trường hợp rất đặc biệt.

3.2 Số học dấu chấm động

3.2.1 Hạn chế của biểu diễn bằng số

Vì các máy tính bằng số sử dụng một số hữu hạn các số nhị phân để biểu diễn một số thực nên chúng chỉ có thể biểu diễn một tập con hữu hạn của các số thực (hoặc số phức). Hạn chế này đưa ra hai khó khăn. Đầu tiên, các số được biểu diễn không thể lớn hoặc nhỏ tùy ý. Thứ hai, phải có các khoảng trống giữa chúng.

Các máy tính hiện đại biểu diễn các số đủ lớn và nhỏ mà hạn chế đầu tiên hiếm khi đưa ra các khó khăn. Ví dụ, số học chính xác gấp đôi IEEE được sử dụng một cách thưa thớt cho phép các số lớn như 1.79×10^{308} và nhỏ như 2.23×10^{-308}. Mặt khác, *tràn số* và *tràn dưới* thông thường không là nguy cơ quan trọng.

Bằng phản chứng, bài toán các khoảng trống giữa các số được biểu diễn một sự liên quan suốt việc tính toán khoa học. Ví dụ, trong số học độ chính xác bội IEEE, khoảng [1,2] được biểu diễn bằng một tập con rời rạc

$$1, 1 + 2^{-52}, 1 + 2 \times 2^{-52}, 1 + 3 \times 2^{-52}, \ldots, 2. \tag{3.2.1}$$

Khoảng [2,4] được biểu diễn bằng các số bội của 2 ,

$$2, 2 + 2^{-51}, 2 + 2 \times 2^{-51}, 2 + 3 \times 2^{-51}, \ldots, 4.$$

Tổng quát, khoảng $[2^j, 2^{j+1}]$ được biểu diễn bằng (3.2.1) nhân với 2^j. Do đó trong số học độ chính xác bội IEEE, các khoảng trống giữa các số gần kề không bao giờ lớn hơn $2^{-52} \approx 2.22 \times 10^{-16}$.

3.2.2 Các số dấu chấm động

Số học IEEE là một ví dụ của một hệ thống số học được dựa vào sự biểu diễn *dấu chấm động* của các số thực. Trong một hệ thống số dấu chấm động, vị trí của dấu chấm thập phân (hoặc nhị phân) được lưu trữ một cách tách biệt từ các chữ số và các khoảng trống giữa các số gần kề biểu diễn thang các số trong kích thước của các số. Điều này phân biệt với biểu diễn *dấu cố định* nơi mà các khoảng trống có cùng kích thước.

Đặc biệt, ta hãy xét một hệ thống số dấu chấm động được lý tưởng hóa như sau. Hệ thống dấu chấm động của tập con rời rạc $\mathbf{F}$ của các số thực $\mathbb{R}$ xác định bởi một số nguyên $\beta \geqslant 2$ được biết như là *cơ số* hay *radix*(tiêu biểu là 2) và một số nguyên $t \geqslant 1$ được biết như là *độ chính xác* (24 hoặc 53 cho độ chính xác đơn và bội IEEE tương ứng). Các phần tử của $\mathbf{F}$ là số 0 cùng với tất cả các số có dạng

$$x = \pm(m/\beta^t)\beta^e, \tag{3.2.2}$$

với m là số nguyên nằm trong $1 \leqslant m \leqslant \beta^t$ và e là một số nguyên tùy ý. Ta có thể hạn chế thành $\beta^{t-1} \leqslant m \leqslant \beta^t - 1$ và do đó việc chọn m là duy nhất. Khi đó, con số $\pm(m/\beta^t)$ được biết như là *phân số* hoặc *phần định trị* của x, và e là *số mũ*. Hệ thống dấu chấm động của chúng ta được lý tưởng hóa để lờ đi tràn số và tràn dưới. Như một kết quả, $\mathbf{F}$ là một tập vô hạn đếm được và $F = \beta F$.

3.2.3 Machine Epsilon

Lời giải của $\mathbf{F}$ được tóm tắt bởi một số được biết như *machine epsilon*. Tạm thời, ta hãy xác định số này bằng

$$\epsilon_{machine} = \frac{1}{2}\beta^{1-t}. \tag{3.2.3}$$

Số này là phân nửa khoảng cách giữa 1 và số dấu chấm động lớn hơn tiếp theo. $\epsilon_{machine}$ có tính chất theo sau:

$$\text{Với mọi } x \in \mathbb{R}, \text{ tồn tại } x' \in \boldsymbol{F} \text{ sao cho } |x - x'| \leqslant \epsilon_{machine}|x|. \tag{3.2.4}$$

Cho các giá trị của β và t thông thường trong các máy tính khác nhau, $\epsilon_{machine}$ thường nằm giữa 10^{-6} và 10^{-35}. Trong số học độ chính xác đơn và bội IEEE, $\epsilon_{machine}$ được thiết lặp tương ứng là $2^{-24} \approx 5.96 \times 10^{-8}$ và $2^{-53} \approx 1.11 \times 10^{-16}$.

Cho $fl : \mathbb{R} \to \boldsymbol{F}$ là hàm cho xấp xỉ dấu chấm động gần với một số thực nhất, tương đương *được làm tròn* của nó trong hệ thống dấu chấm động. Bất đẳng thức (3.2.4) có

thể được phát biểu liên quan tới fl

$$\text{Với mọi } x \in \mathbb{R}, \text{ tồn tại } \epsilon \text{ với } |\epsilon| \leqslant \epsilon_{machine} \text{ sao cho } fl(x) = x(1+\epsilon). \tag{3.2.5}$$

Sự khác nhau giữa một số thực và xấp xỉ dấu chấm động gần nó nhất thường là lớn hơn $\epsilon_{machine}$ trong các số hạng liên quan.

3.2.4 Số học dấu chấm động

Trong một máy tính, tất cả các phép toán toán học được giảm xuống thành các phép toán số học chủ yếu như $+, -, \times$ và $\div$. Theo toán học, các ký hiệu này biểu diễn các phép toán trong $\mathbb{R}$. Trong một máy tính, chúng là các phép toán trong $\mathbf{F}$ và được ký hiệu bởi $\oplus, \ominus, \otimes$ và $\oslash$.

Cho x và y là các số dấu chấm động tùy ý, nghĩa là $x, y \in \boldsymbol{F}$. Cho $*$ là một trong số những phép toán $+, -, \times$, hoặc $\div$, và cho $\circledast$ là một mô hình dấu chấm động của nó. Khi đó, $x \circledast y$ phải được cho một cách chính xác bởi

$$x \circledast y = fl(x * y). \tag{3.2.6}$$

Nếu tính chất này đúng thì từ (3.2.5) và (3.2.6), máy tính có tính chất đơn giản và mạnh như sau

Tiên đề cơ sở của số học dấu chấm động

Với mọi $x, y \in \boldsymbol{F}$, tồn tại ϵ với $|\epsilon| \leqslant \epsilon_{machine}$ sao cho

$$x \circledast y = (x * y)(1+\epsilon). \tag{3.2.7}$$

Mặt khác, *mọi phép toán số học dấu chấm động chính xác lên tới một sai số tương đối của kích thước nhiều nhất* $\epsilon_{machine}$.

3.2.5 Số học dấu chấm động phức

Các số phức dấu chấm động được biểu diễn như là các cặp của các số thực dấu chấm động, và các phép toán cơ bản trên chúng được tính bằng sự giảm bớt thành các phần thực và phần ảo. Tiên đề (3.2.7) là hợp lý cho số phức như là các số học dấu chấm động thực, ngoại trừ cho $\otimes$ và $\oslash$, $\epsilon_{machine}$ phải được tăng lên từ (3.2.3) bằng các thừa số trong bậc $2^{3/2}$ và $2^{5/2}$ tương ứng. Một khi $\epsilon_{machine}$ được điều chỉnh trong kiểu này thì phân tích sai số làm tròn cho các số phức có thể tiến hành như cho các số thực.

3.3 Tính ổn định

3.3.1 Các thuật toán

Trong mục 3.1, ta xác định *bài toán* toán học như là một hàm $f : X \to Y$ từ không gian vector X của dữ liệu vào một không gian vector Y của các lời giải.

Một *thuật toán* có thể được xem như một ánh xạ khác $\tilde{f} : X \to Y$ giữa hai không gian giống nhau. Cho một bài toán f, hệ thống dấu chấm động tính toán của nó thỏa mãn (3.2.7) (nhưng không nhất thiết thỏa (3.2.6)), một thuật toán cho f và một sự thực thi thuật toán này trong dạng của một chương trình máy tính được cố định. Dữ liệu $x \in X$ được cho, dữ liệu này được làm tròn tới dấu chấm động thỏa mãn (3.2.5) và khi đó được áp dụng như đầu vào của chương trình máy tính. Chạy thuật toán và kết quả là sự tập hợp các số dấu chấm động mà chúng thuộc không gian vector Y (vì thuật toán được thiết kế để giải f). Kết quả tính được này được gọi là $\tilde{f}(x)$.
Như là một mức tối thiểu, $\tilde{f}(x)$ sẽ bị ảnh hưởng bởi các sai số làm tròn. Việc phụ thuộc vào các trường hợp, nó cũng có thể bị ảnh hưởng bởi tất cả các loại phức tạp khác như sự hội tụ cho phép hoặc ngay cả các công việc khác chạy trong máy tính, trong các trường hợp mà phép gán của các tính toán tới các bộ xử lý không được xác định cho tới khi chạy. Do đó, "hàm" $\tilde{f}(x)$ cũng có thể lấy các giá trị khác nhau từ 1 lần chạy tiếp theo nên nó có thể là đa trị. (Thật vậy, bài toán f nên được cho phép là đa trị; điều này cho phép xử lý bằng tay các trường hợp mà một lời giải không duy nhất là có thể chấp nhận được, ví dụ, hai căn bậc hai của một số phức.).
Ký hiệu dấu (˜) là rất thích hợp. $\tilde{f}$ được tính tương tự như f. Ví dụ, lời giải của một hệ thống các phương trình $Ax = b$ có thể được ký hiệu bởi $\tilde{x}$.

3.3.2 Sự đúng đắn

Ngoài các trường hợp tầm thường, $\tilde{f}$ không thể là hàm liên tục. Tuy nhiên, một thuật toán tốt nên xấp xỉ bài toán f được kết hợp. Để làm ý tưởng này, ta xét *sai số tuyệt đối* của một phép tính, $\|\tilde{f}(x) - f(x)\|$, hoặc *sai số tương đối*,

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|}. \tag{3.3.1}$$

Trong sách này, ta chủ yếu sử dụng các con số tương đối và do đó (3.3.1) sẽ là độ đo sai số tiêu chuẩn.
Nếu $\tilde{f}$ là một thuật toán tốt thì nó có thể loại ra sai số tương đối là nhỏ của bậc $\epsilon_{machine}$. Một thuật toán $\tilde{f}$ cho một bài toán f là *đúng đắn* nếu với mỗi $x \in X$,

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(\tilde{x})\|} = O(\epsilon_{machine}). \tag{3.3.2}$$

Ký hiệu $O(\epsilon_{machine})$ trong (3.3.2) nghĩa là "trong bậc của machine epsilon".

3.3.3 Tính ổn định

Ta nói rằng một thuật toán cho bài toán f là *ổn định* nếu với mọi $x \in X$,

$$\frac{\|\tilde{f}(x) - f(\tilde{x})\|}{\|f(\tilde{x})\|} = O(\epsilon_{machine}) \tag{3.3.3}$$

với $\tilde{x}$ nào đó thỏa

$$\frac{\|\tilde{x} - x\|}{\|x\|} = O(\epsilon_{machine}). \tag{3.3.4}$$

Như vậy,

Một thuật toán ổn định cho câu trả lời gần đúng tới câu hỏi gần đúng.

3.3.4 Tính ổn định ngược

Một thuật toán $\tilde{f}$ cho một bài toán f là *ổn định ngược* nếu với mọi $x \in X$,

$$\tilde{f}(x) = f(\tilde{x}) \text{ với } \tilde{x} \text{ nào đó thỏa } \frac{\|\tilde{x} - x\|}{\|x\|} = O(\epsilon_{machine}). \tag{3.3.5}$$

Đây là một việc xiết chặt của định nghĩa tính ổn định mà trong đó $O(\epsilon_{machine})$ trong (3.3.3) đã được thay thế bởi 0. Như vậy,

Một thuật toán ổn định ngược cho câu trả lời chính xác tới câu hỏi gần đúng.

3.3.5 Ý nghĩa của $O(\epsilon_{machine})$

Bây giờ ta giải thích rõ ràng ý nghĩa của "$O(\epsilon_{machine})$" trong (3.3.2) - (3.3.5).
Ký hiệu

$$\varphi(t) = O(\psi(t)) \tag{3.3.6}$$

là một tiêu chuẩn trong toán học, với một định nghĩa rõ ràng. Phương trình này khẳng định rằng tồn tại hằng số dương C nào đó sao cho, với mọi t đủ gần một giới hạn đã biết (ví dụ, $t \to 0$ hoặc $t \to \infty$),

$$|\varphi(t)| \leqslant C\psi(t). \tag{3.3.7}$$

Ví dụ, $\sin^2 t = O(t^2)$ khi $t \to 0$ khẳng định rằng tồn tại một hằng số C sao cho, với mọi t đủ nhỏ, $|\sin^2 t| \leqslant Ct^2$.
Hơn nữa, tiêu chuẩn trong toán học là các phát biểu có dạng

$$\varphi(s,t) = O(\psi(t)) \text{ không thay đổi theo } s, \tag{3.3.8}$$

với φ là hàm phụ thuộc vào biến t và s. Từ "không thay đổi" cho biết tồn tại một hằng số C như trong (3.3.7) mà nó là đúng cho mọi sự lựa chọn của s. Do đó, ví dụ

$$(\sin^2 t)(\sin^2 x) = O(t^2)$$

xem như là không thay đổi khi $t \to 0$, nhưng sự không thay đổi là tổn thất nếu ta thay thế $\sin^2 s$ bằng s^2.
Trong sách này, sử dụng ký hiệu "O" theo sau các định nghĩa tiêu chuẩn này. Đặc biệt, ta thường xác định các kết quả dọc theo các dòng của

$$\|\text{con số được tính}\| = O(\epsilon_{machine}). \tag{3.3.9}$$

Đầu tiên, "‖con số được tính‖" biểu diễn chuẩn của một số nào đó hoặc sự lựa chọn các số được xác định bởi một thuật toán $\tilde{f}$ cho một bài toán f, phụ thuộc vào cả hai dữ liệu $x \in X$ cho f và $\epsilon_{machine}$. Ví dụ sai số tương đối trong (3.3.1). Thứ hai, quá trình giới hạn ẩn là $\epsilon_{machine} \to 0$ (nghĩa là, $\epsilon_{machine}$ là biến tương ứng với t trong (3.3.8)). Thứ ba, "O" áp dụng không thay đổi cho mọi dữ liệu $x \in X$ (nghĩa là, x là biến tương ứng với s). Ta sẽ ít khi đề cập sự không thay đổi tương ứng với $x \in X$ mà nó thường là ẩn.
Phương trình (3.3.9) nói rằng nếu ta chạy thuật toán trong câu hỏi trong các máy tính thỏa (3.2.5) và (3.2.7) cho một chuỗi các giá trị của $\epsilon_{machine}$ giảm xuống thành 0, khi đó ‖con số được tính‖ sẽ được đảm bảo để giảm xuống trong tỷ lệ thức với $\epsilon_{machine}$ hoặc nhanh hơn. Các máy tính lý tưởng này được yêu cầu thỏa mãn (3.2.5) và (3.2.7) nhưng không thỏa yêu cầu khác.

3.3.6 Phụ thuộc vào m và n, không phụ thuộc A và b

Giả sử ta đang xét một thuật toán cho việc giải một hệ thống $m \times m$ không suy biến của phương trình $Ax = b$ cho biến x, và ta khẳng định rằng kết quả tính được $\tilde{x}$ cho thuật toán này thỏa mãn

$$\frac{\|\tilde{x} - x\|}{\|x\|} = O(\kappa(A)\epsilon_{machine}). \tag{3.3.10}$$

Khẳng định này nghĩa là một chặn

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leqslant C\kappa(A)\epsilon_{machine}. \tag{3.3.11}$$

đúng cho một hằng số C, không phụ thuộc ma trận A hoặc vế bên phải b, với mọi $\epsilon_{machine}$ đủ nhỏ.
Nếu mẫu số trong một công thức giống (3.3.11) bằng 0 thì nó được xác định bởi quy ước theo sau. Khi đó, (3.3.11) được viết lại như sau

$$\|\tilde{x} - x\| \leqslant C\kappa(A)\epsilon_{machine}\|x\|. \tag{3.3.12}$$

không có sự khác nhau nếu $\|x\| \neq 0$, nhưng nếu $\|x\| = 0$, (3.3.12) làm rõ ý nghĩa của (3.3.10) là $\|\tilde{x} - x\| = 0$ với mọi $\epsilon_{machine}$ đủ nhỏ.
Mặc dù hằng số C của (3.3.11) hoặc (3.3.12) không phụ thuộc vào A hoặc b mà phụ thuộc vào số chiều m. Đây là một chuỗi định nghĩa của một bài toán trong mục (3.1). Nếu các chiều m hoặc n xác định một bài toán f thay đổi thì các không gian X và Y cũng phải thay đổi, và do đó ta có một bài toán mới, f'. Như một vấn đề thực tiễn, các hiệu quả của các sai số làm tròn trong các thuật toán của phương pháp số trong đại số tuyến tính thường phát triển với m và n. Tuy nhiên, sự phát triển này thường là chậm đủ để nó không đúng đắn. Phụ thuộc vào m hoặc n tiêu biểu là tuyến tính, bậc hai hoặc bậc ba trong trường hợp xấu nhất (số mũ phụ thuộc vào sự lựa chọn của chuẩn tốt như

sự lựa chọn của thuật toán), và các sai số cho hầu hết dữ liệu là nhỏ hơn nhiều trong trường hợp xấu nhất.

3.3.7 Sự độc lập của chuẩn

Định lý 3.3.1 *Cho các bài toán f và các thuật toán $\tilde{f}$ xác định trong các không gian hữu hạn chiều X và Y, các tính chất của sự đúng đắn, tính ổn định và ổn định ngược độc lập với sự lựa chọn các chuẩn trong X và Y.*

Chứng minh. Trong một không gian vector hữu hạn chiều X và Y, tất cả các chuẩn tương đương, nghĩa là nếu $\|\cdot\|$ và $\|\cdot\|'$ là hai chuẩn trong cùng không gian thì tồn tại các hằng số dương C_1 và C_2 sao cho $C_1\|x\| \leqslant \|x\|' \leqslant C_2\|x\|$ với mọi x trong không gian đó. Sự thay đổi của chuẩn có thể làm ảnh hưởng đến kích thước của hằng số C ẩn trong một phát biểu bao gồm $O(\epsilon_{machine})$, nhưng không tồn tại một hằng số như vậy.

3.3.8 Tính ổn định của số học dấu chấm động

Bốn bài toán tính toán đơn giản nhất là $+, -, \times$, và $\div$. Ta sẽ sử dụng các phép toán dấu chấm động $\oplus, \ominus, \otimes$, và $\oslash$ đã cung cấp với máy tính. Các tiên đề (3.2.5) và (3.2.7) cho thấy bốn ví dụ của thuật toán phù hợp với tiêu chuẩn này đều là ổn định ngược.
Ta hãy chứng minh điều này cho phép trừ, vì đó là một phép toán cơ bản mà ta có thể mong đợi là rủi ro lớn nhất của tính không ổn định. Như trong Ví dụ 3.1.4, không gian dữ liệu X là tập hợp các vector 2 chiều ($\mathbb{C}^2$) và không gian lời giải Y là tập hợp các vô hướng ($\mathbb{C}$). Theo Định lý 3.3.1, ta không cần chỉ rõ các chuẩn trong các không gian này. Với dữ liệu $x = (x_1, x_2)^* \in X$, bài toán phép trừ tương ứng với hàm $f(x_1, x_2) = x_1 - x_2$, và thuật toán ta đang xét có thể được viết

$$\tilde{f}(x_1, x_2) = fl(x_1) \ominus fl(x_2).$$

Phương trình này có nghĩa rằng đầu tiên ta làm tròn x_1 và x_2 thành các giá trị dấu chấm động, khi đó áp dụng phép toán $\ominus$. Do (3.2.5), ta có

$$fl(x_1) = x_1(1 + \epsilon_1),\ fl(x_2) = x_2(1 + \epsilon_2)$$

với $|\epsilon_1|, |\epsilon_2| \leqslant \epsilon_{machine}$ bất kì. Do (3.2.7), ta có

$$fl(x_1) \ominus fl(x_2) = (fl(x_1) - fl(x_2))(1 + \epsilon_3)$$

với $|\epsilon_3| \leqslant \epsilon_{machine}$ bất kì. Kết hợp các phương trình này ta được

$$\begin{aligned} fl(x_1) \ominus fl(x_2) &= [x_1(1 + \epsilon_1) - x_2(1 + \epsilon_2)](1 + \epsilon_3) \\ &= x_1(1 + \epsilon_1)(1 + \epsilon_3) - x_2(1 + \epsilon_2)(1 + \epsilon_3) \\ &= x_1(1 + \epsilon_4) - x_2(1 + \epsilon_5) \end{aligned}$$

với $|\epsilon_4|, |\epsilon_5| \leqslant 2\epsilon_{machine} + O(\epsilon^2_{machine})$. Mặt khác, kết quả tính được $\tilde{f}(x) = fl(x_1) \ominus fl(x_2)$ chính xác bằng với hiệu $\tilde{x_1} - \tilde{x_2}$, với $\tilde{x_1}$, và $\tilde{x_2}$ thỏa mãn

$$\frac{|\tilde{x_1} - x_1|}{|x_1|} = O(\epsilon_{machine}),\ \frac{|\tilde{x_2} - x_2|}{|x_2|} = O(\epsilon_{machine}),$$

và $C > 2$ bất kỳ sẽ đủ cho các hằng số ẩn trong các ký hiệu "O". Với sự lựa chọn bất kì của một chuẩn $\|\cdot\|$ trong không gian $\mathbb{C}^2$, điều này kéo theo (3.3.5).

3.3.9 Các ví dụ

Ví dụ 3.3.1. (Tích trong). Giả sử ta được cho các vector $x, y \in \mathbb{C}^m$ và mong muốn tính tích trong $\alpha = x^*y$. Thuật toán rõ ràng là để tính từng cặp tích $\overline{x_i}y_i$ với $\otimes$ và cộng với $\oplus$ để thu được một kết quả tính được $\tilde{\alpha}$. Ta có thể được chứng tỏ rằng thuật toán này là ổn định ngược trong mục 3.5.

Ví dụ 3.3.2. (Tích ngoài). Mặt khác, giả sử ta mong muốn tính tích ngoài hạng 1 $A = xy^*$ với các vector $x \in \mathbb{C}^m, y \in \mathbb{C}^n$. Thuật toán rõ ràng là để tính mn tích $x_i\overline{y_i}$ với $\otimes$ và tập hợp chúng thành một ma trận $\tilde{A}$. Thuật toán này là ổn định nhưng nó không ổn định ngược. Vì ma trận $\tilde{A}$ không có hạng chính xác là 1 nên nó không thể được viết tổng quát dưới dạng $(x + \delta x)(y + \delta y)^*$. Như một qui luật, cho các bài toán mà số chiều của không gian lời giải Y là lớn hơn số chiều của không gian bài toán X, tính ổn định ngược là hiếm.

Ví dụ 3.3.3. Giả sử ta sử dụng $\oplus$ để tính $x + 1$, $x \in \mathbb{C}$ được cho: $\tilde{f}(x) = fl(x) \oplus 1$. Thuật toán này là ổn định nhưng không ổn định ngược. Lý do là cho $x \approx 0$, phép cộng $\oplus$ sẽ đưa ra các sai số tuyệt đối của kích thước $O(\epsilon_{machine})$. Liên quan với kích thước x, các sai số này là không bị chặn, nên chúng không thể được hiểu như nguyên nhân gây ra bởi các nhiễu tương đối nhỏ trong dữ liệu. Nếu bài toán đã được tính $x + y$ cho dữ liệu x và y, khi đó thuật toán sẽ là ổn định ngược.

Ví dụ 3.3.4. Cho cả $\sin x$ và $\cos x$, tính ổn định ngược cũng được thể hiện bên ngoài hàm có đạo hàm bằng 0 tại các điểm nào đó. Ví dụ, giả sử ta tính $f(x) = \sin x$ trong một máy tính với $x = \pi/2 - \delta, 0 < \delta \ll 1$. Giả sử ta đủ may mắn để có được một kết quả được tính như là câu trả lời chính xác, được làm tròn tới hệ thống dấu chấm động: $\tilde{f}(x) = fl(\sin x)$. Vì $f'(x) = \cos x \approx \delta$ nên ta có $\tilde{f}(x) = f(\tilde{x})$ với $\tilde{x}$ nào đó thỏa mãn $\tilde{x} - x \approx (\tilde{f}(x) - f(x))/\delta = O(\epsilon_{machine}/\delta)$. Vì δ có thể là nhỏ tùy ý nên sai số ngược này không phải là của kích thước $O(\epsilon_{machine})$.

3.3.10 Thuật toán không ổn định

Sử dụng đa thức đặc trưng để tìm các trị riêng của một ma trận.

Vì z là một trị riêng của A nếu và chỉ nếu $p(z) = 0$, với $p(z)$ là đa thức đặc trưng của $det(zI - A)$ nên các nghiệm của p là các trị riêng của A. Một phương pháp được đưa ra

cho việc tính toán các trị riêng được đề nghị như sau:

1. Tìm các hệ số của đa thức đặc trưng,

2. Tìm các nghiệm của nó.

Thuật toán này không chỉ là không ổn định ngược mà còn là không ổn định, và nó không nên được sử dụng. Mặc dù trong các trường hợp mà việc trích các trị riêng là bài toán điều kiện tốt, nó có thể đưa ra các sai số tương đối lớn hơn $\epsilon_{machine}$.
Tính không ổn định được phát hiện trong việc tìm nghiệm của bước thứ hai. Như ta thấy trong Ví dụ 3.1.4, bài toán tìm các nghiệm của một đa thức với các hệ số được cho trước, nói chung là điều kiện xấu. Các sai số nhỏ trong các hệ số của đa thức đặc trưng có xu hướng được khuếch đại khi tìm các nghiệm, mặc dù việc tìm nghiệm đã hoàn thành.
Ví dụ, giả sử $A = I$, ma trận đơn vị 2×2. Các trị riêng của A không nhạy với nhiễu của các phần tử, và một thuật toán ổn định có thể tính chúng với các sai số $O(\epsilon_{machine})$. Tuy nhiên, thuật toán được miêu tả ở trên đưa ra các sai số trong bậc của $\sqrt{\epsilon_{machine}}$. Đa thức đặc trưng $x^2 - 2x + 1$, ngay trong Ví dụ 3.1.4. Khi các hệ số trong đa thức này được tính, chúng có thể được mong đợi có các sai số trong bậc của $\epsilon_{machine}$, và điều này có thể là nguyên nhân làm thay đổi các nghiệm bằng bậc $\sqrt{\epsilon_{machine}}$. Ví dụ, nếu $\epsilon_{machine} = 10^{-16}$, các nghiệm của đa thức đặc trưng tính được có thể được làm nhiễu từ các trị riêng hiện tại bởi xấp xỉ 10^{-8}, sự hao hụt 8 số nhị phân của sự đúng đắn.
Nếu sử dụng thuật toán được miêu tả để tính các trị riêng của ma trận đơn vị 2×2 thì ta sẽ thấy rằng không có sai số tại tất cả các trị riêng này. Bởi vì các hệ số và các nghiệm của $x^2 - 2x + 1$ là các số nguyên nhỏ sẽ được biểu diễn một cách chính xác trong một máy tính. Tuy nhiên, nếu sự thực thi được làm trong ma trận được làm nhiễu không đáng kể như ma trận A

$$A = \begin{bmatrix} 1 + 10^{-14} & 0 \\ 0 & 1 \end{bmatrix},$$

thì các trị riêng tính được sẽ phân biệt bởi bậc được mong đợi $\sqrt{\epsilon_{machine}}$.

3.3.11 Sự đúng đắn của thuật toán ổn định ngược

Giả sử ta có một thuật toán ổn định ngược $\tilde{f}$ cho một bài toán $f : X \to Y$. Sự đúng đắn phụ thuộc vào số điều kiện $\kappa = \kappa(x)$ của f. Nếu $\kappa(x)$ là nhỏ thì các kết quả sẽ là đúng đắn trong nghĩa tương đối, nhưng nếu nó là lớn thì sự đúng đắn sẽ cho phép tương đối.

Định lý 3.3.2 *Giả sử một thuật toán ổn định ngược được áp dụng để giải một bài toán $f : X \to Y$ với số điều kiện κ trong một máy tính thỏa các tiên đề (3.2.5) và (3.2.7).*

Khi đó, các sai số tương đối thỏa mãn

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} = O(\kappa(x)\epsilon_{machine}). \tag{3.3.13}$$

Chứng minh. Theo Định nghĩa (3.3.5) của ổn định ngược, ta có $\tilde{f}(x) = f(\tilde{x})$ với $\tilde{x}$ nào đó thỏa mãn

$$\frac{\|\tilde{x} - x\|}{\|x\|} = O(\epsilon_{machine}).$$

Theo Định nghĩa (3.1.5) của $\kappa(x)$, điều này kéo theo

$$\frac{\|\tilde{f}(x) - f(x)\|}{\|f(x)\|} \leqslant (\kappa(x) + o(1))\frac{\|\tilde{x} - x\|}{\|x\|}, \tag{3.3.14}$$

với $o(1)$ ký hiệu con số hội tụ tới 0 khi $\epsilon_{machine} \to 0$. Việc kết hợp các các chặn này thu được (3.3.13).

3.3.12 Phân tích sai số ngược

Quá trình mà ta đã thực hiện trong chứng minh Định lý 3.3.2 được biết như là *phân tích sai số ngược*. Ta thu được một ước lượng đúng đắn bằng hai bước. Bước đầu tiên là nghiên cứu điều kiện của bài toán. Bước còn lại là nghiên cứu tính ổn định của thuật toán. Kết luận của chúng ta là nếu thuật toán ổn định thì sự đúng đắn cuối cùng phản ánh số điều kiện đó.

Theo toán học, điều này là không phức tạp nhưng nó chắc chắn không phải là ý tưởng đầu tiên cho phân tích một thuật toán số. Ý tưởng đầu tiên sẽ là *phân tích sai số tiến*. Các sai số làm tròn đưa ra tại mỗi bước của tính toán được ước lượng, và vì một lý do chưa xác định, một tổng được duy trì như thế nào khi chúng có thể kết hợp từng bước một.

Thực nghiệm đã cho thấy rằng hầu hết các thuật toán của phương pháp số trong đại số tuyến tính, phân tích sai số tiến là khó thực hiện hơn phân tích sai số ngược. Giả sử thuật toán được chứng minh là đúng đắn được sử dụng để giải $Ax = b$ trong một máy tính. Nó là một thiết lập mà các kết quả thu được sẽ được nhất quán nhỏ hơn sự đúng đắn khi A là bài toán điều kiện xấu. Bây giờ, phân tích sai số tiến có thể nắm bắt hiện tượng này như thế nào? Số điều kiện của A cho thấy được nhiều hơn hoặc ít hơn tại mức của các phép toán dấu chấm động không thấy được bao gồm trong việc giải $Ax = b$. Phân tích tiến sẽ phải tìm ra số điều kiện đó nếu nó kết thúc với một kết quả chính xác.

Nói tóm lại, các thuật toán tốt nhất cho hầu hết các bài toán làm không tốt hơn tính các lời giải chính xác cho dữ liệu bị nhiễu nhỏ. Phân tích sai số ngược là một phương pháp lập luận gần phù hợp với thực tế ngược này.

3.4 Tính ổn định của tam giác hóa Householder

3.4.1 Định lý

Tam giác hóa Householder là ổn định ngược cho mọi ma trận A và mọi máy tính thỏa (3.2.5) và (3.2.7).
Kết quả sẽ có dạng

$$\tilde{Q}\tilde{R} = A + \delta A, \tag{3.4.1}$$

với δA nhỏ. Mặt khác, tích của Q với R được tính bằng với một nhiễu nhỏ của ma trận được cho A. Theo $\tilde{R}$, ma trận tam giác trên được xây dựng bằng tam giác hóa Gram - Schmidt trong số học dấu chấm động. Theo $\tilde{Q}$, một ma trận đã biết là *unita một cách chính xác*. Nhắc lại, $Q = Q_1 Q_2 \dots Q_n$ (2.4.8), với Q_k là mặt phản xạ Householder được xác định bởi vector v_k (2.4.5) xác định tại bước thứ k của Thuật toán 2.3. Trong tính toán dấu chấm động, ta thu được một chuỗi các vector $\tilde{v_k}$. Cho $\tilde{Q_k}$ ký hiệu mặt phản xạ *unita một cách chính xác* được xác định bởi vector dấu chấm động $\tilde{v_k}$ (theo toán học, không phải trong máy tính). Xác định

$$\tilde{Q} = \tilde{Q_1}\tilde{Q_2} \dots \tilde{Q_n}. \tag{3.4.2}$$

Ma trận unita một cách chính xác $\tilde{Q}$ này sẽ có vai trò đối với "Q được tính". Trong các ứng dụng, như được thảo luận trong mục 2.4, ma trận Q nói chung không được tạo thành rõ ràng bằng bất kì cách nào nên nó sẽ không hữu ích để xác định một "Q được tính" của nhiều dạng trước. Các vector $\tilde{v_k}$ được hình thành rõ ràng và thiết lập như trong (3.4.2).
Dưới đây là định lý giải thích sự thực thi trong Matlab của chúng.

Định lý 3.4.1 *Cho phân tích QR $A = QR$ của một ma trận $A \in \mathbb{C}^{m\times n}$ được tính bởi tam giác hóa Householder (Thuật toán 2.3) trong một máy tính thỏa các tiên đề (3.2.5) và (3.2.7), và cho các thừa số được tính $\tilde{Q}$ và $\tilde{R}$ được xác định như ở trên. Khi đó, ta có*

$$\tilde{Q}\tilde{R} = A + \delta A, \ \frac{\|\delta A\|}{\|A\|} = O(\epsilon_{machine}) \tag{3.4.3}$$

với $\delta A \in \mathbb{C}^{m\times n}$ bất kì.

Biểu thức $O(\epsilon_{machine})$ trong (3.4.3) đã được thảo luận trong các mục 3.3. Chặn đúng khi $\epsilon_{machine} \to 0$, đồng nhất cho mọi ma trận A có số chiều m và n được cố định bất kì, nhưng không đồng nhất tương ứng với m và n. Bởi vì tất cả các chuẩn trong không gian hữu hạn chiều là tương đương nhau, ta không cần một chuẩn đặc biệt (Định lý 3.3.1).

3.4.2 Phân tích một thuật toán giải phương trình Ax = b

Ta đã thấy rằng tam giác hóa Householder là không ổn định ngược nhưng thường không đúng đắn trong chiều tiến. Phân tích QR nói chung không kết thúc trong chính nó mà để

kết thúc khác như là nghiệm của một hệ thống các phương trình, bài toán bình phương nhỏ nhất, hoặc một bài toán trị riêng. Sự đúng đắn của phân tích QR đủ cho các ứng dụng, hoặc ta cần sự đúng đắn của Q và R là đủ cho hầu hết các mục tiêu. Ta có thể chứng minh điều này bằng các đối số đơn giản.

Ví dụ mà ta sẽ xét là sử dụng tam giác hóa Householder để giải hệ thống tuyến tính $m \times m$ không suy biến $Ax = b$. Ý tưởng này được thảo luận tại phần cuối của mục 2.2. Dưới đây là một phát biểu đầy đủ hơn của thuật toán đó.

Thuật toán 3.1 Giải $Ax = b$ bằng phân tích QR

1: $QR = A$ Phân tích A thành QR bằng Thuật toán 2.3, với Q biểu diễn tích của các mặt phản xạ.
2: $y = Q^*b$ Xây dựng Q^*b bằng Thuật toán 2.4
3: $x = R^{-1}y$ Giải hệ thống tam giác $Rx = y$ bằng phép thế ngược (Thuật toán 3.2).

Thuật toán này là ổn định ngược và việc chứng minh điều này là không phức tạp, mỗi bước trong 3 bước là tự nó ổn định ngược. Ở đây, ta sẽ phát biểu các kết quả ổn định ngược cho 3 bước (không chứng minh).

Bước đầu tiên của Thuật toán 3.1 là phân tích QR của A, dẫn đến các ma trận $\tilde{R}$ và $\tilde{Q}$ được tính. Tính ổn định ngược của quá trình này đã được biểu diễn bởi (3.4.3).

Bước thứ hai là tính Q^*b bằng Thuật toán 2.4. Khi Q^*b được tính bằng Thuật toán 2.4 với việc làm tròn các sai số nên kết quả sẽ không chính xác là $\tilde{Q}^*b$. Thay vào đó, nó sẽ là một vector $\tilde{y}$ nào đó và thỏa ước lượng tính ổn định ngược theo sau:

$$(\tilde{Q} + \delta Q)\tilde{y} = b,\ \|\delta Q\| = O(\epsilon_{machine}). \tag{3.4.4}$$

Giống như (3.4.3), đẳng thức này là chính xác. Mặt khác, kết quả việc áp dụng các mặt phản xạ Householder trong số học dấu chấm động chính xác là tương đương với việc nhân b với một ma trận bị nhiễu nhỏ, $(\tilde{Q} + \delta Q)^{-1}$.

Bước cuối cùng của Thuật toán 3.1 là phép thế ngược để tính $\tilde{R}^{-1}\tilde{y}$. Trong bước này, các sai số làm tròn mới sẽ được đưa ra nhưng nhiều hơn 1 nên tính toán là ổn định ngược. Ước lượng này có dạng

$$(\tilde{R} + \delta R)\tilde{x} = \tilde{y},\ \frac{\|\delta R\|}{\|\tilde{R}\|} = O(\epsilon_{machine}). \tag{3.4.5}$$

Đẳng thức trong vế trái khẳng định rằng kết quả dấu chấm động $\tilde{x}$ là lời giải chính xác của một nhiễu nhỏ của hệ là chính xác $\tilde{R}x = \tilde{y}$.

Định lý 3.4.2 *Thuật toán 3.1 là ổn định ngược, thỏa mãn*

$$(A + \Delta A)\tilde{x} = b,\ \frac{\|\Delta A\|}{\|A\|} = O(\epsilon_{machine}) \tag{3.4.6}$$

với $\Delta A \in \mathbb{C}^{m\times n}$ *bất kì.*

Chứng minh. Kết hợp (3.4.4) và (3.4.5), ta có

$$b = (\tilde{Q} + \delta Q)(\tilde{R} + \delta R)\tilde{x} = [\tilde{Q}\tilde{R} + (\delta Q)\tilde{R} + \tilde{Q}(\delta R) + (\delta Q)(\delta R)]\tilde{x}.$$

Do đó, theo (3.4.3),

$$b = [A + \delta A + (\delta Q)\tilde{R} + \tilde{Q}(\delta R) + (\delta Q)(\delta R)]\tilde{x}.$$

Phương trình này có dạng

$$b = (A + \Delta A)\tilde{x},$$

với ΔA là tổng của 4 số hạng. Để ước lượng (3.4.6), ta phải cho thấy rằng mỗi số hạng này là tương đối nhỏ với A.
Vì $\tilde{Q}\tilde{R} = A + \delta A$ và $\tilde{Q}$ là unita nên ta có

$$\frac{\|\tilde{R}\|}{\|A\|} \leqslant \|\tilde{Q}^*\| \frac{\|A + \delta A\|}{\|A\|} = O(1)$$

khi $\epsilon_{machine} \to 0$, do (3.4.3). Do (3.4.4)nên

$$\frac{\|(\delta Q)\tilde{R}\|}{\|A\|} \leqslant \|\delta Q\| \frac{\|\tilde{R}\|}{\|A\|} = O(\epsilon_{machine})$$

Tương tự,

$$\frac{\|\tilde{Q}(\delta R)\|}{\|A\|} \leqslant \|\tilde{Q}\| \frac{\|\delta R\|}{\|\tilde{R}\|} \frac{\|\tilde{R}\|}{\|A\|} = O(\epsilon_{machine})$$

do (3.4.5). Cuối cùng,

$$\frac{\|(\delta Q)(\delta R)\|}{\|A\|} \leqslant \|\delta Q\| \frac{\|\delta R\|}{\|A\|} = O(\epsilon_{machine}^2).$$

Khi đó, tổng nhiễu ΔA thỏa mãn

$$\frac{\|\Delta A\|}{\|A\|} \leqslant \frac{\|\delta A\|}{\|A\|} + \frac{\|(\delta Q)\tilde{R}\|}{\|A\|} + \frac{\|\tilde{Q}(\delta R)\|}{\|A\|} + \frac{\|(\delta Q)(\delta R)\|}{\|A\|} = O(\epsilon_{machine}),$$

như được yêu cầu.

Kết hợp Định lý 3.1.2, 3.3.2, 3.4.2 và 3.4.2 cho kết quả theo sau về sự đúng đắn của các lời giải $Ax = b$.

Định lý 3.4.3 *Lời giải $\tilde{x}$ được tính bằng Thuật toán 3.1 thỏa mãn*

$$\frac{\|\tilde{x} - x\|}{\|x\|} = O(\kappa(A)\epsilon_{machine}). \tag{3.4.7}$$

3.5 Tính ổn định của phép thế ngược

3.5.1 Hệ thống tam giác

Một hệ phương trình tổng quát $Ax = b$ có thể được giảm xuống thành một hệ thống tam giác trên $Rx = b$ bằng phân tích QR. Các hệ thống tam giác dưới và tam giác trên

cũng xuất hiện trong khử Gauss, trong phân tích Cholesky, và trong các tính toán số khác.

Các hệ thống này dễ dàng được giải bằng một qúa trình của phép thế liên tiếp được gọi là *phép thế tiến* nếu hệ thống là tam giác dưới và *phép thế ngược* nếu hệ thống là tam giác trên. Mặc dù hai trường hợp là đồng nhất nhưng cho tính xác định ta xét phép thế ngược trong mục này.

Giả sử ta mong muốn giải $Rx = b$,

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ & r_{22} & & \\ & & \ddots & \vdots \\ & & & r_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \tag{3.5.1}$$

với $b \in \mathbb{C}^m$ và $R \in \mathbb{C}^{m\times m}$ là các ma trận không suy biến và tam giác trên được cho, và $x \in \mathbb{C}^m$ không được biết. Ta có thể làm điều này bằng việc giải các thành phần của x, bắt đầu với x_m và hoàn thành với x_1. Cho thuận lợi ta viết thuật toán như một chuỗi các công thức hơn là vòng lặp.

Thuật toán 3.2 Phép thế ngược

$x_m = \frac{b_m}{r_{mm}}$

$x_{m-1} = \dfrac{(b_{m-1} - x_m r_{m-1,m})}{r_{m-1,m-1}}$

$x_{m-2} = \dfrac{(b_{m-2} - x_{m-1} r_{m-2,m-1} - x_m r_{m-2,m})}{r_{m-2,m-2}}$

$\vdots$

$x_j = \left(b_j - \sum_{k=j+1}^{m} x_k r_{jk}\right) \Big/ r_{jj}$

Cấu trúc là tam giác, với một phép trừ và một phép nhân tại mỗi vị trí. Do đó, đếm số phép toán là hai lần diện tích tam giác $m \times m$:

$$\text{Phép thế ngược: } \sim m^2 \text{ phép toán dấu chấm động.} \tag{3.5.2}$$

3.5.2 Định lý ổn định ngược

Trong mục trước, phép thế ngược xuất hiện như một trong ba bước trong lời giải của $Ax = b$ bằng phân tích QR. Trong (3.4.3) - (3.4.5) ta khẳng định rằng mỗi bước này là ổn định ngược nhưng ta không chứng minh yêu cầu này.

Định lý 3.5.1 *Cho Thuật toán 3.2 được áp dụng cho bài toán (3.5.1) bao gồm các số dấu chấm động trong một máy tính thỏa (3.2.7). Thuật toán này là ổn định ngược mà lời giải tính được $\tilde{x} \in \mathbb{C}^m$ thỏa mãn*

$$(R + \delta R)\tilde{x} = b \tag{3.5.3}$$

cho tam giác trên $\delta R \in \mathbb{C}^{m\times m}$ *bất kì mà*

$$\frac{\|\delta R\|}{\|R\|} = O(\epsilon_{machine}). \tag{3.5.4}$$

Đặc biệt, với mỗi i, j,

$$\frac{|\delta r_{ij}|}{|r_{ij}|} \leqslant m\epsilon_{machine} + O(\epsilon^2_{machine}). \tag{3.5.5}$$

Trong (3.5.5) và thông qua mục này, ta tiếp tục sử dụng quy ước của (3.3.12) mà nếu mẫu số bằng 0 thì tử số cũng được khẳng định bằng 0 (với mọi $\epsilon_{machine}$ đủ nhỏ).

3.5.3 m=1

Theo (3.5.3), công việc của chúng ta là biểu diễn mọi sai số dấu chấm động như là một nhiễu của dữ liệu đầu vào. Ta hãy bắt đầu với trường hợp đơn giản nhất mà R có số chiều là 1×1. Phép thế ngược trong trường hợp này bao gồm một bước,

$$\tilde{x_1} = b_1 \oslash r_{11}.$$

Tiên đề (3.2.7) cho $\oslash$ đảm bảo rằng lời giải tính được là gần đúng:

$$\tilde{x_1} = \frac{b_1}{r_{11}}(1 + \epsilon_1),\ |\epsilon_1| \leqslant \epsilon_{machine}.$$

Tuy nhiên, ta muốn biểu diễn sai số nếu như nó đưa ra kết quả từ một nhiễu trong R. Để kết thúc điều này, ta đặt $\epsilon_1^{'} = -\dfrac{\epsilon_1}{1+\epsilon_1}$, công thức ở trên trở thành

$$\tilde{x_1} = \frac{b_1}{r_{11}(1 + \epsilon_1^{'})},\ |\epsilon_1^{'}| \leqslant \epsilon_{machine} + O(\epsilon^2_{machine}). \tag{3.5.6}$$

Chú ý rằng, $\epsilon_1^{'}$ bằng $-\epsilon_1$ cộng với một số hạng bậc ϵ_1^2. Ta có thể tự do di chuyển các nhiễu tương đối nhỏ từ các tử số tới các mẫu số hoặc ngược lại, và kết quả thay đổi bằng các số hạng bậc $\epsilon^2_{machine}$.

Trong (3.5.6), đẳng thức là chính xác; phép chia thuộc về toán học chứ không phải là dấu chấm động. Công thức phát biểu rằng phép thế ngược 1×1 là ổn định ngược, với $\tilde{x_1}$ là lời giải một cách chính xác tới một bài toán bị nhiễu, cụ thể

$$(r_{11} + \delta r_{11})\tilde{x_1} = b_1,$$

với $\delta r_{11} = \epsilon_1^{'} r_{11}$. Do đó

$$\frac{|\delta r_{11}|}{|r_{11}|} \leqslant \epsilon_{machine} + O(\epsilon^2_{machine}).$$

3.5.4 m = 2

Giả sử ta có một ma trận tam giác trên $R \in \mathbb{C}^{2\times 2}$ và một vector $b \in \mathbb{C}^2$. Tính $\tilde{x} \in \mathbb{C}^2$ tiến hành trong hai bước. Đầu tiên là giống như trong trường hợp 1×1:

$$\tilde{x_2} = b_2 \oslash r_{22} = \frac{b_2}{r_{22}(1 + \epsilon_1)},\ |\epsilon_1| \leqslant \epsilon_{machine} + O(\epsilon^2_{machine}). \tag{3.5.7}$$

Bước thứ hai được xác định bởi công thức

$$\tilde{x_1} = (b_1 \ominus (\tilde{x_2} \otimes r_{12})) \oslash r_{11}.$$

Để thiết lặp ổn định ngược, ta phải biểu diễn các sai số trong 3 phép toán dấu chấm động này như các nhiễu trong các phần tử r_{ij}.

Phép nhân là dễ. Ta sử dụng tiên đề (3.2.7) để giải thích phép nhân dấu chấm động như là một nhiễu trong r_{12}:

$$\tilde{x_1} = (b_1 \ominus \tilde{x_2} r_{12}(1+\epsilon_2)) \oslash r_{11},\ |\epsilon_2| \leqslant \epsilon_{machine}.$$

Phép chia và phép trừ là khó thấy hơn. Đầu tiên, ta viết công thức với toán học chính xác theo (3.2.7):

$$\tilde{x_1} = (b_1 - \tilde{x_2} r_{12}(1+\epsilon_2))(1+\epsilon_3) \oslash r_{11} \tag{3.5.8}$$

$$= \frac{(b_1 - \tilde{x_2} r_{12}(1+\epsilon_2))(1+\epsilon_3)}{r_{11}}(1+\epsilon_4). \tag{3.5.9}$$

Ở đây (3.5.7) đảm bảo $|\epsilon_3|,\ |\epsilon_4| \leqslant \epsilon_{machine}$. Ta di chuyển các số hạng ϵ_3 và ϵ_4 từ tử số thành mẫu số. Điều này đưa ra

$$\tilde{x_1} = \frac{b_1 - \tilde{x_2} r_{12}(1+\epsilon_2)}{r_{11}(1+\epsilon_3^{'})(1+\epsilon_4^{'})},$$

với $|\epsilon_3^{'}|, |\epsilon_4^{'}| \leqslant \epsilon_{machine} + O(\epsilon_{machine}^2)$, hoặc tương đương

$$\tilde{x_1} = \frac{b_1 - \tilde{x_2} r_{12}(1+\epsilon_2)}{r_{11}(1+2\epsilon_5)}, \tag{3.5.10}$$

với $|\epsilon_5| \leqslant \epsilon_{machine} + O(\epsilon_{machine}^2)$. Công thức này phát biểu rằng $\tilde{x_1}$ sẽ là chính xác nếu r_{22}, r_{12} và r_{11} được làm nhiễu bởi các thừa số $(1+\epsilon_1), (1+\epsilon_2)$ và $(1+2\epsilon_5)$ tương ứng. Các nhiễu này có thể được tóm tắt bằng phương trình

$$(R + \delta R)\tilde{x} = b,$$

với các phần tử δr_{ij} của δR thỏa mãn

$$\begin{bmatrix} |\delta r_{11}|/|r_{11}| & |\delta r_{12}|/|r_{12}| \\ & |\delta r_{22}|/|r_{22}| \end{bmatrix} = \begin{bmatrix} 2|\epsilon_5| & |\epsilon_2| \\ & |\epsilon_1| \end{bmatrix} \leqslant \begin{bmatrix} 2 & 1 \\ & 1 \end{bmatrix} \epsilon_{machine} + O(\epsilon_{machine}^2).$$

Công thức này đảm bảo $\|\delta R\|/\|R\| = O(\epsilon_{machine})$ trong chuẩn ma trận bất kỳ và do đó phép thế ngược 2×2 đó là ổn định ngược.

3.5.5 m =3

Phân tích cho một ma trận 3×3 bao gồm tất cả lý do cần thiết cho trường hợp tổng quát. Đầu tiên, 2 bước là giống như trước:

$$\tilde{x_3} = b_3 \oslash r_{33} = \frac{b_3}{r_{33}(1+\epsilon_1)}, \tag{3.5.11}$$

$$\tilde{x_2} = (b_2 \ominus (\tilde{x_3} \otimes r_{23})) \oslash r_{22} = \frac{b_2 - \tilde{x_3} r_{23}(1+\epsilon_2)}{r_{22}(1+2\epsilon_3)}, \tag{3.5.12}$$

với

$$\begin{bmatrix} 2|\epsilon_3| & |\epsilon_2| \\ & |\epsilon_1| \end{bmatrix} \leqslant \begin{bmatrix} 2 & 1 \\ & 1 \end{bmatrix} \epsilon_{machine} + O(\epsilon^2_{machine}).$$

Bước thứ ba bao gồm tính toán

$$\tilde{x_1} = [(b_1 \ominus (\tilde{x_2} \otimes r_{12})) \ominus (\tilde{x_3} \otimes r_{13})] \oslash r_{11}. \tag{3.5.13}$$

Ta biến đổi hai phép toán $\otimes$ trong (3.5.13) thành phép nhân toán học bằng việc đưa ra các nhiễu ϵ_4 và ϵ_5

$$\tilde{x_1} = [(b_1 \ominus \tilde{x_2} r_{12}(1+\epsilon_4)) \ominus \tilde{x_3} r_{13}(1+\epsilon_5)] \oslash r_{11}.$$

Ta biến đổi các phép toán $\ominus$ thành các phép trừ toán học thông qua các nhiễu ϵ_6 và ϵ_7:

$$\tilde{x_1} = [(b_1 - \tilde{x_2} r_{12}(1+\epsilon_4))(1+\epsilon_6) - \tilde{x_3} r_{13}(1+\epsilon_5)](1+\epsilon_7) \oslash r_{11}.$$

Cuối cùng, $\oslash$ được ước lượng bằng việc sử dụng ϵ_8. Ta hãy thay thế điều này bằng $\epsilon_8^{'}$ với $|\epsilon_8| \leqslant \epsilon_{machine} + O(\epsilon^2_{machine})$ và đặt kết quả trong mẫu số:

$$\tilde{x_1} = \frac{[(b_1 - \tilde{x_2} r_{12}(1+\epsilon_4))(1+\epsilon_6) - \tilde{x_3} r_{13}(1+\epsilon_5)](1+\epsilon_7)}{r_{11}(1+\epsilon_8^{'})}.$$

Ta đổi ϵ_7 thành $\epsilon_7^{'}$ và di chuyển nó tới mẫu số như thường dùng. Số hạng bao gồm ϵ_6 yêu cầu một thủ thuật mới. Ta di chuyển nó thành mẫu số như vậy nhưng để giữ đẳng thức hợp lệ, ta làm cân bằng bằng việc đặt một thừa số mới $(1+\epsilon_6^{'})$ vào số hạng r_{13}. Khi đó

$$\tilde{x_1} = \frac{b_1 - \tilde{x_2} r_{12}(1+\epsilon_4) - \tilde{x_3} r_{13}(1+\epsilon_5)(1+\epsilon_6^{'})}{r_{11}(1+\epsilon_6^{'})(1+\epsilon_7^{'})(1+\epsilon_8^{'})}.$$

Bây giờ, r_{13} có 2 nhiễu của kích thước nhiều nhất là $\epsilon_{machine}$, và r_{11} có ba nhiễu. Trong công thức này, tất cả các sai số trong phép tính đã được biểu diễn như các nhiễu trong các phần tử của R.

Kết quả có thể được tóm tắt như

$$(R + \delta R)\tilde{x} = b,$$

với các phần tử δr_{ij} thỏa mãn

$$\begin{bmatrix} |\delta r_{11}|/|r_{11}| & |\delta r_{12}|/|r_{12}| & |\delta r_{13}|/|r_{13}| \\ & |\delta r_{22}|/|r_{22}| & |\delta r_{23}|/|r_{23}| \\ & & |\delta r_{33}|/|r_{33}| \end{bmatrix} \leqslant \begin{bmatrix} 3 & 1 & 2 \\ & 2 & 1 \\ & & 1 \end{bmatrix} \epsilon_{machine} + O(\epsilon_{machine}^2).$$

3.5.6 m tổng quát

Phân tích trong các trường hợp số chiều cao hơn là tương tự. Ví dụ, trong trường hợp 5×5 ta thu được chặn theo từng thành phần

$$\frac{|\delta R|}{|R|} \leqslant \begin{bmatrix} 5 & 1 & 2 & 3 & 4 \\ & 4 & 1 & 2 & 3 \\ & & 3 & 1 & 2 \\ & & & & 1 \end{bmatrix} \epsilon_{machine} + O(\epsilon_{machine}^2). \tag{3.5.14}$$

Các phần tử của ma trận trong công thức này thu được từ 3 thành phần. Các phép nhân $\tilde{x_k} r_{jk}$ đưa ra các nhiễu $\epsilon_{machine}$ trong dạng

$$\otimes : \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ & 0 & 1 & 1 & 1 \\ & & 0 & 1 & 1 \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}. \tag{3.5.15}$$

Các phép chia cho r_{kk} đưa ra các nhiễu trong dạng

$$\oslash : \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}. \tag{3.5.16}$$

Cuối cùng, phép trừ cũng xuất hiện trong dạng (3.5.15). Do quyết định tính toán từ trái sang phải nên mỗi phép trừ đưa ra một nhiễu trên đường chéo và tại mỗi vị trí tới bên phải. Điều này tăng thêm lên thành dạng

$$\ominus : \begin{bmatrix} 4 & 0 & 1 & 2 & 3 \\ & 3 & 0 & 1 & 2 \\ & & 2 & 0 & 1 \\ & & & 1 & 0 \\ & & & & 0 \end{bmatrix}. \tag{3.5.17}$$

Thêm (3.5.15), (3.5.16) và (3.5.17) đưa ra kết quả trong (3.5.14). Điều này hoàn thành chứng minh của Định lý 3.5.1.

3.6 Điều kiện của các bài toán bình phương nhỏ nhất

3.6.1 Bốn bài toán điều kiện

Trong mục này, ta quay lại bài toán bình phương nhỏ nhất tuyến tính (2.5.2), được minh họa lại như trong Hình 3.2. Giả sử ma trận xác định bài toán có hạng đầy đủ và viết $\|\cdot\| = \|\cdot\|_2$:

Cho $A \in \mathbb{C}^{m\times n}$có hạng đầy đủ $(m \geqslant n), b \in \mathbb{C}^m$, tìm $x \in \mathbb{C}^n$ sao cho $\|b-Ax\|$ được cực tiểu hóa. (3.6.1)

Lời giải x và điểm tương ứng $y = Ax$ là gần b nhất trong $range(A)$ được cho bởi

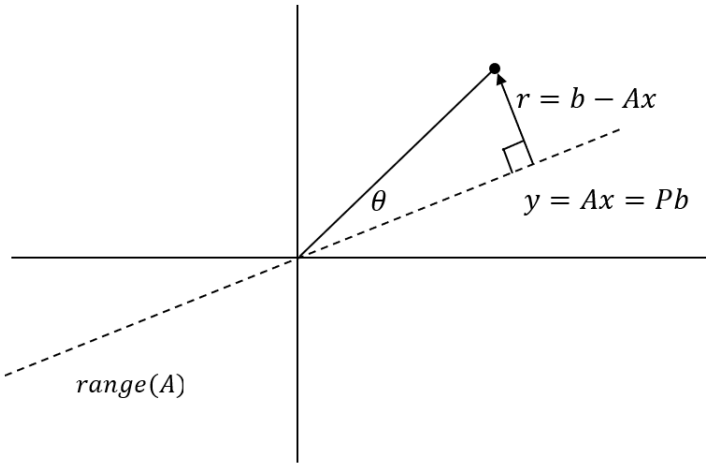
$$x = A^+ b,\ y = Pb, \tag{3.6.2}$$

với $A^+ \in \mathbb{C}^{n\times m}$ là giả nghịch đảo của A (2.5.11) và $P = AA^+ \in \mathbb{C}^{m\times m}$ là phép chiếu trực giao lên trên $range(A)$.

Ta xét điều kiện của (3.6.1) tương ứng với các nhiễu. Điều kiện liên quan với độ nhạy của các lời giải tới các nhiễu trong dữ liệu. Cho (3.6.1), ta sẽ khảo sát hai lựa chọn của mỗi dữ liệu. Dữ liệu cho bài toán là ma trận A có $m \times n$ chiều và vector b có m chiều. Lời giải là vector hệ số x hoặc điểm $y = Ax$ tương ứng. Do đó

$$\text{Dữ liệu:} A, b, \quad \text{Lời giải:} x, y.$$

Đồng thời, hai cặp lựa chọn này xác định bốn câu hỏi điều kiện mà ta sẽ xét, tất cả đều có ứng dụng trong các ngữ cảnh nào đó.



Hình 3.2: Bài toán bình phương nhỏ nhất

3.6.2 Định lý

Đầu tiên là số điều kiện của A. Cho một ma trận vuông, đó là là $\kappa(A) = \|A\|\|A^{-1}\|$, và trong trường hợp hình chữ nhật, định nghĩa tổng quát hóa (3.1.17),

$$\kappa(A) = \|A\|\|A^{+}\| = \frac{\sigma_1}{\sigma_n}. \tag{3.6.3}$$

Thứ hai là góc θ như trong Hình 3.2, một độ đo của tính chính xác của sự điều chỉnh cho vừa:

$$\theta = \cos^{-1}\frac{\|y\|}{\|b\|}. \tag{3.6.4}$$

Thứ ba là độ đo của $\|y\|$ bao nhiêu không đạt được giá trị có thể lớn nhất của nó, $\|A\|$ và $\|x\|$ được cho:

$$\eta = \frac{\|A\|\|x\|}{\|y\|} = \frac{\|A\|\|x\|}{\|Ax\|}. \tag{3.6.5}$$

Các tham số này nằm trong các khoảng

$$1 \leqslant \kappa(A) < \infty,\ 0 \leqslant \theta \leqslant \pi/2,\ 1 \leqslant \eta \leqslant \kappa(A). \tag{3.6.6}$$

Định lý 3.6.1 *Cho $b \in \mathbb{C}^m$ và $A \in \mathbb{C}^{m\times n}$ có hạng đầy đủ được cố định. Bài toán bình phương nhỏ nhất (3.6.1) có các số điều kiện tương đối trong chuẩn 2 theo sau (3.1.5) miêu tả các độ nhạy của y và x tới các nhiễu trong b và A:*

	y	x
b	$\dfrac{1}{\cos\theta}$	$\dfrac{\kappa(A)}{\eta\cos\theta}$
A	$\dfrac{\kappa(A)}{\cos\theta}$	$\kappa(A) + \dfrac{\kappa(A)^2\tan\theta}{\eta}$

Các kết quả trong dòng đầu tiên là chính xác cho các nhiễu δb nào đó và các kết quả trong dòng thứ hai là các chặn trên.

Trong trường hợp đặc biệt $m = n$, (3.6.1) giảm xuống thành một hệ thống các phương trình không suy biến, vuông với $\theta = 0$. Trong trường hợp này, các số trong cột thứ 2 của định lý giảm xuống thành $\kappa(A)/\eta$ và $\kappa(A)$ mà chúng là các kết quả trong (3.1.13) và (3.1.18) suy ra dễ dàng hơn, và số trong vị trí bên trái thấp hơn có thể được thay thế bằng 0.

3.6.3 Biến đổi thành một ma trận đường chéo

Cho A có một SVD $A = U\Sigma V^*$, với Σ là ma trận đường chéo $m \times n$ với các phần tử trên đường chéo dương. Vì các nhiễu là đo được trong chuẩn 2, các kích thước của chúng không bị tác động bởi một thay đổi cơ sở Unita, nên cách xử lý nhiễu của A là giống

như của Σ. Do đó, không mất tính tổng quát, ta có thể giải quyết Σ một cách trực tiếp. Cho phần còn lại của thảo luận, ta giả sử $A = \Sigma$ và viết

$$A = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \\ & & & \end{bmatrix} = \begin{bmatrix} A_1 \\ 0 \end{bmatrix}. \tag{3.6.7}$$

Ở đây A_1 là ma trận đường chéo có $n \times n$ chiều, các phần tử còn lại của A là 0.
Phép chiếu trực giao của b lên trên $range(A)$ bây giờ là không tầm thường. Viết

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

với b_1 chứa n phần tử đầu tiên của b. Khi đó, phép chiếu $y = Pb$ là

$$y = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}.$$

Để tìm x tương ứng ta có thể viết $Ax = y$ như

$$\begin{bmatrix} A_1 \\ 0 \end{bmatrix} x = \begin{bmatrix} b_1 \\ 0 \end{bmatrix},$$

mà nó kéo theo

$$x = A_1^{-1} b_1. \tag{3.6.8}$$

Từ các công thức này, phép chiếu trực giao và giả nghịch đảo là các ma trận khối 2×2 và 1×2.

$$P = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \; A^+ = [A_1^{-1} \quad 0]. \tag{3.6.9}$$

3.6.4 Độ nhạy của y tới các nhiễu trong b

Ta bắt đầu với 4 kết quả điều kiện đơn giản nhất. Do (3.6.2), các quan hệ giữa b và y chỉ là phương trình tuyến tính $y = Pb$. Ánh xạ Jacobi này là P vào chính nó, với $\|P\| = 1$ theo (3.6.9). Do (3.1.6) và (3.6.4), số điều kiện của y tương ứng với các nhiễu trong b là

$$\kappa_{b \mapsto y} = \frac{\|P\|}{\|y\|/\|b\|} = \frac{1}{\cos\theta}.$$

Điều này thiết lặp kết quả ở bên trái phía trên của Định lý 3.6.1. Số điều kiện được thực hiện cho các nhiễu δb với $\|P(\delta b)\| = \|\delta b\|$, mà nó xuất hiện khi δb bằng 0 ngoại trừ n phần tử đầu tiên.

3.6.5 Độ nhạy của x tới các nhiễu trong b

Quan hệ giữa b và x cũng là tuyến tính, $x = A^+b$, với Jacobian A^+. Do (3.1.6), (3.6.4) và (3.6.5), số điều kiện của x tương ứng với các nhiễu trong b là

$$\kappa_{b\mapsto x} = \frac{\|A^+\|}{\|x\|/\|b\|} = \|A^+\|\frac{\|b\|}{\|y\|}\frac{\|y\|}{\|x\|} = \|A^+\|\frac{1}{\cos\theta}\frac{\|A\|}{\eta} = \frac{\kappa(A)}{\eta\cos\theta}.$$

Điều này thiết lặp kết quả ở bên phải phía trên của Định lý 3.6.1. Số điều kiện của x được thực hiện bởi các nhiễu δb thỏa mãn $\|A^+(\delta b)\| = \|A^+\|\|\delta b\| = \|\delta b\|/\sigma_n$, mà nó xuất hiện khi δb bằng 0 ngoại trừ trong phần tử thứ n (hoặc cũng có thể trong các phần tử khác, nếu A có nhiều một giá trị suy biến bằng σ_n).

3.6.6 Độ dốc range của A

Phân tích các nhiễu trong A là bài toán không tuyến tính và tinh vi hơn. Ta sẽ tiếp tục bằng việc tính các Jacobi theo phương pháp đại số. Đầu tiên ta quan sát các nhiễu trong A làm ảnh hưởng tới bài toán bình phương nhỏ nhất trong 2 cách: chúng làm biến dạng ánh xạ của $\mathbb{C}^n$ lên $range(A)$, và chúng làm thay đổi chính $range(A)$.
Ta có thể hình dung các thay đổi nhỏ trong $range(A)$ như "các độ dốc" nhỏ của không gian này. Góc lớn nhất của độ dốc $\delta\alpha$ mà nó có thể tác động bằng một nhiễu nhỏ δA được xác định như sau. Ảnh bên dưới A của quả cầu đơn vị n chiều là một siêu ellip mà nó nằm trong $range(A)$. Để thay đổi $range(A)$ hiệu quả như có thể thực hiện được, ta lấy một điểm $p = Av$ trong siêu ellip (do đó $\|v\| = 1$) và nhích nó trong một phương δp trực giao với $range(A)$. Một ma trận nhiễu mà nó đạt được hầu hết một cách hiệu là $\delta A = (\delta p)v^*$, sao cho $(\delta A)v = \delta p$ với $\|\delta A\| = \|\delta p\|$. Bây giờ rõ ràng rằng để thu được độ dốc lớn nhất với một $\|\delta p\|$ được cho, ta sẽ lấy p là gần với gốc như có thể thực hiện được. Đó là $p = \sigma_n u_n$ với σ_n là giá trị suy biến nhỏ nhất của A và u_n là vector suy biến trái tương ứng. Với A trong dạng đường chéo (3.6.7), p là bằng với cột cuối cùng của A, v^* là vector n chiều $(0, 0, \ldots, 0, 1)$, và δA là một nhiễu của các phần tử của A bên dưới đường chéo trong cột này. Một nhiễu như vậy làm nghiêng $range(A)$ bằng góc $\delta\alpha$ được cho bởi $\tan(\delta\alpha) = \|\delta p\|/\sigma_n$. Vì $\|\delta p\| = \|\delta A\|$ và $\delta\alpha \leqslant \tan(\delta\alpha)$ nên ta có

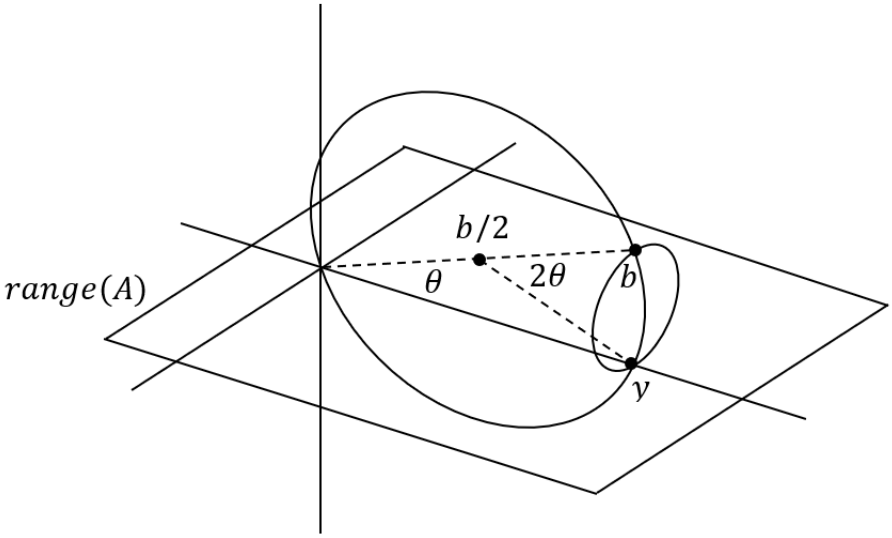
$$\delta\alpha \leqslant \frac{\|\delta A\|}{\sigma_n} = \frac{\|\delta A\|}{\|A\|}\kappa(A), \tag{3.6.10}$$

với các lựa chọn δA là nhỏ vô cùng (để $\delta\alpha = \tan(\delta\alpha)$).

3.6.7 Độ nhạy của y tới các nhiễu trong A

Bây giờ ta được chuẩn bị để suy ra dòng thứ hai của bảng trong Định lý 3.6.1. Ta bắt đầu với phần tử bên trái của nó. Vì y là phép chiếu trực giao của b lên trên $range(A)$ nên nó được xác định bởi b và chỉ có $range(A)$. Do đó, để phân tích độ nhạy của y tới

các nhiễu trong A, ta có thể nghiên cứu một cách đơn giản hiệu quả trong y của độ dốc $range(A)$ bằng góc $\delta\alpha$ nào đó.



Hình 3.3: Hai đường tròn trong hình cầu dọc theo mà y di chuyển khi $range(A)$ thay đổi. Đường tròn lớn, bán kính $\|b\|/2$, tương ứng với độ dốc $range(A)$ trong mặt phẳng $0-b-y$, và đường tròn nhỏ, bán kính $(\|b\|/2)\sin\theta$, tương ứng với độ dốc của nó trong một phương trực giao. Tuy nhiên $range(A)$ bị dốc, y còn lại trong hình cầu bán kính $\|b\|/2$ có tâm tại $b/2$

Một tính chất hình học trở nên rõ ràng khi ta hình dung việc cố định b và xem y biến đổi như $range(A)$ bị nghiêng (Hình 3.3). Cho dù $range(A)$ bị nghiêng như thế nào, vector $y \in range(A)$ thường phải là trực giao với $y-b$. Đó là, đường $y-b$ phải nằm tại góc bên phải với đường $0-y$. Mặt khác, khi $range(A)$ được điều chỉnh, y di chuyển dọc theo quả cầu bán kính $\|b\|/2$ tâm tại điểm $b/2$.

Độ dốc $range(A)$ trong mặt phẳng $0-b-y$ bởi góc $\delta\alpha$ thay đổi thành góc 2θ tại tâm $b/2$ bằng $2\delta\alpha$. Do đó, nhiễu tương ứng δy là cơ sở của một tam giác cân với góc ở giữa $2\delta\alpha$ và độ dài cạnh $\|b\|/2$. Điều này kéo theo $\|\delta y\| = \|b\|\sin(\delta\alpha)$. Độ dốc $range(A)$ trong phương khác bất kỳ đưa ra kết quả trong hình học tương tự trong một mặt phẳng khác và các nhiễu nhỏ hơn bằng một thừa số nhỏ như $\sin\theta$. Do đó, cho các nhiễu bất kỳ bằng một góc $\delta\alpha$ ta có

$$\|\delta y\| \leqslant \|b\|\sin(\delta\alpha) \leqslant \|b\|\delta\alpha. \tag{3.6.11}$$

Do (3.6.4) và (3.6.10), điều này cho chúng ta $\|\delta y\| \leqslant \|\delta A\|\kappa(A)\|y\|/\|A\|\cos\theta$, đó là,

$$\frac{\|\delta y\|}{\|y\|} \Big/ \frac{\|\delta A\|}{\|A\|} \leqslant \frac{\kappa(A)}{\cos\theta}. \tag{3.6.12}$$

Điều này thiết lập kết quả bên trái phía dưới hơn của Định lý 3.6.1.

3.6.8 Độ nhạy của x tới các nhiễu trong A

Bây giờ ta sẵn sàng để phân tích quan hệ thú vị nhất của Định lý 3.6.1: độ nhạy của x tới các nhiễu trong A.

Một nhiễu δA làm nghiêng tự nhiên thành 2 phần: một phần δA_1 trong n dòng đầu tiên của A, và phần khác δA_2 trong $m-n$ dòng còn lại:

$$\delta A = \begin{bmatrix} \delta A_1 \\ \delta A_2 \end{bmatrix} = \begin{bmatrix} \delta A_1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \delta A_2 \end{bmatrix}$$

Đầu tiên, ta hãy xét hiệu quả của các nhiễu δA_1. Một nhiễu như vậy thay đổi ánh xạ của A trong range của nó nhưng không là $range(A)$ vào chính nó hoặc y. Nó làm nhiễu A_1 bằng δA_1 trong hệ thống vuông (3.6.8) không thay đổi b_1. Số điều kiện cho các nhiễu như vậy được cho bởi (3.1.18), có dạng

$$\frac{\|\delta x\|}{\|x\|} \Big/ \frac{\|\delta A_1\|}{\|A\|} \leqslant \kappa(A_1) = \kappa(A). \tag{3.6.13}$$

Tiếp theo ta xét hiệu quả của các nhiễu δA_2 (nhỏ vô cùng). Một nhiễu như vậy làm dốc $range(A)$ không thay đổi việc ánh xạ của A trong không gian này. Điểm y và do đó vector b_1 được làm nhiễu, nhưng không là A_1. Điều này tương ứng với nhiễu b_1 trong (3.6.8) không thay đổi A_1. Số điều kiện cho các nhiễu như vậy được cho bởi (3.1.14), có dạng

$$\frac{\|\delta x\|}{\|x\|} \Big/ \frac{\|\delta b_1\|}{\|b_1\|} \leqslant \frac{\kappa(A_1)}{\eta(A_1; x)} = \frac{\kappa(A)}{\eta}. \tag{3.6.14}$$

Để hoàn thành đối số ta cần liên kết δb_1 với δA_2. Bây giờ vector b_1 là y được biểu diễn trong các tọa độ của $range(A)$. Do đó, chỉ các thay đổi trong y được thực hiện như các thay đổi trong b_1 là nằm song song với $range(A)$; các thay đổi trực giao không có hiệu quả. Đặc biệt, nếu $range(A)$ bị làm dốc bởi môt góc $\delta\alpha$ trong mặt phẳng $0-b-y$, nhiễu kết quả δy không nằm song song với $range(A)$ nhưng tại một góc $\pi/2-\theta$. Do đó, thay đổi trong b_1 thỏa mãn $\|\delta b_1\| = \sin\theta\|\delta y\|$. Theo (3.6.11), do đó ta có

$$\|\delta b_1\| \leqslant (\|b\|\delta\alpha)\sin\theta. \tag{3.6.15}$$

Nếu $range(A)$ được làm dốc trong một phương trực giao với mặt phẳng $0-b-y$, ta thu được chặn giống nhau, nhưng cho một lý do khác. Bây giờ δy là song song với $range(A)$ nhưng nó là một thừa số của $\sin\theta$ nhỏ hơn, như được miêu tả ở trên trong sự kết nối với Hình 3.3. Do đó, ta có $\|\delta y\| \leqslant (\|b\|\delta\alpha)\sin\theta$, và vì $\|\delta b_1\| \leqslant \|\delta y\|$ nên ta được (3.6.15). Vì $\|b_1\| = \|b\|\cos\theta$ nên ta có thể viết lại (3.6.15) như

$$\frac{\|\delta b_1\|}{\|b_1\|} \leqslant (\delta\alpha)\tan\theta. \tag{3.6.16}$$

Liên kết $\delta\alpha$ với $\|\delta A_2\|$ theo (3.6.10) và kết hợp (3.6.14) với (3.6.16), ta được

$$\frac{\|\delta x\|}{\|x\|} \Big/ \frac{\|\delta A_2\|}{\|A\|} \leqslant \frac{\kappa(A_1)^2\tan\theta}{\eta}.$$

Thêm điều này vào (3.6.13) thiết lặp kết quả bên phải ở phía dưới hơn của Định lý 3.6.1.

Bài tập

1. Giả sử A là ma trận 202×202 với $\|A\|_2 = 100$ và $\|A\|_F = 101$. Cho chặn thấp hơn có thể được rõ ràng nhất trong số điều kiện chuẩn 2 κ.

2. Xét đa thức $p(x) = (x-2)^9 = x^9 - 18x^8 + 144x^7 - 672x^6 + 2016x^5 - 4032x^4 = 5376x^3 - 4608x^2 + 2304x - 512$.

   (a) Vẽ $p(x)$ với $x = 1.920, 1.921, 1922, \ldots, 2.080$ và đánh giá p thông qua các hệ số 1, -18, 44, ...

   (b) Vẽ tương tự và đánh giá p thông qua $(x-2)^9$.

3. (a) Chứng minh rằng $(1 + O(\epsilon_{machine}))(1 + O(\epsilon_{machine})) = 1 + O(\epsilon_{machine})$. Ý nghĩa chính xác của phát biểu này là nếu f là một hàm thỏa $f(\epsilon_{machine}) = (1 + O(\epsilon_{machine}))(1 + O(\epsilon_{machine}))$ khi $\epsilon_{machine} \to 0$ thì f cũng thỏa $f(\epsilon_{machine}) = (1 + O(\epsilon_{machine}))$ khi $\epsilon_{machine} \to 0$.

   (b) Chứng minh rằng $(1 + O(\epsilon_{machine}))^{-1} = (1 + O(\epsilon_{machine}))$.

4. Một hệ thống tam giác (3.5.1) được giải bằng phép thế ngược. Định lý 3.5.1 suy ra sai số $\|\tilde{x} - x\|$?

5. Xét một thuật toán cho bài toán tính SVD (đầy đủ) của một ma trận. Dữ liệu cho bài toán này là ma trận A và lời giải là ba ma trận U (Unita), Σ (đường chéo), và V (Unita) sao cho $A = U\Sigma V^*$.

   (a) Giải thích thuật toán này là ổn định ngược.

   (b) Cho một lý do đơn gian mà thuật toán này là không ổn định ngược. Giải thích.

   (c) Các thuật toán tiêu chuẩn cho việc tính toán SVD là ổn định. Giải thích tính ổn định cho 1 thuật toán như vậy.

6. (a) Cho các ma trận Unita $Q_1, \ldots, Q_k \in \mathbb{C}^{m,m}$ được cố định và xét bài toán tính tích $B = Q_k \ldots Q_1 A$, với $A \in \mathbb{C}^{m \times n}$. Tính toán được thực hiện từ trái qua phải bằng các phép toán dấu chấm động trong máy tính thỏa (3.2.5) và (3.2.7). Chứng minh rằng thuật toán là ổn định ngược (A được làm nhiễu, Q_j cố định và không được làm nhiễu).

   (b) Cho một ví dụ để chứng minh kết quả này là không còn đúng nếu các ma trận Unita Q_j được thay bằng các ma trận tùy ý $X_j \in \mathbb{C}^{m \times m}$.

7. Xét

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \\ 1 & 1.0001 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 0.0001 \\ 4.0001 \end{bmatrix}$$

(a) Tính A^+, P.

(b) Tìm các lời giải chính xác của x và $y = Ax$ cho bài toán bình phương nhỏ nhất $Ax \approx b$.

(c) Tính $\kappa(A), \theta$ và η.

(d) Bốn số điều kiện của Định lý 3.6.1.

(e) Cho các ví dụ của các nhiễu δb và δA đạt được một cách xấp xỉ bốn số điều kiện này.

8. Giải thích vì sao số điều kiện của y tương ứng với các nhiễu trong A trở thành 0 trong trường hợp $m = n$.

9. Cho $A \in \mathbb{C}^{m\times n}$ hạng n và $b \in \mathbb{C}^m$, xét hệ thống khối 2×2 của các phương trình

$$\begin{bmatrix} I & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

với I là ma trận đơn vị $m \times m$. Chứng minh hệ thống này có một nghiệm duy nhất $(r, x)^T$ và r, x là thặng dư và lời giải của bài toán bình phương nhỏ nhất (3.6.1).

10. Cho đoạn code trong Matlab như sau

```
[U, V, S] = svd(A);
S = diag(A);
tol = max(size(A))*S(1)*eps;
r = sum(S > tol);
S = diag(ones(r, 1)./S(1:r));
X = V(:, 1:r)*S*U(:, 1:r)';
```

Đoạn code trên trả về kết quả gì?

Chương 4

Hệ phương trình

4.1 Khử Gauss

4.1.1 Phân tích LU

Khử Gauss chuyển một hệ thống tuyến tính đầy đủ thành một hệ thống tam giác trên bằng việc áp dụng các phép biến đổi tuyến tính đơn giản trong vế trái. Nó tương tự tam giác hóa Householder cho việc tính các phân tích QR. Sự khác nhau là các phép biến đổi được áp dụng trong khử Gauss mà không là unita.

Cho $A \in \mathbb{C}^{m\times m}$ là một ma trận vuông. (Thuật toán cũng có thể được áp dụng cho các ma trận hình chữ nhật, nhưng điều này ít được làm trong thực hành) Ý tưởng là để biến đổi A thành một ma trận tam giác trên U có $m \times m$ chiều bằng việc đưa ra các số 0 bên dưới đường chéo, đầu tiên trong cột 1, trong cột 2, ...- như trong tam giác hóa Householder. Điều này được làm bằng việc trừ các bội của mỗi dòng từ các dòng con theo sau. Quá trình "khử" này là tương đương với việc nhân A cho một chuỗi các ma trận tam giác dưới L_k trong vế trái:

$$\underbrace{L_{m-1}\dots L_2L_1}_{L^{-1}}A = U. \tag{4.1.1}$$

Đặt $L = L_1^{-1}L_2^{-1}\dots L_{m-1}^{-1}$ được $A = LU$. Do đó ta thu được một *phân tích* LU của A

$$A = LU, \tag{4.1.2}$$

với U là ma trận tam giác trên và L là ma trận tam giác dưới. Nó đưa ra L là *tam giác dưới đơn vị*, nghĩa là tất cả các phần tử trên đường chéo của nó là bằng 1.

Ví dụ, giả sử ta bắt đầu với một ma trận 4×4. Thuật toán tiến hành trong 3 bước (so

với (2.4.1))

$$\underset{A}{\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{bmatrix}} \xrightarrow{L_1} \underset{L_1A}{\begin{bmatrix} \times & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \end{bmatrix}} \xrightarrow{L_2} \underset{L_2L_1A}{\begin{bmatrix} \times & \times & \times & \times \\ & \times & \times & \times \\ & \mathbf{0} & \times & \times \\ & \mathbf{0} & \times & \times \end{bmatrix}} \xrightarrow{L_3} \underset{L_3L_2L_1A}{\begin{bmatrix} \times & \times & \times & \times \\ & \times & \times & \times \\ & & \times & \times \\ & & \mathbf{0} & \boldsymbol{x} \end{bmatrix}}.$$

Phép biến đổi thứ k L_k đưa các số 0 bên dưới đường chéo trong cột k bằng việc trừ các bội của dòng k từ các dòng $k+1, \ldots, m$. Vì $k-1$ phần tử đầu tiên của dòng k là 0, phép toán này không phá hủy các số 0 bất kì được đưa ra trước đó.
Do đó khử Gauss làm tăng thêm sự phân loại các thuật toán của chúng ta cho việc phân tích một ma trận:

$$\begin{aligned} \text{Gram - Schmidt: } A &= QR \text{ bằng trực giao hóa tam giác,} \\ \text{Householder: } A &= QR \text{ bằng tam giác hóa trực giao,} \\ \text{Khử Gauss: } A &= LU \text{ bằng tam giác hóa tam giác.} \end{aligned}$$

4.1.2 Ví dụ

Giả sử ta bắt đầu với ma trận 4×4

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}. \tag{4.1.3}$$

Bước đầu tiên của khử Gauss là

$$L_1A = \begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{bmatrix}.$$

Ta đã trừ 2 lần dòng thứ nhất từ dòng thứ 2, 4 lần dòng thứ nhất từ dòng thứ 3, và 3 lần dòng thứ nhất từ dòng thứ 4. Bước thứ 2 giống như điều này:

$$L_2L_1A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & -3 & 1 & \\ & -4 & & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & 3 & 5 & 5 \\ & 4 & 6 & 8 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{bmatrix}.$$

Lần này ta đã trừ 3 lần dòng 2 từ dòng 3 và 4 lần dòng 2 từ dòng 4. Cuối cùng, trong bước thứ 3 ta trừ dòng 3 từ dòng 4:

$$L_3L_2L_1A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & 2 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{bmatrix} = U.$$

Bây giờ, để đưa ra phân tích đầy đủ $A = LU$, ta cần tính tích $L = L_1^{-1}L_2^{-1}L_3^{-1}$. Nghịch đảo của L_1 phải là L_1, nhưng với mỗi phần tử bên dưới đường chéo được lấy phủ định:

$$\begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ -4 & & 1 & \\ -3 & & & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & & 1 & \\ 3 & & & 1 \end{bmatrix}. \tag{4.1.4}$$

Tương tự, các nghịch đảo của L_2 và L_3 được thu được bằng việc lấy phủ định các phần tử dưới đường chéo. Cuối cùng, tích $L_1^{-1}L_2^{-1}L_3^{-1}$ cũng là ma trận tam giác dưới đơn vị với các phần tử dưới đường chéo khác 0 của L_1^{-1}, L_2^{-1}, và L_3^{-1} đã đưa vào những chỗ xấp xỉ. Ta có

$$\underset{A}{\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}} = \underset{L}{\begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 4 & 3 & 1 & \\ 3 & 4 & 1 & 1 \end{bmatrix}} \underset{U}{\begin{bmatrix} 2 & 1 & 1 & 0 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 2 \end{bmatrix}} \tag{4.1.5}$$

4.1.3 Công thức tổng quát

Giả sử x_k ký hiệu là cột thứ k của ma trận bắt đầu bước k. Khi đó phép biến đổi L_k phải được chọn sao cho

$$x_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{kk} \\ x_{k+1,k} \\ \vdots \\ x_{mk} \end{bmatrix} \xrightarrow{L_k} L_k x_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Để làm điều này ta mong muốn trừ l_{jk} lần dòng k từ dòng j, với l_{jk} là *số nhân*

$$l_{jk} = \frac{x_{jk}}{x_{kk}} \ (k < j \leqslant m). \tag{4.1.6}$$

Ma trận L_k có dạng

$$L_k = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{mk} & & & 1 \end{bmatrix},$$

với các phần tử dưới đường chéo khác 0 được thay thế trong cột k. Điều này tương tự (2.4.2) cho tam giác hóa Householder.

Trong ví dụ ở trên, L_k đó có thể được đảo ngược bằng việc lấy phủ định các phần tử dưới đường chéo của nó (4.1.4), và L đó có thể được tạo thành bằng việc tập hợp các phần tử l_{jk} trong các nơi xấp xỉ (4.1.5). Ta hãy xác định

$$l_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \vdots \\ l_{m,k} \end{bmatrix}.$$

Khi đó L_k có thể được viết $L_k = I - l_k e_k^*$, với e_k là vector cột với 1 nằm ở vị trí k và 0 nằm ở những vị trí khác. Kiểu thưa thớt của l_k kéo theo $e_k^* l_k = 0$, và do đó $(I - l_k e_k^*)(I + l_k e_k^*) = I - l_k e_k^* l_k e_k^* = I$. Mặt khác, nghịch đảo của L_k là $I + l_k e_k^*$ như trong (4.1.4).

Xét ví dụ, tích $L_k^{-1} L_{k+1}^{-1}$. Từ kiểu thưa thớt của l_{k+1}, ta có $e_k^* l_{k+1} = 0$, và do đó

$$L_k^{-1} L_{k+1}^{-1} = (I + l_k e_k^*)(I + l_{k+1} e_{k+1}^*) = I + l_k e_k^* + l_{k+1} e_{k+1}^*.$$

Do đó $L_k^{-1} L_{k+1}^{-1}$ cũng là ma trân tam giác dưới đơn vị với các phần tử của cả L_k^{-1} và L_{k+1}^{-1} được thêm vào bên dưới đường chéo. Khi ta lấy tích của tất cả các ma trận này để tạo thành dạng L

$$L = L_1^{-1} L_2^{-1} \dots L_{m-1}^{-1} = \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ l_{m1} & l_{m2} & \cdots & l_{m,m-1} & 1 \end{bmatrix}. \tag{4.1.7}$$

Mặc dù ta không đề cập nó trong mục 2.3, xét sự thưa thớt mà chúng dẫn đến (4.1.7) cũng xuất hiện trong giải thích (2.3.10) của Gram - Schmidt được sửa đổi xử lý các phép nhân phải liên tiếp cho các ma trận tam giác R_k.

Khử Gauss trong thực hành, các ma trận L_k không bao giờ được tạo thành và được nhân rõ ràng. Các số nhân l_{jk} được tính và lưu trực tiếp vào L, và khi đó các phép biến đổi L_k được áp dụng:

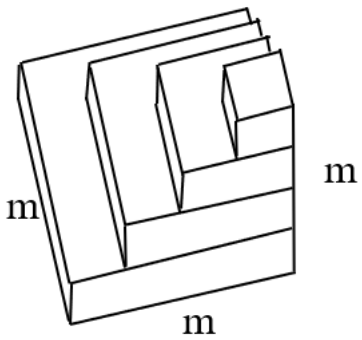
Thuật toán 4.1 Khử Gauss không quay

```
U = A, L = I
for k = 1 to m − 1 do
    for j = k + 1 to m do
        l_jk = u_jk / u_kk
        u_{j,k:m} = u_{j,k:m} − l_jk u_{k,k:m}
    end for
end for
```

4.1.4 Đếm số phép toán

Đếm phép toán tiệm cận của thuật toán này có thể được suy ra từ hình học. Việc làm được chi phối bởi phép toán vector trong vòng lặp bên trong, $u_{j,k:m} = u_{j,k:m} - l_{jk}u_{k,k:m}$, mà nó thực thi một phép nhân vector với vô hướng và một phép trừ vector. Nếu $l = m-k+1$ ký hiệu là chiều dài của các vector dòng đang được thao tác, số phép toán dấu chấm động là $2l$: 2 phép toán dấu chấm động trên phần tử.

Với mỗi giá trị của k, vòng lặp bên trong được lặp lại cho các dòng $k+1, \ldots, m$. Việc làm đã bao gồm sự tương ứng tới một lớp của khối theo sau:



Đây là hình giống với hình mà ta đã đưa ra trong mục 2.3 để biểu diễn tam giác hóa Householder (giả sử $m = n$). Tuy nhiên, mỗi hình lập phương đơn vị biểu diễn 4 phép toán dấu chấm động hơn là 2. Như trước đây, khối hội tụ tới một hình chớp khi $m \to \infty$, với thể tích $\frac{1}{3}m^3$. Tại 2 phép toán dấu chấm động trên một đơn vị thể tích, điều này tăng thêm thành

$$\text{Khử Gauss: } \approx \frac{2}{3}m^3 \text{ phép toán dấu chấm động.} \tag{4.1.8}$$

4.1.5 Giải phương trình Ax =b bằng phân tích LU

Nếu A được phân tích thành L và U thì một hệ thống các phương trình $Ax = b$ được giảm xuống thành dạng $LUx = b$. Khi đó nó có thể được giải bằng việc giải 2 hệ thống tam giác: đầu tiên là $Ly = b$ với biến y (phép thế ngược), khi đó $Ux = y$ với biến x (phép thế ngược). Bước đầu tiên cần $\sim \frac{2}{3}m^3$ phép toán dấu chấm động. Bước 2 và bước 3 thì mỗi bước cần $\sim m^2$ phép toán dấu chấm động. Tổng số là $\sim \frac{2}{3}m^3$ phép toán dấu chấm động, một phần hai của hình là $\sim \frac{4}{3}m^3$ phép toán dấu chấm động (2.4.10) cho một lời giải bằng tam giác hóa Householder (Thuật toán 3.1).

Ví dụ 4.1.1. Giải hệ phương trình $Ax = b$, với

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 2 & -1 \\ 4 & -1 & 6 \end{bmatrix} \text{ và } b = \begin{bmatrix} 3 \\ 0 \\ 11 \end{bmatrix}$$

Ta tính L và U sao cho $A = LU$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & -3 & 1 \end{bmatrix} \text{ và } U = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & -2 \end{bmatrix}$$

Để giải hệ phương trình $Ax = b$, đầu tiên ta giải phương trình $Ly = b$ bằng phép thế tiến và thu được $y = [3, -3, 4]^T$. Tiếp theo, ta giải phương trình $Ux = y$ bằng phép thế lùi và thu được $x = [0, 1, 2]^T$.

4.1.6 Tính không ổn định của khử Gauss không quay

Khử Gauss như được đưa ra không hữu dụng cho việc giải các hệ thống tuyến tính tổng quát vì nó không ổn định ngược. Tính không ổn định có liên hệ với một cái khác, khó khăn rõ ràng hơn. Với các ma trận nào đó, khử Gauss thất bại hoàn toàn bởi vì nó xâm phạm đến việc chia cho 0.

Ví dụ, xét

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Ma trận này có hạng đầy đủ và là điều kiện tốt, với $\kappa(A) = (3 + \sqrt{5})/2 \approx 2.618$ trong chuẩn 2. Tuy nhiên, khử Gauss thất bại tại bước đầu tiên.

Giả sử ta áp dụng khử Gauss cho

$$A = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix}. \tag{4.1.9}$$

Quá trình không thất bại. Thay vì, 10^{20} lần dòng đầu tiên được trừ từ dòng thứ 2, và các thừa số theo sau được đưa ra:

$$L = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix}, \; U = \begin{bmatrix} 10^{-20} & 1 \\ 0 & 1 - 10^{20} \end{bmatrix}.$$

Tuy nhiên, giả sử các tính toán này được thực hiện trong số học dấu chấm động với $\epsilon_{machine} \approx 10^{-16}$. Số $1 - 10^{20}$ sẽ không được biểu diễn chính xác. Nó sẽ được làm tròn thành số dấu chấm động gần nhất. Cho đơn giản, nó chính xác là -10^{20}. Khi đó, các ma trận dấu chấm động được đưa ra bởi thuật toán sẽ là

$$\tilde{L} = \begin{bmatrix} 1 & 0 \\ 10^{20} & 1 \end{bmatrix}, \; \tilde{U} = \begin{bmatrix} 100^{-20} & 1 \\ 0 & -10^{20} \end{bmatrix}.$$

Bậc của việc làm tròn này có thể cho phép được đầu tiên. Sau tất cả, ma trận $\tilde{U}$ là gần với U chính xác liên quan với $\|U\|$. Tuy nhiên, bài toán trở nên rõ ràng khi ta tính tích $\tilde{L}\tilde{U}$:

$$\tilde{L}\tilde{U} = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 0 \end{bmatrix}.$$

Ma trận này không gần với A vì số 1 nằm ở vị trí $(2, 2)$ đã được thay thế bằng 0. Nếu ta giải hệ thống $\tilde{L}\tilde{U}x = b$ thì kết quả sẽ khác lời giải $Ax = b$. Ví dụ, với $b = (1, 0)^*$ ta được $\tilde{x} = (0, 1)^*$, trong khi lời giải chính xác là $x \approx (-1, 1)^*$.

Khử Gauss đã tính toán phân tích LU ổn định: $\tilde{L}$ và $\tilde{U}$ là gần với các thừa số chính xác cho một ma trận gần với A. Nó đã không giải $Ax = b$ ổn định vì phân tích LU là *không ổn định ngược*. Như một quy tắc, nếu một bước của thuật toán là ổn định nhưng không phải là thuật toán ổn định ngược cho việc giải một bài toán con thì tính ổn định trên tất cả phép tính có thể là nguy cơ.

Thật vậy, cho các ma trận A có $m \times m$ chiều tổng quát, một tình huống tệ hơn điều này, khử Gauss không quay là không ổn định hoặc không ổn định ngược như một thuật toán tổng quát cho phân tích LU. Hơn nữa, các ma trận tam giác được sinh ra có các số điều kiện mà chúng có thể là tùy ý hơn là các số điều kiện này của chính ma trận A, việc đưa thêm vào các nguồn không ổn định trong các giai đoạn phép thế tiến và ngược của lời giải $Ax = b$.

4.2 Phép toán quay

Trong mục trước ta thấy rằng khử Gauss trong dạng thuần túy của nó là không ổn định. Tính không ổn định có thể được điều khiển bằng việc làm nhiễu các dòng của ma trận đang được tính trong nó, một phép toán gọi là *quay (pivoting)*. Pivoting đã là một đặc trưng tiêu chuẩn của các tính toán khử Gauss từ những năm 1950.

4.2.1 Các pivot

Tại bước k của khử Gauss, các bội của dòng k được trừ từ các dòng $k+1, \ldots, m$ của ma trận X để đưa các số 0 trong k phần tử của các dòng này. Trong phép toán dòng k, cột k, và đặc biệt phần tử x_{kk} đóng vai trò đặc biệt. Ta gọi x_{kk} là *pivot*. Từ mỗi phần tử trong ma trận con $X_{k+1:m,k:m}$ được trừ tích của một số trong dòng k và một số trong cột k, chia cho x_{kk}:

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ & x_{kk} & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \times & \times & \times & \times \\ & x_{kk} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \end{bmatrix}$$

Tuy nhiên, không có lý do vì sao dòng và cột thứ k phải được chọn cho sự khử. Ví dụ, ta có thể dễ dàng đưa các số 0 vào trong cột k bằng việc cộng thêm các bội của dòng i nào đó với $k < i \leqslant m$ với các dòng khác $k, \ldots, m$. Trong trường hợp này, x_{ik} sẽ là pivot. Dưới đây là một minh họa với $k = 2$ và $i = 4$:

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & x_{kk} & \times & \times & \times \\ & \times & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \times & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \\ & x_{kk} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \end{bmatrix}.$$

Tương tự, ta có thể đưa các số 0 vào trong cột j hơn là cột k. Dưới đây là một minh họa với $k = 2, i = 4, j = 3$:

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & x_{ij} & \times & \times \\ & \times & \times & \times & \times \end{bmatrix} \longrightarrow \begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \mathbf{0} & \times & \times \\ & \times & \mathbf{0} & \times & \times \\ & \times & x_{ij} & \times & \times \\ & \times & \mathbf{0} & \times & \times \end{bmatrix}.$$

Nói chung, ta tự do chọn phần tử bất kì của $X_{k:m,k:m}$ làm pivot, chỉ cần nó khác 0. Khả năng mà một phần tử $x_{kk} = 0$ có thể xuất hiện kéo theo một vài tính linh hoạt của việc chọn pivot đôi khi có thể là cần thiết, ngay cả quan điểm toán học thuần túy. Cho tính ổn định số, nó mong muốn để pivot ngay khi x_{kk} khác 0 nếu có một phần tử lớn hơn có thể. Đặc biệt, nó là phổ biến để lựa chọn như là pivot số lớn nhất giữa một tập các phần tử đang được xét như các ứng viên.

Cấu trúc của quá trình khử nhanh chóng trở nên khó hiểu nếu các số 0 được đưa vào trong các loại tùy ý thông qua ma trận. Để thấy điều này, ta muốn giữ lại cấu trúc tam

giác được miêu tả trong mục cuối cùng chương này, và có một cách dễ dàng để làm điều này. Ta sẽ không nghĩ về pivot x_{ij} như ở vế bên trái trong minh họa ở trên. Thay vì, tại bước k, ta sẽ tưởng tượng các dòng và cột của ma trận đang làm việc được hoán vị để mà di chuyển x_{ij} vào vị trí (k, k). Khi đó sự khử được làm, các số 0 được đưa vào thành các phần tử $k+1, \ldots, m$ của cột k, ngay cả khi trong khử Gauss không quay. Sự đổi chổ cho nhau các dòng và các cột có thể là cái thường được nghĩ như *pivoting*.

Ý tưởng mà các dòng và các cột được đổi chổ cho nhau là một khái niệm bắt buộc. Nếu như nó là một ý tưởng tốt để hoán đổi vị trí chúng trong máy tính là ít rõ ràng. Trong các thực thi bất kỳ, dữ liệu trong bộ nhớ máy tính là được hoán đổi tại mỗi bước pivot. Mặt khác, hiệu quả tương đương được đạt được bởi địa chỉ không trực tiếp với các vector chỉ số được hoán vị. Xấp xỉ là thay đổi tốt nhất từ máy tới máy và phụ thuộc vào nhiều nhân tố.

4.2.2 Quay từng phần

Nếu mọi phần tử của $X_{k:m,k:m}$ được xem xét như là một pivot có thể tại bước k, thì có $O((m-k)^2)$ phần tử được kiểm tra để xác định là lớn nhất. Tính tổng trên m bước, tổng chi phí của việc chọn các pivot là $O(m^3)$ phép toán, cộng thêm chi phí khử Gauss đáng kể, không đề cập tới các khó khăn tiềm năng của sự truyền dữ liệu trong một dãy không thể dự đoán qua tất cả các phần tử của một ma trận . Chiến lược tốn kém này được gọi là *quay đầy đủ*.

Đặc biệt, các pivot tốt tương đương nhau có thể được tìm thấy bằng việc xét một số lớn hơn nhiều số phần tử. Phương pháp tiêu chuẩn cho việc làm này là *quay từng phần*. Ở đây, chỉ các dòng là được hoán đổi vị trí cho nhau. Pivot tại mỗi bước được chọn như là phần tử lớn nhất của $m-k+1$ phần tử trên đường chéo phụ của cột k mà nó có tổng số chi phí là $O(m-k)$ phép toán cho việc chọn pivot tại mỗi bước. Do đó $O(m^2)$ phép toán tất cả. Để đưa pivot thứ k vào vị trí (k, k), không cột nào cần được hoán đổi vị trí; nó đủ để hoán vị dòng k với dòng chứa pivot.

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & x_{ik} & \times & \times & \times \\ & \times & \times & \times & \times \end{bmatrix} \xrightarrow{P_1} \begin{bmatrix} \times & \times & \times & \times & \times \\ & x_{ik} & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & \times & \times & \times & \times \end{bmatrix} \xrightarrow{L_1} \begin{bmatrix} \times & \times & \times & \times & \times \\ & x_{ik} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \\ & \mathbf{0} & \times & \times & \times \end{bmatrix}.$$

Chọn pivot Đổi dòng Khử

Thuật toán này có thể được biểu diễn như một tích ma trận. Ta thấy trong mục cuối cùng mà bước khử tương ứng với phép nhân trái bởi một ma trận tam giác dưới cơ bản L_k. Quay từng phần làm phức tạp các chủ đề bằng việc áp dụng một ma trận hoán vị

P_k trong vế trái của ma trận làm việc trước mỗi sự khử. (Một ma trận hoán vị là một ma trận với 0 hầu khắp nơi trừ một số 1 ở mỗi dòng và cột. Nghĩa là, nó là một ma trận thu được từ ma trận đơn vị bằng việc làm nhiễu các dòng hoặc các cột.) Sau $m-1$ bước, A trở thành một ma trận tam giác dưới U:

$$L_{m-1}P_{m-1}\dots L_2P_2L_1P_1A = U. \tag{4.2.1}$$

4.2.3 Ví dụ

Ta sẽ sử dụng lại ví dụ 4.1.3,

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}. \tag{4.2.2}$$

Với quay từng phần, việc đầu tiên ta làm là đổi chổ dòng đầu tiên và dòng thứ 3 (nhân trái với P_1):

$$\begin{bmatrix} & & 1 & \\ & 1 & & \\ 1 & & & \\ & & & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{bmatrix}.$$

Bước khử đầu tiên (nhân trái với L_1):

$$\begin{bmatrix} 1 & & & \\ -\frac{1}{2} & 1 & & \\ -\frac{1}{4} & & 1 & \\ -\frac{3}{4} & & & 1 \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \end{bmatrix}.$$

Dòng thứ 2 và dòng thứ 4 được đổi chổ (nhân với P_2):

$$\begin{bmatrix} 1 & & & \\ & & & 1 \\ & & 1 & \\ & 1 & & \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \end{bmatrix}.$$

Khi đó bước khử thứ 2 (nhân với L_1):

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & \frac{3}{7} & 1 & \\ & \frac{2}{7} & & 1 \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{2}{7} & \frac{4}{7} \\ & & -\frac{6}{7} & -\frac{2}{7} \end{bmatrix}.$$

Dòng thứ 3 và dòng thứ 4 được đổi chổ (nhân với P_3):

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & 1 \\ & & 1 & \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{2}{7} & \frac{4}{7} \\ & & -\frac{6}{7} & -\frac{2}{7} \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{6}{7} & -\frac{2}{7} \\ & & -\frac{2}{7} & \frac{4}{7} \end{bmatrix}.$$

Bước khử cuối cùng (nhân với L_3):

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & -\frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{6}{7} & -\frac{2}{7} \\ & & -\frac{2}{7} & \frac{4}{7} \end{bmatrix} = \begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{6}{7} & -\frac{2}{7} \\ & & & \frac{2}{3} \end{bmatrix}.$$

4.2.4 Phân tích PA= LU

Ta đã tính phân tích LU của PA, với P là một ma trận hoán vị.

$$\underset{P}{\begin{bmatrix} & & 1 & \\ & & & 1 \\ & 1 & & \\ 1 & & & \end{bmatrix}} \underset{A}{\begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}} = \underset{L}{\begin{bmatrix} 1 & & & \\ \frac{3}{4} & 1 & & \\ \frac{1}{2} & -\frac{2}{7} & 1 & \\ \frac{1}{4} & -\frac{3}{7} & \frac{1}{3} & 1 \end{bmatrix}} \underset{U}{\begin{bmatrix} 8 & 7 & 9 & 5 \\ & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ & & -\frac{6}{7} & -\frac{2}{7} \\ & & & \frac{2}{3} \end{bmatrix}}. \tag{4.2.3}$$

Công thức này nên được so sánh với (4.1.5). Sự có mặt của các số nguyên và các phân số ở đây là không phân biệt. Sự phân biệt ở đây là tất cả các phần tử đường chéo phụ của L có độ dài nhỏ hơn 1, là một hệ quả của tính chất $|x_{kk}| = \max_j |x_{jk}|$ trong (4.1.6) được đưa ra bởi pivoting.

Nó không rõ ràng từ (4.2.3). Quá trình khử đưa về dạng

$$L_3P_3L_2P_2L_1P_1A = U,$$

mà nó không giống tam giác dưới. Nhưng ở đây, sáu phép toán cơ bản này có thể được sắp xếp lại thành dạng

$$L_3P_3L_2P_2L_1P_1 = L_3^{'}L_2^{'}L_1^{'}P_3P_2P_1, \tag{4.2.4}$$

với $L_k^{'}$ bằng với L_k nhưng với các phần tử trên đường chéo phụ đổi chổ cho nhau. Để chính xác, định nghĩa

$$L_3^{'} = L_3,\ L_2^{'} = P_3L_2P_3^{-1},\ L_1^{'} = P_3P_2L_1P_2^{-1}P_3^{-1}.$$

Vì mỗi định nghĩa này chỉ áp dụng các hoán vị P_j với $j > k$ tới L_k nên dễ dàng kiểm tra $L_k^{'}$ có cấu trúc giống như L_k. Việc tính tích của các ma trận $L_k^{'}$ đưa ra

$$L_3^{'}L_2^{'}L_1^{'}P_3P_2P_1 = L_3(P_3L_2P_3^{-1})(P_3P_2L_1P_2^{-1}P_3^{-1})P_3P_2P_1 = L_3P_3L_2P_2L_1P_1$$

như trong (4.2.4).

Tổng quát, cho ma trận $m \times m$, phân tích (4.2.1) được cung cấp bởi khử Gauss với quay từng phần có thể được viết dưới dạng

$$(L_{m-1}^{'} \dots L_2^{'} L_1^{'})(P_{m-1} \dots P_2 P_1)A = U, \tag{4.2.5}$$

với $L_k^{'}$ được xác định bởi

$$L_k^{'} = P_{m-1} \dots P_{k+1} L_k P_{k+1}^{-1} \dots P_{m-1}^{-1}. \tag{4.2.6}$$

Tích của các ma trận $L_k^{'}$ là ma trận tam giác dưới đơn vị và dễ dàng lấy nghịch đảo bằng việc lấy phủ định các phần tử trên đường chéo phụ, như trong khử Gauss không có quay. Viết $L = (L_{m-1}^{'} \dots L_2^{'} L_1^{'})^{-1}$ và $P = P_{m-1} \dots P_2 P_1$, ta có

$$PA = LU. \tag{4.2.7}$$

Tổng quát, ma trận vuông bất kỳ A, suy biến hoặc không suy biến, có một phân tích (4.2.7), với P là một ma trận hoán vị, L là ma trận tam giác dưới đơn vị với các phần tử tam giác dưới nhỏ hơn bằng 1, và U là ma trận tam giác trên.

Công thức nổi tiếng (4.2.7) có một giải thích đơn giản. Khử Gauss với quay từng phần là tương đương với thủ tục theo sau:

1. Hoán vị các dòng của A theo P.
2. Áp dụng khử Gauss không quay cho PA.

Quay từng phần không được tiến hành trong thực hành vì P không được biết trước lúc đầu. Dưới đây là thuật toán.

Thuật toán này yêu cầu số phép toán dấu chấm động giống (4.1.8) như khử Gauss không quay, cụ thể là $\frac{2}{3}m^3$. Như với Thuật toán 4.1, sử dụng bộ nhớ máy tính có thể được cực tiểu hóa nếu được miêu tả bằng việc viết chồng lên U và L thành mảng tương tự lưu trữ A.

Thuật toán 4.2 Khử Gauss với quay từng phần

```
U = A, L = I, P = I
for k = 1 to m − 1 do
    Chọn i ⩾ k để cực đại hóa |u_ik|
    u_{k,1:k−1} ↔ u_{i,k:m} (đổi chổ 2 dòng với nhau)
    l_{k,1:k−1} ↔ l_{i,1:k−1}
    p_{k,:} ↔ p_{i,:}
    for j = k + 1 to m do
        l_jk = u_jk/u_kk
        u_{j,k:m} = u_{j,k:m} − l_jk u_{k,k:m}
    end for
end for
```

Trong thực hành, P không được biểu diễn một cách chính xác như là một ma trận. Các dòng được đổi chổ cho nhau tại mỗi bước, hoặc thông qua một vector hoán vị.

4.2.5 Quay đầy đủ

Trong quay đầy đủ, sự lựa chọn các pivot lấy một số lượng thời gian đáng kể. Trong thực hành, điều này hiếm khi được làm bởi vì sự thực thi trong tính ổn định là mép biên.

Trong dạng ma trận, quay đầy đủ đưa đến mỗi bước khử với một ma trận hoán vị P_k của các dòng được áp dụng trong vế trái và cũng là hoán vị Q_k của các cột được áp dụng trong vế phải:

$$L_{m-1}P_{m-1}\dots L_2P_2L_1P_1AQ_1Q_2\dots Q_{m-1} = U. \tag{4.2.8}$$

Nếu $L_k^{'}$ được xác định như trong (4.2.6) (các hoán vị cột không được bao gồm) thì

$$(L_{m-1}^{'}\dots L_2^{'}L_1^{'})(P_{m-1}\dots P_2P_1)A(Q_1Q_2\dots Q_{m-1}) = U. \tag{4.2.9}$$

Đặt $L = (L_{m-1}^{'}\dots L_2^{'}L_1^{'})^{-1}, P = P_{m-1}\dots P_2P_1$, và $Q = Q_1Q_2\dots Q_{m-1}$, ta được

$$PAQ = LU. \tag{4.2.10}$$

4.3 Tính ổn định của khử Gauss

4.3.1 Tính ổn định và kích thước của L và U

Phân tích tính ổn định của khử Gauss với quay từng phần là phức tạp và nó đã là một điểm khó trong giải tích số từ những năm 1950.

Trong (4.1.9), ta cho một ví dụ với ma trận 2×2 mà trong đó khử Gauss không quay là không ổn định. Trong ví dụ đó, thừa số L có một phận tử kích thước 20^{20}. Cố gắng để giải một hệ thống các phương trình dựa vào L đưa ra các sai số làm tròn tương đối bậc $\epsilon_{machine}$. Do đó sai số tuyệt đối bậc $\epsilon_{machine} \times 10^{20}$.

Tính không ổn định trong khử Gauss (quay hoặc không quay) chỉ có thể xuất hiện nếu một hoặc cả hai thừa số L và U là tương đối lớn gần với kích thước của A. Do đó, mục đích của quay là để chắc chắn rằng L và U là không quá lớn. Miễn là tất cả các số lượng trung gian xuất hiện trong suốt sự khử là kích thước dễ sử dụng, nghĩa là các sai số làm tròn mà chúng phát sinh là rất nhỏ, và thuật toán là ổn định ngược.

Định lý theo sau làm ý tưởng này rõ ràng.

Định lý 4.3.1 *Cho phân tích $A = LU$ của một ma trận không suy biến $A \in \mathbb{C}^{m\times m}$ được tính toán bởi khử Gauss không quay (Thuật toán 4.1) trong một máy tính thỏa các tiên đề (3.2.5) và (3.2.7). Nếu A có một phân tích LU thì với mọi $\epsilon_{machine}$ đủ nhỏ, phân tích hoàn thành một cách đầy đủ trong số học dấu chấm động (không có các pivot 0), và các ma trận tính được $\tilde{L}$ và $\tilde{U}$ thỏa mãn*

$$\tilde{L}\tilde{U} = A + \delta A, \ \frac{\|\delta A\|}{\|L\|\|U\|} = O(\epsilon_{machine}) \tag{4.3.1}$$

với $\delta A \in \mathbb{C}^{m\times n}$ *bất kì.*

Nếu $\|L\|\|U\| = O(\|A\|)$ thì (4.3.1) khẳng định rằng khử Gauss là ổn định ngược. Nếu $\|L\|\|U\| \neq O(\|A\|)$ thì ta phải mong đợi không ổn định ngược.

Cho khử Gauss không quay, cả L và U có thể là lớn không giới hạn. Thuật toán đó là không ổn định bởi chuẩn bất kì.

4.3.2 Các thừa số tăng

Xét khử Gauss với quay từng phần. Bởi vì mỗi sự lựa chọn pivot bao gồm tối đa hóa trên một cột nên thuật toán này đưa ra ma trận L với các phần tử bên dưới đường chéo có giá trị tuyệt đối nhỏ hơn hoặc bằng 1. Điều này kéo theo $\|L\| = O(1)$ trong chuẩn bất kì. Do đó, cho khử Gauss với quay từng phần, (4.3.1) giảm thành điều kiện $\frac{\|\delta A\|}{\|U\|} = O(\epsilon_{machine})$. Ta kết luận rằng thuật toán là ổn định ngược đưa ra $\|U\| = O(\|A\|)$.

Khử Gauss giảm một ma trận đầy đủ A thành một ma trận U tam giác trên. Đặc biệt, cho *thừa số tăng* cho A được xác định như là tỉ số

$$\rho = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|}. \tag{4.3.2}$$

Nếu ρ bậc 1 thì quá trình khử là ổn định. Nếu ρ lớn hơn bậc 1 thì ta phải mong đợi tính không ổn định. Đặc biệt, vì $\|L\| = O(1)$ và (4.3.2) kéo theo $\|U\| = O(\rho\|A\|)$ nên kết quả theo sau là một hệ quả của Định lý 4.3.1.

Định lý 4.3.2 *Cho phân tích* $PA = LU$ *của một ma trận* $A \in \mathbb{C}^{m\times m}$ *được tính bởi khử Gauss với quay từng phần (Thuật toán 4.2) trong một máy tính thỏa mãn các tiên đề (3.2.5) và (3.2.7). Khi đó, các ma trận tính được* $\tilde{P}, \tilde{L}$ *và* $\tilde{U}$ *thỏa mãn*

$$\tilde{L}\tilde{U} = \tilde{P}A + \delta A, \ \frac{\|\delta A\|}{\|A\|} = O(\rho\epsilon_{machine}) \tag{4.3.3}$$

với $\delta A \in \mathbb{C}^{m\times n}$ *bất kì,* ρ *là thừa số tăng của* A*. Nếu* $|l_{ij}| < 1$ *với* $i > j$ *(không phụ thuộc vào sự lựa chọn các pivot trong số học chính xác) thì* $\tilde{P} = P$ *với mọi* $\epsilon_{machine}$ *đủ nhỏ.*

Theo Định lý 4.3.2 và định nghĩa (3.3.5) của tính ổn định ngược, khử Gauss ổn định ngược nếu $\rho = O(1)$ với tất cả các ma trận có số chiều được cho là m, và trường hợp khác là không.

4.3.3 Tính không ổn định trong trường hợp xấu nhất

Cho các ma trận A nào đó, mặc dù các hiệu quả thuận lợi của quay, ρ đưa ra là lớn. Ví dụ, giả sử A là ma trận

$$A = \begin{bmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}. \tag{4.3.4}$$

Tại bước đầu tiên, không quay diễn ra, nhưng các phần tử $2, 3, \ldots, m$ trong cột cuối cùng được làm gấp đôi từ 1 tới 2. Sự xuất hiện làm gấp đôi khác tại mỗi bước khử sau đó xảy ra. Tại bước cuối ta có

$$U = \begin{bmatrix} 1 & & & & 1 \\ & 1 & & & 2 \\ & & 1 & & 4 \\ & & & 1 & 8 \\ & & & & 16 \end{bmatrix}. \tag{4.3.5}$$

Cuối cùng, phân tích $PA = LU$

$$\begin{bmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ -1 & -1 & 1 & & \\ -1 & -1 & -1 & 1 & \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & 1 \\ & 1 & & & 2 \\ & & 1 & & 4 \\ & & & 1 & 8 \\ & & & & 16 \end{bmatrix}. \tag{4.3.6}$$

Cho ma trận 5×5 này, thừa số tăng là $\rho = 16$. Cho ma trận $m \times m$ của dạng tương tự, $\rho = 2^{m-1}$.

Thừa số tăng bậc 2^m tương ứng với sự hao hụt trong bậc m bits của độ chính xác, mà nó là thê thảm cho một tính toán thực tế. Vì một máy tính điển hình biểu diễn số dấu chấm động chỉ với 64 bit nên sự hao hụt m bits của độ chính xác là không chấp nhận được cho các tính toán thực tế.

Điều này đưa chúng ta tới điểm khó khăn. Ở đây, trong thảo luận khử Gauss với quay các định nghĩa của tính ổn định đưa ra trong mục 3.3 thất bại.

Theo các định nghĩa, tất cả các vấn đề đó trong việc xác định tính ổn định hoặc tính ổn định ngược là tồn tại một chặn nào đó đều có thể dùng được tới tất cả các ma trận *cho mỗi chiều được cố định* m.

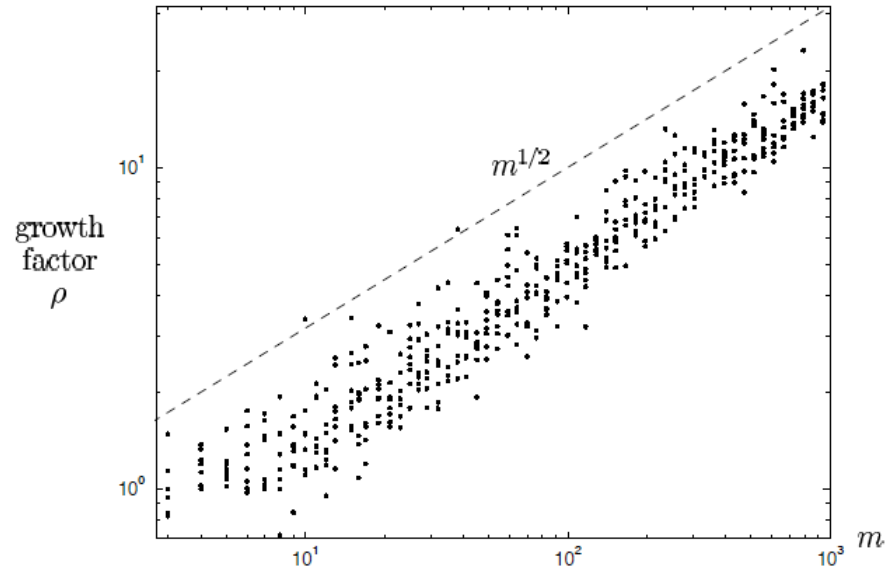
Ở đây, với mỗi m, ta có một chặn đều bao gồm hằng số 2^{m-1}. Do đó, theo các định nghĩa của chúng ta, khử Gauss là ổn định ngược.

Định lý 4.3.3 *Theo các định nghĩa trong mục 3.3, khử Gauss với quay từng phần là ổn định ngược.*

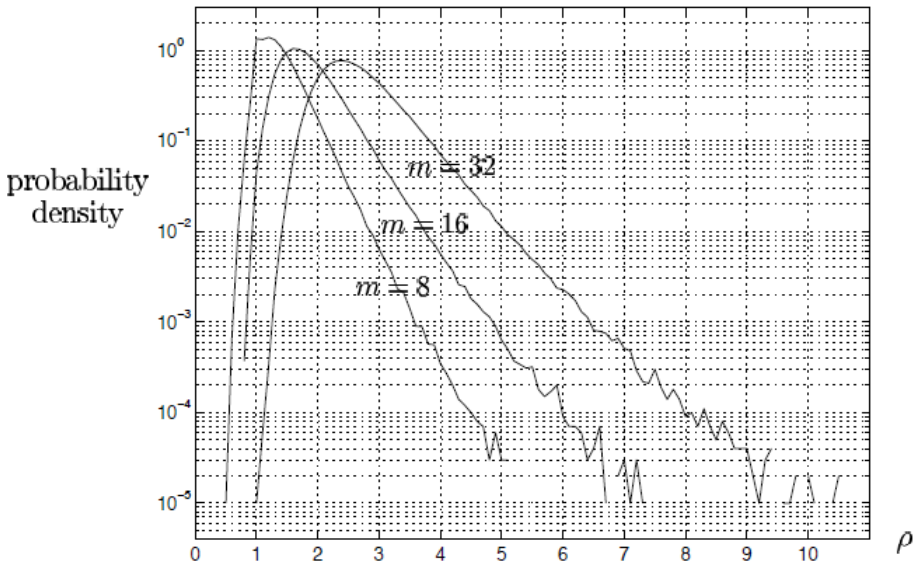
Khử Gauss cho các ma trận nào đó là không ổn định, khi ta có thể được thừa nhận bởi các thực thi số với Matlab, Linpack, Lapack, hoặc các gói phần mềm nổi tiếng hoàn hảo.

4.3.4 Tính ổn định trong thực hành

Mặc dù các ví dụ giống (4.3.4), khử Gauss với quay từng phần là hoàn toàn ổn định trong thực hành. Các thừa số lớn U giống (4.3.5) dường như không bao giờ xuất hiện trong các ứng dụng thực tế. Trong 50 năm tính toán, không bài toán ma trận nào kích thích tính không ổn định được biết đã xuất hiện dưới các trường hợp tự nhiên.
Ta có thể học nhiều hơn về hiện tượng này bằng việc xét các ma trận ngẫu nhiên. Chúng có tất cả các loại tính chất đặc biệt, và nếu ta cố gắng để miêu tả chúng như các ví dụ ngẫu nhiên từ phân phối bất kì nên nó sẽ phải là một phân bố lạ lùng. Dĩ nhiên nó sẽ là vô lý để mong đợi rằng phân phối đặc biệt nào đó của các ma trận ngẫu nhiên nên phù hợp với cách xử lý của các ma trận xuất hiện trong thực hành trong cách định lượng gần.
Tuy nhiên, hiện tượng được giải thích không là vấn đề của các con số rõ ràng. Các ma trận với các thừa số tăng lớn là hiếm khi loại trừ nhau trong các ứng dụng. Nếu ta có thể cho thấy rằng chúng là hiếm khi loại trừ nhau giữa các ma trận ngẫu nhiên trong lớp định nghĩa tốt nào đó thì các kỹ thuật được bao gồm phải chắc chắn là giống nhau. Đối số không phụ thuộc vào một độ đo của việc thỏa thuận "loại trừ lẫn nhau" với cái khác tới thừa số đặc biệt bất kì như 2 hoặc 10 hoặc 100.
Hình 4.1 và hình 4.2 đưa ra các thực thi với các ma trận ngẫu nhiên: mỗi phần tử là một mẫu độc lập từ phân phối chuẩn thực tế với trung vị 0 và phương sai chuẩn $m^{1/2}$. Trong Hình 4.1, một sự tập hợp của các ma trận ngẫu nhiên của các số chiều thay đổi đã được phân tích và các thừa số tăng đưa ra như một đồ thị khuếch tán. Chỉ 2 trong số các ma trận cho một thừa số tăng lớn như là $m^{-1/2}$. Trong Hình 4.2, các kết quả của việc phân tích một triệu ma trận mỗi chiều $m = 8, 16, 32$. Và được cho thấy ở đây, các thừa số tăng đã được tập hợp trong các ngăn bề rộng 0.2 và các dữ liệu kết quả vẽ như một phân bố trù mật xác suất. Trù mật xác suất của các thừa số tăng xuất hiện để giảm các lũy thừa với kích thước. Giữa 3 triệu ma trận này, vì thừa số tăng cao nhất có thể đã là 2,147,483,648 nên số lớn nhất được đếm một cách chính xác là 11.99.
Các kết quả tương tự đạt được với các ma trận ngẫu nhiên được xác định bởi các phân phối xác suất khác, như là các phần tử được phân phối đều trong [-1,1]. Nếu bạn lấy 1 tỷ ma trận ngẫu nhiên thì bạn sẽ hầu hết chắc chắn không tìm thấy một ma trận mà



Hình 4.1: Các thừa số tăng cho khử Gauss với quay từng phần được áp dụng cho 496 ma trận ngẫu nhiên (các phần tử độc lập với nhau và được phân phối chuẩn) của các số chiều thay đổi. Kích thước của ρ là bậc $m^{1/2}$, ít hơn nhiều giá trị có thể lớn nhất 2^{m-1}.



Hình 4.2: Các phân phối trù mật xác suất cho các thừa số tăng của các ma trận ngẫu nhiên có số chiều $m = 8, 16, 32$, được dựa vào các kích thước mẫu của một triệu con số cho mỗi chiều. Sự trù mật xuất hiện để giảm số mũ với ρ. Hình răng cưa gần cuối mỗi đường cong là một thành phần lạ của các kích thước mẫu hữu hạn.

khử Gauss là không ổn định.

4.3.5 Giải thích

Nếu $PA = LU$ thì $U = L^{-1}PA$. Nếu khử Gauss là không ổn định khi được áp dụng tới ma trận A thì điều này kéo theo rằng ρ là lớn, khi đó L^{-1} cũng phải là quá lớn. Bây giờ, khi nó xảy ra, các ma trận tam giác ngẫu nhiên hướng tới các ma trận khả nghịch lớn, lũy thừa lớn như là một hàm của số chiều m. Đặc biệt, điều này là đúng cho các ma trận tam giác ngẫu nhiên của dạng được phát biểu bởi khử Gauss với quay từng phần, với 1 nằm trên đường chéo và các phần tử có trị tuyệt đối nhỏ hơn bằng 1 bên dưới đường chéo.

Khi khử Gauss được áp dụng tới các ma trận ngẫu nhiên A thì các thừa số kết quả L là bất kì nhưng ngẫu nhiên. Các sự tương quan xuất hiện giữa các dấu của các phần tử của L mà chúng đưa ra các ma trận điều kiện tốt đặc biệt. Một phần tử đặc trưng của L^{-1}, xa hơn là các lũy thừa lớn, thường là nhỏ hơn 1 trong dấu giá trị tuyệt đối. Hình 4.3 đưa ra sự rõ ràng của hiện tượng này được dựa vào một ma trận đơn (nhưng đặc trưng) có số chiều $m = 128$.

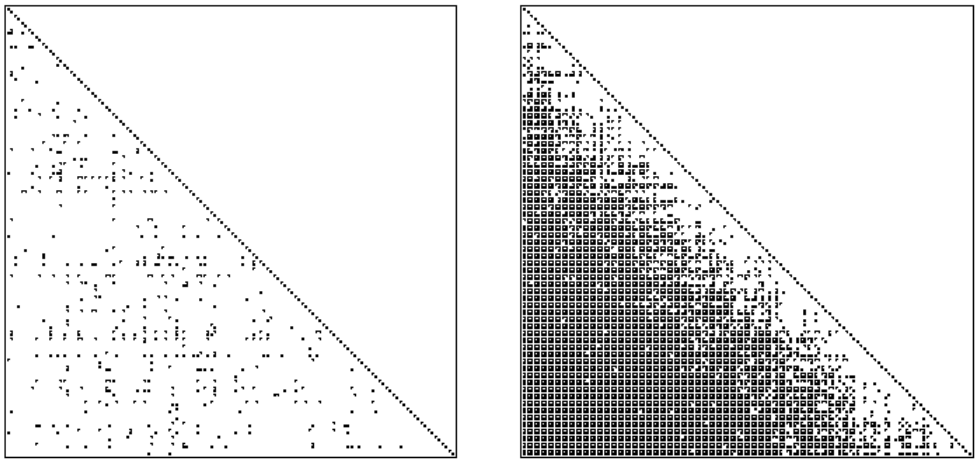
Do đó ta đến câu hỏi: vì sao các ma trận L được đưa ra bằng khử Gauss hầu hết không bao giờ có ma trận nghịch đảo lớn?

Câu trả lời nằm trong sự xem xét các không gian cột. Vì U là ma trận tam giác trên và $PA = LU$ nên không gian cột của PA và L là giống nhau. Nghĩa là cột đầu tiên của không gian sinh PA giống cột đầu tiên của L, 2 cột đầu tiên của không gian sinh PA giống với 2 cột đầu tiên của L, ... Nếu A là ngẫu nhiên thì các không gian cột của nó được định hướng một cách ngẫu nhiên, và nó theo sau sự tương tự phải đúng là các không gian cột của $P^{-1}L$. Tuy nhiên, điều kiện này là không tương thích với L^{-1} lớn. Nó có thể được cho thấy rằng nếu L^{-1} là lớn khi đó các không gian cột của L, hoặc của các hoán vị bất kì $P^{-1}L$ phải được làm xuyên đi trong một mô hình mà nó xa hơn ngẫu nhiên.

Hình 4.4 cho sự rõ ràng của điều này. Hình này cho thấy "nơi năng lượng là" trong các không gian cột liên tiếp của 2 ma trận giống nhau như trong Hình 4.3. Thiết bị cho việc làm này là một *điển hình* Q, được xác định bởi các lệnh của Matlab

$$[Q, R] = qr(A),\ spy(abs(Q)) > 1/sqrt(m)). \tag{4.3.7}$$

Đầu tiên các lệnh này tính phân tích QR của ma trận A, khi đó đồ thị một dấu chấm tại mỗi vị trí của Q tương ứng với một phần tử lớn hơn phương sai chuẩn, $m^{-1/2}$. Hình miêu tả cho một ma trận ngẫu nhiên A, mặc dù sau đó các hoán đổi dòng với nhau thành dạng PA, các không gian cột được định hướng một cách ngẫu nhiên, trong khi với ma trận A cho một thừa số tăng lớn, các định hướng là rất xa từ ngẫu nhiên. Nó



Hình 4.3: Cho A là ma trận ngẫu nhiên 128×128 với phân tích $PA = LU$. Trong vế trái, L^{-1} được cho thấy rằng: các phần tử biểu diễn các dấu chấm với độ dài lớn hơn 1. Trong vế phải hình tương tự cho $\tilde{L}^{-1}$, với $\tilde{L}$ là giống như L ngoại trừ các dấu của các phần tử trên đường chéo phụ của nó đã được ngẫu nhiên. Khử Gauss hướng đến đưa ra các ma trận L mà chúng là điều kiện tốt đặc biệt.

giống như là bằng việc xác định số lượng đối số này, nó có thể được chứng minh rằng các thừa số tăng lớn hơn bậc $m^{1/2}$ là hiếm giữa các ma trận ngẫu nhiên trong ý nghĩ mà với $\alpha > 1/2$ bất kì và $M > 0$, xác suất của sự kiện $\rho > m^{\alpha}$ là nhỏ hơn m^{-M} với mọi m đủ lớn. Tuy nhiên, một định lý như vậy không được chứng minh.

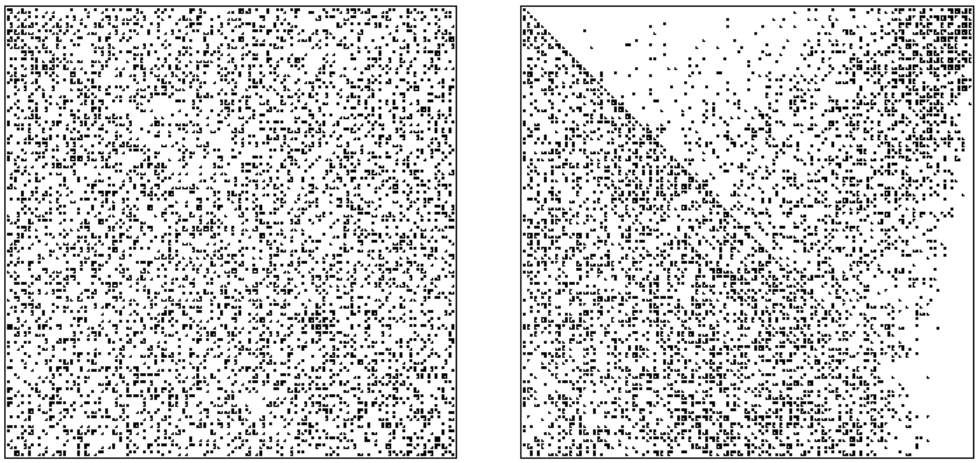
Ta hãy tóm tắt tính ổn định của khử Gauss với quay từng phần. Thuật toán này là không ổn định cao cho các ma trận A bất kì. Tuy nhiên, cho tính không ổn định xuất hiện thì các không gian cột của A phải được làm xuyên đi trong dạng rất đặc biệt, mà nó là hiếm trong ít nhất một lớp các ma trận ngẫu nhiên. Các thập kỉ của thực nghiệm tính toán đã đề nghị rằng các ma trận mà các không gian cột của chúng được làm xuyên đi trong dạng này xuất hiện rất hiếm trong các ứng dụng.

4.4 Phân tích Cholesky

Các ma trận xác định dương Hermit có thể được phân tích thành các thừa số tam giác nhanh như các ma trận tổng quát. Thuật toán tiêu chuẩn cho điều này là phân tích Cholesky, một biến thể của khử Gauss mà nó tính toán trong cả bên trái và bên phải của ma trận A trong 1 lần, lưu trữ và khai thác ma trận đối xứng.

4.4.1 Các ma trận xác định dương Hermit

Một ma trận thực $A \in \mathbb{R}^{m \times m}$ là *đối xứng* nếu nó có các phần tử giống nhau bên dưới đường chéo cũng như bên trên đường chéo: $a_{ij} = a_{ji}$ với mọi i, j, do đó $A = A^T$. Một ma



Hình 4.4: Trong vế trái, ma trận ngẫu nhiên A sau khi hoán vị thành dạng PA, hoặc tương đương, thừa số L. Trong vế phải, ma trận $\tilde{L}$ với các dấu được ngẫu nhiên. Các không gian cột của $\tilde{L}$ được làm xuyên đi trong kiểu không giống lũy thừa để xuất hiện trong các loại đặc trưng của các ma trận ngẫu nhiên.

trận như vậy thỏa mãn $x^T Ay = y^T Ax$ với mọi vector $x, y \in \mathbb{R}^m$.

Với một ma trận phức $A \in \mathbb{C}^{m\times m}$, tính chất tương tự A là *hermit*. Một ma trận hermit có các phần tử bên dưới đường chéo là các liên hợp phức của các phần tử ở bên trên đường chéo: $a_{ij} = \overline{a_{ij}}$, do đó $A = A^*$. Chú ý rằng điều này có nghĩa là các phần tử trên đường chéo của một ma trận hermit phải là thực.

Một ma trận hermit A thỏa mãn $x^*Ay = \overline{y^*Ax}$ với mọi $x, y \in \mathbb{C}^m$. Đặc biệt điều này nghĩa là với $x \in \mathbb{C}^m$ bất kì, x^*Ax là thực. Nếu thêm $x^*Ax > 0$ với mọi $x \neq 0$, thì A được nói là *xác định dương hermit* (hoặc đôi khi là *xác định dương*). Nhiều ma trận xuất hiện trong các hệ thống vật lý là xác định dương hermit do các luật vật lý cơ bản.

Nếu A là một ma trận xác định dương hermit $m \times m$ và X là một ma trận $m \times n$ hạng đầy đủ với $m \geqslant n$ thì ma trận X^*AX cũng là xác định dương hermit. Nó là hermit bởi vì $(X^*AX)^* = X^*A^*X = X^*AX$. Nó là xác định dương bởi vì, cho vector $x \neq 0$ bất kì, ta có $Xx \neq 0$ và do đó $x^*(X^*AX)x = (Xx)^*A(Xx) > 0$. Bằng việc chọn X để làm một ma trận $m \times n$ với 1 trong mỗi cột và 0 ở những nơi khác, ta có thể viết ma trận con chính $n \times n$ bất kì của A trong dạng X^*AX. Do đó, ma trận con chính bất kì của A phải là xác định dương. Đặc biệt, mọi phần tử đường chéo của A là một số thực dương.

Các trị riêng của một ma trận xác định dương hermit cũng là các số thực dương. Nếu $Ax = \lambda x$ với $x \neq 0$ thì ta có $x^*Ax = \lambda x^*x > 0$ và do đó $\lambda > 0$. Ngược lại, nó có thể được chứng minh rằng nếu một ma trận hermit có tất cả các trị riêng dương thì nó là xác định dương.

Các vector riêng tương ứng với các trị riêng phân biệt của một ma trận hermit là trực

giao. Giả sử $Ax_1 = \lambda_1 x_1$ và $Ax_2 = \lambda_2 x_2$ với $\lambda_1 \neq \lambda_2$. Khi đó

$$\lambda_2 x_1^* x_2 = x_1^* A x_2 = \overline{x_2^* A x_1} = \overline{\lambda_1 x_2^* x_1} = \lambda_1 x_1^* x_2,$$

nên $(\lambda_1 - \lambda_2) x_1^* x_2 = 0$. Vì $\lambda_1 \neq \lambda_2$, ta có $x_1^* x_2 = 0$.

4.4.2 Khử Gauss đối xứng

Bây giờ ta trở lại bài toán phân tích một ma trận xác định dương hermit thành các thừa số tam giác. Để bắt đầu, xét một bước đơn của khử Gauss được áp dụng cho một ma trận hermit A với 1 nằm ở vị trí bên trái trên:

$$A = \begin{bmatrix} 1 & w^* \\ w & K \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ w & I \end{bmatrix} \begin{bmatrix} 1 & w^* \\ 0 & K - ww^* \end{bmatrix}.$$

Như được miêu tả trong mục 4.1, các số 0 đã được đưa vào trong cột đầu tiên của ma trận bằng một phép toán tam giác dưới cơ bản trong vế trái mà nó trừ các bội của dòng đầu tiên từ các dòng sau đó.

Bây giờ khử Gauss sẽ tiếp tục giảm thành dạng tam giác bằng việc đưa các số 0 trong cột thứ hai. Tuy nhiên, để giữ sự đối xứng, đầu tiên phân tích Cholesky đưa các số 0 trong dòng đầu tiên để phù hợp với số 0 vừa được đưa ra trong cột đầu tiên. Ta có thể làm điều này bằng phép toán tam giác trên bên phải mà nó trừ các bội của cột đầu tiên từ các cột sau đó:

$$\begin{bmatrix} 1 & w^* \\ 0 & K - ww^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & K - ww^* \end{bmatrix} \begin{bmatrix} 1 & w^* \\ 0 & I \end{bmatrix}.$$

Chú ý rằng phép toán tam giác trên này một cách chính xác là phụ hợp của phép toán tam giác dưới mà ta thường đưa các số 0 trong cột đầu tiên.

Việc kết hợp với các phép toán ở trên, ta thấy rằng ma trận A đã được phân tích thành 3 số hạng:

$$A = \begin{bmatrix} 1 & w^* \\ w & K \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ w & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & K - ww^* \end{bmatrix} \begin{bmatrix} 1 & w^* \\ 0 & I \end{bmatrix}. \tag{4.4.1}$$

Ý tưởng của phân tích Cholesky là để tiếp tục quá trình này, việc đưa các số 0 trong một cột và một dòng của A một cách đối xứng cho tới khi nó được giảm xuống thành ma trận đơn vị.

4.4.3 Phân tích Cholesky

Cho $a_{11} > 0$ bất kì, nhưng không phải $a_{11} = 1$. Tổng quát hóa của (4.4.1) được thực hiện bằng việc hiệu chỉnh một vài phần tử của R_1 bằng một thừa số của $\sqrt{a_{11}}$. Cho $\alpha = \sqrt{a_{11}}$

và quan sát:

$$A = \begin{bmatrix} a_{11} & w^* \\ w & K \end{bmatrix}$$
$$= \begin{bmatrix} \alpha & 0 \\ w/\alpha & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & K - ww^*/a_{11} \end{bmatrix} \begin{bmatrix} \alpha & w^*/\alpha \\ 0 & I \end{bmatrix} = R_1^* A_1 R_1.$$

Đây là bước cơ sở mà nó được áp dụng lặp lại trong phân tích Cholesky. Nếu phần tử bên trái phía trên của ma trận con $K - ww^*/a_{11}$ là dương thì công thức tương tự có thể được sử dụng để phân tích nó; Khi đó ta có $A_1 = R_2^* A_2 R_2$ và do đó $A = R_1^* R_2^* A_2 R_2 R_1$. Quá trình được tiếp tục giảm thành góc dưới cùng bên phải, cuối cùng cho chúng ta một phân tích

$$A = \underbrace{R_1^* R_2^* \dots R_m^*}_{R^*} \underbrace{R_m \dots R_2 R_1}_{R}. \tag{4.4.2}$$

Phương trình này có dạng

$$A = R^* R,\ r_{ij} > 0, \tag{4.4.3}$$

với R là ma trận tam giác trên. Sự giảm loại này của một ma trận xác định dương hermit được biết như là môt *phân tích Cholesky.*

Cho A là ma trận xác định dương

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} r_{11} & 0 & 0 & \dots & 0 \\ r_{12} & r_{22} & 0 & \dots & 0 \\ r_{13} & r_{23} & r_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1n} & r_{2n} & r_{3n} & \dots & r_{nn} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ 0 & r_{22} & r_{23} & \dots & r_{2n} \\ 0 & 0 & r_{33} & \dots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & r_{nn} \end{bmatrix}$$

Phân tích Cholesky có thể được biểu diễn trong dạng như sau

$$r_{11} = \sqrt{a_{11}}$$
$$r_{1j} = \frac{a_{1j}}{r_{11}} \quad j = 2, \dots, n$$
$$r_{ii} = +\sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2}$$
$$r_{ji} = \frac{\left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}\right)}{r_{ii}} \quad j = i+1, \dots, n.$$

Ví dụ 4.4.1. Cho

$$A = \begin{bmatrix} 4 & -2 & 4 & 2 \\ -2 & 10 & -2 & -7 \\ 4 & -2 & 8 & 4 \\ 2 & -7 & 4 & 7 \end{bmatrix}$$

Ta có A là ma trận thực, đối xứng và thỏa mãn tính chất

$$x^T Ax > 0$$

với $x \in \mathbb{R}^n$. Do đó, A xác định dương và tính thừa số Cholesky R như sau

$$\begin{aligned}
r_{11} &= \sqrt{a_{11}} = 2 \\
r_{12} &= \frac{a_{12}}{r_{11}} = \frac{-2}{2} = -1 \\
r_{13} &= \frac{a_{13}}{r_{11}} = \frac{4}{2} = 2 \\
r_{14} &= \frac{a_{14}}{r_{11}} = \frac{2}{2} = 1 \\
r_{22} &= \sqrt{a_{22} - r_{12}^2} = \sqrt{10 - (-1)^2} = 3 \\
r_{23} &= \frac{(a_{23} - r_{12}r_{13}}{r_{22}} = \frac{-2 - (-1)2}{3} = 0 \\
r_{24} &= \frac{(a_{24} - r_{12}r_{14}}{r_{22}} = \frac{-7 - (-1)1}{3} = -2 \\
r_{33} &= \sqrt{a_{33} - r_{13}^2 - r_{23}^2} = \sqrt{8 - (2)^2 - 0} = 2 \\
r_{34} &= \frac{(a_{34} - r_{13}r_{14} - r_{23}r_{24}}{r_{33}} = \frac{4 - (2)(1) - 0(-2)}{2} = 1 \\
r_{44} &= \sqrt{a_{44} - r_{14}^2 - r_{24}^2 - r_{34}^2} = \sqrt{7 - (1)^2 - (-2)^2 - (1)^2} = 1
\end{aligned}$$

Vậy

$$R = \begin{bmatrix} 2 & -1 & 2 & 1 \\ 0 & 3 & 0 & -2 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Định lý 4.4.1 *Mọi ma trận xác định dương hermit $A \in \mathbb{C}^{m\times m}$ có một phân tích Cholesky (4.4.3) duy nhất.*

Chứng minh. Phân tích tồn tại vì thuật toán không thể bị phá hủy. Thật vậy, thuật toán cũng thiết lập sự duy nhất. Tại mỗi bước (4.4.2), giá trị $\alpha = \sqrt{a_{11}}$ được xác định bởi dạng của phân tích R^*R, và α được xác định, dòng đầu tiên của R_1^* cũng được xác định như vậy. Vì các con số tương tự được xác định tại mỗi bước của sự giảm nên phân tích hoàn toàn là duy nhất.

4.4.4 Thuật toán

Khi phân tích Cholesky được thực thi, chỉ phân nửa ma trận đang được tính toán cho các nhu cầu được trình bày một cách rõ ràng. Sự rút gọn này cho phép phân nửa phép

toán số học được cho phép. Ma trận đầu vào A biểu diễn một nửa siêu đường chéo của ma trận xác định dương hermit $m \times m$ được phân tích. (Trong phần mềm thực hành, hệ thống lưu trữ được nén có thể được sử dụng để tránh việc tàn phá phân nửa phần tử của một mảng vuông.) Ma trận đầu ra R biểu diễn thừa số tam giác trên cho $A = R^*R$. Mỗi bước lặp bên ngoài tương ứng với phân tích cơ bản đơn: phần tam giác trên của ma trận con $R^*_{k:m,k:m}$ đưa ra phần siêu đường chéo của ma trận hermit được lưu trữ tại bước thứ k.

Thuật toán 4.3 Phân tích Cholesky

```
R = A
for k = 1 to m do
    for j = k + 1 to m do
        R_{j,j:m} = R_{j,j:m} − R_{k,j:m} conj(R_{kj}) / R_{kk}
    end for
    R_{k,k:m} = R_{k,k:m} / sqrt(R_{kk})
end for
```

4.4.5 Đếm số phép toán

Số học được làm trong phân tích Cholesky được chi phối bởi vòng lặp bên trong. Sự thực thi đơn của dòng lệnh

$$R_{j,j:m} = R_{j,j:m} - R_{k,j:m}\overline{R_{kj}}/R_{kk}$$

cần 1 phép chia, $m - j + 1$ phép nhân và $m - j + 1$ phép trừ, nên tổng $\sim 2(m - j)$ phép toán dấu chấm động. Sự tính toán này được lặp lại một lần với mỗi j từ $k+1$ tới m, và vòng lặp đó được lặp lại với mỗi k từ 1 tới m. Tổng là

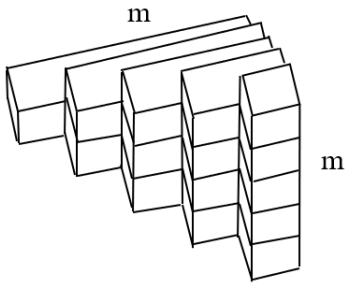
$$\sum_{k=1}^{m}\sum_{j=k+1}^{m} 2(m-j) \sim 2\sum_{k=1}^{m}\sum_{j=1}^{k} j \sim \sum_{k=1}^{m} k^2 \sim \frac{1}{3}m^3 \text{ phép toán dấu chấm động.}$$

Do đó, phân tích Cholesky chỉ bao gồm một nửa phép toán như khử Gauss, mà nó sẽ yêu cầu $\sim \frac{2}{3}m^3$ phép toán dấu chấm động để phân tích cùng ma trận.

Như thường lệ, đếm số phép toán cũng có thể được xác định bằng đồ thị. Với mỗi k, 2 phép toán dấu chấm động được thực hiện (một phép nhân và một phép trừ) tại mỗi vị trí của một lớp tam giác. Thuật toán hoàn toàn tương ứng với việc xếp chồng lên m lớp:

Khi $m \to \infty$, khối hội tụ về một tứ diện với thể tích $\frac{1}{6}m^3$. Vì mỗi đơn vị thể tích tương ứng với 2 phép toán dấu chấm động, ta thu được

$$\text{Phân tích Cholesky } \sim \frac{1}{3}m^3 \text{ phép toán dấu chấm động} \tag{4.4.4}$$



4.4.6 Tính ổn định

Thuật toán này thường là ổn định. Bằng trực giác, lý do mà các thừa số R không bao giờ có thể tăng lớn. Trong chuẩn 2, ví dụ ta có $\|R\| = \|R^*\| = \|A\|^{1/2}$ (chứng minh: SVD), và trong chuẩn p khác với $1 \leqslant p \leqslant \infty$, $\|R\|$ không thể khác $\|A\|^{1/2}$ bởi nhiều hơn một thừa số của $\sqrt{m}$. Do đó, số phần tử lớn hơn nhiều số phần tử của A có thể không bao giờ xuất hiện.

Chú ý rằng tính ổn định của phân tích Cholesky đạt được không cần cho quay bất kì. Bằng trực giác, ta có thể quan sát thấy rằng điều này là có liên quan tới tác dụng của ma trận xác định dương hermit là trên đường chéo. Ví dụ, không khó để cho thấy rằng phần tử lớn nhất phải xuất hiện trên đường chéo, và tính chất này đưa vào các ma trận con xác định dương được xây dựng trong quá trình qui nạp (4.4.2).

Phân tích tính ổn định của quá trình Cholesky dẫn tới kết quả ổn định ngược như sau.

Định lý 4.4.2 *Cho $A \in \mathbb{C}^{m\times m}$ là xác định dương hermit, và cho một phân tích Cholesky của A được tính bởi Thuật toán 4.3 trong một máy tính thỏa mãn (3.2.5) và (3.2.7). Với $\epsilon_{machine}$ đủ nhỏ, quá trình này được bảo đảm để chạy hoàn toàn (nghĩa là, các phần tử r_{kk} khác 0 và âm sẽ không xuất hiện), sinh ra một thừa số được tính $\tilde{R}$ mà nó thỏa mãn*

$$\tilde{R}^*\tilde{R} = A + \delta A, \ \frac{\|\delta A\|}{\|A\|} = O(\epsilon_{machine}) \tag{4.4.5}$$

với $\delta A \in \mathbb{C}^{m\times m}$ bất kì.

Giống như nhiều thuật toán khác, thuật toán này sẽ trông tệ hơn nhiều nếu ta cố gắng thực hiện phân tích sai số tiến hơn là sai số ngược. Nếu A là điều kiện xấu thì $\tilde{R}$ sẽ không gần với R; Tốt nhất ta có thể nói là $\|\tilde{R} - R\|/\|R\| = O(\kappa(A)\epsilon_{machine})$. (Mặc khác, phân tích Cholesky tổng quát là bài toán điều kiện xấu.) Tích $\tilde{R}^*\tilde{R}$ thỏa mãn chặn sai số tốt hơn nhiều (4.4.5). Do đó các sai số được đưa ra trong $\tilde{R}$ bằng việc làm tròn là lớn nhưng "tương quan ranh mãnh" như ta thấy trong mục 3.4 cho phân tích QR.

4.4.7 Giải phương trình Ax=b

Nếu A là xác định dương hermit thì cách tiêu chuẩn để giải một hệ thống các phương trình $Ax = b$ là bằng phân tích Cholesky. Thuật toán 4.3 giảm hệ thống thành $R^*Rx = b$,

và khi đó ta giải 2 hệ thống tam giác liên tiếp: đầu tiên $R^*y = b$ với biến y không được biết, khi đó $Rx = y$ với biến x không được biết. Mỗi lời giải tam giác cần $\sim m^2$ phép toán dấu chấm động nên việc làm có tổng cộng là $\sim \frac{1}{3}m^3$ phép toán dấu chấm động.

Ví dụ 4.4.2. Cho

$$A = \begin{bmatrix} 4 & -2 & 4 & 2 \\ -2 & 10 & -2 & -7 \\ 4 & -2 & 8 & 4 \\ 2 & -7 & 4 & 7 \end{bmatrix} \text{ và } b = \begin{bmatrix} 8 \\ 2 \\ 16 \\ 6 \end{bmatrix}$$

sử dụng phân tích Cholesky để giải phương trình $Ax = b$. Trong Ví dụ **??**, ta sử dụng phân tích Cholesky tìm R như sau

$$R = \begin{bmatrix} 2 & -1 & 2 & 1 \\ 0 & 3 & 0 & -2 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Để giải phương trình $Ax = b$, đầu tiên ta giải phương trình $R^T y = b$ bằng phép thế tiến và thu được $y = [4, 2, 4, 2]^T$. Khi đó, ta giải phương trình $Rx = b$ bằng phép thế ngược và thu được $x = [1, 2, 1, 2]^T$.

Định lý 4.4.3 *Nghiệm của các hệ thống xác định dương hermit $Ax = b$ thông qua phân tích Cholesky (Thuật toán 4.3) là ổn định ngược, sinh ra một nghiệm được tính $\tilde{x}$ thõa mãn*

$$(A + \Delta A)\tilde{x} = b, \ \frac{\|\Delta A\|}{\|A\|} = O(\epsilon_{machine}) \tag{4.4.6}$$

với $\Delta A \in \mathbb{C}^{m\times m}$ bất kì.

Bài tập

1. Cho $A \in \mathbb{C}^{m\times m}$ là không suy biến. Chứng minh rằng A có một phân tích LU nếu và chỉ nếu với mọi k, $1 \leqslant k \leqslant m$, khối ở trên bên trái $A_{1:k,1:k}$ cấp $k \times k$ là không suy biến. Chứng minh phân tích LU này là duy nhất.

2. Giả sử $A \in \mathbb{C}^{m\times m}$ thỏa điều kiện của Bài tập 1 và được phân dải với băng thông $2p + 1$, nghĩa là $a_{ij} = 0$ với $|i - j| > p$. Các mô hình thưa thớt của các phân tích L và U của A?

3. Giả sử ma trận A cấp $m \times m$ được viết thành dạng khối

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

với A_{11} là ma trận cấp $n \times n$ và A_{22} là ma trận cấp $(m-n) \times (m-n)$. Giả sử A thỏa các điều kiện của Bài tập 1.

(a) Kiểm tra công thức

$$\begin{bmatrix} I & \\ -A_{21}A_{11}^{-1}I & \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

cho sự khử khối A_{21}. Ma trận $A_{22} - A_{21}A_{11}^{-1}A_{12}$ là *phần bù Schur* của A_{11} trong A.

(b) Giả sử A_{21} được khử dòng bằng n bước của khử Gauss. Chứng minh rằng khối ở trên bên phải cấp $(m-n) \times (m-n)$ của kết quả là $A_{22} - A_{21}A_{11}^{-1}A_{12}$.

4. Cho A là ma trận 4×4

$$A = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}.$$

a) Xác định $detA$ từ (4.1.5).

b) Xác định $detA$ từ (4.2.3).

c) Miêu tả khử Gauss với quay từng phần có thể được sử dụng để tìm định thức của 1 ma trận vuông tổng quát.

5. Xét khử Gauss tiến hành với pivoting bởi các cột thay vì các dòng, đưa ra một phân tích $AQ = LU$, với Q là ma trận hoán vị.

a) Chứng minh rằng nếu A không suy biến thì phân tích như vậy luôn tồn tại.

b) Chứng minh rằng nếu A suy biến thì phân tích như vậy không tồn tại.

6. Khử Gauss có thể được sử dụng để tính A^{-1} của một ma trận không suy biến $A \in \mathbb{C}^{m \times m}$.

a) Miêu tả một thuật toán tính A^{-1} bằng việc giải hệ thống m phương trình và chứng minh rằng đếm số phép toán tiệm cận của nó là $8m^3/3$ phép toán dấu chấm động.

b) Miêu tả một biến thể của thuật toán mà nó giảm số phép toán xuống còn $2m^3$ phép toán dấu chấm động.

c) Giả sử ta mong muốn giải hệ thống n phương trình $Ax_j = b_j$ hoặc $AX = B$, với $B \in \mathbb{C}^{m \times n}$. Đếm số phép toán tiệm cận (hàm của m và n) cho việc làm này từ phân tích LU và sự tính toán của A^{-1}?

7. Chứng minh rằng khử Gauss với quay từng phần được áp dụng cho ma trận bất kì $A \in \mathbb{C}^{m\times m}$ thì thừa số tăng (4.3.2) thỏa $\rho \leqslant 2^{m-1}$.

8. Giả sử $PA = LU$ (phân tích LU với quay từng phần) và $A = QR$ (phân tích QR). Miêu tả mối quan hệ giữa dòng cuối của L^{-1} và cột cuối của Q.

9. Cài đặt thuật toán khử Gauss với quay từng phần (Thuật toán 4.2).

10. Kiểm tra các ma trận sau

$$A = \begin{bmatrix} 9 & 3 & 3 \\ 3 & 10 & 7 \\ 3 & 5 & 9 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & 2 & 6 \\ 2 & 2 & 5 \\ 6 & 5 & 29 \end{bmatrix}$$

$$C = \begin{bmatrix} 4 & 4 & 8 \\ 4 & -4 & 1 \\ 8 & 1 & 6 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

có phải là ma trận xác định dương không?

11. Cài đặt Thuật toán Cholesky (Thuật toán 4.3).

12. Cho

$$A = \begin{bmatrix} 16 & 4 & 8 & 4 \\ 4 & 10 & 8 & 4 \\ 8 & 8 & 12 & 10 \\ 4 & 4 & 10 & 12 \end{bmatrix} \text{ và } b = \begin{bmatrix} 32 \\ 26 \\ 38 \\ 30 \end{bmatrix}$$

a) Chứng minh A là ma trận xác định dương.

b) Tính thừa số Cholesky của A.

c) Giải phương trình $Ax = b$.

13. Cho A là ma trận vuông không suy biến và cho $A = QR$ và $A^*A = U^*U$ lần lượt là phân tích QR và phân tích Cholesky, với sự chuẩn hóa thông thường $r_{jj}, u_{jj} > 0$. $R = U$?

Tài liệu tham khảo

[1] Lloyd N. Trefethen, and David Bau, III (1997), Numerical Linear Algebra, *Society for Industrial and Applied Mathematics.*

[2] Gene H. Golub, and Charles F. Van Loan (2013), Matrix Computations, *The Johns Hopkins University.*

[3] Kirk Baker (2013), Singular Value Decomposition Tutorial, *The Ohio State University.*

[4] Andrew Lounsbury (2018), Singular Value Decomposition, *Tennessee Technological University.*

[5] David S. Watkins (1991), Fundamentals of matrix computations, Wiley, New York.