Chapter 11
Matrix Factorizations and Their Applications

The techniques described below are often called *direct solving methods*.

The direct solution (by Cramer's method, see Section 3.3.2) of a linear system $Mx = b$, when $M \in \mathbf{GL}_n(k)$ ($b \in k^n$) is computationally expensive, especially if one wishes to solve the system many times with various values of b. In the next chapter we study iterative methods for the case $k = \mathbb{R}$ or $\mathbb{C}$. Here we concentrate on a simple idea: to decompose M as a product PQ in such a way that the resolution of the intermediate systems $Py = b$ and $Qx = y$ is "cheap". In general, at least one of the matrices is triangular. For example, if P is lower-triangular ($p_{ij} = 0$ if $i < j$), then its diagonal entries p_{ii} are nonzero, and one may solve the system $Py = b$ step by step:

$$\begin{aligned} y_1 &= \frac{b_1}{p_{11}}, \\ &\vdots \\ y_i &= \frac{b_i - p_{i1}y_1 - \cdots - p_{i,i-1}y_{i-1}}{p_{ii}}, \\ &\vdots \\ y_n &= \frac{b_n - p_{n1}y_1 - \cdots - p_{n,n-1}y_{n-1}}{p_{nn}}. \end{aligned}$$

The computation of y_i needs $2i - 1$ operations and the final result is obtained in n^2 operations. This is not expensive if one notices that even computing the product $x = M^{-1}b$ (assuming that M^{-1} is computed once and for all, an expensive task) needs $2n^2 - n$ operations in general, and still n^2 in the triangular case.

Another example of easily invertible matrices is that of orthogonal matrices: if $Q \in \mathbf{O}_n$ (or $Q \in \mathbf{U}_n$), then $Qx = y$ is equivalent to $x = Q^T y$ (or $x = Q^* y$), which provides x in $O(n^2)$ operations.

D. Serre, *Matrices,* Graduate Texts in Mathematics 216,
DOI 10.1007/978-1-4419-7683-3_11,

11.1 The *LU* Factorization

Let k be a field.

Definition 11.1 *Let $M \in \mathbf{GL}_n(k)$ be given. We say that M admits an LU factorization if there exist in $\mathbf{GL}_n(k)$ two matrices L (lower-triangular with 1s on the diagonal) and U (upper-triangular) such that $M = LU$.*

Remarks

- The diagonal entries of U are not equal to 1 in general. The *LU* factorization is thus asymmetric with respect to L and U.
- The letters L and U recall the shape of the matrices: L for *lower* and U for *upper*.
- If there exists an *LU* factorization (which is unique, as we show below), then it can be computed by induction on the size of the matrix. The algorithm is provided in the proof of the next theorem. Indeed, if $N^{(p)}$ denotes the matrix extracted from N by keeping only the first p rows and columns, we have easily

$$M^{(p)} = L^{(p)}U^{(p)},$$

  which is nothing but the *LU* factorization of $M^{(p)}$.

Definition 11.2 *The* leading principal minors *of M are the determinants of the matrices $M^{(p)}$ defined above, for $1 \leq p \leq n$.*

Theorem 11.1 *The matrix $M \in \mathbf{GL}_n(k)$ admits an LU factorization if and only if its leading principal minors are nonzero. When this condition is fulfilled, the LU factorization is unique.*

Proof. Let us begin with uniqueness: if $LU = L'U'$, then $(L')^{-1}L = U'U^{-1}$, which reads $L'' = U''$, where L'' and U'' are triangular of opposite types, the diagonal entries of L'' being 1s. We deduce $L'' = U'' = I_n$; that is, $L' = L$, $U' = U$.

We next prove the necessity. Let us assume that M admits an *LU* factorization. Then $\det M^{(p)} = \det L^{(p)} \det U^{(p)} = \prod_{1 \leq j \leq p} u_{jj}$, which is nonzero because U is invertible.

Finally, we prove the sufficiency, that is, the existence of an *LU* factorization. We proceed by induction on the size of the matrices. It is clear if $n = 1$. Otherwise, let us assume that the statement is true up to the order $n-1$ and let $M \in \mathbf{GL}_n(k)$ be given, with nonzero leading principal minors. We look for L and U in the blockwise form

$$L = \begin{pmatrix} L' & 0 \\ X^T & 1 \end{pmatrix}, \quad U = \begin{pmatrix} U' & Y \\ 0 & u \end{pmatrix},$$

with $L', U' \in \mathbf{M}_{n-1}(k)$, and so on. We likewise obtain the description

$$M = \begin{pmatrix} M' & R \\ S^T & m \end{pmatrix}.$$

Multiplying blockwise, we obtain the equations

$$L'U' = M', \quad L'Y = R, \quad (U')^T X = S, \quad u = m - X^T Y.$$

By assumption, the leading principal minors of M' are nonzero. The induction hypothesis guarantees the existence of the factorization $M' = L'U'$. Then Y and X are the unique solutions of (triangular) Cramer systems. Finally, u is explicitly given. □

Let us evaluate the number of operations needed in the computation of L and U. We pass from a factorization in $\mathbf{GL}_{n-1}(k)$ to a factorization in $\mathbf{GL}_n(k)$ by means of the computations of X (in $(n-1)(n-2)$ operations), Y (in $(n-1)^2$ operations) and u (in $2(n-1)$ operations), for a total of $(n-1)(2n-1)$ operations. Finally, the computation *ex nihilo* of an *LU* factorization costs

$$P(n) = 3 + 10 + \cdots + (n-1)(2n-1) = \frac{2}{3}n^3 + O(n^2)$$

operations.

Proposition 11.1 *The LU factorization is computable in* $\frac{2}{3}n^3 + O(n^2)$ *operations.*

One says that the *complexity* of the *LU* factorization is $\frac{2}{3}n^3$.

Remark

When all leading principal minors but the last (the determinant of M) are nonzero, the proof above furnishes a factorization $M = LU$, in which U is not invertible; that is, $u_{nn} = 0$.

11.1.1 Block Factorization

One can likewise perform a *blockwise LU* factorization. If $n = p_1 + \cdots + p_r$ with $p_j \geq 1$, the matrices L and U are block-triangular. The diagonal blocks are square, of respective sizes $p_1, \ldots, p_r$. Those of L are of the form I_{p_j}, whereas those of U are invertible. A necessary and sufficient condition for such a factorization to exist is that the leading principal minors of M, of orders $p_1 + \cdots + p_j$ ($j \leq r$), be nonzero. As above, we may allow that $\det M \neq 0$, with the price that the last diagonal block of U be singular.

Such a factorization is useful for the resolution of the linear system $MX = b$ if the diagonal blocks of U are easily inverted, for instance if their sizes are small enough (say $p_j \approx \sqrt{n}$). Another favorable situation is when most of the diagonal blocks are equal to each other, because then one has to invert only a few blocks.

We have performed this blockwise factorization in Section 3.3.1 when $r = 2$. Recall that if

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \tag{11.1}$$

where the diagonal blocks are square and A is invertible, then

$$M = LU \qquad \text{with} \qquad L = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix}, \quad U = \begin{pmatrix} A & B \\ 0 & D - CA^{-1}B \end{pmatrix}. \tag{11.2}$$

From this, we see that if M is nonsingular too, then

$$M^{-1} = U^{-1}L^{-1} = \begin{pmatrix} A^{-1} & \cdot \\ 0 & (D - CA^{-1}B)^{-1} \end{pmatrix} \cdot \begin{pmatrix} I & 0 \\ \cdot & I \end{pmatrix} = \begin{pmatrix} \cdot & \cdot \\ \cdot & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

When all the blocks have the same size, a similar analysis yields Banachiewicz' formula

Corollary 11.1 *Let $M \in \mathbf{GL}_n(k)$, with $n = 2m$, read blockwise*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad A, B, C, D \in \mathbf{GL}_m(k).$$

Then

$$M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & (C - DB^{-1}A)^{-1} \\ (B - AC^{-1}D)^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

Proof. We can verify the formula by multiplying by M. The only point to show is that the inverses are meaningful: $A - BD^{-1}C, \ldots$ are invertible. Because of the symmetry of the formulæ, it is enough to check it for a single term, namely $D - CA^{-1}B$. Schur's complement formula gives $\det(D - CA^{-1}B) = \det M / \det A$, which is nonzero by assumption. □

11.1.2 Complexity of Matrix Inversion

We can now show that the complexity of inverting a matrix is not higher than that of matrix multiplication, at equivalent sizes. This fact is due independently to Boltz, Banachiewicz, and to Strassen We assume here that $k = \mathbb{R}$ or $\mathbb{C}$.

Notation 11.1 *We denote by J_n the number of operations in k used in the inversion of a typical $n \times n$ matrix, and by P_n the number of operations (in k) used in the product of two $n \times n$ matrices.*

Of course, the number J_n must be understood for generic matrices, that is, for matrices within a dense open subset of $\mathbf{M}_n(k)$. More important, J_n, P_n also depend on the algorithm chosen for inversion or for multiplication. In the sequel we wish to adapt the inversion algorithm to the one used for multiplication.

Let us examine first of all the matrices whose size n has the form 2^k.

We decompose the matrices $M \in \mathbf{GL}_n(k)$ blockwise as in (11.1), with blocks of size $n/2 \times n/2$. The condition $A \in \mathbf{GL}_{n/2}(k)$ defines a dense open set, because $M \mapsto \det A$ is a nonzero polynomial. Suppose that we are given an inversion algorithm for generic matrices in $\mathbf{GL}_{n/2}(k)$ in $j_{k-1} = J_{2^{k-1}}$ operations. Then blockwise *LU* factorization and the formula $M^{-1} = U^{-1}L^{-1}$, where

$$L^{-1} = \begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix}, \quad U = \begin{pmatrix} A^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ 0 & (D-CA^{-1}B)^{-1} \end{pmatrix},$$

furnish an inversion algorithm for generic matrices in $\mathbf{GL}_n(k)$. We can then bound j_k by means of j_{k-1} and the number $\pi_{k-1} = P_{2^{k-1}}$ of operations used in the computation of the product of two matrices of size $2^{k-1} \times 2^{k-1}$. We also denote by $\sigma_k = 2^{2k}$ the number of operations involved in the computation of the sum of matrices in $\mathbf{M}_{2^k}(k)$.

To compute M^{-1}, we first compute A^{-1}, then CA^{-1}, which gives us L^{-1} in $j_{k-1} + \pi_{k-1}$ operations. Then we compute $(D-CA^{-1}B)^{-1}$ (this amounts to $\sigma_{k-1} + \pi_{k-1} + j_{k-1}$ operations) and $A^{-1}B(D-CA^{-1}B)^{-1}$ (at cost $2\pi_{k-1}$), which furnishes U^{-1}. The computation of $U^{-1}L^{-1}$ is done with the expense of $\sigma_{k-1} + 2\pi_{k-1}$ operations. Finally,

$$j_k \leq 2j_{k-1} + 2\sigma_{k-1} + 6\pi_{k-1}.$$

In other words,

$$2^{-k}j_k - 2^{1-k}j_{k-1} \leq 2^{k-1} + 3 \cdot 2^{1-k}\pi_{k-1}. \tag{11.3}$$

The complexity of the product in $\mathbf{M}_n(k)$ obeys the inequalities

$$n^2 \leq P_n \leq n^2(2n-1).$$

The first inequality is due to the number of data ($2n^2$) and the fact that each operation involves only two of them. The second is given by the naive algorithm that consists in computing n^2 scalar products.

Lemma 19. *If $P_n \leq c_\alpha n^\alpha$ (with $2 \leq \alpha \leq 3$), then $j_\ell \leq C_\alpha \pi_\ell$, where $C_\alpha = 1 + 3c_\alpha/(2^{\alpha-1}-1)$.*

Proof. It is enough to sum (11.3) from $k = 1$ to l and use the inequality $1 + q + \cdots + q^{l-1} \leq q^\ell/(q-1)$ for $q > 1$. □

When n is not a power of 2, we obtain M^{-1} by computing the inverse of a block-diagonal matrix $\operatorname{diag}(M, I)$, whose size N satisfies $n \leq N = 2^\ell < 2n$. We obtain $J_n \leq j_\ell \leq C_\alpha \pi_\ell$. This is the first part of the following result.

Proposition 11.2 *If the complexity P_n of the product in $\mathbf{M}_n(\mathbb{C})$ is bounded by $c_\alpha n^\alpha$, then the complexity J_n of inversion in $\mathbf{GL}_n(\mathbb{C})$ is bounded by $d_\alpha n^\alpha$, where*

$$d_\alpha = \left(1 + \frac{3c_\alpha}{2^{\alpha-1}-1}\right) 2^\alpha.$$

Conversely, if the complexity of inversion in $\mathbf{GL}_n(\mathbb{C})$ is bounded by $\delta_\alpha n^\alpha$, then the complexity of the product in $\mathbf{M}_n(\mathbb{C})$ is bounded by $\gamma_\alpha n^\alpha$, where

$$\gamma_\alpha = 3^\alpha \delta_\alpha.$$

That can be summarized as follows:

Those who know how to multiply also know how to invert.

Proof. There remains to prove the second part. We notice that if $A, B \in \mathbf{M}_n(\mathbb{C})$ are given, then the matrix

$$M = \begin{pmatrix} I_n & -A & 0_n \\ 0_n & I_n & -B \\ 0_n & 0_n & I_n \end{pmatrix} \in \mathbf{M}_{3n}(\mathbb{C})$$

is invertible, with inverse

$$M^{-1} = \begin{pmatrix} I_n & A & AB \\ 0_n & I_n & B \\ 0_n & 0_n & I_n \end{pmatrix}.$$

Given A and B, we compute M^{-1}, thus AB, in $\delta_\alpha(3n)^\alpha$ operations at most (and certainly much less). □

11.1.3 Complexity of the Matrix Product

The ideas that follow apply to the product of rectangular matrices, but for the sake of simplicity, we present only the case of square matrices.

As we have seen above, the complexity P_n of matrix multiplication in $M_n(k)$ is $O(n^3)$. However, better algorithms allow us to improve the exponent 3. The simplest and oldest one is Strassen's algorithm, which uses a recursion argument. It is based upon a way of computing the product of two 2×2 matrices by means of 7 multiplications and 18 additions. Two features of Strassen's formula are essential. First, the number of multiplications that it involves is strictly less than that (eight) of the naive algorithm. The second is that the method is valid when the matrices have entries in a *noncommutative* ring, and so it can be employed for two matrices $M, N \in \mathbf{M}_n(k)$, considered as elements of $\mathbf{M}_2(A)$, with $A := \mathbf{M}_{n/2}(k)$. This trick yields

$$P_n \leq 7P_{n/2} + 18\left(\frac{n}{2}\right)^2.$$

For $n = 2^\ell$, we infer

$$7^{-\ell}\pi_\ell - 7^{1-\ell}\pi_{\ell-1} \leq \frac{9}{2}\left(\frac{4}{7}\right)^\ell,$$

which, after summation from $k = 0$ to ℓ, gives

$$7^{-\ell}\pi_\ell \leq \frac{9}{2} \times \frac{1}{1-4/7},$$

because of $\frac{4}{7} < 1$. Finally,

$$\pi_\ell \leq \frac{21}{2}\, 7^\ell.$$

When n is not a power of two, one chooses ℓ such that $n \leq 2^\ell < 2n$ and we obtain the following result.

Proposition 11.3 *The complexity of the multiplication of $n \times n$ matrices is $O(n^\alpha)$, with $\alpha = \log 7/\log 2 = 2.807\ldots$ More precisely,*

$$P_n \leq \frac{147}{2}\, n^{\log 7/\log 2}.$$

The exponent α can be improved, at the cost of greater complication and a larger constant c_α. The best exponent known in 2009, due to Coppersmith and Winograd [11], is $\alpha = 2.376\ldots$ It is fifteen years old, whereas Strassen's is forty years old. A rather complete analysis can be found in the book by Bürgisser, Clausen, and Shokrollahi [7].

Here is Strassen's formula [37]. Let $M, N \in \mathbf{M}_2(A)$, with

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad N = \begin{pmatrix} x & y \\ z & t \end{pmatrix}.$$

One first forms the expressions $x_1 = (a+d)(x+t)$, $x_2 = (c+d)x$, $x_3 = a(y-t)$, $x_4 = d(z-x)$, $x_5 = (a+b)t$, $x_6 = (c-a)(x+y)$, $x_7 = (b-d)(z+t)$. Each one involves one multiplication and either one or two addition(s). Then the product is given by eight more additions:

$$MN = \begin{pmatrix} x_1 + x_4 - x_5 + x_7 & x_3 + x_5 \\ x_2 + x_4 & x_1 - x_2 + x_3 + x_6 \end{pmatrix}.$$

Remark

The use of a fast method for matrix multiplication does reduce the complexity of many algorithms. Let us consider for instance the calculation of the characteristic polynomial $P(A)$ in the form improved by Preparata and Sawarte (see Section 3.10.2). If matrix multiplication is done in $O(n^\alpha)$ operations, then P_A is obtained in $O(n^\beta)$ operations, with $\beta = \max\{\alpha + \frac{1}{2}, 3\}$. If one has a not too cumbersome method with some $\alpha \leq 2.5$, it is thus useless to try to reduce α.

11.2 Cholesky Factorization

In this section $k = \mathbb{R}$, and we consider symmetric positive-definite matrices.

Theorem 11.2 *Let $M \in \mathbf{SPD}_n$. Then there exists a unique lower-triangular matrix $L \in \mathbf{M}_n(\mathbb{R})$, with strictly positive diagonal entries, satisfying $M = LL^T$.*

We warn the reader that, because of the symmetry between the lower- and upper-triangular factors, the diagonal entries of the matrix L are not units in general.

Proof. Uniqueness. If L_1 and L_2 have the properties stated above, then $I_n = LL^T$, for $L = L_2^{-1}L_1$, which still has the same form. In other words, $L = L^{-T}$, where both sides are triangular matrices, but of opposite types (lower and upper). This equality shows that L is actually diagonal, with $L^2 = I_n$. Because its diagonal is positive, we obtain $L = I_n$; that is, $L_2 = L_1$.

We give two constructions of L.

First method. The matrix $M^{(p)}$ is positive-definite (test the quadratic form induced by M on the linear subspace $\mathbb{R}^p \times \{0\}$). The leading principal minors of M are thus nonzero and there exists an LU factorization $M = L_0U_0$. Let D be the diagonal of U_0, which is invertible. Then $U_0 = DU_1$, where the diagonal entries of U_1 equal 1. By transposition, we have $M = U_1^T D_0 L_0^T$. From uniqueness of the LU factorization, we deduce $U_1 = L_0^T$ and $M = L_0 D L_0^T$. Then $L = \sqrt{D}L_0$ satisfies the conditions of the theorem. Observe that $D > 0$ because $D = PMP^T$ with $P = L_0^{-1}$, and thus D is positive-definite.

Second method. We proceed by induction over n. The statement is clear if $n = 1$. Otherwise, we seek an L of the form

$$L = \begin{pmatrix} L' & 0 \\ X^T & \ell \end{pmatrix},$$

knowing that

$$M = \begin{pmatrix} M' & R \\ R^T & m \end{pmatrix}.$$

The matrix L' is obtained by Choleski factorization of M', which belongs to $\mathbf{SPD}_{n-1}$. Then X is obtained by solving $L'X = R$. Finally, ℓ is a square root of $m - \|X\|^2$. Because $0 < \det M = (\ell \det L')^2$, we see that $m - \|X\|^2 > 0$; we thus choose $\ell = \sqrt{m - \|X\|^2}$. This method again shows uniqueness.

□

Remark

Choleski factorization extends to Hermitian positive-definite matrices. In that case, L has complex entries, but its diagonal entries are still real and positive.

11.3 The *QR* Factorization

We turn to the situation where one factor is triangular, and the other one is unitary.

Proposition 11.4 *Let $M \in \mathbf{GL}_n(\mathbb{C})$ be given. Then there exist a unitary matrix Q and an upper-triangular matrix R; the diagonal entries of the latter real positive, such that $M = QR$. This factorization is unique.*

We observe that the condition on the numbers r_{jj} is essential for uniqueness. In fact, if D is diagonal with $|d_{jj}| = 1$ for every j, then $Q' := Q\bar{D}$ is unitary, $R' := DR$ is upper-triangular, and $M = Q'R'$, which gives an infinity of factorizations "QU". Even in the real case, where Q is orthogonal, there are 2^n "QU" factorizations.

Proof. Uniqueness. If (Q_1, R_1) and (Q_2, R_2) give two factorizations, then $Q = R$ with $Q := Q_2^{-1}Q_1$ and $R := R_2R_1^{-1}$. Because Q is unitary, that is, $Q^* = Q^{-1}$, this implies $R^* = R^{-1}$. Because the inverse of a triangular matrix is a triangular matrix of the same type, whereas R^* is of opposite type, this tells us that R is diagonal. In additional, its diagonal part is strictly positive. Then $R^2 = R^*R = Q^*Q = I_n$ gives $R = I_n$. Finally, $R_2 = R_1$ and consequently, $Q_2 = Q_1$.

Existence. It follows from that of Choleski factorization. If $M \in \mathbf{GL}_n(\mathbb{C})$, the matrix M^*M is Hermitian positive-definite, and hence it admits a Choleski factorization R^*R, where R is upper-triangular with real positive diagonal entries. Defining $Q := MR^{-1}$, we have

$$Q^*Q = R^{-*}M^*MR^{-1} = R^{-*}R^*RR^{-1} = I_n;$$

hence Q is unitary. Finally, $M = QR$ by construction.

□

The method used above is unsatisfactory from a practical point of view, because one can compute Q and R directly, at a lower cost, without computing M^*M or its Choleski factorization. Moreover, the direct method, which we present now, is based on a theoretical observation: the QR factorization is nothing but the Gram–Schmidt orthonormalization procedure in $\mathbb{C}^n$, with respect to the canonical scalar product $\langle \cdot, \cdot \rangle$. In fact, giving M in $\mathbf{GL}_n(\mathbb{C})$ amounts to giving a basis $\{V^1, \ldots, V^n\}$ of $\mathbb{C}^n$, where $V^1, \ldots, V^n$ are the column vectors of M. If $Y^1, \ldots, Y^n$ denote the column vectors of Q, then $\{Y^1, \ldots, Y^n\}$ is an orthonormal basis. If $M = QR$, then

$$V^k = \sum_{j=1}^{k} r_{jk} Y^j.$$

Denoting by E_k the linear subspace spanned by $Y^1, \ldots, Y^k$, of dimension k, one sees that $V^1, \ldots, V^k$ are in E_k. Hence $\{V^1, \ldots, V^k\}$ is a basis of E_k. The columns of Q are therefore obtained by the Gram–Schmidt procedure, applied to the columns of M : Y^k is a unitary vector in E_k, orthogonal to E_{k-1}, where $E_k := \mathrm{Span}(V^1, \ldots, V^k)$.

The practical computation of Q and R is done by induction on k. If $k = 1$, then

$$r_{11} = \|V^1\|, \quad Y^1 = \frac{1}{r_{11}} V^1.$$

If $k > 1$, and if $Y^1, \ldots, Y^{k-1}$ are already known, one looks for Y^k and the entries r_{jk} ($j \leq k$). For $j < k$, we have

$$r_{jk} = \langle V^k, Y^j \rangle.$$

Then

$$r_{kk} = \|Z_k\|, \quad Y^k = \frac{1}{r_{kk}} Z^k,$$

where

$$Z^k := V^k - \sum_{j=1}^{k-1} r_{jk} Y^j.$$

Let us examine the complexity of the procedure described above. To pass from the step $k-1$ to the step k, one computes $k-1$ scalar products, then Z^k, its norm, and finally Y^k. This requires $(4n-1)k+3n$ operations. Summing from $k=1$ to n yields $2n^3 + O(n^2)$ operations. This method is not optimal, as we show in Section 13.3.3.

The interest of this construction lies also in giving a more complete statement than Proposition 11.4.

Theorem 11.3 *Let $M \in \mathbf{M}_n(\mathbb{C})$ be a matrix of rank p. There exists $Q \in \mathbf{U}_n$ and an upper-triangular matrix R, with $r_{\ell\ell} \in \mathbb{R}^+$ for every ℓ and $r_{jk} = 0$ for $j > p$, such that $M = QR$.*

Remark

The QR factorization of a singular matrix (i.e., a noninvertible one) is not unique. There exists, in fact, a QR factorization for rectangular matrices in which R is a "quasi-triangular" matrix.

11.4 Singular Value Decomposition

As we show in Exercise 14 below (see also Exercise 11 of Chapter 7), the eigenvalues of the matrix $H = \sqrt{M^*M}$, the Hermitian factor in the polar decomposition of a nonsingular matrix $M \in \mathbf{M}_n(\mathbb{C})$, are of some practical importance. They are called the *singular values* of M. These are the square roots of the eigenvalues of M^*M, thus one may even speak of the singular values of an arbitrary matrix, neither an invertible, nor even a square one. Recalling that (see Exercise 14 in Chapter 3) when M is $n \times m$, M^*M and MM^* have the same nonzero eigenvalues, counting them with multiplicities, one may even speak of the singular values of a rectangular matrix, up to an ambiguity concerning the multiplicity of the eigenvalue 0.

The main result of this section is the following.

Theorem 11.4 *Let $M \in \mathbf{M}_{n\times m}(\mathbb{C})$ be given. There exist two unitary matrices $U \in \mathbf{U}_n$, $V \in \mathbf{U}_m$ and a quasi-diagonal matrix*

$$D = \begin{pmatrix} s_1 & & & & \\ & \ddots & & & \\ & & s_r & & \\ & & & 0 & \\ & & & & \ddots \end{pmatrix},$$

with $s_1, \ldots, s_r \in (0, +\infty)$ and $s_1 \geq \cdots \geq s_r$, such that $M = UDV$. The numbers r and $s_1, \ldots, s_r$ are uniquely defined; they are respectively the rank and the nonzero singular values of M.

If $M \in \mathbf{M}_{n\times m}(\mathbb{R})$, one may choose U, V to be real orthogonal.

We notice that although D is uniquely defined, the other factors U and V are not unique. For instance, $M = I_n$ yields $D = I_n$ and $V = U^*$, where U can be an arbitrary unitary matrix.

Proof. To begin with, let us recall the following facts. We have

$$\mathbb{C}^n = R(M) \oplus^{\perp} \ker M^*, \qquad \ker MM^* = \ker M^*, \qquad R(M) = R(MM^*), \quad (11.4)$$

and on the other hand

$$\mathbb{C}^m = \ker(M) \oplus^{\perp} R(M^*), \qquad R(M^*) = R(M^*M), \qquad \ker M = \ker M^*M. \quad (11.5)$$

Inasmuch as MM^* is positive-semidefinite, we may write its eigenvalues as

$$s_1^2, \ldots, s_r^2, 0, \ldots,$$

where the s_js, the singular values of M, are positive real numbers arranged in decreasing order. The spectrum of M^*M has the same form, except for the multiplicity of 0. The index r is the rank of MM^*, that is, that of M, or as well that of M^*. The multiplicities of 0 as an eigenvalue of M^*M and MM^*, respectively, differ by $n-m$, whereas the multiplicities of other eigenvalues are the same for both matrices. We set $S = \operatorname{diag}(s_1, \ldots, s_r)$.

Because MM^* is hermitian, there exists an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ of $\mathbb{C}^n$ consisting of eigenvectors associated with the s_j^2s, followed by vectors of $\ker M^*$ (because of (11.4)). Let us form the unitary matrix

$$U = (\mathbf{u}_1 | \ldots | \mathbf{u}_n).$$

Written blockwise, we have $U = (U_R, U_K)$, where

$$MM^*U_R = U_R S^2, \quad M^*U_K = 0.$$

Let us define $V_R := M^*U_R S^{-1}$. From above, we have

$$V_R^* V_R = S^{-1} U_R^* MM^* U_R S^{-1} = I_r.$$

This means that the columns $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of V_R constitute an orthonormal family. Obviously, it is included in $R(M^*)$.

Because $\dim R(M^*) = r$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ form an orthonormal basis of this space and can be extended to an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ of $\mathbb{C}^m$, where $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_m$ belong to $\ker M$ (because of (11.5)). Let $V =: (V_R, V_K)$ be the unitary matrix whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

We compute blockwise the product U^*MV. From $MV_K = 0$ and $M^*U_K^* = 0$, we get

$$U^*MV = \begin{pmatrix} U_R^*MV_R & 0 \\ 0 & 0 \end{pmatrix}.$$

Finally, we obtain

$$U_R^*MV_R = U_R^*MM^*U_RS^{-1} = U_R^*U_RS = S.$$

□

11.5 The Moore–Penrose Generalized Inverse

The resolution of a general linear system $Ax = b$, where A may be singular and may even not be square, is a delicate question, whose treatment is made much simpler by the use of the Moore–Penrose generalized inverse.

Theorem 11.5 *Let $A \in M_{n\times m}(\mathbb{C})$ be given. There exists a unique matrix $A^\dagger \in M_{m\times n}(\mathbb{C})$, called the Moore–Penrose* generalized inverse*, satisfying the following four properties.*

1. $AA^\dagger A = A$.
2. $A^\dagger AA^\dagger = A^\dagger$.
3. $AA^\dagger \in H_n$.
4. $A^\dagger A \in H_m$.

Finally, if A has real entries, then so has $A^\dagger$.

When $A \in \mathbf{GL}_n(\mathbb{C})$, $A^\dagger$ coincides with the standard inverse A^{-1}, because the latter obviously satisfies the four properties. More generally, if A is onto, then Property 1 shows that $AA^\dagger = I_n$, (i.e. $A^\dagger$ is a right inverse of A). Likewise, if A is one-to-one, then $A^\dagger A = I_m$, (i.e. $A^\dagger$ is a left inverse of A).

Proof. We first remark that if X satisfies these four properties, and if $U \in \mathbf{U}_n$, $V \in \mathbf{U}_m$, then V^*XU^* is a generalized inverse of UAV. Therefore, existence and uniqueness need to be proved for only a single representative D of the equivalence class of A modulo unitary multiplications on the right and the left. From Theorem 11.4, we may choose a quasi-diagonal matrix D, with given $s_1, \ldots, s_r$, the nonzero singular values of A.

Let $D^\dagger$ be any generalized inverse of D, which we write blockwise as

$$D^\dagger = \begin{pmatrix} G & H \\ J & K \end{pmatrix}$$

with $G \in \mathbf{M}_r(\mathbb{C})$. From Property 1, we obtain $S = SGS$, where $S := \mathrm{diag}(s_1, \ldots, s_r)$. Inasmuch as S is nonsingular, we obtain $G = S^{-1}$. Next, Property 3 implies $SH = 0$, that is, $H = 0$. Likewise, Property 4 gives $JS = 0$, that is, $J = 0$. Finally, Property 2 yields $K = JSH = 0$. We see, then, that $D^\dagger$ must equal (uniqueness)

$$\begin{pmatrix} S^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

One easily checks that this matrix solves our problem (existence). □

Some obvious properties are stated in the following proposition. We warn the reader that, contrary to what happens for the standard inverse, the generalized inverse of AB does not equal $B^\dagger A^\dagger$ in general.

Proposition 11.5 *The following equalities hold for the generalized inverse:*

$$(\lambda A)^\dagger = \frac{1}{\lambda} A^\dagger \quad (\lambda \neq 0), \quad \left(A^\dagger\right)^\dagger = A, \quad \left(A^\dagger\right)^* = (A^*)^\dagger.$$

If $A \in GL_n(\mathbb{C})$, then $A^\dagger = A^{-1}$.

Because $(AA^\dagger)^2 = AA^\dagger$, the matrix $AA^\dagger$ is a projector, which can therefore be described in terms of its range and kernel. Because $AA^\dagger$ is Hermitian, these subspaces are orthogonal to each other. Obviously, $R(AA^\dagger) \subset R(A)$. But because $AA^\dagger A = A$, the reverse inclusion holds too. Finally, we have

$$R(AA^\dagger) = R(A),$$

and $AA^\dagger$ is the orthogonal projector onto $R(A)$. Likewise, $A^\dagger A$ is an orthogonal projector. Obviously, $\ker A \subset \ker A^\dagger A$, and the identity $AA^\dagger A = A$ implies the reverse inclusion, so that

$$\ker A^\dagger A = \ker A.$$

Finally, $A^\dagger A$ is the orthogonal projector onto $(\ker A)^\perp$.

11.5.1 Solutions of the General Linear System

Given a matrix $M \in M_{n\times m}(\mathbb{C})$ and a vector $b \in \mathbb{C}^n$, let us consider the linear system

$$Mx = b. \tag{11.6}$$

In (11.6), the matrix M need not be square, even not of full rank. From Property 1, a necessary condition for the solvability of (11.6) is $MM^\dagger b = b$. Obviously, this is also sufficient, because it ensures that $x_0 := M^\dagger b$ is a solution. Hence, the generalized

inverse plays one of the roles of the standard inverse, namely to provide one solution of (11.6) when it is solvable. To catch every solution of that system, it remains to solve the homogeneous problem $My = 0$. From the analysis done in the previous section, $\ker M$ is nothing but the range of $I_m - M^\dagger M$. Therefore, we may state the following proposition:

Proposition 11.6 *The system* (11.6) *is solvable if and only if* $b = MM^\dagger b$. *When it is solvable, its general solution is* $x = M^\dagger b + (I_m - M^\dagger M)z$, *where* z *ranges* $\mathbb{C}^m$. *Finally, the special solution* $x_0 := M^\dagger b$ *is the one of least Hermitian norm.*

There remains to prove that x_0 has the smallest norm among the solutions. That comes from the Pythagorean theorem and from the fact that $R(M^\dagger) = R(M^\dagger M) = (\ker M)^\perp$.

Exercises

1. Assume that there exists an algorithm for multiplying two $N \times N$ matrices with entries in a noncommutative ring by means of K multiplications and L additions. Show that the complexity of the multiplication in $\mathbf{M}_n(k)$ is $O(n^\alpha)$, with $\alpha = \log K / \log N$.
2. What is the complexity of Choleski factorization?
3. Let $M \in \mathbf{SPD}_n$ be also tridiagonal. What is the structure of L in the Choleski factorization? More generally, what is the structure of L when $m_{ij} = 0$ for $|i - j| > r$? (When $r \ll n$ we say that M is a *band matrix*.)
4. (Continuation of Exercise 3)
   For $i \leq n$, denote by $\phi(i)$ the smallest index j such that $m_{ij} \neq 0$. In Choleski factorization, show that $l_{ij} = 0$ for every pair (i, j) such that $j < \phi(i)$.
5. In the QR factorization, show that the map $M \mapsto (Q, R)$ is continuous on $\mathbf{GL}_n(\mathbb{C})$.
6. Let H be an $n \times n$ Hermitian matrix, that blockwise reads

$$H = \begin{pmatrix} A & B^* \\ B & C \end{pmatrix}.$$

   Assume that $A \in \mathbf{HPD}_{n-k}$ $(1 \leq k \leq n-1)$.
   Find a matrix T of the form

$$T = \begin{pmatrix} I_{n-k} & 0 \\ \cdot & I_k \end{pmatrix}$$

   such that THT^* is block-diagonal. Deduce that if $W \in \mathbf{H}_k$, then

$$H - \begin{pmatrix} 0 & 0 \\ 0 & W \end{pmatrix}$$

is positive-(semi)definite if and only if $S-W$ is so, where S is the Schur complement of A in H.

7. (Continuation of Exercise 6) Fix the size k. We keep $A \in \mathbf{HPD}_{n-k}$ constant and let vary B and C. We denote by $S(H)$ the Schur complement of A. Using the previous exercise, show that if $\lambda \in [0,1]$:

   a. $S(\lambda H + (1-\lambda)H') - \lambda S(H) - (1-\lambda)S(H')$ is positive-semidefinite.
   b. If $H - H'$ is positive-semidefinite, then so is $S(H) - S(H')$.

   In other words, $H \mapsto S$ is "concave nondecreasing" from the affine subspace formed of those matrices of $\mathbf{H}_n$ with prescribed $A \in \mathbf{HPD}_{n-k}$, into the ordered set $\mathbf{H}_k$.
8. In Proposition 11.4, find an alternative proof of the uniqueness part, by inspection of the spectrum of the matrix $Q := Q_2^{-1}Q_1 = R_2R_1^{-1}$.
9. Identify the generalized inverse of row matrices and column matrices.
10. What is the generalized inverse of an orthogonal projector, that is, an Hermitian matrix P satisfying $P^2 = P$? Deduce that the description of $AA^\dagger$ and $A^\dagger A$ as orthogonal projectors does not characterize $A^\dagger$ uniquely.
11. Given a matrix $B \in \mathbf{M}_{p\times q}(\mathbb{C})$ and a vector $a \in \mathbb{C}^p$, let us form the matrix $A := (B,a) \in \mathbf{M}_{p\times(q+1)}(\mathbb{C})$.

    a. Let us define $d := B^\dagger a$, $c := a - Bd$, and

    $$b := \begin{cases} c^\dagger, & \text{if } c \neq 0, \\ (1+|d|^2)^{-1}d^*B^\dagger, & \text{if } c = 0. \end{cases}$$

    Prove that

    $$A^\dagger = \begin{pmatrix} B^\dagger - db \\ b \end{pmatrix}.$$

    b. Deduce an algorithm (Greville's algorithm) in $O(pq^2)$ operations for the computation of the generalized inverse of a $p \times q$ matrix. **Hint:** To get started with the algorithm, use Exercise 9.

12. Let $A \in \mathbf{M}_n(\mathbb{C})$ be given, with eigenvalues λ_j and singular values σ_j, $1 \leq j \leq n$ (we include zeroes in this list if A is singular). We choose the decreasing orders:

    $$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|, \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n.$$

    Recall that the σ_js are the square roots of the eigenvalues of A^*A.
    We wish to prove the inequality

    $$\prod_{j=1}^{k} |\lambda_j| \leq \prod_{j=1}^{k} \sigma_j, \quad 1 \leq k \leq n.$$

    a. Directly prove the case $k = 1$. Show the equality in the case $k = n$.
    b. Working within the exterior algebra (see Chapter 4), we define an endomorphism $A^{\wedge p}$ over $\Lambda^p(\mathbb{C}^n)$ by

$$A^{\wedge p}(x_1 \wedge \cdots \wedge x_p) := (Ax_1) \wedge \cdots \wedge (Ax_p), \qquad \forall x_1, \ldots, x_p \in \mathbb{C}^n.$$

Prove that the eigenvalues of $A^{\wedge p}$ are the products of p terms λ_j with pairwise distinct indices. Deduce the value of the spectral radius.

c. Let $\{\mathbf{e}^{i_1}, \ldots, \mathbf{e}^{i_n}\}$ be the canonical basis of $\mathbb{C}^n$. We endow $\Lambda^p(\mathbb{C}^n)$ with the natural Hermitian norm in which the canonical basis made of $\mathbf{e}^{i_1} \wedge \cdots \wedge \mathbf{e}^{i_p}$ with $i_1 < \cdots < i_p$, is orthonormal. We denote by $\langle \cdot, \cdot \rangle$ the scalar product in $\Lambda^p(\mathbb{C}^n)$.
   i. If $x_1, \ldots, x_p, y_1, \ldots, y_p \in \mathbb{C}^n$, prove that

$$\langle x_1 \wedge \cdots \wedge x_p, y_1 \wedge \cdots \wedge y_p \rangle = \det(x_i^* y_j)_{1 \le i,j \le p}.$$

   ii. For $M \in \mathbf{M}_n(\mathbb{C})$, show that the Hermitian adjoint of $M^{\wedge p}$ is $(M^*)^{\wedge p}$.
   iii. If $U \in \mathbf{U}_n$, show that $U^{\wedge p}$ is unitary.
   iv. Deduce that the norm of $A^{\wedge p}$ equals $\sigma_1 \cdots \sigma_p$.

d. Conclude.

13. Let A, B, C be complex matrices of respective sizes $n \times r$, $s \times m$, and $n \times m$. Prove that the equation

$$AXB = C$$

is solvable if and only if

$$AA^\dagger CB^\dagger B = C.$$

In this case, verify that every solution is of the form

$$A^\dagger CB^\dagger + Y - A^\dagger AYBB^\dagger,$$

where Y is an arbitrary $r \times s$ matrix. We recall that $M^\dagger$ is the Moore–Penrose inverse of M.

14. The deformation of an elastic body is represented at each point by a square matrix $F \in \mathbf{GL}_3^+(\mathbb{R})$ (the sign $+$ expresses that $\det F > 0$). More generally, $F \in \mathbf{GL}_n^+(\mathbb{R})$ in space dimension n. The density of elastic energy is given by a function $F \mapsto W(F) \in \mathbb{R}^+$.

   a. The principle of frame indifference says that $W(QF) = W(F)$ for every $F \in \mathbf{GL}_n^+(\mathbb{R})$ and every rotation Q. Show that there exists a map $w : \mathbf{SPD}_n \to \mathbb{R}^+$ such that $W(F) = w(H)$, where $F = QH$ is the polar decomposition.
   b. When the body is isotropic, we also have $W(FQ) = W(F)$, for every $F \in \mathbf{GL}_n^+(\mathbb{R})$ and every rotation Q. Show that there exists a map $\phi : \mathbb{R}^n \to \mathbb{R}^+$ such that $W(F) = \phi(h_1, \ldots, h_n)$, where the h_j are the entries of the characteristic polynomial of H. In other words, $W(F)$ depends only on the singular values of F.

15. A matrix $A \in \mathbf{M}_n(\mathbb{R})$ is called a *totally positive* matrix when all minors

$$A \begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix}$$

with $1 \leq p \leq n$, $1 \leq i_1 < \cdots < i_p \leq n$ and $1 \leq j_1 < \cdots < j_p \leq n$ are positive.

a. Prove that the product of totally positive matrices is totally positive.
b. Prove that a totally positive matrix admits an LU factorization and that every "nontrivial" minor of L and U is positive. Here, "nontrivial" means

$$L \begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix}$$

with $1 \leq p \leq n$, $1 \leq i_1 < \cdots < i_p \leq n$, $1 \leq j_1 < \cdots < j_p \leq l$, and $i_s \geq j_s$ for every s, because every other minor vanishes trivially. For U, read $i_s \leq j_s$ instead. **Note**: One says that L and U are *triangular totally positive*.
c. Show that a Vandermonde matrix (see Exercise 17 of Chapter 3) is totally positive whenever $0 < a_1 < \cdots < a_n$.