Numerical Linear Algebra
Chap. 1: Basic Concepts from Linear Algebra

Heinrich Voss

voss@tu-harburg.de

Hamburg University of Technology
Institute for Numerical Simulation

TUHH
Technische Universität Hamburg-Harburg

Vectors

The vector space $\mathbb{R}^n$ is defined by

$$\mathbb{R}^n := \{(x_1, \ldots, x_n)^T : x_j \in \mathbb{R},\ j = 1, \ldots, n\},$$

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}, \quad \alpha \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{pmatrix}$$

and $\mathbb{C}^n$ correspondingly.

A subset $X \subset \mathbb{R}^n$ is a subspace of $\mathbb{R}^n$ if it is closed with respect to addition and multiplication by scalars, i.e.

$$x, y \in X \quad \Rightarrow \quad x + y \in X,$$

$$\alpha \in \mathbb{R},\ x \in \mathbb{R}^n \quad \Rightarrow \quad \alpha x \in X.$$

Subspaces

A set of vectors $\{a^1, \ldots, a^m\} \subset \mathbb{C}^n$ is linearly independent if

$$\sum_{j=1}^{m} \alpha_j a^j = 0 \quad \Rightarrow \quad \alpha_j = 0,\ j = 1, \ldots, m.$$

Otherwise, a nontrivial combination of the a^j is zero, and $\{a^1, \ldots, a^m\}$ is said to be linearly dependent.

Given vectors $a^1, \ldots, a^m$ the set of all linear combinations of these vectors is a subspace referred to as the span of $a^1, \ldots, a^m$:

$$\operatorname{span}\{a^1, \ldots, a^m\} = \left\{ \sum_{j=1}^{m} \alpha_j a^j \ : \ \alpha_j \in \mathbb{C} \right\}.$$

Subspaces ct.

If $\{a^1, \ldots, a^m\}$ is linearly independent and $b \in \text{span}\{a^1, \ldots, a^m\}$, then b has a unique representation as linear combination of the a^j.

If $S_1, \ldots, S_k$ are subspace of $\mathbb{C}^n$ then their sum

$$S := S_1 + \cdots + S_k := \left\{ \sum_{j=1}^{k} a^j \ : \ a^j \in S_j, \ j = 1, \ldots, k \right\}$$

is also a subspace of $\mathbb{C}^n$. S is said to be the direct sum if each $x \in S$ has a unique representattion $x = a^1 + \cdots + a^k$ with $a^j \in S_j$. In this case we write

$$S = S_1 \oplus \cdots \oplus S_k.$$

The intersection of the subspaces $S_1, \ldots, S_k$ is also a subspace,

$$S = S_1 \cap \cdots \cap S_k.$$

Dimension

The subset $\{a^{i_1}, \ldots, a^{i_k}\}$ is a maximal linearly independent subset of $\{a^1, \ldots, a^m\}$ if it is linearly independent and is not contained properly in any linearly independent subset of $\{a^1, \ldots, a^m\}$.
If $\{a^{i_1}, \ldots, a^{i_k}\}$ is a maximal linearly independent subset then

$$\operatorname{span}\{a^{i_1}, \ldots, a^{i_k}\} = \operatorname{span}\{a^1, \ldots, a^m\}.$$

If $S \subset \mathbb{C}^n$ is a subspace it is always possible to find a maximal linearly independent subset $\{a^1, \ldots, a^k\}$. Then $S = \operatorname{span}\{a^1, \ldots, a^m\}$, and $\{a^1, \ldots, a^m\}$ is called a basis of S.

All bases for a subspace S have the same number of elements. This number is the dimension of S, and it is denoted by $\dim(S)$.

Linear map

A map $A : \mathbb{C}^n \to \mathbb{C}^m$ is called linear, if

$$A(x+y) = Ax+Ay \quad \text{for every } x, y \in \mathbb{C}^n \quad \text{and} \quad A(\lambda x) = \lambda Ax \quad \text{for every } x \in \mathbb{C}^n,$$

For $j = 1, \ldots, n$ let $e_j \in \mathbb{C}^n$ be the j-th canonical unit vector having a 1 in its j-th component and zeros elsewhere. Then for $x = (x_j)_{j=1,\ldots,n} \in \mathbb{C}^n$

$$Ax = A\Big(\sum_{j=1}^{n} x_j e_j\Big) = \sum_{j=1}^{n} A(x_j e_j) = \sum_{j=1}^{n} x_j Ae_j =: \sum_{j=1}^{n} x_j a_j$$

Hence, the images $a_j := Ae_j$ of the canonical basis vectors characterize the linear map A. We therefore identify A with the $m \times n$ matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{2m} & \ldots & a_{mn} \end{pmatrix}.$$

Matrix-vector product

For $A = (a_{jk}) \in \mathbb{C}^{m\times n}$ and $x = (x_k) \in \mathbb{C}^n$ we have

$$Ax = \left(\sum_{k=1}^{n} a_{jk} x_k \right)_{j=1,\ldots,m} =: b \in \mathbb{C}^m.$$

The vector b is called matrix-vector product of A and x.

For every $x \in \mathbb{C}^n$ the matrix-vector product $b = Ax$ is a linear combination of the columns a_j of the matrix A.

Matrix-matrix product

Let $A : \mathbb{C}^n \to \mathbb{C}^m$ and $B : \mathbb{C}^m \to \mathbb{C}^p$. Then the composition

$$B \circ A : \mathbb{C}^n \to \mathbb{C}^p, \ (B \circ A)x = B(Ax)$$

is linear as well, and

$$BAx = B(Ax) = B\Big(\sum_{k=1}^{n} a_{jk}x_k\Big) = \sum_{j=1}^{m} b_{ij}\Big(\sum_{k=1}^{n} a_{jk}x_k\Big) = \sum_{k=1}^{n}\Big(\sum_{j=1}^{m} b_{ij}a_{jk}\Big)x_k.$$

Hence, the composit map of B and A is represented by the matrix-matrix product $C := BA \in \mathbb{C}^{p\times n}$ with elements

$$c_{ik} = \sum_{j=1}^{m} b_{ij}a_{jk}, \ i = 1,\ldots,p, \ k = 1,\ldots,n.$$

Notice that the matrix-matrix product of B and A is only defined if the number of columns of B equals the number of rows of A.

Range of a matrix

The range of a matrix A, written range(A), is the set of vectors that can be expressed as Ax for some x. The formula $b = Ax$ leads naturally to the following characterization of range(A).

Theorem 1
range(A) is the space spanned by the columns of A.

Proof
$Ax = \sum_{j=1}^{n} a_j x_j$ is a linear combination of the columns a_j of A.

Conversely, any vector y in the space spanned by the columns of A can be written as a linear combination of the columns, $y = \sum_{j=1}^{n} a_j x_j$. Forming a vector x out of the coefficients x_j, we have $y = Ax$, and thus y is in the range of A.

In view of Theorem 1, the range of a matrix A is also called the column space of A.

Nullspace of A

The nullspace of $A \in \mathbb{C}^{m\times n}$, written null($A$), is the set of vectors x that satisfy $Ax = 0$, where 0 is the 0-vector in $\mathbb{C}^n$.

The entries of each vector $x \in$ null(A) give the coefficients of an expansion of zero as a linear combination of columns of A:

$$0 = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n.$$

The column rank of a matrix is the dimension of its column space.

Similarly, the row rank of a matrix is the dimension of the space spanned by its rows.

Row rank always equals column rank (among other proofs, this is a corollary of the singular value decomposition, discussed later), so we refer to this number simply as the rank of a matrix.

Full rank

An $m \times n$ matrix of full rank is one that has the maximal possible rank (the lesser of m and n).

This means that a matrix of full rank with $m \geq n$ must have n linearly independent columns. Such a matrix can also be characterized by the property that the map it defines is one-to-one.

Theorem 2 A matrix $A \in \mathbb{C}^{m \times n}$ with $m \geq n$ has full rank if and only if it maps no two distinct vectors to the same vector.

Proof If A is of full rank, its columns are linearly independent, so they form a basis for range(A). This means that every $b \in$ range(A) has a unique linear expansion in terms of the columns of A, and therefore, every $b \in$ range(A) has a unique linear expansion in terms of the columns of A, and thus, every $b \in$ range(A) has a unique x such that $b = Ax$.

Conversely, if A is not of full rank, its columns a_j are dependent, and there is a nontrivial linear combination such that $\sum_{j=1}^{n} c_j a_j = 0$. The nonzero vector c formed from the coefficients c_j satisfies $Ac = 0$. But then A maps distinct vectors to the same vector since, for any x it holds that $Ax = A(x + c)$.

Inverse matrix

A nonsingular or invertible matrix is a square matrix of full rank.

Note that the m columns of a nonsingular $m \times m$ matrix A form a basis for the whole space $\mathbb{C}^m$. Therefore, we can uniquely express any vector as a linear combination of them.

In particular, the canonical unit vector e_j, can be expanded:

$$e_j = \sum_{i=1}^{m} z_{ij} a_j.$$

Let $Z \in \mathbb{C}^{m \times m}$ be the matrix with entries z_{ij}, and let z_j denote the jth column of Z. Then it holds $e_j = Az_j$, and putting these vectos together

$$AZ = (e_1, \ldots, e_m) =: I$$

where I ist the $m \times m$ identity. Z is the inverse of A, and is denoted by $Z =: A^{-1}$.

Gaussian elimination

The simplest way to solve a linear system (by hand or on a computer) is Gaussian elimination.

It transforms a linear system to an equivalent one with upper-triangular system matrix by applying simple linear transformations.

Let $A \in \mathbb{C}^{m\times n}$ be given. The idea is to transform A into an upper-triangular matrix by introducing zeros below the diagonal, first in column 1, then in column 2 , etc. This is done by subtracting suitable multiples of each row from the subsequent ones.

This eleimination process is equivalent to multiplying A by a sequence of lower triangular matrices L_j on the left:

$$L_{n-1} L_{n-2} \cdots L_1 A = U.$$

LU factorization

Setting $L := L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1}$ gives $A = LU$. Thus we obtain an LU factorization of A

$$A = LU,$$

where U is upper-triangular, and L is (as a product of lower-triangular matrices) lower-triangular.

It turns out that L can be chosen such that all diagonal entries are equal to 1. A matrix with this property is called unit lower-triangular.

Example

$$A = \begin{pmatrix} 2 & 1 & 3 & 4 \\ -2 & 1 & -1 & -2 \\ 4 & 4 & 5 & 11 \\ -2 & 1 & -7 & -1 \end{pmatrix}$$

The first step of Gaussian elimination looks like this: The first row is added to the second one, twice the first row is subtracted from the third one, and the first row is added to third on. This can be written as

$$L_1 A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ -2 & 1 & -1 & -2 \\ 4 & 4 & 5 & 11 \\ -2 & 1 & -7 & -1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & 2 & -4 & 3 \end{pmatrix}$$

Next we subtract the second row from the third and the fourth row:

$$L_2 L_1 A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & 2 & -4 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & -3 & 1 \\ 0 & 0 & -6 & 1 \end{pmatrix}$$

Example ct.

$$L_2L_1A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & 2 & -4 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & -3 & 1 \\ 0 & 0 & -6 & 1 \end{pmatrix}$$

Finally we subtract twice the third row from the fourth row

$$L_3L_2L_1A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & -3 & 1 \\ 0 & 0 & -6 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & -3 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

To exhibit the full factorization $A = LU$ we need to compute the product $L = L_1^{-1}L_2^{-1}L_3^{-1}$.

Surprisingly, this turns out to be trivial. The inverse of L_j, $j = 1, 2, 3$ is just L_j itself, but with each entry below the diagonal negated:

Example ct.

$$L_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}, \ldots$$

The product $L_1^{-1}L_2^{-1}L_3^{-1}$ is just the unit lower-triangular matrix with the nonzero subdiagonal entries of L_1^{-1}, L_2^{-1} and L_3^{-1} inserted in the appropriate places:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ -1 & 1 & 2 & 1 \end{pmatrix}.$$

Together we have

$$A = \begin{pmatrix} 2 & 1 & 3 & 4 \\ -2 & 1 & -1 & -2 \\ 4 & 4 & 5 & 11 \\ -2 & 1 & -7 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ -1 & 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & -3 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix} = LU.$$

General case

Suppose x_k is the kth column of the matrix at the beginning of step k. Then the transformation has to be chosen such that

$$x_k = \begin{pmatrix} x_{1k} & \\ \vdots & x_{kk} \\ x_{k,k+1} & \\ \vdots & \\ x_{mk} & \end{pmatrix} \rightarrow \begin{pmatrix} x_{1k} & \\ \vdots & x_{kk} \\ 0 & \\ \vdots & \\ 0 & \end{pmatrix}.$$

To this end we subtract ℓ_{jk} times row k from row j with $\ell_{jk} = x_{jk}/x_{kk}$, $k < j \leq m$, which is performed multiplying by

$$L_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -\ell_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -\ell_{mk} & & 1 \end{pmatrix}.$$

General case ct.

In the numerical example, we noted that L_k can be inverted by negating its subdiagonal entries, and that L can be formed by collecting the entries ℓ_{jk} in the appropriate places. These observations are true in the general case.

With

$$\ell_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \ell_{k+1,k} \\ \vdots \\ \ell_{mk} \end{pmatrix}$$

the matrix L_k can be written $L_k = I - \ell_k e_k^H$, where e_k is the kth standard unit vector. The sparsity pattern of ℓ_k implies $e_k^H \ell_k = 0$, and therefore

$$(I - \ell_k e_k^H)(I + \ell_k e_k^H) = I - \ell_k e_k^H \ell_k e_k^H = I.$$

In other words, the inverse of L_k is $I + \ell_k e_k^H$.

General case ct.

That $L = L_1^{-1} \cdots L_m^{-1}$ can be formed by collecting the entries ℓ_{jk} in the appropriate places is proved by induction.

Assume that

$$L_1^{-1} \cdots L_k^{-1} = I + \sum_{j=1}^{k} \ell_j e_j^H.$$

Then it follws from $\ell_j e_j^H \ell_{k+1} = 0$ for $j = 1, \ldots, k$

$$\begin{aligned} L_1^{-1} \cdots L_k^{-1} L_{k+1}^{-1} &= (I + \sum_{j=1}^{k} \ell_j e_j^H)(I + \ell_{k+1} e_{k+1}^H \\ &= I + \sum_{j=1}^{k+1} \ell_j e_j^H + \sum_{j+1}^{k} \ell_j e_j^H \ell_{k+1} e_{k+1}^H. = I + \sum_{j=1}^{k+1} \ell_j e_j^H. \end{aligned}$$

Gaussian elimination

In practical Gaussian elimination, the matrices L_k are never formed and multiplied explicitly. The multipliers ℓ_k are computed and stored directly into L, and the transformations L_k are then applied implicitly.

Gaussian elimination without pivoting

$U = A$, $L = I$
for k=1:m-1 **do**
  for j=k+1:m **do**
    $\ell_{jk} = u_{jk} / u_{kk}$
    $u_{j,k:m} = u_{j,k:m} - \ell_{jk} u_{k,k_m}$
  end for
end for

Three matrices A, L, U are not really needed; to minimize memory use on the computer, both L and U can be written into the same array as A.

Linear systems

If A is factored into L and U, a system of equations $Ax = b$ is reduced to the form $LUx = b$.

Thus it can be solved by solving two triangular systems: first $Ly = b$ for the unknown y (forward substitution), then $Rx = y$ for the unknown x (back substitution).

This is particularly advantageous, if several linear systems with the same system matrix have to be solved.

Failure of Gaussian elimination

Unfortunately, Gaussian elimination as presented so far is unusable for solving general linear systems, for it is not stable.

The instability is related to another, more obvious difficulty. For certain matrices, Gaussian elimination fails entirely, because it attempts division by zero. For example, consider

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

This matrix has full rank. Nevertheless, Gaussian elimination fails at the first step.

Pivoting

At step k of Gaussian elimination, multiples of row k are subtracted from rows $k+1, \ldots, m$ of the working matrix X in order to introduce zeros in entry k of these rows.
In this operation row k, column k, and especially the entry x_{kk} play special roles. We call x_{kk} the pivot.

From every entry in the submatrix $X_{k+l:m,k:m}$ is subtracted the product of a number in row k and a number in column k, divided by x_{kk}

However, there is no reason why the kth row and column must be chosen for the elimination. For example, we could just as easily introduce zeros in column k by adding multiples of some row i with $k < i < m$ to the other rows.

Similarly, we could introduce zeros in column j rather than column k to eliminate in linear system with system matrix the unknown x_j from all remaining equations but one.

Pivoting ct.

All in all, we are free to choose any entry of $X_{k:m,k:m}$ as the pivot, as long as it is nonzero. The possibility that an entry $X_{kk} = 0$ might arise implies that some flexibility of choice of the pivot may sometimes be necessary, even from a pure mathematical point of view.

For numerical stability, however, it is desirable to pivot even when x_{kk} is nonzero if there is a larger element available. In practice, it is common to pick as pivot the largest number among a set of entries being considered as candidates.

The structure of the elimination process quickly becomes confusing if zeros are introduced in arbitrary patterns through the matrix. To see what is going on, we want to retain the triangular structure, and there is an easy way to do this. We shall not think of the pivot x_{ij} as left in place. Instead, at step k, we shall imagine that the rows and columns of the working matrix are permuted so as to move x_{ij} into the (k, k) position. Then, when the elimination is done, zeros are introduced into entries $k + 1, \ldots, m$ of column k, just as in Gaussian elimination without pivoting. This interchange of rows and perhaps columns is what is usually thought of as pivoting.

Partial pivoting

If every entry of $X_{k:m,k:m}$ is considered as a possible pivot at step k, there are $(m-k)^2$ entries to be examined to determine the largest. This expensive strategy is called complete pivoting.

In practice, equally good pivots can be found by considering a much smaller number of entries. The standard method for doing this is partial pivoting. Here, only rows are interchanged.

The pivot at each step is chosen as the largest of the $m-k+1$ subdiagonal entries in column k. To bring the kth pivot into the (k,k) position, no columns need to be permuted; only row k is swapped with the row containing the pivot.

As usual in numerical linear algebra, this algorithm can be expressed as a matrix product. We saw in the last lecture that an elimination step corresponds to left-multiplication by an elementary lower-triangular matrix L_k. Partial pivoting complicates matters by applying a permutation matrix P_k on the left of the working matrix before each elimination. After $m-1$ steps, A becomes an upper-triangular matrix U:

$$L_{m-1}P_{m-1}L_{m-2}\cdots L_2P_2L_1P_1A = U.$$

Example

$$A = \begin{pmatrix} 2 & 1 & 3 & 4 \\ -2 & 1 & -1 & -2 \\ 4 & 4 & 5 & 11 \\ -2 & 1 & -7 & -1 \end{pmatrix}$$

Since $|a_{31}| \geq |a_{j1}|$ for $j = 1, 2, 3, 4$ we interchange rows three and one:

$$P_1 A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ -2 & 1 & -1 & -2 \\ 4 & 4 & 5 & 11 \\ -2 & 1 & -7 & -1 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 5 & 11 \\ -2 & 1 & -1 & -2 \\ 2 & 1 & 3 & 4 \\ -2 & 1 & -7 & -1 \end{pmatrix}$$

Next we eliminate the subdiagonal elements of the first column

$$L_1 P_1 A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ -1/2 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 4 & 5 & 11 \\ -2 & 1 & -1 & -2 \\ 2 & 1 & 3 & 4 \\ -2 & 1 & -7 & -1 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & -1 & 1/2 & -3/2 \\ 0 & 3 & -9/2 & 9/2 \end{pmatrix}$$

Example ct.

$|x_{22}|$ is already maximal in the second column; no permutation is necessary

$$L_2L_1P_1A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1/3 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & -1 & 1/2 & -3/2 \\ 0 & 3 & -9/2 & 9/2 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & 0 & 1 & -1/3 \\ 0 & 0 & -6 & 1 \end{pmatrix}$$

P_3 permutes rows four and three

$$P_3L_2L_1P_1A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & 0 & 1 & -1/3 \\ 0 & 0 & -6 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & 0 & -6 & 1 \\ 0 & 0 & 1 & -1/3 \end{pmatrix}$$

And the final elimination step yields

$$P_3L_2L_1P_1A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/6 & 1 \end{pmatrix} \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & 0 & -6 & 1 \\ 0 & 0 & 1 & -1/3 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & 0 & -6 & 1 \\ 0 & 0 & 0 & -1/6 \end{pmatrix}$$

$PA = LU$

The matrix equation $L_{m-1}P_{m-1}L_{m-2}\cdots L_2P_2L_1P_1A = U$ in our example reads

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 1/2 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 1/ & -1/3 & 1/6 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ -2 & 1 & -1 & -2 \\ 4 & 4 & 5 & 11 \\ -2 & 1 & -7 & -1 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & 0 & -6 & 1 \\ 0 & 0 & 0 & -1/6 \end{pmatrix}$$

If we knew in advance the permutations of the rows which are performed in the course of the Gaussian elimination we could apply these permutations first (which corresponds to one matrix multiplication of A by a permutation matrix P) and determine the LU factorization of PA without pivoting:

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 & 4 \\ -2 & 1 & -1 & -2 \\ 4 & 4 & 5 & 11 \\ -2 & 1 & -7 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 \\ -1/1 & 1 & 1 & 0 \\ 1/ & -1/3 & -1/6 & 1 \end{pmatrix} \begin{pmatrix} 4 & 4 & 5 & 11 \\ 0 & 3 & 3/2 & 7/2 \\ 0 & 0 & -6 & 1 \\ 0 & 0 & 0 & -1/6 \end{pmatrix}$$

$PA = LU$ ct.

The factorization $PA = LU$ can be determined in the course of the Gaussian elimination not knowing the permutations in advance.

Gaussian elimination generates the decomposition

$$L_{m-1}P_{m-1}L_{m-2}P_{m-2}L_{m-3}\cdots L_2P_2L_1P_1A = U.$$

The lefthand side can be rewritten as

$$\begin{aligned}
&L_{m-1}P_{m-1}L_{m-2}P_{m-1}^{-1}P_{m-1}P_{m-2}L_{m-3}\cdots L_2P_2L_1P_1A\\
&= L_{m-1}L'_{m-2}P_{m-1}P_{m-2}L_{m-3}\cdots L_2P_2L_1P_1A\\
&=: L_{m-1}L'_{m-2}P_{m-1}P_{m-2}L_{m-3}P_{m-2}^{-1}P_{m-1}^{-1}P_{m-1}P_{m-2}P_{m-3}L_{m-4}\cdots L_2P_2L_1P_1A\\
&=: L_{m-1}L'_{m-2}L'_{m-3}P_{m-1}P_{m-2}P_{m-3}L_{m-4}\cdots L_2P_2L_1P_1A = \cdots\\
&= (L_{m-1}L'_{m-2}\cdots L'_1)(P_{m-1}P_{m-2}\cdots P_1)A = U
\end{aligned}$$

with

$$L'_k = P_{m-1}P_{m-2}\cdots P_{k+1}L_kP_{k+1}^{-1}\cdots L_{m-2}^{-1}L_{m-1}^{-1},\ k = 1,\ldots,m-2.$$

$PA = LU$ ct.

Multiplying L_k by P_{k+1} on the left exchanges rows $k+1$ and ℓ for some *ell* $> k+1$, and multiplying by P_{k+1}^{-1} on the right exchanges columns $k+1$ and ℓ. Hence, $P_{k+1} L_k P_{k+1}^{-1}$ has the same structure as L_k, and this structure is kept when multiplying with further permutations $P_{k+2}, \ldots, P_{m-1}$ and their inverses on the left and right, respectively.

The matrices L'_k are unit lower-triangular and easily invertible by negating the subdiagonal entries, just as in Gaussian elimination without pivoting.

Writing $L = (L_{m-1} L'_{m-2} \cdots L'_1)^{-1}$ and $P = P_{m-1} \cdots P_2 P_1$, we have

$$PA = LU.$$

Gaussian elimination

Gaussian elimination with partial pivoting

$U = A$, $L = I$, $P = I$
for k=1:m-1 **do**
  selct $i \geq k$ to maximize $|u_{ik}|$
  $u_{k,k:m} \leftrightarrow u_{i,k:m}$
  $\ell_{k,1:k-1} \leftrightarrow \ell_{i,1:k-1}$
  $p_{k,:} \leftrightarrow p_{i,:}$
  for j=k+1:m **do**
    $\ell_{jk} = u_{jk} / u_{kk}$
    $u_{j,k:m} = u_{j,k:m} - \ell_{jk} u_{k,k_m}$
  end for
end for

Adjoint matrix

The complex conjugate of a scalar $z \in \mathbb{C}$, written $\overline{z}$ or z^H, is obtained by negating its imaginary part. For real $z \in \mathbb{R}$, we have $\overline{z} = z$.

The Hermitian conjugate or adjoint of an $m \times n$ matrix $A \in \mathbb{C}^{m \times n}$, written A^H, is the $n \times m$ matrix whose (i, j) entry is the complex conjugate of the (j, i) entry of A, i.e.

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \quad \Rightarrow \quad A^H = \begin{pmatrix} \overline{a_{11}} & \overline{a_{21}} \\ \overline{a_{12}} & \overline{a_{22}} \\ \overline{a_{13}} & \overline{a_{23}} \end{pmatrix}$$

If $A = A^H$, then A is called Hermitian. By definition, a Hermitian matrix must be square.

For real A, the adjoint simply interchanges the rows and columns of A. In this case, the adjoint is also known as the transpose, and is written A^T. If a real matrix is Hermitian, that is, $A = A^T$, then it is also said to be symmetric.

Inner product

The inner product of two column vectors $x, y \in \mathbb{C}^n$ is the product of the adjoint of x by y:

$$x^H y := \sum_{j=1}^{n} \overline{x_j} y_j.$$

The Euklidean length of of a vector $x \in \mathbb{C}^n$ is written $\|x\|$, and can be defined as the square root of the inner product of x with itself:

$$\|x\| = \sqrt{x^H x} = \sqrt{\sum_{j=1}^{n} |x_j|^2}.$$

The cosine of the angle ϕ between x and y can be expressed in terms of the inner product as

$$\cos\phi = \frac{x^H y}{\|x\| \cdot \|y\|}.$$

Orthogonal vectors

A pair of vectors x and y are orthogonal if $x^H y = 0$. If x and y are real, this means they lie at right angles to each other in $\mathbb{R}^n$.

Two sets of vectors X and Y are orthogonal (also stated "X is orthogonal to Y") if every $x \in X$ is orthogonal to every $y \in Y$.

A set of nonzero vectors S is orthogonal if its elements are pairwise orthogonal, i.e.,

$$x, y \in S,\ x \neq y \quad \Rightarrow \quad x^H y = 0.$$

A set of vectors is orthonormal if it is orthogonal and, in addition, every $x \in S$ has $\|x\| = 1$.

Orthogonal vectors ct.

Theorem
The vectors in an orthogonal set S are linearly independent.

Proof: For $v_1, \ldots, v_k \in S$ let

$$\sum_{j=1}^{k} c_j v_j = 0.$$

Multiplying by $v_i \in S$, $i \in \{1, \ldots, j\}$ one gets

$$0 = v_i^H \sum_{j=1}^{k} c_j v_j = \sum_{j=1}^{k} c_j v_i^H v_j = c_i v_i^H v_i = c_i \|v_i\|^2 \quad \Rightarrow \quad c_i = 0$$

which implies the linear independence of S

As a corollary of the Theorem it follows that if an orthogonal set $S \subset \mathbb{C}^m$ contains m vectors, then it is a basis for $\mathbb{C}^m$.

Representation by orthonormal basis

Given a vector $b \in \mathbb{C}^m$ and a basis $\{q_1, \ldots, q_m\}$ of $\mathbb{C}^m$ one usually has to solve a linear system to obtain the representation $b = \sum_{j=1}^m \beta_j q_j$ with respect to this basis, namely

$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ & & \ddots & \\ q_{m1} & q_{m2} & \cdots & q_{mm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \quad \text{where } q_j = \begin{pmatrix} q_{1j} \\ q_{2j} \\ \vdots \\ q_{mj} \end{pmatrix}.$$

If $\{q_1, \ldots, q_m\}$ is an orthonormal basis, i.e. $q_i^H q_j = \delta_{ij}$ where δ_{ij} is the Kronecker symbol equal 1 if $i = j$ and 0 if $i \neq j$, then

$$q_i^H b = q_i^H(\beta_1 q_1 + \beta_2 q_2 + \cdots + \beta_m q_m) = \sum_{j=1}^m \beta_j q_i^H q_j = \beta_i,$$

and the representation of b is given by

$$b = \sum_{j=1}^m (q_j^H b) q_j = \sum_{j=1}^m (q_j q_j^H) b.$$

Representation by orthonormal basis ct.

$$b = \sum_{j=1}^{m} (q_j^H b) q_j = \sum_{j=1}^{m} (q_j q_j^H) b.$$

contains two different ways to represent b, once with $(q_j^H b) q_j$, and again with $(q_j q_j^H) b$.

These expressiosn are equal, but they have different interpretations.

In the first case, we view b as a sum of coefficients $(q_j^H b)$ times vectors q_j.

In the second, we view b as a sum of orthogonal projections of b onto the various directions q_j. The jth projection operation is achieved by the rank-one matrix $q_j^H q_j$.

Unitary matrices

A square matrix $Q \in \mathbb{C}^{m\times m}$ is unitary (in the real case, we also say orthogonal) if $Q^H = Q^{-1}$, i.e, if $Q^H Q = I$.

In other words, the columns q_j of a unitary matrix form an orthonormal basis of $\mathbb{C}^m$.

$$q_i^H q_j = \delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}$$

δ_{ij} is called Kronecker delta.

Vector norms

Norms serve the purpose to measure the length of vectors.

A vector norm on $\mathbb{C}^n$ is a function

$$\|\cdot\| : \mathbb{C}^n \to \mathbb{R}_+ := \{\alpha \in \mathbb{R} : \alpha \geq 0\}$$

that satisfies the following properties:

(i) $\|x\| = 0 \quad \Leftrightarrow \quad x = 0$

(ii) $\|\alpha x\| = |\alpha| \cdot \|x\|$ for every $x \in \mathbb{C}^n$ and $\alpha \in \mathbb{C}$

(iii) $\|x + y\| \leq \|x\| + \|y\|$ for every $x, y \in \mathbb{C}^n$

Example: $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$ called p-norm.

$$\begin{aligned} \|x\|_1 &= |x_1| + \cdots + |x_n| \\ \|x\|_2 &= \sqrt{|x_1|^2 + \cdots + |x_n|^2} \\ \|x\|_\infty &= \max_{j=1,\ldots,n} |x_j| \end{aligned}$$

Properties of vector norms

Hölder's inequality

$$|x^H y| \leq \|x\|_p \cdot \|y\|_q \quad \text{where } \frac{1}{p} + \frac{1}{q} = 1.$$

Important special case: Cauchy–Schwarz inequality

$$|x^H y| \leq \|x\|_2 \cdot \|y\|_2.$$

All norms on $\mathbb{C}^n$ are equivalent, i.e. if $\|\cdot\|$ and $\|\cdot\|'$ are two norms on $\mathbb{C}^n$ then there exist positive constants C_1 and C_2 such that

$$C_1\|x\| \leq \|x\|' \leq C_2\|x\| \quad \text{for every } x \in \mathbb{C}^n.$$

$$\begin{array}{ccccc} \|x\|_2 & \leq & \|x\|_1 & \leq & \sqrt{n}\|x\|_2 \\ \|x\|_\infty & \leq & \|x\|_2 & \leq & \sqrt{n}\|x\|_\infty \\ \|x\|_\infty & \leq & \|x\|_1 & \leq & n\|x\|_\infty \end{array}$$

Errors

Suppose that $\hat{x}$ is an approximation to x. For a given vector norm $\|\cdot\|$

$$\epsilon_{\text{abs}} := \|x - \hat{x}\|$$

is the absolute error of $\hat{x}$.

If $x \neq 0$ then

$$\epsilon_{\text{rel}} := \frac{\|x - \hat{x}\|}{\|x\|}$$

is the relative error of $\hat{x}$.

If

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \approx 10^{-p}$$

then the largest component of $\hat{x}$ has approximately p correct significant digits.

If $x = (9.876\ ,\ 0.0543)^T$ and $\hat{x} = (9.875\ ,\ 0.0700)^T$, then $\|x - \hat{x}\|_\infty / \|x\|_\infty \approx 1.6e-3 \approx 10^{-3}$, and the first component has about 3 correct leading digits whereas the second component has no correct significant digit.

Matrix norms

Let $A \in \mathbb{C}^{m\times n}$, let $\|\cdot\|_n$ be a vector norm in $\mathbb{C}^n$ and $\|\cdot\|_m$ be a vector norm in $\mathbb{C}^m$. Then

$$\|A\|_{m,n} := \sup_{x\neq 0} \frac{\|Ax\|_m}{\|x\|_n}$$

is the matrix norm subordinate to the vector norms $\|\cdot\|_n$ and $\|\cdot\|_m$.

From

$$\frac{\|Ax\|_m}{\|x\|_n} = \left\| A\left(\frac{x}{\|x\|_n}\right)\right\|_m$$

it follows that

$$\|A\|_{m,n} = \max\{\|Ax\|_m \,:\, \|x\|_n = 1\}.$$

In particular, this observation guarantees that the maximum is attained by some $x \in \mathbb{C}^n$ since the mapping $x \mapsto \|Ax\|_m$ is continuous and $\{x \,:\, \|x\|_n = 1\}$ is compact.

Properties of matrix norms

$$\begin{aligned} \|A\|_{m,n} = 0 &\iff \|Ax\|_m = 0 \text{ for every } x \in \mathbb{C}^n \\ &\iff Ax = 0 \text{for every } x \in \mathbb{C}^n \\ &\iff A = O \end{aligned}$$

$$\begin{aligned} \|\alpha A\|_{m,n} &= \max\{\|\alpha Ax\|_m \,:\, \|x\|_n = 1\} \\ &= \max\{|\alpha| \cdot \|Ax\|_m \,:\, \|x\|_n = 1\} \\ &= |\alpha| \cdot \|A\|_{m,n} \end{aligned}$$

$$\begin{aligned} \|A+B\|_{m,n} &= \max\{\|Ax+Bx\|_m \,:\, \|x\|_n = 1\} \\ &\le \max\{\|Ax\|_m + \|Bx\|_m \,:\, \|x\|_n = 1\} \\ &\le \max\{\|Ax\|_m \,:\, \|x\|_n = 1\} + \max\{\|Bx\|_m \,:\, \|x\|_n = 1\} \\ &= \|A\|_{m,n} + \|B\|_{m,n} \end{aligned}$$

Hence $\|\cdot\|_{m,n}$ is a vector norm on the vector space $\mathbb{C}^{m\times n}$

Submultiplicativity of matrix norms

$$\|Ax\|_m \leq \|A\|_{m,n} \cdot \|x\|_n \quad \text{for every } x \in \mathbb{C}^n \text{ and every } A \in \mathbb{C}^{m\times n}$$

follows immediately from the definition of the matrix norm

For every $A \in \mathbb{C}^{m\times n}$ and every $B \in \mathbb{C}^{n\times p}$ it holds

$$\|AB\|_{m,p} \leq \|A\|_{m,n} \cdot \|B\|_{n,p}.$$

For every $x \in \mathbb{C}^p$

$$\|ABx\|_m = \|A(Bx\|_m \leq \|A\|_{m,n} \cdot \|Bx\|_n \leq \|A\|_{m,n} \cdot \|B\|_{n,p} \cdot \|x\|_p,$$

and therefore

$$\|AB\|_{m,p} = \max\{\|ABx\|_m \,:\, \|x\|_p = 1\} \leq \|A\|_{m,n} \cdot \|B\|_{n,p}.$$

Geometric interpretation

The matrix norm $\|A\|_{m,n}$ is the smallest nonnegative number μ such that

$$\|Ax\|_m \leq \mu \cdot \|x\|_n \quad \text{for every } x \in \mathbb{C}^n.$$

Hence, $\|A\|_{m,n}$ is the maximum elongation of a vector x by the mapping $x \mapsto Ax$ with respect to the norm $\|\cdot\|_n$ in the domain $\mathbb{C}^n$ and $\|\cdot\|_m$ in the range $\mathbb{C}^m$

From now on we only consider the case that the same (type of) norm is used in the domain and in the range (even if the two spaces are of different dimensions), and we denote the matrix norm by the same symbol that is used for the vector norm.

Hence, if $A \in \mathbb{C}^{5\times 9}$ then $\|A\|_\infty$ denotes the matrix norm of A with respe<ct to the maximum norm in the domain $\mathbb{C}^9$ and in the range $\mathbb{C}^5$.

Matrix ∞-norm

$$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$$

For every $x \in \mathbb{C}^n$ it holds

$$\begin{aligned} \|Ax\|_\infty &= \max_{i=1,\dots,m} |\sum_{j=1}^{n} a_{ij} x_j| \le \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| \\ &\le \|x\|_\infty \cdot \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|. \end{aligned}$$

Thus

$$\|A\|_\infty \le \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}| \qquad (*).$$

Matrix ∞-norm ct.

Let $k \in \{1, \ldots, m\}$ such that

$$\sum_{j=1}^{n} |a_{ij}| \leq \sum_{j=1}^{n} |a_{kj}| \quad \text{for every } i = 1, \ldots, n$$

and define $x \in \mathbb{C}^n$ by $x_j := 1$, if $a_{kj} = 0$, and $x_j := \overline{a_{kj}}/|a_{kj}|$, otherwise.

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

$$\begin{aligned} \|Ax\|_\infty &= \max_{i=1,\ldots,m} |\sum_{j=1}^{n} a_{ij} x_j| \geq |\sum_{j=1}^{n} a_{kj} x_j| \\ &= \left| \sum_{j=1}^{n} |a_{kj}| \right| = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|. \end{aligned}$$

Hence

$$\|A\|_\infty = \max\{\|Ay\|_\infty \,:\, \|y\|_\infty = 1\} \geq \|Ax\|_\infty \geq \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|,$$

which together with inequality (*) yields the proposition.

Matrix 1-norm and 2-norm

Analogously the 1-norm of a matrix $A \in \mathbb{C}^{m\times N}$ is easily shown to be

$$\|A\|_1 := \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$$

The matrix 2-norm, called spectral norm of $A \in \mathbb{C}^{m\times n}$ is is the square root of the largest eigenvalue of $A^H A$. This follows from Rayleigh's principle:

$$\frac{\|Ax\|_2^2}{\|x\|^2} = \frac{x^H A^H A x}{x^H x} \le \max\{\lambda \,:\, A^H A x = \lambda x\}.$$

Hence, $\|A\|_2$ =square root of maximum eigenvalue of $A^H A$.

The spectral norm can be easily bounded by

$$\|A\|_2 \le \sqrt{\|A\|_1 \cdot \|A\|_\infty}$$

Let $z \neq 0$ such that $A^H A z = \|A\|_2^2 z$. Then

$$\|A\|_2^2 \|z\|_1 = \|A^H A z\|_1 \le \|A^H\|_1 \|A\|_1 \|z\|_1 = \|A\|_\infty \|A\|_1 \|z\|_1.$$

Frobenius norm

For every $A \in \mathbb{C}^{m\times n}$ it holds

$$\|A\|_2 \le \|A\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

$\|A\|_F$ is called Frobenius norm or Schur norm

$\|\cdot\|_F$ is a vector norm on $\mathbb{C}^{m\times n}$ (= euclidian norm on $\mathbb{C}^{m\cdot n}$), but it is not matrix norm subordinate to a vector norm, since in this case it would hold for $n = m$ for the unit matrix I

$$\|I\| = \max\{\|Ix\| \ : \ \|x\| = 1\} = 1 \quad \text{whereas } \|I\|_F = \sqrt{n}.$$

Frobenius norm ct.

From the Cauchy Schwarz inequality it follows for every $x \in \mathbb{C}^n$

$$\|Ax\|_2^2 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} x_j \right|^2 \leq \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right) \sum_{j=1}^{n} |x_j|^2 = \|A\|_F^2 \cdot \|x\|_2^2,$$

and hence $\|A\|_2 \leq \|A\|_F$.

Singular value decomposition

The Singular value decomposition (SVD) is motivated by the following geometric fact:
The image of the unit sphere under any $m \times n$ matrix is a hyperellipse.

The SVD is applicable to both real and complex matrices. However, in describing the geometric interpretation, we assume as usual that the matrix is real.

The term "hyperellipse" may be unfamiliar, but this is just the m-dimensional generalization of an ellipse. We may define a hyperellipse in $\mathbb{R}^m$ as the surface obtained by stretching the unit sphere in $\mathbb{R}^m$ by some factors $\sigma_1, \ldots, \sigma_m$ (possibly zero) in some orthogonal directions $u_1, \ldots, u_m \in \mathbb{R}^m$.

For convenience, let us take the u_i to be unit vectors, i.e., $\|u_i\|_2 = 1$. The vectors $\{\sigma_i u_i\}$ are the principal semiaxes of the hyperellipse, with lengths $\sigma_1, \ldots, \sigma_m$.

If A has rank r, exactly r of the lengths σ_i will turn out to be nonzero, and in particular, if $m > n$, at most n of them will be nonzero.

SVD ct.

Let S be the unit sphere in $\mathbb{R}^n$, and take any $A \in \mathbb{R}^{m\times n}$ with $m >_n$. For simplicity, suppose for the moment that A has full rank n.

The image AS is a hyperellipse in $\mathbb{R}^m$. We now define some properties of A in terms of the shape of AS. First, we define the n singular values of A. These are the lengths of the n principal semiaxes of AS, written as $\sigma_1, \ldots, \sigma_n$. It is conventional to assume that the singular values are numbered in descending order, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

Next, we define the n left singular vectors of A. These are the unit vectors $u_l, u_2, \ldots, u_n$ oriented in the directions of the principal semiaxes of AS, numbered to correspond with the singular values. Thus the vector $\sigma_i u_i$ is the ith largest principal semiaxis of AS.

Finally, we define the n right singular vectors of A. These are the unit vectors $\{v_l, v_2, \ldots, v_n\} \subset S$ that are the preimages of the principal semiaxes of AS, numbered so that $Av_j = \sigma_j u_j$.

Reduced SVD

The equations relating right singular vectors $\{v_j\}$ and left singular vectors $\{u_j\}$ can be written as

$$Av_j = \sigma_j u_j, \quad j = 1, \ldots, n.$$

This collection of vector equations can be expressed as a matrix equation,

$$A[v_1, \ldots, v_n] = [u_1, \ldots, u_n] \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$$

or more compactly $AV = \hat{U}\hat{\Sigma}$ where $\hat{\Sigma} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with positive entries, $\hat{U} \in \mathbb{R}^{m\times n}$ and $V \in \mathbb{R}^{n\times n}$ have orthonormal columns.

Multiplying on the right by V^H (notice that V is unitary!) one gets

$$A = \hat{U}\hat{\Sigma}V^H$$

which is called the reduced singular value decomposition or reduced SVD of A.

Full SVD

In most application the SVD is used in exactly the form just described. However, this is not the way in which the idea of an SVD is usually formulated in textbooks. We have introduced the term "reduced" and the hats on U and Σ in order to distinguish the factorization from the more standard "full" SVD.

The columns of U are n orthonormal vectors in the m-dimensional space $\mathbb{C}^m$. Unless $m = n$, they do not form a basis of $\mathbb{C}^m$, nor is U a unitary matrix. However, by adjoining an additional $m - n$ orthonormal columns, $\hat{U}$ can be extended to a unitary matrix. Let us do this in an arbitrary fashion, and call the result U.

If $\hat{U}$ is replaced by U, then $\hat{\Sigma}$ will have to change too. For the product to remain unaltered, the last $m - n$ columns of U should be multiplied by zero. Accordingly, let Σ be the $m \times n$ matrix consisting of Σ in the upper $n \times n$ block together with $m - n$ rows of zeros below.

We now have a new factorization, the full SVD of A:

$$A = U\Sigma V^H$$

where U is $m \times m$ and unitary, V is $n \times n$ and unitary, and Σ is $m \times n$ and diagonal with positive real entries.

Full SVD ct.

Having described the full SVD, we can now discard the simplifying assumption that A has full rank.

If A is rank-deficient, the factorization $A = U\Sigma V^H$ is still appropriate. All that changes is that now not n but only r of the left singular vectors of A are determined by the geometry of the hyperellipse.

To construct the unitary matrix U, we introduce $m - r$ instead of just $m - n$ additional arbitrary orthonormal columns. The matrix V will also need $n - r$ arbitrary orthonormal columns to extend the r columns determined by the geometry. The matrix Σ will now have r positive diagonal entries, with the remaining $n - r$ equal to zero.

By the same token, the reduced SVD also makes sense for matrices A of less than full rank. One can take $\hat{U}$ to be $m \times n$, with $\hat{\Sigma}$ of dimensions $n \times n$ with some zeros on the diagonal, or further compress the representation so that $\hat{U}$ is $m \times r$ and $\hat{\Sigma}$ is $r \times r$ and strictly positive on the diagonal.

Formal definition of SVD

Let m and n be arbitrary (not necessarily $m \geq n$). Given $A \in \mathbb{C}^{m\times n}$, a singular value decomposition (SVD) of A is a factorization

$$A = U\Sigma V^H$$

where $U \in \mathbb{C}^{m\times m}$ is unitary, $V \in \mathbb{C}^{n\times n}$ is unitary, and $\Sigma \in \mathbb{C}^{m\times n}$ is diagonal.

In addition, it is assumed that the diagonal entries σ_j of Σ are nonnegative and in nonincreasing order; that is, $\sigma_l \geq \sigma_2 \geq \ldots \sigma_p$, where $p = \min(m, n)$.

Note that the diagonal matrix Σ has the same shape as A even when A is not square, but U and V are always square unitary matrices.

It is clear that the image of the unit sphere in $\mathbb{R}^n$ under a map $A = U\Sigma V^H$ must be a hyperellipse in $\mathbb{R}^m$. The unitary map V^H preserves the sphere, the diagonal matrix Σ stretches the sphere into a hyperellipse aligned with the canonical basis, and the final unitary map U rotates or reflects the hyperellipse without changing its shape.
Thus, if we can prove that every matrix has an SVD, we shall have proved that the image of the unit sphere under any linear map is a hyperellipse.

Existence and Uniqueness

Theorem
Every matrix $A \in \mathbb{C}^{m\times n}$ has a singular value decomposition $A = U\Sigma V^H$.

Furthermore, the singular values σ_j are uniquely determined, and, if A is square and the σ_j are distinct, the left and right singular vectors $\{u_j\}$ and $\{v_j\}$ are uniquely determined up to complex signs (i.e., complex scalar factors of absolute value 1).

Proof: To prove existence of the SVD, we isolate the direction of the largest action of A, and then proceed by induction on the dimension of A.

Set $\sigma_1 = \|A\|_2$. By a compactness argument, there must be a vector $v_1 \in \mathbb{C}^n$ with $\|v_1\|_2 = 1$ and $\|u_1\|_2 = \sigma_1$, where $u_l = Av_l$.

Consider any extensions of v_1 to an orthonormal basis $\{v_j\}$ of $\mathbb{C}^n$ and of u_l to an orthonormal basis $\{u_j\}$ of $\mathbb{C}^m$, and let U_l and V_l denote the unitary matrices with columns u_j and v_j, respectively.

Proof ct.

$$U_1^H A V_1 = \begin{pmatrix} u_1^H \\ \vdots \\ u_m^H \end{pmatrix} A \begin{pmatrix} v_1 & \dots & v_n \end{pmatrix} = \begin{pmatrix} u_1^H \\ \vdots \\ u_m^H \end{pmatrix} \begin{pmatrix} Av_1 & \dots & Av_n \end{pmatrix}$$

$$= \begin{pmatrix} u_1^H \\ \vdots \\ u_m^H \end{pmatrix} \begin{pmatrix} \sigma_1 u_1 & \dots & Av_n \end{pmatrix} = \begin{pmatrix} \sigma_1 & w^H \\ 0 & B \end{pmatrix} =: S$$

where 0 is a column vector of dimension $m-1$, w^H is a row vector of dimension $n-1$, and $B \in \mathbb{C}^{m-1 \times n-1}$.

$$\left\| \begin{pmatrix} \sigma_1 & w^H \\ 0 & B \end{pmatrix} \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2 \geq \sigma_1^2 + w^H w = \sqrt{\sigma_1^2 + w^H w} \left\| \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2$$

implying $\|S\|_2 \geq \sqrt{\sigma_1^2 + w^H w}$.

Since U_1 and V_1 are unitary, it follows that $\|S\|_2 = \|A\|_2 = \sigma_1$, so this implies $w = 0$.

Proof ct.

If $n = 1$ or $m = 1$, we are done.

Otherwise, the submatrix B describes the action of A on the subspace orthogonal to v_l. By the induction hypothesis, B has an SVD $B = U_2 \Sigma_2 V_2^H$.

Now it is easily verified that

$$A = U_1 \begin{pmatrix} 1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V_2 \end{pmatrix}^H V_1^H$$

is an SVD of A, completing the existence proof.

For the uniqueness claim, the geometric justification is straightforward: if the semiaxis lengths of a hyperellipse are distinct, then the semiaxes themselves are determined by the geometry, up to signs.