Applied Matrix Theory, Math 464/514, Fall 2023

Jens Lorenz

September 13, 2023

Department of Mathematics and Statistics,
UNM, Albuquerque, NM 87131

Contents

1 Gaussian Elimination and LU–Factorization

Consider a linear system of equations $Ax = b$ where $A \in \mathbb{C}^{n\times n}$ is a square matrix, $b \in \mathbb{C}^n$ is a given vector, and $x \in \mathbb{C}^n$ is unknown. Gaussian elimination remains one of the most basic and important algorithms to compute the solution x. If the algorithm does not break down and one ignores round–off errors, then the solution x is computed in $\mathcal{O}(n^3)$ arithmetic operations.

For simplicity, we describe Gaussian elimination first without pivoting, i.e., without exchanges of rows and columns. We will explain that the elimination process (if it does not break down) leads to a matrix factorization, $A = LU$, the so–called LU–factorization of A. Here L is unit–lower triangular and U is upper triangular.

The triangular matrices L and U with $A = LU$ are computed in $\mathcal{O}(n^3)$ steps. Once the factorization $A = LU$ is known, the solution of the system

$$Ax = LUx = b$$

can be computed in $\mathcal{O}(n^2)$ steps. This observation is important if one wants to solve linear systems $Ax = b$ with the same matrix A, but different right–hand sides b. For example, if one wants to solve a nonlinear system $Ax = b + \varepsilon F(x)$ by an iterative process

$$Ax^{(j+1)} = b + \varepsilon F(x^{(j)}), \quad j = 0, 1, 2, \ldots$$

then the LU–factorization of A is very useful.

Gaussian elimination without pivoting may break down for very simple invertible systems. An example is

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

with unique solution

$$x_1 = x_2 = 1 \ .$$

We will introduce permutations and permutation matrices and then describe Gaussian elimination with row exchanges, i.e., with partial pivoting. It corresponds to a matrix factorization $PA = LU$ where P is a permutation matrix, L is unit lower triangular and U is upper triangular. The algorithm is practically and theoretically important. On the theoretical side, it leads to Fredholm's alternative for any system $Ax = b$ where A is a square matrix.

On the practical side, partial pivoting is recommended even if the algorithm without pivoting does not break down. Partial pivoting typically leads to better numerical stability.

1.1 Gaussian Elimination Without Pivoting

Example 1.1 Consider the system

$$\begin{pmatrix} 2 & 1 & 1 \\ 6 & 2 & 1 \\ -2 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 7 \end{pmatrix} \tag{1.1}$$

which we abbreviate as $Ax = b$. The usual elimination process leads to the equivalent systems

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -4 \\ 8 \end{pmatrix} \tag{1.2}$$

and

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -4 \\ -4 \end{pmatrix} \tag{1.3}$$

The transition from (1.1) to (1.2) can be described as follows: Equation 1 multiplied by -3 is added to equation 2 and equation 1 multiplied by 1 is added to equation 3. These two steps eliminate x_1 from the second and third equation. Similarly, the transition from (1.2) to (1.3) can be described as follows: Equation 2 multiplied by 3 is added to equation 3. This step eliminates x_2 from the third equation.

The diagonal elements 2 in (1.1) and -1 in (1.2) are called the pivots of the elimination process.

In matrix form, the two steps of the elimination process can be written as

$$E_2 E_1 A x = E_2 E_1 b$$

with

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

and

$$E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}$$

Here the elimination matrices E_j are unit–lower triangular and, below the diagonal in column j, the matrix E_j contains the multipliers of the elimination process. The multipliers in the first step are -3 and 1. The multiplier in the second step is 3. The last system, $Ux = \tilde{b}$, can be solved by backward substitution:

$$x_3 = 1, \quad x_2 = 2, \quad x_1 = -1 \ .$$

We note that the elimination process leads to the factorization

$$E_2 E_1 A = U \ ,$$

which we also can write as

$$A = E_1^{-1} E_2^{-1} U = LU$$

with

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -1 & -3 & 1 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{pmatrix}$$

Note that L is unit lower triangular and contains the negatives of the multipliers below the diagonal. The first two diagonal entries of U are the pivots of the elimination process.

It is not difficult to generalize the example. Gaussian elimination consists of two processes, an elimination process and a back-substitution process. (In the elimination process, some variables are successively eliminated from some equations.)

Lemma 1.1 *a) Let E_k denote an elimination matrix, containing multipliers m_j for $k+1 \le j \le n$ in its k-th column below the diagonal. Then E_k^{-1} is obtained from E_k by changing the signs of the multipliers, i.e., by replacing m_j with $-m_j$.*

b) If E_k and E_l are elimination matrices and $k < l$, then

$$Q = E_k^{-1} E_l^{-1}$$

is obtained from E_k^{-1} and E_l^{-1} in a very simple way: Q is unit–lower triangular and contains the entries of E_k^{-1} in its k–th column, the entries of E_l^{-1} in its l–th column.

Proof: a) Let

$$E_k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & m_{k+1} & \ddots & & \\ & & \vdots & & \ddots & \\ & & m_n & & & 1 \end{pmatrix} \tag{1.4}$$

and let F_k denote the corresponding matrix where each m_j is replaced by $-m_j$. Application of E_k to a vector $x \in \mathbb{C}^n$ yields

$$E_k \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} + m_{k+1}x_k \\ \vdots \\ x_n + m_n x_k \end{pmatrix} .$$

It then follows that

$$F_k E_k x = x \quad \text{for all} \quad x \in \mathbb{C}^n .$$

This implies that $F_k E_k = I$, i.e., $F_k = E_k^{-1}$.

b) Let $k < l$ and consider the matrices F_k and F_l. We write these as

$$F_k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & \beta_{k+1} & \ddots & & \\ & & \vdots & & \ddots & \\ & & \beta_n & & & 1 \end{pmatrix} , \quad F_l = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & \alpha_{l+1} & \ddots & & \\ & & \vdots & & \ddots & \\ & & \alpha_n & & & 1 \end{pmatrix} .$$

Applying $F_k F_l$ to any vector $x \in \mathbb{C}^n$ yields

$$F_k F_l x = F_k \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_l \\ x_{l+1} + \alpha_{l+1} x_l \\ \vdots \\ x_n + \alpha_n x_l \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} + \beta_{k+1} x_k \\ \vdots \\ x_l + \beta_l x_k \\ x_{l+1} + \alpha_{l+1} x_l + \beta_{l+1} x_k \\ \vdots \\ x_n + \alpha_n x_l + \beta_n x_k \end{pmatrix}$$

If we now denote by Q the matrix which contains the β_j in its k-th column and the α_j in its l–th column, then we obtain that

$$F_k F_l x = Qx \quad \text{for all} \quad x \in \mathbb{C}^n .$$

This implies that $F_k F_l = Q$. $\diamond$

2nd Proof of a) (more formal) We have

$$E_k = I + \sum_{j=k+1}^{n} m_j e^j e^{kT} = I + S$$

and set

$$F_k = I - \sum_{j=k+1}^{n} m_j e^j e^{kT} = I - S \ .$$

We now multiply:

$$F_k E_k = (I - S)(I + S) = I - S^2$$

with

$$S^2 = \Big(\sum_{j=k+1}^{n} m_j e^j e^{kT} \Big) (\sum_{l=k+1}^{n} m_l e^l e^{kT} \Big) \ .$$

Here

$$e^j e^{kT} e^l e^{kT} = 0$$

since

$$e^{kT} e^l = 0 \quad \text{for} \quad l \neq k \ .$$

2nd Proof of b) (more formal) Let $1 \leq k < l \leq n$ and let

$$\begin{aligned} F_k &= I + \sum_{j=k+1}^{n} \beta_j e^j e^{kT} = I + M_k \\ F_l &= I + \sum_{i=l+1}^{n} \alpha_i e^i e^{lT} = I + M_l \end{aligned}$$

Then we have

$$F_k F_l = (I + M_k)(I + M_l) = I + M_k + M_l + M_k M_l \ .$$

Here

$$M_k M_l = \sum_{j=k+1}^{n} \sum_{i=l+1}^{n} \alpha_i \beta_j e^j e^{kT} e^i e^{lT} \ .$$

Since $i \geq l + 1 > k$ we have

$$e^{kT} e^i = 0 \ ,$$

thus $M_k M_l = 0$ and

$$F_k F_l = I + M_k + M_l \ .$$

This completes the 2nd proof of the lemma. $\diamond$

The process of elimination described above applied to a system

$$Ax = b$$

can be written in the form

$$E_{n-1} \dots E_1 A x = E_{n-1} \dots E_1 b \; .$$

Here

$$E_{n-1} \dots E_1 A =: U$$

is upper triangular. One obtains that

$$L := E_1^{-1} \dots E_{n-1}^{-1}$$

is unit lower triangular. Thus one has the factorization

$$A = LU \; .$$

The matrix L contains the multipliers of the elimination process multiplied by -1.

The system $Ax = b$ can be written as $LUx = b$. This system can be solved by solving first $Ly = b$ (forward substitution) and then $Ux = y$ (backward substitution).

Summary: Assume that the Gaussian elimination process can be applied to the system $Ax = b$ without occurances of a zero pivot. Then one obtains a factorization $A = LU$ where L is unit–lower triangular and U is upper triangular. The diagonal elements

$$u_{11}, \dots, u_{n-1n-1}$$

are the pivots of the elimination process, which are different from zero by assumption. (Otherwise the process breaks down.) If also $u_{nn} \neq 0$ then the system $Ax = b$ has a unique solution x. The solution can be obtained from $LUx = b$ by forward and backward substitution: First solve the system $Ly = b$ for y by forward substitution, then solve $Ux = y$ for x by backward substitution.

1.2 Application to $Ax = b + \varepsilon F(x)$

We explain here why it is interesting that Gaussian elimination corresponds to the matrix factorization $A = LU$.

Operation Count: Suppose we have computed the factors L and U of the factorization $A = LU$. Then the system $Ax = b$ can be written as $LUx = b$ and we can solve $Ly = b$ for y and then $Ux = y$ for x by forward and backward substitution, respectively. This costs $\mathcal{O}(n^2)$ operations. To compute the factorization $A = LU$ costs $\mathcal{O}(n^3)$ operations. Thus, if n is large, the numerical work for solving $LUx = b$ is negligible compared with the work for computing the factorization.

Application: In the following application one has to solve many linear systems $Ax = b^{(j)}$ with the same matrix A, but with many different right–hand sides $b^{(j)}$. The right–hand sides are not all known in advance.

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ denote a smooth nonlinear map. The system

$$Ax = b + \varepsilon F(x)$$

can be treated by the fixed point iteration

$$Ax^{j+1} = b + \varepsilon F(x^j), \quad j = 0, 1, \ldots \tag{1.5}$$

where $Ax^0 = b$. In each step one has to solve a linear system with the same matrix A. One computes the (expensive) LU–factorization of A only once, of course.

Remark on convergence: Let $\| \cdot \|$ denote a vector norm on $\mathbb{R}^n$ and assume that $F : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz bounded with Lipschitz constant L:

$$\|F(x) - F(y)\| \le L\|x - y\| \quad \text{for all} \quad x, y \in \mathbb{R}^n \ .$$

Define

$$\Phi(x) = A^{-1}b + \varepsilon A^{-1}F(x), \quad x \in \mathbb{R}^n \ .$$

The iteration (1.5) is equivalent to the fixed point iteration

$$x^{j+1} = \Phi(x^j)$$

but note that, in practice, we do not compute A^{-1} because that would be too expensive in terms of effort. Instead, we solve the systems occurring in (1.5).

If $\|A^{-1}\|$ denotes the corresponding matrix norm (see Chapter 2) then

$$\|\Phi(x) - \Phi(y)\| \le |\varepsilon|\|A^{-1}\|L\|x - y\| \ .$$

Therefore, if

$$|\varepsilon|\|A^{-1}\|L < 1 \ ,$$

then, by the contraction mapping theorem, the iteration sequence x^j defined by (1.5) converges to the unique solution x^* of the nonlinear system $Ax = b + \varepsilon F(x)$.

Remarks on Newton's Iteration: It is interesting to compare the above fixed point iteration with Newton's iteration. We must solve

$$Q(x) \equiv Ax - \varepsilon F(x) - b = 0 \ .$$

Denote the unknown solution by x^* and let x^0 be a starting vector. Let $x^* = x^0 + h$ where $h \in \mathbb{R}^n$ is assumed to be a vector with a small norm. We have

$$0 = Q(x^*) = Q(x^0 + h) = Q(x^0) + Q'(x^0)h + \mathcal{O}(\|h\|^2) \ .$$

Neglecting the $\mathcal{O}(\|h\|^2)$ term one obtains the linear system

$$Q'(x^0)h = -Q(x^0)$$

for h. The vector $x^1 = x^0 + h$ is the next iterate. Then, solving

$$Q'(x^1)h = -Q(x^1)$$

for h and setting $x^2 = x^1 + h$ one obtains x^2.

In general, solve the linear system

$$Q'(x^j)h = -Q(x^j)$$

for h and the set

$$x^{j+1} = x^j + h \quad \text{for} \quad j = 1, 2, \ldots$$

Note that

$$Q'(x^j) = A - \varepsilon F'(x^j) \ .$$

Typically, Newton's method converges faster than the fixed point iteration

$$Ax^{j+1} = b + \varepsilon F(x^j) \ ,$$

but the matrix $Q'(x^j)$ in the system $Q'(x^j)h = -Q(x^j)$ changes in each iteration step.

Note that the system $Q'(x^j)h = -Q(x^j)$ reads

$$\Big(A - \varepsilon F'(x^j)\Big)h = b^j \quad \text{with} \quad b^j = b - Ax^j + \varepsilon F(x^j) \ .$$

Equivalently, the system to be solved for h is

$$Ah = b^j + \varepsilon F'(x^j)h \ .$$

This suggests that one may try the iteration

$$Ah^0 = b^j, \quad Ah^{l+1} = b^j + \varepsilon F'(x^j)h^l \quad \text{for} \quad l = 0, 1, \ldots$$

to obtain a vector h^l which may be a good approximation of the solution h of the linear system

$$\Big(A - \varepsilon F'(x^j)\Big)h = b^j \ .$$

In this way one can take advantage of the LU–factorization of A.

1.3 Initial Boundary Value Problems

There are other situations where many systems $Ax = b^j$ with the same matrix A have to be solved. As an example, consider the heat equation. Let $\Omega \subset \mathbb{R}^3$ denote a bounded domain with boundary $\partial\Omega$. We want to determine the solution $u(x,t)$ of the initial–boundary value problem

$$\begin{aligned}
u_t(x,t) &= \Delta u(x,t) \quad \text{for} \quad x \in \Omega, \quad t \geq 0 \ , \\
u(x,t) &= 0 \quad \text{for} \quad x \in \partial\Omega, \quad t \geq 0 \ , \\
u(x,0) &= f(x) \quad \text{for} \quad x \in \partial\Omega \ .
\end{aligned}$$

Let $\Delta t > 0$ denote a time step and let Ω_h denote a spatial grid in Ω. Let $u_h(\cdot, j\Delta t)$ denote the grid function, which we want to compute and which will approximate $u(\cdot, j\Delta t)$. For simplicity, we write u^j for the grid function $u_h(\cdot, j\Delta t)$. If Ω_h has n grid–points then

$$u^j = u_h(\cdot, j\Delta t) \in \mathbb{R}^n \ .$$

A discrete version of the above IBV problem is

$$\frac{1}{\Delta t}(u^{j+1} - u^j) = \Delta_h\Big(\frac{1}{2}(u^j + u^{j+1})\Big) \quad \text{for} \quad j = 0, 1, \ldots$$

with $u^0 = f_h$. Here Δ_h is a discrete version of the Laplacian Δ.

One obtains the following system for the grid function u^{j+1}:

$$\Big(I - \frac{\Delta t}{2}\Delta_h\Big)u^{j+1} = \Big(I + \frac{\Delta t}{2}\Delta_h\Big)u^j \quad \text{for} \quad j = 0, 1, \ldots$$

If one wants to compute the solution for 1,000 time steps, say, one has to solve 1,000 linear systems $Au^{j+1} = b^j$ with the same matrix A. The LU–factorization of A is very useful.

Remarks: It is typically not a good idea to use the explicit–in–time discretization

$$\frac{1}{\Delta t}(u^{j+1} - u^j) = \Delta_h u^j, \quad j = 0, 1, \ldots$$

because it requires a very small time–step Δt to avoid instabilities. Note that

$$u^{j+1} = (I + \Delta t \Delta_h)u^j$$

and assume that $-\lambda < 0$ is an eigenvalue of Δ_h. The matrix $I + \Delta t \Delta_h$ has the eigenvalue $1 - \lambda\Delta t$. Instabilities occur if Δ_h has an eigenvalue $-\lambda < 0$ with

$$1 - \lambda\Delta t < -1 \ .$$

Thus, stability requires that the time step $\Delta t > 0$ is so small that

$$\lambda \Delta t \leq 2$$

for all eigenvalues λ of Δ_h. If, for example, Δ_h has the eigenvalue $-\lambda = -10^{10}$ then the time step $\Delta t > 0$ must satisfy

$$\Delta t \leq 2 * 10^{-10}$$

to avoid instabilities of the explicit–in–time discretization.

1.4 Partial Pivoting and the Effect of Rounding

Consider the following system $Ax = b$:

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} .$$

It is clear that the algorithm that we have described above breaks down since $a_{11} = 0$. However, if we exchange the two rows of the above system we obtain

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

with solution

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

What happens if a_{11} is not zero, but very small in absolute value? Consider the system

$$\begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{where} \quad 0 < |\varepsilon| << 1 . \tag{1.6}$$

We first compute the exact solution. After exchanging rows we obtain

$$\begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

and elimination leads to

$$\begin{pmatrix} 1 & 1 \\ 0 & 1-\varepsilon \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1-2\varepsilon \end{pmatrix} . \tag{1.7}$$

The exact solution is

$$x_2 = \frac{1-2\varepsilon}{1-\varepsilon} = 1 - \frac{\varepsilon}{1-\varepsilon}, \quad x_1 = 1 + \frac{\varepsilon}{1-\varepsilon} ,$$

thus

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{1-\varepsilon} \begin{pmatrix} \varepsilon \\ -\varepsilon \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \mathcal{O}(\varepsilon) .$$

Note that the ε–perturbation term in the matrix A leads to an $\mathcal{O}(\varepsilon)$ perturbation of the exact solution, which is reasonable.

We now want to discuss the effect of rounding when the system (1.6) is solved numerically. In practice, computations are done most often in floating point arithmetic. For example, in MATLAB machine epsilon is[1]

$$\varepsilon_M \sim 2 * 10^{-16} \; .$$

Here $1+\varepsilon_M > 1$ but, after rounding, $1+\varepsilon =_R 1$ if $|\varepsilon| < \varepsilon_M$.

In the given system (1.6) assume that $\varepsilon = 10^{-17}$, for example, and $\varepsilon_M = 2 * 10^{-16}$. After rounding, the system (1.7) becomes

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \; .$$

The solution is

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_{num} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \; ,$$

which is precisely the solution of the given system if we set $\varepsilon = 0$. Thus, if we first pivot and then eliminate with rounding, we obtain a reasonable result.

Now assume again that $\varepsilon = 10^{-17}$, but we do not pivot. In exact arithmetic, the eliminations process starting with (1.6) yields

$$\begin{pmatrix} \varepsilon & 1 \\ 0 & 1-1/\varepsilon \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2-1/\varepsilon \end{pmatrix} \; .$$

Assuming that $\varepsilon = 10^{-17}$ this system becomes after rounding

$$\begin{pmatrix} 10^{-17} & 1 \\ 0 & -10^{17} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -10^{17} \end{pmatrix} \; .$$

The numerical solution which one obtains is

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_{num} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \; .$$

This numerical solution differs from the correct solution by $\mathcal{O}(1)$.

The example shows that it may not be a good idea to work with a pivot which is very small in absolute value.

1.5 First Remarks an Permutations and Permutation Matrices

A permutation of n elements is a one–to–one map of the set $\{1, 2, \ldots, n\}$ onto itself. Any such permutation can be described by a matrix

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma_1 & \sigma_2 & \ldots & \sigma_n \end{pmatrix} \tag{1.8}$$

which encodes the map

[1]Typing *eps* into MATLAB yields $ans = 2.22 * 10^{-16}$ for machine epsilon, ε_M. Here, by definition, ε_M is the smallest positive number that, when added to 1, creates a number greater than 1 on the computer.

$$\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$$

where $j \to \sigma_j$ for $1 \leq j \leq n$. One often identifies this map σ with the matrix (1.8).

The simplest permutations are the identity, *id*, and transpositions. Any transposition exchanges exactly two elements of the set $\{1, 2, \ldots, n\}$ and leaves all other elements of the set fixed.

Here

$$id \widehat{=} \begin{pmatrix} 1 & 2 & \ldots & n \\ 1 & 2 & \ldots & n \end{pmatrix}$$

and an example of a transposition is

$$T_{12} \widehat{=} \begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ 2 & 1 & 3 & \ldots & n \end{pmatrix} .$$

The transposition T_{12} maps 1 to 2, maps 2 to 1, and leaves all other elements of the set $\{1, 2, \ldots, n\}$ fixed.

With S_n one denotes the group of all permutations of n elements. It is easy to show that S_n has $n!$ elements. If σ and τ are elements of S_n then their product $\sigma\tau = \sigma \circ \tau$ is defined by

$$(\sigma\tau)(j) = (\sigma \circ \tau)(j) = \sigma(\tau(j)), \quad 1 \leq j \leq n .$$

Definition: An $n \times n$ matrix P is called a permutation matrix if every row and every column of P contains exactly one entry equal to one and all other entries of P are zero.

Relation between permutations and permutation matrices. Let e^j denote the standard j–th basis vector of $\mathbb{R}^n$. For example,

$$e^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{etc.}$$

If $\sigma \in S_n$ then let

$$P_\sigma = (e^{\sigma_1}, \ldots, e^{\sigma_n})$$

denote the associated permutation matrix, i.e., the k–th column of P_σ is the vector e^{σ_k}. Clearly,

$$P_\sigma e^k = e^{\sigma_k} .$$

Therefore,

$$\begin{aligned} P_\sigma P_\tau e^k &= P_\sigma e^{\tau(k)} \\ &= e^{\sigma(\tau(k))} \\ &= e^{(\sigma\circ\tau)(k))} \\ &= P_{\sigma\tau} e^k \end{aligned}$$

which implies that

$$P_\sigma P_\tau = P_{\sigma\tau} \ . \tag{1.9}$$

Every permutation matrix $P \in \mathbb{R}^{n\times n}$ has the form

$$P = P_\sigma = \Big(e^{\sigma_1}, \ldots, e^{\sigma_n}\Big)$$

for some $\sigma \in S_n$. Since P_σ is orthogonal we have

$$P_\sigma^T P_\sigma = I$$

and

$$P_{\sigma^{-1}} = (P_\sigma)^{-1} = P_\sigma^T \ .$$

Row and column exchanges: Let $A \in \mathbb{C}^{n\times n}$, thus

$$\begin{aligned} A &= \sum_{l=1}^n \sum_{k=1}^n a_{lk} e^j e^{kT} \\ &= \sum_{l=1}^n e^l \Big(\sum_{k=1}^n a_{lk} e^{kT}\Big) \end{aligned}$$

Here the term in brackets is row_l of A. We have

$$P_\sigma A = \sum_{l=1}^n e^{\sigma(l)} \Big(\sum_{k=1}^n a_{lk} e^{kT}\Big) \ .$$

Thus, if A is multiplied by P_σ , then row_l of A becomes $row_{\sigma(l)}$ of the product $P_\sigma A$.

Columns: Consider AP_σ. We have

$$(AP_\sigma)^T = P_\sigma^T A^T = P_{\sigma^{-1}} A^T \ ,$$

thus

$$AP_\sigma = \Big(P_{\sigma^{-1}} A^T\Big)^T \ .$$

If A^T is multiplied by $P_{\sigma^{-1}}$ from the left then row_l of A^T becomes $row_{\sigma-1(l)}$ of $P_{\sigma^{-1}} A^T$. In other words, if A is multiplied by P_σ from the right, then $column_l$ of A becomes $column_{\sigma^{-1}(l)}$ of AP_σ.

Transpositions. Let $1 \le i < j \le n$. The permutation which exchanges i and j and leaves all other elements of the set $\{1, 2, \ldots, n\}$ fixed, is a transposition, which we denote by T_{ij}. It is then clear that

$$T_{ij} T_{ij} = id \ .$$

If P is the permutation matrix corresponding to T_{ij} then the rule (1.9) implies that

$$PP = I \ .$$

Thus, if P corresponds to a transposition, then

$$P^{-1} = P \ .$$

It is not difficult to show that for any permutation matrix P we have

$$P^T P = I \ ,$$

i.e., the relation $P^T = P^{-1}$ holds for every permutation matrix. For a transposition, the corresponding permutation matrix P is symmetric, $P^T = P$.

Elimination Matrices and Transpositions. Let E_k denote an elimination matrix as defined in (1.4) and let

$$1 \le k < i < j \le n \ .$$

Denote by P the permutation matrix corresponding to the transposition T_{ij}. We want to understand the matrix PE_kP. Taking an arbitrary $x \in \mathbb{C}^n$ we have

$$PE_kPx = PE_kP \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \\ x_n \end{pmatrix} = PE_k \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_j \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = P \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} + m_{k+1}x_k \\ \vdots \\ x_j + m_i x_k \\ \vdots \\ x_i + m_j x_k \\ \vdots \\ x_n + m_n x_k \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} + m_{k+1}x_k \\ \vdots \\ x_i + m_j x_k \\ \vdots \\ x_j + m_i x_k \\ \vdots \\ x_n + m_n x_k \end{pmatrix} .$$

It is not difficult to show that the last vector agrees with $\tilde{E}_k x$ where the matrix $\tilde{E}_k$ is obtained from E_k by exchanging the multipliers m_i and m_j, and leaving all other matrix elements unchanged. This yields that $PE_kPx = \tilde{E}_k x$ for all $x \in \mathbb{C}^n$ and, therefore,

$$PE_kP = \tilde{E}_k \ .$$

Another Proof of the Formula $PE_kP = \tilde{E}_k$: Let $1 \le k < i < j \le n$ and let $\sigma \in S_n$ denote the transpositon exchanging i and j, thus

$$\sigma(i) = j, \quad \sigma(j) = i, \quad \sigma(l) = l \quad \text{if} \quad l \ne i \quad \text{and} \quad l \ne j \ .$$

Let $P = T_{ij}$ denote the corresponding permutation matrix. Let

$$\begin{aligned} E_k &= I + \sum_{l=k+1}^n m_l e^l e^{kT} = I + M_k \\ \tilde{E}_k &= I + \sum_{l=k+1}^n m_l e^{\sigma(l)} e^{kT} = I + \tilde{M}_k \end{aligned}$$

We will to show that

$$T_{ij} M_k T_{ij} = \tilde{M}_k \ .$$

For all $x \in \mathbb{C}^n$ we have

$$\begin{aligned} T_{ij} M_k x &= T_{ij} \Big(\sum_{l=k+1}^n m_l e^l x_k \Big) \\ &= x_k \sum_{l=k+1}^n m_l T_{ij} e^l \\ &= x_k \sum_{l=k+1}^n m_l e^{\sigma(l)} \end{aligned}$$

Also, since $1 \le k < i < j \le n$ we have

$$(T_{ij} x)_k = x_k \ .$$

Therefore,

$$T_{ij} M_k T_{ij} x = x_k \sum_{l=k+1}^n m_l e^{\sigma(l)} \ .$$

The right–hand side agrees with

$$\tilde{M}_k x \ ,$$

and one obtains that

$$T_{ij} M_k T_{ij} x = \tilde{M}_k x$$

for all $x \in \mathbb{C}^n$. This proves that

$$T_{ij} E_k = \tilde{E}_k T_{ij} \quad \text{for} \quad 1 \le k < i < j \le n \ .$$

1.6 Formal Description of Gaussian Elimination with Partial Pivoting

Gaussian elimination with partial pivoting can be written as

$$E_{n-1}P_{n-1}\dots E_1P_1Ax = E_{n-1}P_{n-1}\dots E_1P_1b \ .$$

Here the P_j are permutation matrices corresponding to transpositions and the E_j are elimination matrices, as above. Essentially, the P_j almost commute with the E_i; one only has to permute the multipliers.

As an example, consider

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ \beta & 0 & 1 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \ .$$

We have

$$P_2E_1 = (P_2E_1P_2)P_2$$

and

$$P_2E_1P_2 = \begin{pmatrix} 1 & 0 & 0 \\ \beta & 1 & 0 \\ \alpha & 0 & 1 \end{pmatrix} =: \tilde{E}_1 \ .$$

Thus,

$$P_2E_1 = \tilde{E}_1P_2 \ .$$

In other words, moving the multiplier P_2 from the left side of E_1 to the right side results in permuting the multipliers.

This generalizes. One obtains

$$PA = LU$$

where L is unit lower triangular. The permutations have been applied to the multipliers (multiplied by -1) that are collected in the matrix L.

1.7 Fredholm's Alternative for Linear Systems $Ax = b$

One can use the factorization process leading to $PA = LU$ to prove the following important result.

Theorem 1.1 *Consider an $n \times n$ matrix A (over any field F). Either the system*

$$Ax = b$$

has a unique solution $x \in F^n$ for every $b \in F^n$; or the homogeneous equation $Ax = 0$ has a nontrivial solution $x \in F^n, x \neq 0$.

Proof: There are two cases:

Case 1: Gaussian elimination with partial pivoting can be carried out and leads to a factorization

$$PA = LU$$

where P is a permutation matrix, L is unit lower triangular, and U is upper triangular with

$$u_{jj} \neq 0 \quad \text{for} \quad j = 1, 2, \ldots n \ .$$

In this case, the system $Ax = b$ is equivalent to

$$LUx = Pb \ .$$

This system is uniquely solvable. In fact, one can construct the unique solution by first solving

$$Ly = Pb$$

for $y \in F^n$ (forward substitution) and then solving

$$Ux = y$$

for $x \in F^n$ (backward substitution).

Case 2: Gaussian elimination breaks down or leads to a factorization $PA = LU$ with $u_{nn} = 0$. In both cases one obtains an invertible matrix

$$H = E_k P_k \cdots E_1 P_1$$

so that

$$HA = \begin{pmatrix} u_{11} & * & * & * & \cdots & * \\ & \ddots & * & * & \cdots & * \\ 0 & & u_{kk} & * & \cdots & * \\ 0 & \cdots & 0 & 0 & * & * \\ \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & * & * \end{pmatrix} =: U$$

with

$$u_{11} \neq 0, \ldots, u_{kk} \neq 0 \quad \text{where} \quad k < n \ .$$

One then can construct a non–zero vector

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

with $Ux = 0$. The numbers $x_k, \ldots, x_1$ are obtained by back substitution. Since $HA = U$ and since H is invertible, one obtains that $Ax = 0$. $\diamond$

Fredholm's alternative is an important result for linear systems

$$Ax = b$$

where $A \in F^{n \times n}$ is a square matrix. Denote by

$$N(A) = \{x \in F^n \ : \ Ax = 0\}$$

the nullspace of A. We can formulate **Fredholm's alternative** as follows: There are precisely two cases:

Case 1: $N(A) = \{0\}$. In this case, for every $b \in F^n$ the system $Ax = b$ is uniquely solvable.

Case 2: $N(A) \neq \{0\}$. Then, if $b \in F^n$ is a given right–hand side, we have either

Case 2a: The system $Ax = b$ is not solvable;

or

Case 2b: The solution of the system $Ax = b$ is not unique.

Gaussian elimination with partial pivoting gives a constructive proof of Fredholm's alternative.

Remarks: Fredholm's Alternative is named after the Swedish mathematician Erik Ivar Fredholm (1866–1927), a professor at Stockholm University. He also worked as an actuary at an insurance company, which used his Fredholm equations to calculate buy–back prices for policies. Fredholm established the alternative for certain integral equations. In functional analysis, one proves the following result: If U is a Banach space and $K : U \to U$ is a compact operator, then Fredholm's alternative holds for the equation

$$(\lambda I - K)u = b \tag{1.10}$$

if λ is any non–zero scalar. Thus, if $\lambda \neq 0$ is not an eigenvalue of K, then the above equation has a unique solution $u \in U$ for any right–hand side $b \in U$. If $\lambda \neq 0$ is an eigenvalue of K then either (1.10) has no solution (Case 2a) or the solution is not unique (Case 2b).

1.8 Application to Strictly Diagonally Dominant Matrices

A matrix $A \in \mathbb{C}^{n\times n}$ is called strictly diagonally dominant if

$$|a_{jj}| > \sum_{k\neq j} |a_{jk}| \quad \text{for} \quad j = 1, \ldots, n \; .$$

Lemma 1.2 *If $A \in \mathbb{C}^{n\times n}$ is strictly diagonally dominant, then the homogeneous system $Ax = 0$ has only the trivial solution, $x = 0$. Therefore, for any $b \in \mathbb{C}^n$, the system $Ax = b$ is uniquely solvable.*

Proof: Let $Ax = 0$ and assume that

$$|x_j| = \max_k |x_k| > 0 \; .$$

We have

$$\begin{aligned} 0 &= (Ax)_j \\ &= a_{jj}x_j + \sum_{k\neq j} a_{jk}x_k \end{aligned}$$

thus

$$a_{jj}x_j = -\sum_{k\neq j} a_{jk}x_k \; .$$

Taking absolute values one finds that

$$\begin{aligned} |a_{jj}||x_j| &\leq \sum_{k\neq j} |a_{jk}||x_k| \\ &\leq \sum_{k\neq j} |a_{jk}||x_j| \end{aligned}$$

If one divides by $|x_j|$ one obtains that

$$|a_{jj}| \leq \sum_{k\neq j} |a_{jk}|$$

which contradicts the assumption that A is strictly diagonally dominant. $\diamond$

1.9 Application of MATLAB

The command

$$[L, U, P] = lu(A)$$

returns a unit lower triangular matrix L, an upper triangular matrix U, and a permutation matrix P so that

$$PA = LU \ .$$

Example: For

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

the factorization $PA = LU$ becomes

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{1}{3} & 1 \end{pmatrix} \begin{pmatrix} 3 & 4 \\ 0 & \frac{2}{3} \end{pmatrix} \ .$$

2 Conditioning of Linear Systems

We consider linear systems $Ax = b$ where $A \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^n$ are given. The unknown exact solution $x \in \mathbb{C}^n$ is assumed to be unique. If one applies a numerical algorithm to solve the system, one typically obtains an inexact solution $x + \tilde{x}$, which solves a perturbed system

$$(A + \tilde{A})(x + \tilde{x}) = b + \tilde{b} \; .$$

We consider the pair (A, b) as the given exact data and the pair $(\tilde{A}, \tilde{b})$ as perturbations of the exact data and ask the following question: If the perturbations $(\tilde{A}, \tilde{b})$ are small, will the perturbation $\tilde{x}$ of the exact solution x also be small?

Roughly speaking, one calls the given system $Ax = b$ well–conditioned if small perturbations $(\tilde{A}, \tilde{b})$ of the data (A, b) lead to small perturbations $\tilde{x}$ of the solution x. On the other hand, if small perturbations of (A, b) may lead to large perturbations of the solution, then the system $Ax = b$ is called ill–conditioned.

To make the question of conditioning precise, we must measure the sizes of vectors and matrices by vector norms and matrix norms. We will then prove that the condition number of the matrix A, i.e., the number

$$\kappa = \|A\|\|A^{-1}\| \; ,$$

describes how the relative solution error

$$\frac{\|\tilde{x}\|}{\|x\|}$$

is related to the relative data error

$$\frac{\|\tilde{A}\|}{\|A\|} + \frac{\|\tilde{b}\|}{\|b\|} \; .$$

2.1 Vector Norms and Induced Matrix Norms

As before, $\mathbb{C}^n$ denotes the vector space of all column vectors

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{with} \quad x_j \in \mathbb{C} \; .$$

A function

$$\| \cdot \| : \begin{cases} \mathbb{C}^n & \to \quad [0, \infty) \\ x & \to \quad \|x\| \end{cases} \tag{2.1}$$

is called a vector norm on $\mathbb{C}^n$ if the following conditions hold:

1. $$\|x\| = 0 \quad \text{if and only if} \quad x = 0 \; ;$$

2.
$$\|\alpha x\| = |\alpha|\|x\| \quad \text{for all} \quad \alpha \in \mathbb{C} \quad \text{and} \qquad \text{for all} \quad x \in \mathbb{C}^n \ ;$$

3.
$$\|x + y\| \le \|x\| + \|y\| \quad \text{for all} \quad x, y \in \mathbb{C}^n \ .$$

Examples of Vector Norms: The most common norms on $\mathbb{C}^n$ are the following:

$$\begin{array}{rcll} |x|_\infty & = & \max_j |x_j| & \text{the maximum norm} \\ |x|_1 & = & \sum_j |x_j| & \text{the one–norm} \\ |x| & = & (\sum_j |x_j|^2)^{1/2} & \text{the Euclidean norm} \end{array}$$

Here the Euclidean vector norm $|x|$ is associated to the Euclidean inner product defined by

$$\langle x, y \rangle = \sum \bar{x}_j y_j = x^* y \ .$$

The relation is simple:

$$|x| = \langle x, x \rangle^{1/2} \ .$$

We also note the Cauchy–Schwarz inequality:

Lemma 2.1
$$|\langle x, y \rangle| \le |x||y| \quad \textit{for all} \quad x, y \in \mathbb{C}^n \ .$$

Proof: We may assume that $y \neq 0$. For all $\lambda \in \mathbb{C}$ we have:

$$\begin{aligned} 0 & \le |x + \lambda y|^2 \\ & = \langle x + \lambda y, x + \lambda y \rangle \\ & = |x|^2 + \bar{\lambda}\langle y, x \rangle + \lambda \langle x, y \rangle + |\lambda^2||y|^2 \end{aligned}$$

Set

$$\lambda = -\frac{\langle y, x \rangle}{|y|^2}$$

and obtain that

$$0 \le |x|^2 - \frac{|\langle y, x \rangle|^2}{|y|^2} - \frac{|\langle x, y \rangle|^2}{|y|^2} + \frac{|\langle y, x \rangle|^2}{|y|^2} \ ,$$

thus

$$0 \le |x|^2 - \frac{|\langle x, y \rangle|^2}{|y|^2} \ .$$

The Cauchy–Schwarz inequality follows. $\diamond$

For any real p with $1 \le p < \infty$ the vector p–norm is given by

$$|x|_p = \Big(\sum_{j=1}^n |x_j|^p \Big)^{1/p} \quad \text{for} \quad x \in \mathbb{C}^n \ .$$

Homework: Prove that $|x|_p \to |x|_\infty$ as $p \to \infty$.

Induced Matrix Norms: Given any vector norm $\| \cdot \|$ on $\mathbb{C}^n$ and given any matrix $A \in \mathbb{C}^{n \times n}$ one defines the induced matrix norm by

$$\|A\| := \max\{ \|Ax\| \ : \ x \in \mathbb{C}^n, \ \|x\| \le 1 \} \ .$$

The following lemma gives a useful characterization of the number $\|A\|$.

Lemma 2.2 *a) For all $x \in \mathbb{C}^n$ the estimate*

$$\|Ax\| \le \|A\| \|x\|$$

holds.

b) If $C \ge 0$ is a constant and if

$$\|Ax\| \le C \|x\| \quad \textit{for all} \quad x \in \mathbb{C}^n \tag{2.2}$$

then $C \ge \|A\|$.

A simple consequence of the lemma is the formula

$$\|A\| = \min\{ C \ge 0 \ : \ (2.2) \text{ holds} \} \ .$$

In other words, the number $\|A\|$ is the smallest constant C for which the estimate (2.2) holds.

It is a good exercise to compute the matrix norms corresponding to the most common vector norms.

Lemma 2.3 *We have*

$$\begin{aligned}
|A|_\infty &= \max_j \sum_k |a_{jk}| \quad (\textit{maximal row sum}) \\
|A|_1 &= \max_k \sum_j |a_{jk}| \quad (\textit{maximal column sum}) \\
|A| &= \sigma_1 \\
|A|^2 &= \rho(A^*A)
\end{aligned}$$

where σ_1 is the largest singular value of A and where $\rho(A^*A)$ is the spectral radius of the Hermitian matrix A^*A.

Proof: The proofs of the formulas for $|A|_\infty$ and $|A|_1$ are elementary. The formulas for $|A|$ use mathematical tools that we will learn later.

1. Proof of the formula for $|A|_\infty$: Set

$$C := \max_j \sum_k |a_{jk}| = \sum_k |a_{lk}| \ .$$

Here $1 \le l \le n$ denotes an index for which the equation holds.

a) For every $x \in \mathbb{C}^n$ we have the following estimates:

$$\begin{aligned} |Ax|_\infty &= \max_j |(Ax)_j| \\ &\le \max_j \sum_k |a_{jk}||x_k| \\ &\le \max_j \sum_k |a_{jk}| \, |x|_\infty \\ &= C|x|_\infty \end{aligned}$$

This proves that $|A|_\infty \le C$.

b) We now prove that the estimate cannot be improved if required for all x. Choose $x \in \mathbb{C}^n$ so that $|x_k| = 1$ and

$$a_{lk} x_k = |a_{lk}| \ .$$

(If $a_{lk} = re^{i\alpha}$ then let $x_k = e^{-i\alpha}$.)

Then we have

$$\begin{aligned} |Ax|_\infty &\ge |(Ax)_l| \\ &= |\sum_k a_{lk} x_k| \\ &= \sum_k |a_{lk}| \\ &= C \\ &= C|x|_\infty \end{aligned}$$

This shows that

$$|A|_\infty \ge C \ .$$

2. Proof of the formula $|A| = \sigma_1$: Let

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

denote a singular value decomposition of A, i.e., U and V are unitary matrices and Σ is a diagonal matrix with diagonal entries σ_j where

$$\sigma_1 \ge \ldots \ge \sigma_n \ge 0 \ .$$

The numbers σ_j are unique. They are the singular values of A. In the following, it is important to note that

$$|Wy| = |y|$$

for any unitary matrix $W \in \mathbb{C}^{n\times n}$ and any $y \in \mathbb{C}^n$.

a) For every $x \in \mathbb{C}^n$ we have the following:

$$\begin{aligned}
|Ax| &= |U\Sigma V^* x| \\
&= |\Sigma V^* x| \\
&\leq \sigma_1 |V^* x| \\
&= \sigma_1 |x|
\end{aligned}$$

This proves that

$$|A| \leq \sigma_1 \ .$$

b) To show that the estimate cannot be improved, choose

$$x = Ve^1 \ .$$

Note that $|x| = 1$. We have

$$\begin{aligned}
Ax &= U\Sigma V^* V e^1 \\
&= U\Sigma e^1 \\
&= \sigma_1 U e^1
\end{aligned}$$

Here Ue^1 is the first column of the unitary matrix U, thus $|Ue^1| = 1$. Therefore,

$$|A||x| \geq |Ax| = \sigma_1 = \sigma_1 |x| \ .$$

This shows that

$$|A| \geq \sigma_1 \ .$$

3. Proof of the formula $|A|^2 = \rho(A^*A)$:
If $A = U\Sigma V^*$ then

$$A^*A = V\Sigma U^* U \Sigma V^* = V \Sigma^2 V^* \ .$$

This shows that the matrix A^*A has the eigenvalues σ_j^2. In particular,

$$\rho(A^*A) = \sigma_1^2 = |A|^2 \ .$$

$\diamond$

Homework: Prove the formula of Lemma 2.3 for $|A|_1$.

2.2 The Condition Number

Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and let $\|\cdot\|$ denote a vector norm on $\mathbb{C}^n$. The induced matrix norm is also denoted by $\|\cdot\|$. The number

$$\kappa = \kappa(A, \|\cdot\|) = \|A\|\,\|A^{-1}\|$$

is called the condition number of A with respect to $\|\cdot\|$.

Remark: The condition number of a nonsingular matrix A with respect to the Euclidean norm is

$$\kappa_2 = \kappa(A, |\cdot|) = \frac{\sigma_1}{\sigma_n}$$

where

$$\sigma_1 \geq \ldots \geq \sigma_n > 0$$

are the singular values of A. The number κ_2 is computed by MATLAB,

$$\kappa_2 = cond(A) \ .$$

It turns out (but this is not easy to make precise) that the condition number describes the sensitivity of the solution x of the system $Ax = b$ with respect to small changes of the data, (A, b). Here one must consider *relative* data errors, as given by

$$\frac{\|\tilde{A}\|}{\|A\|} + \frac{\|\tilde{b}\|}{\|b\|} \ ,$$

and relative solution errors,

$$\frac{\|\tilde{x}\|}{\|x\|} \ .$$

Example 2.1: (a well–conditioned system) Let

$$A = \begin{pmatrix} -\varepsilon & 1 \\ 1 & 1 \end{pmatrix}, \quad A^{-1} = \frac{1}{1+\varepsilon}\begin{pmatrix} -1 & 1 \\ 1 & \varepsilon \end{pmatrix}$$

and consider the system

$$\begin{pmatrix} -\varepsilon & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

with $0 < \varepsilon << 1$. In this case

$$|A|_\infty = 2, \quad |A^{-1}|_\infty = 2/(1+\varepsilon) \ .$$

It follows that

$$\kappa = \frac{4}{1+\varepsilon} \sim 4 \ .$$

The system is well–conditioned. (Recall that Gaussian elimination with partial pivoting had no difficulty with the system, whereas the algorithm without pivoting leads to a wrong solution if $0 < |\varepsilon| < \frac{1}{2}\varepsilon_M$.)

Example 2.2: (an ill–conditioned system) Consider the system

$$\begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \varepsilon \end{pmatrix}$$

with $0 < \varepsilon << 1$. The exact solution is

$$x = \begin{pmatrix} 1 \\ -1 \end{pmatrix} .$$

In this case

$$|A|_\infty = 2 + \varepsilon$$

and

$$A^{-1} = \frac{1}{-\varepsilon} \begin{pmatrix} 1 & -1 \\ -1-\varepsilon & 1 \end{pmatrix} .$$

Therefore,

$$|A^{-1}|_\infty = \frac{2+\varepsilon}{\varepsilon} .$$

The condition number is

$$\kappa = \frac{(2+\varepsilon)^2}{\varepsilon} \sim \frac{4}{\varepsilon} .$$

In this case, if we perturb the right–hand side

$$b = \begin{pmatrix} 0 \\ \varepsilon \end{pmatrix}$$

to

$$b + \tilde{b} = \begin{pmatrix} \delta \\ \varepsilon \end{pmatrix}$$

then the solution is perturbed by

$$\tilde{x} = A^{-1} \begin{pmatrix} \delta \\ 0 \end{pmatrix} = \frac{\delta}{\varepsilon} \begin{pmatrix} -1 \\ 1+\varepsilon \end{pmatrix} .$$

For example, if

$$\varepsilon = 10^{-20} \quad \text{and} \quad \delta = 10^{-10} ,$$

then $\delta/\varepsilon = 10^{10}$. A perturbation of the right–hand side of the system of size $\delta = 10^{-10}$ leads to a rather large change of the solution: The size of the change is approximately 10^{10}. The system is ill–conditioned.

2.3 The Perturbed System $A(x + \tilde{x}) = b + \tilde{b}$

In this section, we only perturb the right–hand side of the system $Ax = b$, but leave the matrix A unperturbed.

Let $\|\cdot\|$ be any fixed norm on $\mathbb{C}^n$ and let $A \in \mathbb{C}^{n\times n}$ be nonsingular. Let $\kappa = \|A^{-1}\|\|A\|$ denote the condition number of A.

We consider the unperturbed system $Ax = b$ with solution $x = A^{-1}b$ and the perturbed system

$$A(x + \tilde{x}) = b + \tilde{b}$$

with solution $x + \tilde{x}$. Thus, $\tilde{x} = A^{-1}\tilde{b}$ is the solution error.

We try to find a bound of the form

$$\frac{\|\tilde{x}\|}{\|x\|} \le C\frac{\|\tilde{b}\|}{\|b\|} \tag{2.3}$$

where C is *realistic*. In other words, the bound (2.3) should hold, but it should not be much too pessimistic.

We first show that the bound (2.3) holds with $C = \kappa$, the condition number. Note that (2.3) is equivalent to

$$\frac{\|b\|}{\|x\|} \cdot \frac{\|\tilde{x}\|}{\|\tilde{b}\|} \le C \tag{2.4}$$

or

$$\frac{\|Ax\|}{\|x\|} \cdot \frac{\|A^{-1}\tilde{b}\|}{\|\tilde{b}\|} \le C \ . \tag{2.5}$$

Here,

$$\begin{aligned} \|Ax\| &\le \|A\|\|x\| & (2.6)\\ \|A^{-1}\tilde{b}\| &\le \|A^{-1}\|\|\tilde{b}\| & (2.7) \end{aligned}$$

Therefore,

$$\frac{\|Ax\|}{\|x\|} \cdot \frac{\|A^{-1}\tilde{b}\|}{\|\tilde{b}\|} \le \|A\|\|A^{-1}\| = \kappa \ . \tag{2.8}$$

Lemma 2.4 *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. If $Ax = b$ and $A(x + \tilde{x}) = b + \tilde{b}$ then the bound*

$$\frac{\|\tilde{x}\|}{\|x\|} \le C\frac{\|\tilde{b}\|}{\|b\|} \tag{2.9}$$

holds with $C = \kappa$. Furthermore, if we require the bound (2.9) with a constant C which depends only on A, but neither on x nor $\tilde{b}$, then the choice $C = \kappa$ is best possible.

Proof: We have shown that (2.9) holds with $C = \kappa$. We have only made the estimates (2.6) and (2.7). These estimates cannot be improved if required for all x and all $\tilde{b}$. $\diamond$

Remark: In many applications, in particular to discretizations of differential equations, the estimate

$$\|Ax\| \leq \|A\|\|x\|$$

is too pessimistic (see Section 2.4.). One might therefore believe that the condition number κ is not a realistic measure for the sensitivity of the system $Ax = b$. However, when analyzing computations in floating point arithmetic, it turns out that one also must analyze perturbations of A. We will show that if perturbations of A occur, the condition number κ is a realistic measure of the sensitivity of the system $Ax = b$. As preparation, we will discuss the Neumann series in Section 2.5.

2.4 Example: A Discretized 4–th Order Boundary–Value problem

We give an example of a system $Ax = b$ where the estimate $\|Ax\| \leq \|A\|\|x\|$ is too pessimistic.

Consider the ODE

$$u^{IV}(t) = f(t), \quad 0 \leq t \leq 1$$

with boundary conditions

$$u(0) = u''(0) = u(1) = u''(1) = 0 \ .$$

Here $f(t), 0 \leq t \leq 1$, denotes a given smooth function.

Let $h = 1/(n+1)$ denote a grid size and let

$$t_j = jh, \quad j = -1, 0, \ldots, n+2 \ ,$$

denote grid points. The discretized boundary conditions are

$$\begin{aligned} u_0 = u_{n+1} &= 0 \\ u_{-1} - 2u_0 + u_1 &= 0 \\ u_n - 2u_{n+1} + u_{n+2} &= 0 \end{aligned}$$

and the discretized ODE is

$$h^{-4}\Big(u_{j-2} - 4u_{j-1} + 6u_j - 4u_{j+1} + u_{j+2}\Big) = f(t_j), \quad j = 1, 2, \ldots, n \ .$$

Using the discretized boundary conditions, one can eliminate the unknowns

$$u_{-1}, u_0, u_{n+1}, u_{n+2}$$

and obtain a system

$$A_h u_h = f_h$$

with

$$A_h \in \mathbb{R}^{n\times n}, \quad u_h, f_h \in \mathbb{R}^n \ .$$

Note: The conditions $u_0 = 0$ and $u_{-1} - 2u_0 + u_1 = 0$ yield that $u_{-1} = -u_1$. Therefore, the difference equation

$$h^{-4}\Big(u_{-1} - 4u_0 + 6u_1 - 4u_2 + u_3\Big) = f(t_1)$$

becomes

$$h^{-4}\Big(5u_1 - 4u_2 + u_3\Big) = f(t_1) \ .$$

One obtains:

$$A_h = \frac{1}{h^4}\begin{pmatrix} 5 & -4 & 1 & & & & 0 \\ -4 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 6 & -4 \\ 0 & & & & 1 & -4 & 5 \end{pmatrix}, \quad u_h = \begin{pmatrix} u_1 \\ \vdots \\ \\ \\ \vdots \\ u_n \end{pmatrix}, \quad f_h = \begin{pmatrix} f(h) \\ \vdots \\ \\ \\ \vdots \\ f(1-h) \end{pmatrix} .$$

We have

$$|A_h|_\infty = 16h^{-4} \ .$$

If $h = 0.01$, for example, then

$$|A_h|_\infty = 1.6 * 10^9 \ .$$

This is a rather large number. However, if $f(t)$ is a smooth function with maximum norm of order one, then the exact solution $u(t)$ of the BVP will also be a smooth function with maximum norm of order one, and we can expect that

$$|u_h|_\infty = \mathcal{O}(1), \quad |f_h|_\infty = \mathcal{O}(1) \ .$$

This holds since the error $u - u_h$ is of order $\mathcal{O}(h^2)$ in maximum norm.

Therefore, the estimate

$$|f_h|_\infty = |A_h u_h|_\infty \le |A_h|_\infty \, |u_h|_\infty$$

is too pessimistic. This may suggest that the condition number of the system

$$A_h u_h = f_h$$

is not a good tool for analyzing the effect of perturbations. The condition number of a matrix A turns out to be the correct tool, however, if one not only considers perturbations of the right–hand side of the system $Ax = b$, but also considers perturbations of A.

Summary: Consider a system

$$Ax = b$$

and assume that $\|x\| \sim \|b\|$ for the solution we are interested in. (Such systems often occur as discretizations of BVPs.) Now perturb the right–hand side. The perturbed system is

$$A(x + \tilde{x}) = b + \tilde{b} \ .$$

Clearly,

$$A\tilde{x} = \tilde{b} \ .$$

We obtain

$$\begin{aligned}
\frac{\|\tilde{x}\|}{\|x\|} &= \frac{\|A^{-1}\tilde{b}\|}{\|x\|} \\
&\sim \frac{\|A^{-1}\tilde{b}\|}{\|b\|} \\
&\leq \|A^{-1}\| \frac{\|\tilde{b}\|}{\|b\|}
\end{aligned}$$

Thus, if we have $\|x\| \sim \|b\|$, then the size of $\|A\|$ does not matter when we estimate the relative error of the solution by the relative error of the data.

2.5 The Neumann Series

Let $Q \in \mathbb{C}^{n \times n}$ and let $\| \cdot \|$ be a vector norm on $\mathbb{C}^n$. As before, $\|Q\|$ denotes the associate matrix norm.

We recall from analysis the geometric series for complex numbers q:

$$\sum_{j=0}^{\infty} q^j = \frac{1}{1-q} \quad \text{for} \quad |q| < 1 \ .$$

It is interesting that one can generalize the result to matrices.

Theorem 2.1 *Assume that $\|Q\| < 1$. Then $I - Q$ is an invertible matrix and*

$$\sum_{j=0}^{\infty} Q^j = (I - Q)^{-1} \ .$$

Proof: Let

$$S_k = \sum_{j=0}^{k} Q^j$$

denote the k–th partial sum of the series $\sum_{j=0}^{\infty} Q^j$. We have

$$S_k(I-Q) = (I + Q + \ldots + Q^k)(I-Q) = I - Q^{k+1} \tag{2.10}$$

Here $Q^{k+1} \to 0$ as $k \to \infty$ since $\|Q^{k+1}\| \le \|Q\|^{k+1}$ and $\|Q\| < 1$. Also,

$$\|S_l - S_k\| \le \sum_{j=k+1}^{l} \|Q\|^j < \varepsilon \quad \text{for} \quad l > k \ge N(\varepsilon) \ .$$

This implies that the sequence S_k converges,

$$S_k \to S = \sum_{j=0}^{\infty} Q^j \ ,$$

and (2.10) yields that

$$S(I-Q) = I \ .$$

This proves the theorem. $\diamond$

Remark: The series expression $\sum_{j=0}^{\infty} Q^j$ for $(I-Q)^{-1}$, called a Neumann series, generalizes to bounded linear operators $Q : U \to U$, where U is a Banach space if $\|Q\| < 1$.

Neumann series and Neumann boundary conditions are named after Carl Neumann (1832–1925). Carl Neumann studied physics with his father and spent most of his career studying mathematical problems arising from physics. He taught at multiple universities and in 1868 was a founder of the journal *Mathematische Annalen*.

A simple application of the Neumann series is the following: Let $P \in \mathbb{C}^{n \times n}$ denote a matrix and let $\varepsilon \in \mathbb{C}$ with

$$|\varepsilon| \|P\| < 1 \ .$$

Then the matrix $I + \varepsilon P$ is nonsingular and

$$(I + \varepsilon P)^{-1} = I - \varepsilon P + \mathcal{O}(\varepsilon^2) \ .$$

Here $\mathcal{O}(\varepsilon^2)$ stands for a matrix term obeying an estimate

$$\|\mathcal{O}(\varepsilon^2)\| \le C|\varepsilon|^2 \quad \text{for} \quad |\varepsilon| \le 1$$

with a constant C independent of ε.

2.6 Data Error and Solution Error

Let $Ax = b$ be a given linear system. We assume that $A \in \mathbb{C}^{n \times n}$ is nonsingular and denote the solution of the system by $x = A^{-1}b$. If we apply an algorithm such as Gaussian elimination with partial pivoting and use floating point arithmetic, then we obtain a numerical solution x_{num} which solves a nearby system

$$(A + \tilde{A})x_{num} = b + \tilde{b} \ .$$

(Estimates for $\tilde{A}$ and $\tilde{b}$ can be demonstrated using backward error analysis.)

For simplicity, let $\tilde{b} = 0$. Consider a system

$$(A + \tilde{A})(x + \tilde{x}) = b$$

and assume that the perturbation term $\tilde{A}$ is so small that

$$\|A^{-1}\tilde{A}\| << 1 \ .$$

We set

$$Q = -A^{-1}\tilde{A}$$

and rewrite the system

$$(A + \tilde{A})(x + \tilde{x}) = b$$

as follows:

$$\begin{aligned}
A(I + A^{-1}\tilde{A})(x + \tilde{x}) &= b \\
A(I - Q)(x + \tilde{x}) &= b \\
(I - Q)(x + \tilde{x}) &= x \\
x - Qx + (I - Q)\tilde{x} &= x \\
(I - Q)\tilde{x} &= Qx
\end{aligned}$$

One obtains:

$$\tilde{x} = \sum_{j=0}^{\infty} Q^{j+1} x \ .$$

This yields the estimate

$$\begin{aligned}
\|\tilde{x}\| &\leq \sum_{j=0}^{\infty} \|Q\|^{j+1} \, \|x\| \\
&= \frac{\|Q\|}{1 - \|Q\|} \, \|x\|
\end{aligned}$$

Since

$$\|Q\| \leq \|A^{-1}\| \|\tilde{A}\|$$

one obtains that

$$\frac{\|\tilde{x}\|}{\|x\|} \leq \frac{1}{1 - \|Q\|} \|A^{-1}\| \|\tilde{A}\| ,$$

thus

$$\frac{\|\tilde{x}\|}{\|x\|} \leq \frac{\|A^{-1}\| \|A\|}{1 - \|Q\|} \frac{\|\tilde{A}\|}{\|A\|} .$$

If $\|Q\| << 1$ this yields, essentially,

$$\frac{\|\tilde{x}\|}{\|x\|} \leq \kappa \frac{\|\tilde{A}\|}{\|A\|}$$

where κ is the condition number,

$$\kappa = \|A^{-1}\| \|A\| .$$

This analysis shows the significance of the condition number κ for the analysis of the perturbed system

$$(A + \tilde{A})(x + \tilde{x}) = b .$$

Summary and Rule of Thumb: Consider a linear system

$$Ax = b$$

where $A \in \mathbb{C}^{n \times n}$ is nonsingular and where $b \in \mathbb{C}^n$ is given. We denote the exact solution by

$$x = A^{-1} b .$$

A numerically computed solution x_{num} satisfies a perturbed system

$$(A + \tilde{A}) x_{num} = b + \tilde{b} .$$

If one uses a good algorithm, then one can use backward error analysis to obtain bounds for $\tilde{A}$ and $\tilde{b}$, but this is nontrivial.

Write

$$x_{num} = x + \tilde{x}, \quad Q = -A^{-1} \tilde{A}$$

and obtain

$$A(I - Q)(x + \tilde{x}) = b + \tilde{b} .$$

We assume

$$\|Q\| << 1, \quad \|\tilde{b}\| << \|b\| .$$

Therefore,

$$(I-Q)^{-1} \sim I+Q \ .$$

Obtain

$$\begin{aligned}
A(I-Q)(x+\tilde{x}) &= b+\tilde{b} \\
(I-Q)(x+\tilde{x}) &= x+A^{-1}\tilde{b} \\
x+\tilde{x} &\sim x+Qx+A^{-1}\tilde{b} \\
\tilde{x} &\sim Qx+A^{-1}\tilde{b}
\end{aligned}$$

Therefore,

$$\|\tilde{x}\| \sim \|A^{-1}\tilde{A}x\| + \|A^{-1}\tilde{b}\| \ .$$

The relative error has two terms,

$$\frac{\|\tilde{x}\|}{\|x\|} \sim \frac{\|A^{-1}\tilde{A}x\|}{\|x\|} + \frac{\|A^{-1}\tilde{b}\|}{\|x\|} \ .$$

Here the matrix $\tilde{A}$ is unstructured and one expects

$$\|A^{-1}\tilde{A}x\| \sim \|A^{-1}\|\|\tilde{A}\|\|x\| = \kappa \frac{\|\tilde{A}\|}{\|A\|}\|x\|$$

where we used the condition number

$$\kappa = \|A^{-1}\|\|A\| \ .$$

This yields

$$\frac{\|A^{-1}\tilde{A}x\|}{\|x\|} \sim \kappa \frac{\|\tilde{A}\|}{\|A\|} \ .$$

Also, $Ax = b$ implies

$$\|b\| \leq \|A\|\|x\| \ ,$$

thus

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|} \ ,$$

thus

$$\frac{\|A^{-1}\tilde{b}\|}{\|x\|} \leq \kappa \frac{\|\tilde{b}\|}{\|b\|} \ .$$

One obtains

$$\frac{\|\tilde{x}\|}{\|x\|} \sim \kappa\Big(\frac{\|\tilde{A}\|}{\|A\|} + \frac{\|\tilde{b}\|}{\|b\|}\Big) \ .$$

A reasonable (somewhat optimistic) estimate is

$$\frac{\|\tilde{A}\|}{\|A\|} + \frac{\|\tilde{b}\|}{\|b\|} \sim \varepsilon_{mach}$$

where ε_{mach} is machine epsilon, and it is assumed that a good algorithm is used to compute x_{num}.

One obtains the rule of thumb

$$\frac{\|\tilde{x}\|}{\|x\|} \sim \kappa\, \varepsilon_{mach} \ . \tag{2.11}$$

The relative error is about the condition number times machine epsilon.

3 Examples of Linear Systems: Discretization Error and Conditioning

ODEs and PDEs often cannot be solved analytically. Using a discretization process (for example, finite differences or finite elements) one replaces the differential equation (plus boundary conditions) by a finite dimensional system. If the differential problem is linear, one typically arrives at a matrix system

$$A_h u_h = b_h$$

where the index h indicates dependency on a step size h. If u is the solution of the differential problem, then the error

$$u - u_h$$

(in some norm or on some grid) is the discretization error. This error occurs because the ODE or PDE is replaced by a discrete system. As discussed in the previous chapter, another error occurs since in floating point arithmetic the solution u_h cannot be computed exactly.

Ideally, one can estimate the condition number of A_h and one can also estimate the discretization error. If this is the case, then one can choose a step–size h for which both errors are of the same order of magnitude. In the next section, we discuss two simple examples.

We will also discuss the Hilbert matrix, an example of an ill–conditioned matrix.

3.1 Difference Approximations of Boundary Value Problems

Difference Operators: Let $h > 0$ denote a step size and let

$$G_h = \{s_j = jh \; : \; j \in \mathbb{Z}\}$$

denote the corresponding one–dimensional grid. A function $u : G_h \to \mathbb{R}$ is called a grid function. One often writes

$$u(s_j) = u(jh) = u_j, \quad j \in \mathbb{Z} \; .$$

Define the shift operator E, acting on grid functions, by

$$(Eu)_j = u_{j+1}, \quad j \in \mathbb{Z} \; .$$

The powers E^ν of E are

$$(E^\nu u)_j = u_{j+\nu}, \quad j \in \mathbb{Z} \; ,$$

for $\nu \in \mathbb{Z}$. We write $I = E^0$ for the identity.

Then the forward divided difference operator D_h is defined by

$$(D_h u)_j = \frac{1}{h}(u_{j+1} - u_j) = \frac{1}{h}(E - I)u_j, \quad j \in \mathbb{Z} \; ,$$

thus

$$D_h = \frac{1}{h}(E - I) \ .$$

We have

$$\begin{aligned}
D_h^2 &= h^{-2}(E-I)^2 \\
&= h^{-2}(E^2 - 2E + I) \\
D_h^3 &= h^{-3}(E^3 - 3E^2 + 3E - I) \\
D_h^4 &= h^{-4}(E^4 - 4E^3 + 6E^2 - 4E + I)
\end{aligned}$$

etc. Centered divided difference operators are

$$\begin{aligned}
D_h^2 E^{-1} &= h^{-2}(E - 2I + E^{-1}) \\
D_h^4 E^{-2} &= h^{-4}(E^2 - 4E + 6I - 4E^{-1} + E^{-2})
\end{aligned}$$

For example,

$$(D_h^2 E^{-1})u_j = h^{-2}(u_{j+1} - 2u_j + u_{j-1}) \ .$$

One can use Taylor's formula to derive the order of approximation of difference operators. For example, let $u \in C^4[-1,1]$. We have, for small $h > 0$:

$$\begin{aligned}
u(h) &= u(0) + hDu(0) + \frac{h^2}{2}D^2u(0) + \frac{h^3}{6}D^3u(0) + \frac{h^4}{24}D^4u(\xi_1) \\
u(-h) &= u(0) - hDu(0) + \frac{h^2}{2}D^2u(0) - \frac{h^3}{6}D^3u(0) + \frac{h^4}{24}D^4u(\xi_2)
\end{aligned}$$

Adding these equations yields

$$u(h) + u(-h) = 2u(0) + h^2D^2u(0) + R(h)$$

with

$$|R(h)| \le \frac{h^4}{12}|D^4u|_\infty \ .$$

Therefore,

$$D^2u(0) = h^{-2}(u(h) - 2u(0) + u(-h)) + \mathcal{O}(h^2) \ .$$

Here the error term is bounded by $\frac{h^2}{12}|D^4u|_\infty$.

A Second–Order BVP: Let $p, f \in C[0,1]$ be given functions and let $\alpha, \beta \in \mathbb{R}$ be given numbers. We want to find a function $u \in C^2[0,1]$ with

$$-u''(s) + p(s)u(s) = f(s) \quad \text{for} \quad 0 \le s \le 1, \quad u(0) = \alpha, \quad u(1) = \beta \ .$$

One can give conditions on p and f which guarantee that the BVP has a unique solution.[2] We denote it by $u_{bvp}(s)$.

Let $n \in \mathbb{N}$ and let $h = 1/(n+1)$ denote a step size. For example, if $n = 99$ then $h = 0.01$. Let $s_j = jh, j = 0, 1, \ldots, n+1$ denote the grid with step size h in $[0, 1]$:

$$s_0 = 0 < s_1 < s_2 < \ldots < s_{n+1} = 1 \ .$$

For $j = 1, \ldots, n$ we replace the derivative operator $-u''(s_j)$ by the second–order divided difference

$$h^{-2}(-u_{j-1} + 2u_j + u_{j+1}) \ .$$

Let $p_j = p(s_j), f_j = f(s_j)$. If $u_h = (u_0, u_1, \ldots, u_{n+1})^T$ then one obtains a a matrix system

$$A_h u_h = b_h$$

with

$$b_h = (\alpha, f_1, \ldots f_n, \beta)^T \ .$$

Under reasonable assumptions, the system $A_h u_h = b_h$ has a unique solution u_h and

$$|u_{bvp} - u_h|_\infty := \max_{j=0,\ldots,n+1} |u_{bvp}(s_j) - u_j| \leq Ch^2$$

where C is a constant independent of the step size h. (Such results are shown in a numerical analysis course.) The error $|u_{bvp} - u_h|_\infty$ is called the discretization error. This error is due to replacing the BVP by a discrete problem.

We have

$$|A_h|_\infty \sim 4h^{-2}$$

and, under suitable conditions,

$$|A_h^{-1}|_\infty \sim 1 \ .$$

This implies that the conditions number is $\kappa \sim h^{-2}$. Our rule of thumb (2.11) yields

$$|u_h - u_{num}|_\infty \sim \varepsilon_M h^{-2} \ .$$

The error $|u_h - u_{num}|_\infty$ is due to solving the system $A_h u_h = b_h$ inexactly, using floating point arithmetic instead of exact arithmetic.

For example, if $h = 10^{-2}$ then the discretization error is

$$|u_{bvp} - u_h|_\infty \sim Ch^2 \sim 10^{-4} \ .$$

[2] For example, if $p, f \in C[0,1]$ and $p(s) \geq 0$ for $0 \leq s \leq 1$, then the BVP has a unique solution in $C^2[0,1]$.

The error due to round–off is

$$|u_h - u_{num}|_\infty \sim \varepsilon_M h^{-2} \sim 10^{-16} \cdot 10^4 = 10^{-12} \ .$$

We obtain that the discretization error dominates the error due to round–off if we choose the rather large step–size $h = 0.01$.

Suppose we want to reduce the discretization error and work with a much smaller step size, $h = 10^{-6}$. Now the discretization error becomes

$$|u_{bvp} - u_h|_\infty \sim Ch^2 \sim 10^{-12} \ .$$

The error due to round–off becomes

$$|u_h - u_{num}|_\infty \sim \varepsilon_M h^{-2} \sim 10^{-16} \cdot 10^{12} = 10^{-4} \ .$$

We obtain that the error due to round–off becomes dominant.

Which step size h is optimal, i.e., leads to the smallest total error? Let us assume that the discretization error is

$$\eta_{discrete} = Ch^2$$

and that the error due to floating point arithmetic is

$$\eta_{arith} = \varepsilon_M h^{-2} \ .$$

Then the total error becomes

$$\eta_{total} = Ch^2 + \varepsilon_M h^{-2} \ .$$

The two error terms are equal if

$$Ch^2 = \varepsilon_M h^{-2} \ ,$$

i.e.,

$$h = \Big(\frac{\varepsilon_M}{C} \Big)^{1/4} \ .$$

For $C = 1$ and $\varepsilon_M = 10^{-16}$ one obtains

$$h = 10^{-4}, \quad \eta_{total} \sim 10^{-8} \ .$$

A Fourth–Order BVP: Let $p, f \in C[0,1]$ be given functions. Consider the BVP

$$u^{IV}(s) + p(s)u(s) = f(s) \quad \text{for} \quad 0 \le s \le 1, \quad u(0) = u''(0) = u(1) = u''(1) = 0 \ .$$

One can give conditions on p and f which guarantee that the BVP has a unique solution in $C^4[0,1]$. We then denote it by $u_{bvp}(s)$. As above, let $h = 1/(n+1)$ denote a step size and let $s_j = jh, j = -1, 0, \ldots, n+2$. We replace $u^{IV}(s_j)$ by

$$h^{-4}(u_{j-2} - 4u_{j-1} + 6u_j - 4u_{j+1} + u_{j+2}) \ .$$

This is used for $j = 1, \ldots, n$. Using the discretized boundary conditions

$$u_0 = 0, \quad u_{-1} - 2u_0 + u_1 = 0$$

we eliminate u_{-1} and u_0 from the system. Similarly, we eliminate u_{n+1} and u_{n+2}. One then obtains a matrix equation

$$A_h u_h = b_h \quad \text{for} \quad u_h = (u_1, \ldots, u_n)^T \ .$$

Here

$$|A_h|_\infty \sim 16h^{-4} \ .$$

Under suitable assumptions,

$$|A_h^{-1}|_\infty \sim 1 \ .$$

The condition number is

$$\kappa \sim 16h^{-4} \ .$$

For the discretization error one obtains as above,

$$|u_{bvp} - u_h|_\infty \sim Ch^2 \ .$$

For the error due to round–off,

$$|u_h - u_{num}|_\infty \sim \varepsilon_M \cdot \kappa \sim 10^{-16} \cdot 16 \cdot h^{-4} \ .$$

The total error becomes

$$\eta_{total} \sim Ch^2 + 16 * 10^{-16} h^{-4} \ .$$

Assuming that $C = 1$ the two error terms become equal if

$$h^6 = 16 * 10^{-16}, \quad h = 0.0034 \ .$$

The total error becomes

$$\eta_{total} \sim 10^{-5} \ .$$

Comment: Given an analytic problem $Au = b$, one often uses a discretization technique to replace it by a matrix problem $A_h u_h = b_h$ with step size $h > 0$. The discretization error can be made arbitrarily small by sending the step size h to zero. However, if h is very small, then the error due to solving the system $A_h u_u = b_h$ in floating point arithmetic cannot be neglected. To estimate this error, the condition number of A_h is important.

3.2 An Approximation Problem and the Hilbert Matrix

The $n \times n$ Hilbert matrix $H^{(n)}$ is

$$H^{(n)} = (h_{ij})_{0 \le i,j \le n-1} \quad \text{with} \quad h_{ij} = \frac{1}{i+j+1} \ .$$

For example,

$$H^{(3)} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}$$

The Hilbert matrix $H^{(n)}$ is notoriously ill–conditioned unless n is quite small. For example, for $n = 10$ MATLAB gives[3]

$$cond(hilb(10)) \sim 1.6 \cdot 10^{13} \ .$$

(This is the condition number with respect to the Euclidean norm, computed via SVD.)

An Approximation Problem. We want to show how the Hilbert matrix comes up if one wants to solve a polynomial approximation problem. On the space $U = C[0,1]$ of continuous real–valued functions define the L_2–inner–product by

$$(u,v) = \int_0^1 u(s)v(s)\, ds \ .$$

Then

$$\|u\| = \sqrt{(u,u)}, \quad u \in U \ ,$$

denotes the corresponding L_2–norm.

Let

$$P_3 = span\{1, s, s^2, s^3\} \subset U$$

denote the subspace of all polynomial of degree ≤ 3. Let $f \in U$ be a given function. We want to determine a polynomial

$$p(s) = \sum_{j=0}^{3} \alpha_j s^j \in P_3$$

so that the error

$$\|f - p\|$$

becomes minimal, i.e., we want to determine $p \in P_3$ so that

$$\|f - p\| < \|f - q\| \quad \text{for all} \quad q \in P_3, \quad q \neq p \ . \tag{3.1}$$

[3]In Wikipedia it is claimed that the condition number of $H^{(n)}$ grows like $\mathcal{O}\Big((1+\sqrt{2})^{4n}/\sqrt{n}\Big)$ as $n \to \infty$. For example, for $n = 10$ one has $(1+\sqrt{2})^{4n}/\sqrt{n} \sim 6 * 10^{14}$.

Lemma 3.1 *The polynomial $p \in P_3$ is the best least squares approximation to $f \in C[0,1]$ if and only if*

$$(s^i, f-p) = 0, \quad i = 0,1,2,3 \ .$$

I.e., $p \in P_3$ is the best approximation to f with respect to the L_2–norm if and only if the error $f-p$ is orthogonal to the space P_3.

Proof: Assume first that $p \in P_3$ satisfies (3.1). Let $q \in P_3$ be arbitrary and consider

$$\eta(\varepsilon) = \|f - (p + \varepsilon q)\|^2 = \int_0^1 (f - p - \varepsilon q)^2 \, ds \ .$$

One obtains that

$$0 = \eta'(0) = -2(q, f-p) \ ,$$

This shows that the error $f-p$ is orthogonal to P_3.

Second, assume that $f-p$ is orthogonal to P_3. Then, for all $\delta \in P_3, \delta \neq 0$:

$$\begin{aligned} \|f-p-\delta\|^2 &= (f-p-\delta, f-p-\delta) \\ &= \|f-p\|^2 + \|\delta\|^2 \\ &> \|f-p\|^2 \end{aligned}$$

which proves (3.1). $\diamond$

The polynomial

$$p(s) = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \alpha_3 s^3 \in P_3$$

is the best approximation to f if and only if

$$(s^i, f-p) = 0, \quad i = 0,1,2,3 \ .$$

This requires

$$\sum_{j=0}^{3} \alpha_j (s^i, s^j) = (s^i, f), \quad i = 0,1,2,3 \ .$$

Here

$$(s^i, s^j) = \int_0^1 s^{i+j} \, ds = \frac{1}{i+j+1} = h_{ij} \ .$$

One obtains the system

$$H^{(4)} \alpha = b \quad \text{with} \quad b_i = (s^i, f), \quad i = 0,1,2,3$$

for the coefficient vector α of the optimal polynomial $p \in P_3$.

The Hilbert matrix $H^{(n+1)} \in \mathbb{R}^{(n+1)\times(n+1)}$ is the matrix with elements

$$h_{ij} = (s^i, s^j) = \frac{1}{i+j+1} \quad \text{for} \quad 0 \le i, j \le n \ .$$

We claim that $H^{(n+1)}$ is nonsingular: Let $H^{(n+1)}\alpha = 0$. Set

$$p(s) = \sum_{j=0}^{n} \alpha_j s^j \in P_n \ .$$

For $0 \le i \le n$ we have

$$\begin{aligned} 0 &= \sum_{j=0}^{n} h_{ij}\alpha_j \\ &= \sum_{j=0}^{n} (s^i, s^j)\alpha_j \\ &= (s^i, \sum_{j=0}^{n} \alpha_j s^j \\ &= (s^i, p(s)) \end{aligned}$$

Therefore,

$$0 = (p,p) = \int_0^1 (p(s))^2 \, ds \ ,$$

and $p \equiv 0$ follows.

Theorem 3.1 *Let $f \in C[0,1]$. There exists a unique $p \in P_n$ with*

$$\|f - p\| < \|f - q\| \quad \textit{for all} \quad q \in P_n, \quad q \ne p \ .$$

Proof: Let $\phi \in \mathbb{R}^{n+1}$ denote the vector with entries

$$\phi_j = (s^j, f) \quad \text{for} \quad j = 0, 1, \ldots, n$$

and let

$$H^{(n+1)}\alpha = \phi \ .$$

Set

$$p(s) = \sum_{j=0}^{n} \alpha_j s^j \ .$$

Then we have for $i = 0, 1, \ldots, n$:

$$\begin{aligned}(s^i, f-p) &= (s^i, f) - (s^i, p) \\ &= \phi_i - \sum_{j=0}^{n} h_{ij}\alpha_j \\ &= 0\end{aligned}$$

This proves that $f - p$ is orthogonal to P_n, and the claim follows. $\diamond$

Remarks: In MATLAB the n–th Hilbert matrix can be obtained by

$$A = hilb(n) \ .$$

The condition number k of A (with respect to the matrix norm corresponding to the Euclidean vector norm) can be obtained by

$$k = cond(A) \ .$$

For $n = 10$ MATLAB gives the condition number $k_{10} = 1.6 * 10^{13}$. For $n = 20$ MATLAB gives the condition number $k_{20} = 1.8 * 10^{20}$.

4 Rectangular Systems: The Four Fundamental Subspaces of a Matrix

If W is a vector space and U and V are subspaces of W, then the set

$$U + V = \{u + v \ : \ u \in U, v \in V\}$$

is again a subspace of W, the (algebraic) sum of U and V. If the subspaces U and V intersect only trivially, i.e., $U \cap V = \{0\}$, then every $w \in U + V$ has a *unique* representation of the form

$$w = u + v \quad \text{with} \quad u \in U \quad \text{and} \quad v \in V \ .$$

In this case, one writes

$$U + V = U \oplus V \ ,$$

and calls the sum $U \oplus V$ the **direct sum** of U and V.

In this chapter, F denotes an arbitrary field and $A \in F^{m \times n}$ denotes a matrix with transpose $A^T \in F^{n \times m}$. As usual, the matrices A and A^T determine linear maps, again denoted by A and A^T,

$$A : F^n \to F^m, \quad A^T : F^m \to F^n \ .$$

The nullspace of A,

$$N(A) = \{x \in F^n \ : \ Ax = 0\}$$

and the range of A^T,

$$R(A^T) = \{x \in F^n \ : \ \text{there exists } y \in F^m \text{ with } x = A^T y\}$$

are subspaces of F^n. Similarly, the nullspace of A^T,

$$N(A^T) = \{y \in F^m \ : \ A^T y = 0\}$$

and the range of A,

$$R(A) = \{y \in F^m \ : \ \text{there exists } x \in F^n \text{ with } y = Ax\}$$

are subspaces of F^n. The basic subject of this chapter is to study how the four fundamental subspaces of A,

$$N(A), \quad R(A), \quad N(A^T), \quad R(A^T)$$

are related to each other.

An important result will be the direct sum decompositions

$$N(A^T) \oplus R(A) = \mathbb{R}^m, \quad N(A) \oplus R(A^T) = \mathbb{R}^n$$

if F is the field of real numbers.

4.1 Dimensions of Ranges and Rank

Let F denote an arbitrary field and let $A \in F^{m\times n}$. The matrices A and A^T determine linear maps, which we again denote by A and A^T,

$$A : F^n \to F^m, \quad A^T : F^m \to F^n \ .$$

The subspace

$$N(A) = \{x \in F^n \ : \ Ax = 0\} \subset F^n$$

is called the nullspace or the kernel of A. The subspace

$$R(A) = \{y \in F^m \ : \ \text{there exists } x \in F^n \text{ with } y = Ax\} \subset F^m$$

is called the range of A. The four fundamental subspaces of A are

$$N(A), \quad R(A), \quad N(A^T), \quad R(A^T) \ ,$$

where

$$N(A) + R(A^T) \subset F^n \quad \text{and} \quad N(A^T) + R(A) \subset F^m \ .$$

Conservation of dimension, proved in Section 4.2, yields that

$$dim\, N(A) + dim\, R(A) = n \quad \text{and} \quad dim\, N(A^T) + dim\, R(A^T) = m \ . \tag{4.1}$$

Another remarkable result is that

$$dim\, R(A) = dim\, R(A^T) \ , \tag{4.2}$$

which we prove in Section 4.5 using the row echelon form of A.

Definition: The number defined by (4.2) is called the rank of the matrix A,

$$dim\, R(A) = dim\, R(A^T) =: rank(A) \ . \tag{4.3}$$

From (4.1) and (4.2) it follows that

$$dim\, N(A) + dim\, R(A^T) = n \quad \text{and} \quad dim\, N(A^T) + dim\, R(A) = m \tag{4.4}$$

where

$$N(A) + R(A^T) \subset F^n \quad \text{and} \quad N(A^T) + R(A) \subset F^m \ . \tag{4.5}$$

If F is any of the fields $\mathbb{Q}$ or $\mathbb{R}$ then one can show, in addition, that

$$N(A) \cap R(A^T) = \{0\} \quad \text{and} \quad N(A^T) \cap R(A) = \{0\} \ .$$

Together with (4.5) one then obtains the important direct sum decompositions

$$N(A) \oplus R(A^T) = F^n \quad \text{and} \quad N(A^T) \oplus R(A) = F^m \ . \tag{4.6}$$

If F is $\mathbb{Q}$ or $\mathbb{R}$ then these are orthogonal direct sums, i.e.,

$$N(A) \perp R(A^T) \quad \text{and} \quad N(A^T) \perp R(A) \ .$$

If $F = \mathbb{C}$ and one replaces A^T by $A^* = \bar{A}^T$ then one also obtains that

$$N(A) \oplus R(A^*) = \mathbb{C}^n \quad \text{and} \quad N(A^*) \oplus R(A) = \mathbb{C}^m \ . \tag{4.7}$$

The decompositions are again orthogonal.

An important implication of the orthogonal decomposition

$$N(A^*) \oplus R(A) = \mathbb{C}^m \quad \text{where} \quad N(A^*) \perp R(A)$$

(for $F = \mathbb{C}$) is the following:

Theorem 4.1 *Let $A \in \mathbb{C}^{m \times n}$ denote a complex matrix and let $b \in \mathbb{C}^m$ denote a given vector. The linear system*

$$Ax = b$$

has a solution $x \in \mathbb{C}^n$ if and only if the right–hand side b is orthogonal to every vector $\xi \in \mathbb{C}^m$ with $A^* \xi = 0$.

For a general field F the equations in (4.6) do not always hold. For example, if $F = K_2 = \{0, 1\}$ and

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

then

$$N(A) = R(A^T) = \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} .$$

4.2 Conservation of Dimension

Let U and V denote vector spaces over the field F and let $A : U \to V$ denote a linear map.

By definition, the nullspace of A is

$$N(A) = \{ u \in U \ : \ Au = 0 \}$$

and the range of A is

$$R(A) = \{ v \in V \ : \ \text{there exists } w \in U \text{ with } Aw = v \} \ .$$

It is easy to see that $N(A)$ is a subspace of U and $R(A)$ is a subspace of V. The following theorem is called *conservation of dimension*.

Theorem 4.2 *(conservation of dimension) Let U and V denote vector spaces and let $A : U \to V$ denote a linear operator. If $dim\, U < \infty$ then*

$$dim\, N(A) + dim\, R(A) = dim\, U \ .$$

Proof: We first assume that $R(A)$ has finite dimension. Let $u_1, \ldots, u_l$ denote a basis of $N(A)$ and let $v_1, \ldots, v_k$ denote a basis of $R(A)$. There exist $w_1, \ldots, w_k \in U$ with $Aw_j = v_j$. We claim that the $l + k$ vectors

$$u_1, \ldots, u_l, w_1, \ldots, w_k$$

form a basis of U.

a) (linear independence) Assume that

$$\alpha_1 u_1 + \ldots + \alpha_l u_l + \beta_1 w_1 + \ldots + \beta_k w_k = 0 \ .$$

Applying A we find that

$$\beta_1 v_1 + \ldots + \beta_k v_k = 0 \ .$$

This implies that $\beta_j = 0$ for all j, and then $\alpha_j = 0$ follows.

b) (the span is U) Let $u \in U$ be arbitrary. Then $Au \in R(A)$, thus there exist scalars $\beta_1, \ldots, \beta_k \in F$ with

$$Au = \beta_1 v_1 + \ldots + \beta_k v_k = \beta_1 Aw_1 + \ldots + \beta_k Aw_k \ .$$

Set

$$w = \beta_1 w_1 + \ldots + \beta_k w_k \ .$$

The above equation implies that

$$A(u - w) = 0 \ ,$$

thus $u - w \in N(A)$, thus

$$u - w = \alpha_1 u_1 + \ldots + \alpha_l u_l \ .$$

We have shown that

$$u = \alpha_1 u_1 + \ldots + \alpha_l u_l + \beta_1 w_1 + \ldots + \beta_k w_k \ .$$

The two arguments given above prove the formula $dim\, N(A) + dim\, R(A) = dim\, U$ under the assumption that $R(A)$ has finite dimension. If $dim\, R(A) = \infty$ then choose k so large that

$$l + k > dim\, U$$

where $l = dim\, N(A)$. If $v_1, \ldots, v_k$ are linear independent vectors in $R(A)$ and $Aw_j = v_j$, then the above argument shows that the $l + k$ vectors

$$u_1, \ldots, u_l, w_1, \ldots, w_k$$

are linearly independent, a contradiction to $l + k > dim\, U$. $\diamond$

4.3 On the Transpose A^T

In this section, F denotes an arbitrary field and we use the notation

$$\langle x, y\rangle_n = \sum_{j=1}^n x_j y_j \quad \text{for} \quad x, y \in F^n \; .$$

Lemma 4.1 *Let $x \in F^n$ and assume that*

$$\langle x, y\rangle_n = 0 \quad \textit{for all} \quad y \in F^n \; .$$

Then $x = 0$.

Proof: Taking $y = (1, 0, \ldots, 0)^T$ one obtains that $x_1 = 0$, etc. $\diamond$

Lemma 4.2 *For all $A \in F^{m\times n}, x \in F^m, y \in F^n$ the formula*

$$\langle x, Ay\rangle_m = \langle A^T x, y\rangle_n$$

holds.

Proof: We have

$$\begin{aligned}
\langle x, Ay\rangle_m &= \sum_{i=1}^m x_i (Ay)_i \\
&= \sum_{i=1}^m x_i \sum_{j=1}^n a_{ij} y_j \\
&= \sum_{j=1}^n \Big(\sum_{i=1}^m a_{ij} x_i \Big) y_j \\
&= \sum_{j=1}^n (A^T x)_j y_j \\
&= \langle A^T x, y\rangle_n
\end{aligned}$$

$\diamond$

Lemma 4.3 *Let $A \in F^{m\times n}, B \in F^{n\times m}$. If the equation*

$$\langle x, Ay\rangle_m = \langle Bx, y\rangle_n$$

holds for all $x \in F^m$ and all $y \in F^n$, then $B = A^T$.

Proof: By the previous lemma we have

$$\langle A^T x, y\rangle_m = \langle Bx, y\rangle_n \quad \text{for all} \quad x \in F^m, y \in F^n \; .$$

Therefore, by Lemma 4.1, $A^T x = Bx$ for all $x \in F^m$. This implies that $B = A^T$.
$\diamond$

Lemma 4.4 *Let $A \in F^{m\times n}, B \in F^{n\times l}$. Then*

$$(AB)^T = B^T A^T \ .$$

Proof: We have

$$\begin{aligned} \langle x, ABy \rangle &= \langle A^T x, By \rangle \\ &= \langle B^T A^T x, y \rangle \end{aligned}$$

and

$$\langle x, ABy \rangle = \langle (AB)^T x, y \rangle \ ,$$

thus

$$\langle B^T A^T x, y \rangle = \langle (AB)^T x, y \rangle \ .$$

The equation $B^T A^T = (AB)^T$ follows. $\diamond$

We know that a square matrix $A \in F^{n\times n}$ is invertible if and only if $Ax = 0$ implies $x = 0$.

Lemma 4.5 *A matrix $A \in F^{n\times n}$ is invertible if and only if A^T is invertible.*

Proof: Assume that A is invertible and let $A^T y = 0$ for some $y \in F^n$. Given any $b \in F^n$ there exists a unique $x \in F^n$ with $Ax = b$. This yields

$$\begin{aligned} \langle b, y \rangle &= \langle Ax, y \rangle \\ &= \langle x, A^T y \rangle \\ &= 0 \end{aligned}$$

It follows that $y = 0$, thus A^T is invertible. Conversely, if one assumes that A^T is invertible, then $(A^T)^T = A$ is invertible. $\diamond$

4.4 Reduction to Row Echelon Form: An Example

Let

$$E = \begin{pmatrix} e_{11} & e_{12} & \dots & e_{1n} \\ \vdots & \vdots & & \vdots \\ e_{m1} & e_{m2} & \dots & e_{mn} \end{pmatrix}$$

denote a matrix in $F^{m\times n}$. We denotes its rows by

$$E_i = (e_{i1}, \dots, e_{in}) \quad \text{for} \quad i = 1, \dots, m \ .$$

We give a general definition, which may be difficult to comprehend.

Definition: The matrix $E \in F^{m\times n}$ has row–echelon form if the following two conditions hold:

1) If $E_i = 0$ and $i < m$ then $E_{i+1} = 0$.

2) If $E_i \neq 0$ then let d_i denote the smallest index j with $e_{ij} \neq 0$. If E has k non–zero rows, then

$$d_1 < d_2 < \ldots < d_k \ .$$

The indices $d_1, d_2, \ldots, d_k$ are called the pivot indices of the matrix E.

Example of a Matrix that has Row Echelon Form: The matrix

$$E = \begin{pmatrix} 0 & 1 & * & * & * \\ 0 & 0 & 0 & 1 & * \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{4\times 5}$$

has row echelon form. Here $*$ stands for an arbitrary scalar. The pivot indices are

$$2,\ 4,\ 5\ .$$

By a process somewhat similar to Gaussian elimination and LU–factorization, one can transform any matrix $A \in F^{m\times n}$ to row echelon form. We first give an example.

Example of Reduction to Row Echelon Form: Let

$$A = \begin{pmatrix} 1 & 2 & 1 & 3 & 3 \\ 2 & 4 & 0 & 4 & 4 \\ 1 & 2 & 3 & 5 & 5 \\ 2 & 4 & 0 & 4 & 7 \end{pmatrix} \in \mathbb{R}^{4\times 5}\ . \tag{4.8}$$

We can construct 4×4 elimination matrices E_1, E_2 and a 4×4 permutation matrix P so that

$$PE_2E_1A = E = \begin{pmatrix} 1 & 2 & 1 & 3 & 3 \\ 0 & 0 & -2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.9}$$

has row echelon form. In fact,

$$E_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -2 & 0 & 0 & 1 \end{pmatrix}, \quad E_1A = \begin{pmatrix} 1 & 2 & 1 & 3 & 3 \\ 0 & 0 & -2 & -2 & -2 \\ 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & -2 & -2 & 1 \end{pmatrix},$$

$$E_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}, \quad E_2E_1A = \begin{pmatrix} 1 & 2 & 1 & 3 & 3 \\ 0 & 0 & -2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix},$$

and

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} .$$

The product

$$H := PE_2E_1$$

is nonsingular. The matrix E has row–echelon form. The number of non–zero rows of E equals 3, which is the rank of A. The pivots of E are in columns

$$d_1 = 1, \quad d_2 = 3, \quad d_3 = 5 .$$

Construction of Bases for the Example: Consider the matrix A given in (4.8). We have constructed an invertible matrix $H \in \mathbb{R}^{4\times 4}$ so that

$$HA = E$$

has row echelon form. See (4.9). We now show how one can construct bases for the four fundamental subspaces

$$N(A), \quad R(A), \quad R(A^T), \quad N(A^T) .$$

Basis of $N(A)$: The system

$$Ax = 0$$

is equivalent to

$$Ex = 0 .$$

Therefore,

$$N(A) = N(E) .$$

We can rewrite the system $Ex = 0$ as

$$\begin{pmatrix} 1 & 1 & 3 \\ 0 & -2 & -2 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ x_5 \end{pmatrix} = -x_2 \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} - x_4 \begin{pmatrix} 3 \\ -2 \\ 0 \end{pmatrix} .$$

The variables x_1, x_3, x_5 are called the basic variables. The variables x_2 and x_4 are called the free variables. If one gives any values to the free variables, x_2 and x_4, then one can solve uniquely for the basic variables. In this way one can obtain a basis of $N(A)$. In the example, we have

$$dim N(A) = 2 = 5 - 3, \quad 3 = rank(A) .$$

One can choose

$$x_2 = 1, \quad x_4 = 0$$

to obtain the basis vector $x^{(1)}$ of $N(A)$ and

$$x_2 = 0, \quad x_4 = 1$$

to obtain the basis vector $x^{(2)}$ of $N(A)$.

Basis of $R(A)$**:** Let $a^{(1)}, \ldots, a^{(5)}$ denote the column vectors of A and let $e^{(1)}, \ldots, e^{(5)}$ denote the column vectors of E. Recall that the pivot indices are

$$d_1 = 1, \quad d_2 = 3, \quad d_3 = 5 \; .$$

We claim that the corresponding columns

$$a^{(1)}, \quad a^{(3)}, \quad a^{(5)}$$

of A form a basis of $R(A)$.

a) **Linear independence:** It is clear that the corresponding columns of E,

$$e^{(1)}, e^{(3)}, e^{(5)} \; ,$$

are linearly independent. Since $HA = E$ we have

$$Ha^{(j)} = e^{(j)} \quad \text{for all} \quad j \; .$$

If

$$\alpha_1 a^{(1)} + \alpha_3 a^{(3)} + \alpha_5 a^{(5)} = 0$$

then we apply H and obtain that

$$\alpha_1 e^{(1)} + \alpha_3 e^{(3)} + \alpha_5 e^{(5)} = 0 \; .$$

Therefore,

$$\alpha_1 = \alpha_3 = \alpha_5 = 0 \; .$$

b) **The span:** Any vector $b \in R(A)$ has the form

$$b = A\alpha = \sum_{j=1}^{5} \alpha_j a^{(j)} \; .$$

Note that

$$e^{(2)}, e^{(4)} \in span\Big(e^{(1)}, e^{(3)}, e^{(5)}\Big) \; .$$

We have, for suitable scalars γ_j,

$$
\begin{aligned}
Hb &= HA\alpha \\
&= E\alpha \\
&= \sum_{j=1}^{5} \alpha_j e^{(j)} \\
&= \gamma_1 e^{(1)} + \gamma_3 e^{(3)} + \gamma_5 e^{(5)} \\
&= \gamma_1 H a^{(1)} + \gamma_3 H a^{(3)} + \gamma_5 H a^{(5)}
\end{aligned}
$$

This implies that

$$b = \gamma_1 a^{(1)} + \gamma_3 a^{(3)} + \gamma_5 a^{(5)} \ .$$

Basis of $R(A^T)$: The equation

$$HA = E$$

implies that

$$E^T = A^T H^T \ .$$

Since H^T is nonsingular, one obtains that

$$R(A^T) = R(E^T) \ .$$

It is then clear that the first three columns of E^T form a basis of $R(A^T)$.

In particular, one obtains that

$$dim\, R(A^T) = dim\, R(A) = 3 \ .$$

Basis of $N(A^T)$: The equation

$$HA = E$$

implies that

$$E^T = A^T H^T \ .$$

If we denote the i–th row of H by

$$h^{(i)T}$$

then H^T has the columns

$$h^{(1)}, \ldots, h^{(4)} \ .$$

Since

$$A^T h^{(j)}$$

is the j–th column of E^T and the last column of E^T is 0, we obtain that

$$h^{(4)} \in N(A^T) \; .$$

We also know from the conservation of dimension that

$$dim\, N(A^T) + dim\, R(A^T) = 4 \; .$$

Since

$$\dim R(A^T) = 3$$

it follows that the vector $h^{(4)}$ forms a basis for $N(A^T)$.

4.5 The Row Echelon Form and Bases of the Four Fundamental Subspaces

Let $A \in F^{m\times n}$. It is not difficult to generalize the above example and to show the following: There exist permutation matrices $P_1, \ldots, P_k \in F^{m\times m}$ and elimination matrices $E_1, \ldots, E_k \in F^{m\times m}$ so that

$$E_k P_k \ldots E_1 P_1 A =: E$$

has row echelon form. Here $0 \leq k \leq \min\{m, n\}$ and E has k non–zero rows with pivots in columns

$$d_1 < \ldots < d_k \; .$$

Set

$$H = E_k P_k \ldots E_1 P_1 \; .$$

Then H and H^T are invertible matrices.

1. A Basis of $N(A)$**:** We have $N(A) = N(E)$. In the system

$$Ex = 0$$

the variables $x_{d_1}, \ldots, x_{d_k}$ are the basic variables whereas the other $n-k$ variables x_j are the free variables. We collect these in

$$x^{II} \in F^{n-k} \; .$$

We the choose

$$x^{II} = (1, 0, \ldots, 0)^T$$

etc. and solve for the basic variables to obtain a solution of $Ex = 0$. In this way we obtain $n-k$ vectors forming a basis of $N(A)$.

2. A Basis of $R(A)$**:**

Denote the columns of A by $a^{(1)}, \ldots, a^{(n)}$ and the columns of E by $e^{(1)}, \ldots, e^{(n)}$. We have $HA = E$, thus $Ha^{(j)} = e^{(j)}$ for $j = 1, \ldots, n$. The vectors

$$e^{(d_1)}, \dots, e^{(d_k)}$$

are linearly independent and their span is $R(E)$.

Since $Ha^{j)} = e^{(j)}$ the vectors

$$a^{(d_1)}, \dots, a^{(d_k)}$$

are linearly independent.

Let $b \in R(A)$ be arbitrary, thus $Hb \in R(E)$. We have

$$Hb = \sum_{l=1}^{k} \gamma_l e^{(d_l)} = \sum_{l=1}^{k} \gamma_l H a^{(d_l)}$$

and

$$b = \sum_{l=1}^{k} \gamma_l a^{(d_l)} \ .$$

It follows that the vectors

$$a^{(d_1)}, \dots, a^{(d_k)}$$

form a basis of $R(A)$.

3. A Basis of $R(A^T)$**:** Since $A^T H^T = E^T$ we have

$$R(A^T) = R(E^T) \ .$$

The k non–zero columns of E^T form a basis of $R(A^T)$. In particular, we note that

$$dim\, R(A) = dim\, R(A^T) \ .$$

4. A Basis of $N(A^T)$**:** Since $E^T = A^T H^T$ and since the last $m-k$ columns of E^T are zero, the last $m-k$ columns of H^T form a basis of $N(A^T)$.

5 Direct Sums and Projectors

5.1 Complementary Subspaces and Projectors

Direct Sum Decomposition: Let W denote a vector space and let U and V denote subspaces of W. One says that W is the direct sum of U and V, written

$$W = U \oplus V \ ,$$

if for every $w \in W$ there are unique vectors $u \in U$ and $v \in V$ with

$$w = u + v \ .$$

If $W = U \oplus V$ one says that U and V are complementary subspaces of W.

Motivation: One reason to write a vector space W as a direct sum, $W = U \oplus V$, is the following: Let $A : W \to W$ denote a linear operator and suppose we can find two subspaces U and V of W which are invariant under A, i.e.,

$$A(U) \subset U \quad \text{and} \quad A(V) \subset V \ .$$

If, in addition, $W = U \oplus V$ then the operator A is completely determined by the two restrictions,

$$A|_U : U \to U \quad \text{and} \quad A|V : V \to V \ .$$

To study the operator $A : W \to W$ it then suffices to study the two restrictions $A|_U$ and $A|_V$ separately, which may be easier. (Divide and conquer.)

Projectors and Direct Sums: A map $P : W \to W$ is called a projector if P is linear and $P^2 = P$.

If $W = U \oplus V$ then the assignment

$$P : \left\{ \begin{array}{ccc} W & \to & W \\ w & \to & u \end{array} \right. \quad \text{where} \quad w = u + v \quad \text{with} \quad u \in U, v \in V$$

defines a linear map $P : W \to W$ with $P^2 = P$. One calls P the projector onto U along V. It is not difficult to show that $Q = I - P$ is the projector onto V along U. Thus, any decomposition $W = U \oplus V$ determines two projectors, P and $Q = I - P$.

Conversely, one can start with a projector $P : W \to W$. If one then sets

$$U = R(P), \quad V = N(P)$$

then

$$W = U \oplus V \ ,$$

and P is the projector onto U along V.

Proof: a) (Existence of the decomposition of w) Let $w \in W$. Set

$$u = Pw, \quad v = w - Pw \ .$$

Then $w = u + v$ and $u \in R(P), v \in N(P)$. This shows that $W = U + V$.

b) (Uniqueness of the decomposition of w) Let $w \in W$ and assume that

$$w = \tilde{u} + \tilde{v}, \quad \tilde{u} \in R(P), \quad \tilde{v} \in N(P) \ .$$

Since $\tilde{u} \in R(P)$ there exists $x \in W$ with $\tilde{u} = Px$. We then have (since $P\tilde{v} = 0$)

$$u = Pw = P(\tilde{u} + \tilde{v}) = P\tilde{u} = P^2 x = Px = \tilde{u} \ .$$

The equation $\tilde{v} = v$ follows since $w = u + v = \tilde{u} + \tilde{v}$.

Let us summarize: Every decomposition of a vector space W,

$$W = U \oplus V \ ,$$

determines a project P onto U along V and a projector $Q = I - P$ onto V along U. Conversely, every projector P determines the decomposition

$$W = R(P) \oplus N(P) \ .$$

So far, there were no restrictions on the dimension of the space W. In the following, we assume that W has the finite dimension n.

Lemma 5.1 *Let $dim\, W = n < \infty$ and let U and V denote subspaces of W. We then have*

$$W = U \oplus V$$

if and only if
a) $U \cap V = \{0\}$
and
b) $dim\, U + dim\, V = dim\, W$.

Proof: 1) First assume that $W = U \oplus V$. We will prove that a) and b) hold.

a) If $w \in U \cap V$ and $w \neq 0$ then

$$w = w + 0 = 0 + w$$

would give two different decompositions, a contradiction.

b) Let $u_1, \ldots, u_l$ be a basis of U and let $v_1, \ldots, v_k$ be a basis of V. If $w \in W$ is given then there are unique α_j and β_j with

$$w = \sum \alpha_j u_j + \sum \beta_j v_j \ .$$

This shows that

$$u_1, \ldots, u_l, v_1, \ldots, v_k$$

is a basis of W. Therefore, $l + k = n$.

2) Second, assume that a) and b) hold for two subspaces U and V of W. We will prove that $W = U \oplus V$.

Let $u_1, \ldots, u_l$ denote a basis of U and let $v_1, \ldots, v_k$ denote a basis of V. By assumption, $l + k = n = dim\, W$.

Suppose that

$$\sum \alpha_j u_j + \sum \beta_j v_j = 0 \ ,$$

thus

$$\sum \alpha_j u_j = - \sum \beta_j v_j \in U \cap V = \{0\} \ .$$

It follows that

$$\alpha_j = \beta_j = 0 \ .$$

Therefore, the vectors

$$u_1, \ldots, u_l, v_1, \ldots, v_k$$

are linearly independent and, since $l + k = n$, the above vectors form a basis of W.

It follows that any $w \in W$ can be written in the form

$$w = \sum \alpha_j u_j + \sum \beta_j v_j$$

where the coefficients α_j and β_j are uniquely determined. This proves the existence and uniqueness of the decomposition

$$w = u + v \quad \text{with} \quad u \in U \quad \text{and} \quad v \in V \ .$$

$\diamond$

5.2 The Matrix Representation of a Projector

In this section let $W = \mathbb{C}^n$ and let

$$\mathbb{C}^n = U \oplus V \ ,$$

i.e., $\mathbb{C}^n$ is the direct sum of the subspaces U and V. We will derive a matrix representation of the projector $P \in \mathbb{C}^{n \times n}$ onto U along V. Let $u_1, \ldots, u_k$ denote a basis of U and let $v_1, \ldots, v_l$ denote a basis of V. By the previous lemma we have $k + l = n$ and

$$u_1, \ldots, u_k, v_1, \ldots, v_l$$

is a basis of $\mathbb{C}^n$. We place these vectors as column vectors in the matrix T,

$$T = \Big(u_1 \ldots u_k v_1 \ldots v_l \Big) \in \mathbb{C}^{n \times n} \ . \tag{5.1}$$

For any given $w \in \mathbb{C}^n$ there exists a unique $\alpha \in \mathbb{C}^k$ and a unique $\beta \in \mathbb{C}^l$ with

$$
\begin{aligned}
w &= \sum_{j=1}^{k} \alpha_j u_j + \sum_{j=1}^{l} \beta_j v_j \\
&= T \begin{pmatrix} \alpha \\ \beta \end{pmatrix}
\end{aligned}
$$

We have $Pw = \sum_{j=1}^{k} \alpha_j u_j$ and

$$
\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = T^{-1} w \ .
$$

One obtains that

$$
\begin{aligned}
Pw &= T \begin{pmatrix} \alpha \\ 0 \end{pmatrix} \\
&= T \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\
&= T \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} T^{-1} w
\end{aligned}
$$

This shows that the projector P onto U along V has the matrix form

$$
P = T \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} T^{-1} \ . \tag{5.2}
$$

If $P \in \mathbb{C}^{n\times n}$ is any projector and we set $U = R(P), V = N(P)$ then

$$
\mathbb{C}^n = U \oplus V \ .
$$

The argument given above shows that P has the form (5.2) where $k = dim R(P)$. Conversely, is is clear that the matrix P defined by (5.2) always is a projector. If T has the form (5.1) then P is the projector onto

$$
U = span\{u_1, \ldots, u_k\}
$$

along

$$
V = span\{v_1, \ldots, v_l\} \ .
$$

5.3 Orthogonal Complements in $\mathbb{C}^n$

For a subspace $U \subset \mathbb{C}^n$, denote its orthogonal complement by

$$
U^\perp = \{v \in \mathbb{C}^n \ : \ \langle u, v\rangle = 0 \quad \text{for all} \quad u \in U\} \ .
$$

Lemma 5.2 *If $U \subset \mathbb{C}^n$ is a subspace, then*

$$
dim\, U + dim\, U^\perp = n \ .
$$

Proof: Let $u_1, \ldots, u_l$ denote a basis of U and define the matrix A with columns u_j:

$$A = (u_1 \ldots u_l) \in \mathbb{C}^{n \times l} \ .$$

A vector $v \in \mathbb{C}^n$ lies in $U^\perp$ if and only if

$$u_j^* v = 0 \quad \text{for} \quad j = 1, \ldots, l \ .$$

Therefore, $v \in U^\perp$ if and only if $A^* v = 0$. Therefore,

$$U^\perp = N(A^*) \ .$$

Since $A^* : \mathbb{C}^n \to \mathbb{C}^l$ and since

$$dim\, R(A^*) = l \ ,$$

it follows from Theorem 4.2 (conservation of dimension) that

$$dim\, N(A^*) = n - l \ .$$

$\diamond$

Lemma 5.3 *Let $U \subset \mathbb{C}^n$ denote a subspace. Then we have*

$$\mathbb{C}^n = U \oplus U^\perp \quad (orthogonally) \ .$$

Proof: It is clear that

$$U \cap U^\perp = \{0\} \ .$$

By the previous lemma we have

$$dim\, U + dim\, U^\perp = n$$

and the claim follows from Lemma 5.1. $\diamond$

Lemma 5.4 *For any subspace $U \subset \mathbb{C}^n$ we have*

$$(U^\perp)^\perp = U \ .$$

5.4 The Four Fundamental Subspaces of $A \in \mathbb{C}^{m \times n}$

Let $A \in \mathbb{C}^{m \times n}$ have $rank\, A = k$. We have

$$\begin{aligned} dim\, R(A) &= dim\, R(A^*) = k \\ dim\, N(A) &= n - k \\ dim\, N(A^*) &= m - k \end{aligned}$$

Two subspaces U, V of $\mathbb{C}^n$ are called orthogonal if

$$\langle u, v\rangle = 0 \quad \text{for all} \quad u \in U, v \in V \ .$$

One then writes

$$U \perp V$$

and calls U and V orthogonal subspaces. If $U \perp V$ then $U \cap V = \{0\}$.

Lemma 5.5 *The subspaces $R(A)$ and $N(A^*)$ are orthogonal subspaces of $\mathbb{C}^m$,*

$$R(A) \perp N(A^*) \ .$$

Proof: Let $b = Ax \in R(A)$ and let $\phi \in N(A^*)$. Then we have

$$\begin{aligned} \langle b, \phi\rangle &= \langle Ax, \phi\rangle \\ &= \langle x, A^*\phi\rangle \\ &= 0 \end{aligned}$$

$\diamond$

Let $A \in \mathbb{C}^{m\times n}$ and consider the subspace

$$U = R(A) \subset \mathbb{C}^m \ .$$

If $rank\, A = k$ then

$$dim\, U = k \ .$$

By Lemma 5.2 we have

$$dim\, U^\perp = m - k \ .$$

We also know from Lemma 5.5 that

$$N(A^*) \subset U^\perp = R(A)^\perp \ .$$

Since $dim\, R(A^*) = k$ Theorem 4.2 (conservation of dimension) implies that

$$dim\, N(A^*) = m - k \ .$$

From

$$N(A^*) \subset U^\perp$$

and

$$dim\, N(A^*) = m - k = dim\, U^\perp$$

we conclude that

$$N(A^*) = U^\perp = R(A)^\perp \ .$$

We have proved the following result:

Theorem 5.1 *Let $A \in \mathbb{C}^{m\times n}$. Then the two subspaces*

$$R(A) \quad \textit{and} \quad N(A^*)$$

are orthogonal complementary subspaces of $\mathbb{C}^m$:

$$\mathbb{C}^m = R(A) \oplus N(A^*) \quad (\textit{orthogonally}) \ .$$

Therefore, the system $Ax = b$ is solvable if and only if $\langle b, \phi \rangle = 0$ for all $\phi \in \mathbb{C}^m$ with $A^*\phi = 0$.

Summary: Let $A \in \mathbb{C}^{m\times n}$. Then A defines a mapping from $\mathbb{C}^n$ to $\mathbb{C}^m$ and A^* defines a mapping from $\mathbb{C}^m$ to $\mathbb{C}^n$. The four fundamental subspaces of A are

$$N(A), \quad R(A), \quad N(A^*), \quad R(A^*) \ .$$

These lead to the following decompositions:

$$R(A^*) \oplus N(A) = \mathbb{C}^n \underset{A^*}{\overset{A}{\rightleftarrows}} \mathbb{C}^m = R(A) \oplus N(A^*)$$

Both sums,

$$R(A^*) \oplus N(A) = \mathbb{C}^n \quad \text{and} \quad \mathbb{C}^m = R(A) \oplus N(A^*) \ ,$$

are direct and orthogonal.

5.5 Orthogonal Projectors

A matrix $A \in \mathbb{R}^{n\times n}$ is called orthogonal if $A^T A = I$. If $P \in \mathbb{R}^{n\times n}$ is a projector satisfying $P^T P = I$ then $R(P) = \mathbb{R}^n$, thus $P = I$, a trivial projector. Thus, the only matrix $A \in \mathbb{R}^{n\times n}$ which is orthogonal and which is a projector is the identity, $A = I$. For this reason, it is not a good idea to call a projector P orthogonal if P is an orthogonal matrix.

An orthogonal projector is defined as follows:

Definition: A projector $P \in \mathbb{C}^{n\times n}$ is called an orthogonal projector if $R(P) \perp N(P)$.

The following theorem characterizes those projectors P which are orthogonal.

Theorem 5.2 *Let $P \in \mathbb{C}^{n\times n}$ denote a projector. The following two conditions are equivalent:*

a) $P^* = P$;

b) $R(P) \perp N(P)$.

Thus, a projector $P \in \mathbb{C}^{n\times n}$ is an orthogonal projector if and only if the matrix P is Hermitian.

Proof: a) implies b): If $P^* = P$ then, trivially, $N(P^*) = N(P)$. Since $R(P) \perp N(P^*)$ the condition b) follows.

b) implies a): Set $U = R(P), V = N(P)$. Assumption b) yields that $\mathbb{C}^n = U \oplus V$. For arbitrary vectors $w, \tilde{w} \in \mathbb{C}^n$ let

$$w = u + v, \quad \tilde{w} = \tilde{u} + \tilde{v}$$

with

$$u, \tilde{u} \in U, \quad v, \tilde{v} \in V \ .$$

We have

$$\begin{aligned} \langle \tilde{w}, Pw \rangle &= \langle \tilde{u} + \tilde{v}, u \rangle \\ &= \langle \tilde{u}, u \rangle \\ \langle P\tilde{w}, w \rangle &= \langle \tilde{u}, u + v \rangle \\ &= \langle \tilde{u}, u \rangle \end{aligned}$$

Thus,

$$\langle \tilde{w}, Pw \rangle = \langle P\tilde{w}, w \rangle$$

for all $w, \tilde{w} \in \mathbb{C}^n$. This implies that $P = P^*$.

Example of an orthogonal projector: Let $u \in \mathbb{C}^n, |u| = 1$. Then

$$P = uu^*$$

is an orthogonal projector. It is clear that

$$\begin{aligned} R(P) &= \mathit{span}\,\{u\} \\ N(P) &= \mathit{hyperplane} \perp u \end{aligned}$$

Thus, P is the projector onto $\mathit{span}\,\{u\}$ along the hyperplane orthogonal to u.

For later reference, we note the following:

Lemma 5.6 *Let $A \in \mathbb{C}^{n \times n}$ be a normal matrix, i.e., $AA^* = A^*A$. Then $N(A) = N(A^*)$.*

Proof: If $Ax = 0$ then

$$\begin{aligned} 0 &= \langle Ax, Ax \rangle \\ &= \langle A^*Ax, x \rangle \\ &= \langle AA^*x, x \rangle \\ &= \langle A^*x, A^*x \rangle \end{aligned}$$

thus $A^*x = 0$. $\diamond$

6 Variational Problems with Equality Constraints

If $F : \mathbb{R}^n \to \mathbb{R}$ is a smooth function and $x^0 \in \mathbb{R}^n$ is a local maximum or minimum of F, then

$$\nabla F(x^0) = 0 \; .$$

Here

$$\nabla F(x) = \Big(\frac{\partial F}{\partial x_1}(x), \ldots, \frac{\partial F}{\partial x_n}(x) \Big)$$

denotes the gradient of F at x.

One says that the equation $\nabla F(x^0) = 0$ is a necessary first order condition for a local extremum of F.

In this chapter we want to maximize or minimize F locally, but also require that the solution $x^0 \in \mathbb{R}^n$ satisfies m equality constraints,

$$c_i(x^0) = 0 \quad \text{for} \quad i = 1, 2, \ldots, m \; .$$

Here $c : \mathbb{R}^n \to \mathbb{R}^m$ is a given smooth function and $m < n$. If $x^0 \in \mathbb{R}^n$ is a solution of this variational problem and if the Jacobian

$$A = c'(x^0) \in \mathbb{R}^{m \times n}$$

has full rank, then the direct sum decomposition

$$N(A) \oplus R(A^T) = \mathbb{R}^n$$

will be important to understand the Lagrange function and Lagrange multipliers of the variational problem.

6.1 First Order Conditions

Let

$$F : \mathbb{R}^n \to \mathbb{R} \quad \text{and} \quad c : \mathbb{R}^n \to \mathbb{R}^m \quad \text{with} \quad m < n$$

denote smooth functions.

We want to maximize (or minimize) the function $F(x)$ subject to the constraint $c(x) = 0$. Denote the constraint manifold by

$$M = \{ x \in \mathbb{R}^n \; : \; c(x) = 0 \} \; .$$

For $\varepsilon > 0$ and $x^0 \in \mathbb{R}^n$ let

$$B_\varepsilon(x^0) = \{ x \in \mathbb{R}^n \; : \; |x - x^0| < \varepsilon \}$$

denote the open ball of radius ε centered at x^0.

A precise formulation of the **variational problem with constraints** is the following: Find

$$x^0 \in \mathbb{R}^n \quad \text{with} \quad c(x^0) = 0 \quad \text{so that, for some } \varepsilon > 0 \ , \tag{6.1}$$

$$F(x^0) \geq F(x) \quad \text{for all} \quad x \in B_\varepsilon(x^0) \cap M \ . \tag{6.2}$$

One defines the Lagrange function

$$L(x, \mu) = F(x) - \sum_{i=1}^m \mu_i c_i(x) \quad \text{for} \quad (x, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \ .$$

The parameters μ_i in the above formula are called Lagrange multipliers.

The gradient of $L(x, \mu)$ with respect to x is

$$\nabla_x L(x, \mu) = \nabla F(x) - \sum_{i=1}^m \mu_i \nabla c_i(x) \ .$$

Here, by convention, the gradients are row vectors. We introduce the Jacobian of the constraint function $c(x)$:

$$c'(x) = \begin{pmatrix} \nabla c_1(x) \\ \vdots \\ \nabla c_m(x) \end{pmatrix} \in \mathbb{R}^{m \times n} \ .$$

In column form, the vector $(\nabla_x L(x, \mu))^T$ can be rewritten as

$$(\nabla_x L(x, \mu))^T = (\nabla F(x))^T - (c'(x))^T \mu \ .$$

If we do not have any constraints and x^0 is a local maximum of $F(x)$, then

$$\nabla F(x^0) = 0 \ .$$

These are n scalar equations for the n components $x_j^0, j = 1, \ldots, n$, of the unknown vector x^0.

For the case with constraints, the following holds:

Theorem 6.1 *Assume that $x^0 \in \mathbb{R}^n$ solves the variational problem (6.1), (6.2) and assume that the Jacobian*

$$c'(x^0) =: A \in \mathbb{R}^{m \times n}$$

has full rank, i.e., rank $A = m$. Then there exists a unique vector

$$\mu^0 = \begin{pmatrix} \mu_1^0 \\ \vdots \\ \mu_m^0 \end{pmatrix} \in \mathbb{R}^m$$

so that

$$\nabla_x L(x^0, \mu^0) = 0 \ . \tag{6.3}$$

Remarks: Theorem 6.1 says that a solution $x^0 \in \mathbb{R}^n$ of the variational problem (6.1), (6.2) and the corresponding vector $\mu^0 \in \mathbb{R}^m$ of Lagrange multipliers solve the following system of equations:

$$\begin{aligned} (\nabla F(x))^T - (c'(x))^T \mu &= 0 \qquad (6.4) \\ c(x) &= 0 \qquad (6.5) \end{aligned}$$

This is a system of $n+m$ equations for the unknown vector

$$\begin{pmatrix} x \\ \mu \end{pmatrix} \in \mathbb{R}^{n+m} \; .$$

Variants of Newton's method can be applied to solve this system numerically. This is an important subject of numerical optimization.

Before proving Theorem 6.1 we consider a simple example. We write (x, y) instead of (x_1, x_2).

Example: Find the extrema of $F(x, y) = 3x + y$ on the unit circle,

$$x^2 + y^2 = 1 \; .$$

The constraint function is

$$c(x, y) = x^2 + y^2 - 1$$

and the Lagrangian is

$$L(x, y, \mu) = 3x + y - \mu(x^2 + y^2 - 1) \; .$$

The Lagrange equations (6.3) become

$$\begin{aligned} L_x(x, y, \mu) = 3 - 2\mu x &= 0 \\ L_y(x, y, \mu) = 1 - 2\mu y &= 0 \end{aligned}$$

One obtains that

$$x = \frac{3}{2\mu}, \quad y = \frac{1}{2\mu}$$

and the constraint $x^2 + y^2 = 1$ yields that

$$\frac{9}{4\mu^2} + \frac{1}{4\mu^2} = 1 \; .$$

One obtains the two solutions

$$\mu_{1,2} = \pm\sqrt{10/4} \; .$$

The extrema of F on the unit circle are attained at

$$P = \Big(\frac{3}{2\mu_1}, \frac{1}{2\mu_1}\Big) = \frac{1}{\sqrt{10}} \, (3, 1)$$

and at
$$Q = \Big(\frac{3}{2\mu_2}, \frac{1}{2\mu_2} \Big) = -\frac{1}{\sqrt{10}}\ (3,1)\ .$$
It is easy to check that
$$F(P) = \sqrt{10}, \quad F(Q) = -\sqrt{10}\ .$$
Thus, the maximum of F on the unit circle is attained at P and the minimum at Q.

Of course, this simple problem can also be solved without the Lagrangian approach: The unit circle has the parameterization
$$(x(t), y(t)) = (\cos t, \sin t), \quad 0 \le t \le 2\pi\ ,$$
which leads us to consider the function
$$\phi(t) = F(x(t), y(t)) = 3\cos t + \sin t, \quad 0 \le t \le 2\pi\ .$$
Extrema can only occur at t–values with
$$0 = \phi'(t) = -3\sin t + \cos t\ .$$
This leads to the t–values with
$$\tan t = \frac{1}{3}\ ,$$
i.e.,
$$t_{1,2} = \arctan(1/3)\ .$$
Since
$$\cos t_1 = 3/\sqrt{10} \quad \text{and} \quad \sin t_1 = 1/\sqrt{10}$$
one obtains the same points P and Q as in the Lagrangian approach.

In general, the advantage of the Lagrangian approach is that it does not require a parameterization of the constraint manifold.

Proof of Theorem 6.1: Roughly, the proof proceeds as follows: If T_{x^0} denotes the tangent space to the constraint manifold M at x^0 then $\nabla F(x^0)$ is orthogonal to T_{x_0}. We have
$$T_{x^0} = N(A)\ ,$$
and the orthogonality relation
$$(\nabla F(x^0))^T \perp N(A)$$
implies that
$$(\nabla F(x^0))^T \in R(A^T)\ .$$

Therefore, there exists $\mu^0 \in \mathbb{R}^m$ with

$$(\nabla F(x^0))^T = A^T \mu^0 \; . \tag{6.6}$$

By assumption, the columns of A^T are linearly independent, and therefore the vector μ^0 is unique. The above equation (6.6) is equivalent to (6.3).

Details: By definition, the tangent space T_{x^0} to the constraint manifold M at the point $x^0 \in M$ consists of all vectors $p \in \mathbb{R}^n$ for which there exists a parameterized curve

$$v : [-\varepsilon, \varepsilon] \to M \quad (\text{for some } \varepsilon > 0)$$

with $v(0) = x^0$ and $v'(0) = p$. From $c_i(v(t)) \equiv 0$ we obtain that

$$\nabla c_i(v(t)) \cdot v'(t) \equiv 0 \quad \text{for} \quad |t| \le \varepsilon \; ,$$

thus (at $t = 0$):

$$\nabla c_i(x^0) \cdot p = 0 \; .$$

Since this holds for $i = 1, \ldots, m$ we obtain that

$$Ap = 0 \; .$$

So far, the arguments show that

$$T_{x_0} \subset N(A) \; .$$

Conversely, if $p \in N(A)$ is arbitrary, then there exists a curve $v(t)$ as above. This can be shown rigorously using the implicit function theorem (see below.) One then obtains that $T_{x_0} = N(A)$.

Since the function $t \to F(v(t))$ has a local maximum at $t = 0$ we have

$$0 = \nabla F(v(0)) \cdot v'(0) = \nabla F(x^0) \cdot p \; .$$

We thus have shown that for all $p \in T_{x^0}$ the orthogonality relation

$$(\nabla F(x^0))^T \perp p$$

holds. Since

$$T_{x^0} = N(A)$$

it follows that

$$(\nabla F(x^0))^T \in R(A^T) \; .$$

The theorem is proved. $\diamond$

To be rigorous, we have to show the following:

Theorem 6.2 *Let $c : \mathbb{R}^n \to \mathbb{R}^m$ denote a C^2–function and let $m < n$. Let $x^0 \in \mathbb{R}^n$ and assume that*

$$c(x^0) = 0, \quad A := c'(x^0) \in \mathbb{R}^{m\times n}, \quad rank\, A = m \ .$$

If $p \in N(A)$ is given then, for some $\varepsilon > 0$, there exists a C^1–function $v : [-\varepsilon, \varepsilon] \to \mathbb{R}^n$ with

$$v(0) = x^0, \quad v'(0) = p, \quad c(v(t)) \equiv 0 \ .$$

Proof: We use the following ansatz for $v(t)$:

$$v(t) = x^0 + tp + tA^T\beta(t), \quad \beta(t) \in \mathbb{R}^m \ ,$$

with

$$\beta(0) = 0 \ .$$

The equation

$$c(x^0 + tp + tA^T\beta(t)) \equiv 0$$

consists of m equations for m variables $\beta_j(t)$.

Let us write

$$c(x^0 + h) = Ah + R(h) \quad \text{where} \quad |R(h)| \le C|h|^2 \quad \text{for} \quad |h| \le 1 \ .$$

The above equation becomes

$$\begin{aligned} 0 &= c(v(t)) \\ &= tAp + tAA^T\beta(t) + R(tp + tA^T\beta(t)) \\ &= tAA^T\beta(t) + R(tp + tA^T\beta(t)) \end{aligned}$$

If $t \neq 0$ we divide by t and obtain the equation

$$0 = AA^T\beta(t) + \frac{1}{t}\, R(tp + tA^T\beta(t)) \quad \text{for} \quad \beta(t) \in \mathbb{R}^m \ .$$

To apply the implicit function theorem, we define $\Phi : \mathbb{R}^m \times [-1, 1] \to \mathbb{R}^m$ by

$$\Phi(\beta, t) = \begin{cases} AA^T\beta, & t = 0 \\ AA^T\beta + \frac{1}{t}\, R(tp + tA^T\beta), & t \neq 0 \end{cases}$$

We have

$$\Phi(0, 0) = 0, \quad \Phi_\beta(0, 0) = AA^T \ ,$$

where AA^T is nonsingular. Since $R(h) \le C|h|^2$ it follows that $\Phi(\beta, t)$ is C^1. By the implicit function theorem, there exist $\varepsilon > 0$ and $\delta > 0$ so that the equation

$$\Phi(\beta, t) = 0$$

has a unique solution $\beta(t) \in B_\delta(0)$ for $|t| < \varepsilon$. The function $t \to \beta(t)$ satisfies $\beta(0) = 0$ and is C^1. $\diamond$

6.2 An Application of Lagrange Multipliers to a Quadratic Form

Let $Q \in \mathbb{R}^{n \times n}$ be symmetric and consider the quadratic form

$$F(x) = x^T Q x, \quad x \in \mathbb{R}^n \ .$$

We want to maximize $F(x)$ over the unit sphere

$$\{x \in \mathbb{R}^n \ : \ |x| = 1\} \ .$$

Set

$$c(x) = |x|^2 - 1, \quad x \in \mathbb{R}^n \ .$$

We have[4]

$$\nabla F(x) = 2(Qx)^T, \quad \nabla c(x) = 2x^T \ .$$

We want to apply Theorem 6.1 with $m = 1$. We have for every $x \in \mathbb{R}^n$ with $c(x) = 0$:

$$c'(x) = 2x^T \neq 0 \ ,$$

thus $c'(x)$ has full rank (equal to one). If x^0 solves the variational problem, then by Theorem 6.1 there exists $\mu_0 \in \mathbb{R}$ with

$$2(Qx^0)^T - 2\mu_0 x^{0T} = 0 \ .$$

Thus,

$$Qx^0 = \mu_0 x^0 \ .$$

In other words, the maximum of $F(x)$ is attained at an eigenvector x^0 of the matrix Q. Since

$$F(x^0) = x^{0T} Q x^0 = \mu_0$$

we also obtain that the maximal value of $F(x)$ on the unit sphere is an eigenvalue of Q.

In this example, the system (6.4), (6.5) becomes

$$Qx - \mu x = 0, \quad |x|^2 - 1 = 0 \ .$$

[4] See Lemma 6.1 below.

The bad news is that *every* scaled eigenvector x^0 with corresponding eigenvalue μ_0 solves this system,

$$Qx^0 = \mu x^0, \quad |x^0| = 1 \ .$$

In other words, the first order conditions (6.4), (6.5) are necessary, but not sufficient for a local extremum. In the next section, we will discuss second order conditions.

6.3 Second Order Conditions for a Local Minimum

Recall from calculus:

Theorem 6.3 *Let $f : \mathbb{R} \to \mathbb{R}$ denote a C^2–function.*
a) If $x^0 \in \mathbb{R}$ is a local minimum of f then

$$f'(x^0) = 0 \quad \textit{and} \quad f''(x^0) \geq 0 \ . \tag{6.7}$$

b) If $x^0 \in \mathbb{R}$ satisfies

$$f'(x^0) = 0 \quad \textit{and} \quad f''(x^0) > 0 \tag{6.8}$$

then x^0 is a local minimum of f.

Thus, the conditions (6.7) are necessary and the conditions (6.8) are sufficient for a local minimum.

The following is a generalization where x varies in $\mathbb{R}^n$. We denote with

$$F''(x) = \Big(D_i D_j F(x) \Big)_{1 \leq i,j \leq n}$$

the Hessian of F.

Theorem 6.4 *Let $F : \mathbb{R}^n \to \mathbb{R}$ denote a C^2–function.*
a) If $x^0 \in \mathbb{R}^n$ is a local minimum of F then

$$\nabla F(x^0) = 0 \quad \textit{and} \quad p^T F''(x^0) p \geq 0 \quad \textit{for all} \quad p \in \mathbb{R}^n \ . \tag{6.9}$$

b) If $x^0 \in \mathbb{R}^n$ satisfies

$$\nabla F(x^0) = 0 \quad \textit{and} \quad p^T F''(x^0) p > 0 \quad \textit{for all} \quad p \in \mathbb{R}^n \setminus \{0\} \tag{6.10}$$

then x^0 is a local minimum of F.

Proof: a) The function $f(t) = F(x^0 + tp)$ has a local minimum at $t = 0$ and we have

$$\begin{aligned} f'(t) &= \nabla F(x^0 + tp) p \\ f''(t) &= p^T F''(x^0 + tp) p \end{aligned}$$

thus

$$\begin{aligned} f'(0) &= \nabla F(x^0)p \\ f''(t) &= p^T F''(x^0)p \end{aligned}$$

Here $p \in \mathbb{R}^n$ is arbitrary. The conditions (6.9) follow.

b) Assume that (6.10) holds. Since $\nabla F(x^0) = 0$ we have by Taylor expansion

$$\begin{aligned} F(x^0 + \varepsilon p) &= F(x^0) + \frac{1}{2}\,\varepsilon^2 p^T F''(x^0)p + \mathcal{O}(\varepsilon^3|p|^3) \\ &\geq F(x^0) + \frac{1}{2}\,\alpha\varepsilon^2|p|^2 + \mathcal{O}(\varepsilon^3|p|^3) \end{aligned}$$

where $\alpha > 0$ is the smallest eigenvalue of the Hessian $F''(x^0)$. This implies that x^0 is a local minimum of F. $\diamond$

Remark: In the above proof the error term $\mathcal{O}(\varepsilon^3|p|^3)$ is correct if $F \in C^3$. If $F \in C^2$ only, then the error term should be replaced by $o(\varepsilon^2|p|^2)$.

We now derive second order conditions for a variational problem with equality constraints.

Variational Problem $\mathbf{VP_{min}}$: Let $F : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^m \to \mathbb{R}^m$ denote C^2–functions where $m < n$. Find $x^0 \in \mathbb{R}^n$ which minimizes F (locally) subject to the constraint $c(x) = 0$.

Let

$$M = \{x \in \mathbb{R}^n \; : \; c(x) = 0\}$$

denote the manifold of all x satisfying the constraint. Let us first assume that x^0 is a solution of VP_{min} and that the matrix $A = c'(x^0)$ has rank m. As we have shown above, the tangent space T_{x^0} of M at x^0 is

$$T_{x^0} = N(A) \; .$$

We set

$$g = (\nabla F(x^0))^T \; .$$

As we have shown above, there is a unique vector $\mu^0 \in \mathbb{R}^m$ of Lagrange multipliers with

$$g = A^T \mu^0 \; .$$

We can also write this as

$$\nabla F(x^0) = g^T = \mu^{0T} A = \sum_{i=1}^m \mu_i^0 \nabla c_i(x^0) \; . \tag{6.11}$$

In other words, $\nabla F(x^0)$ is a linear combination of the gradients of the constraint functions $c_i(x)$ at $x = x^0$.

Let $p \in T_{x^0} = N(A)$ be arbitrary and consider a function $v(t)$ defined for $-\varepsilon \leq t \leq \varepsilon$ with

$$v(0) = x^0, \quad v'(0) = p, \quad c(v(t)) = 0 \quad \text{for} \quad -\varepsilon \le t \le \varepsilon \ .$$

(The existence of $v(t)$ has been proved above using the implicit function theorem.) Set

$$f(t) = F(v(t)) \ .$$

Since x^0 solves VP_{min}, the function $f(t)$ has a local minimum at $t = 0$. Therefore,

$$f'(0) = 0 \quad \text{and} \quad f''(0) \ge 0 \ .$$

We have

$$\begin{aligned}
f'(t) &= F'(v(t))v'(t) \\
f''(t) &= (v'(t))^T F''(v(t))v'(t) + F'(v(t))v''(t) \\
f'(0) &= F'(x^0)v'(0) = g^T p = 0 \\
f''(0) &= p^T F''(x^0)p + g^T v''(0)
\end{aligned}$$

Therefore,

$$0 \le f''(0) = p^T F''(x^0)p + g^T v''(0) \ . \tag{6.12}$$

We also have

$$\begin{aligned}
c_i(v(t)) &= 0 \\
\nabla c_i(v(t))v'(t) &= 0 \\
(v'(t))^T(c_i''(v(t)))v'(t) + \nabla c_i(v(t))v''(t) &= 0
\end{aligned}$$

and, setting $t = 0$:

$$p^T(c_i''(x_0))p + \nabla c_i(x_0)v''(0) = 0 \ .$$

This yields

$$\nabla c_i(x_0)v''(0) = -p^T(c_i''(x_0))p \ . \tag{6.13}$$

Substituting the expression from (6.11) for g^T into (6.12) gives us

$$0 \le f''(0) = p^T F''(x^0)p + \sum_{i=1}^m \mu_i^0 \nabla c_i(x^0)v''(0) \ . \tag{6.14}$$

If we now use (6.13) we obtain that

$$0 \le f''(0) = p^T\Big(F''(x^0) - \sum_{i=1}^m \mu_i^0 c_i''(x^0)\Big)p \ . \tag{6.15}$$

To summarize, we have shown that

$$0 \le p^T \Big(F''(x^0) - \sum_{i=1}^m \mu_i^0 c_i''(x^0) \Big) p \quad \text{for all} \quad p \in N(A)$$

if x^0 solves VP_{min}. Let $p^{(1)}, \ldots, p^{(n-m)}$ denote a basis of $N(A)$ and set

$$Z = \Big(p^{(1)}, \ldots, p^{(n-m)} \Big) \in \mathbb{R}^{n \times (n-m)} \ .$$

Thus, the columns of Z form a basis of $N(A)$. Then

$$p = Z\alpha, \quad \alpha \in \mathbb{R}^{n-m} \ ,$$

is the general element of $N(A)$. One obtains from (6.15):

$$0 \le \alpha^T Z^T \Big(F''(x^0) - \sum_{i=1}^m \mu_i^0 c_i''(x^0) \Big) Z\alpha \quad \text{for all} \quad \alpha \in \mathbb{R}^{n-m} \ . \tag{6.16}$$

Definition: Let $H \in \mathbb{R}^{k \times k}, H^T = H$. The matrix H is called positive semidefinite if

$$\alpha^T H \alpha \ge 0 \quad \text{for all} \quad \alpha \in \mathbb{R}^k \ .$$

The matrix H is called positive definite if

$$\alpha^T H \alpha > 0 \quad \text{for all} \quad \alpha \in \mathbb{R}^k \setminus \{0\} \ .$$

The above considerations lead to the following result about local minima under equality constraints.

Theorem 6.5 *Let $F : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ denote C^2–functions where $m < n$. Let $x^0 \in \mathbb{R}^n$ and assume that the Jacobian $A = c'(x^0)$ has rank m. Further, assume that the columns of the matrix $Z \in \mathbb{R}^{n \times (n-m)}$ form a basis of $N(A)$.*

a) If x^0 is a solution of VP_{min} then there exists a unique $\mu^0 \in \mathbb{R}^m$ with

$$\nabla F(x^0) = \sum_{i=1}^m \mu_i^0 \nabla c_i(x^0) \ . \tag{6.17}$$

Furthermore, the matrix

$$H = Z^T \Big(F''(x^0) - \sum_{i=1}^m \mu_i^0 c_i''(x^0) \Big) Z \tag{6.18}$$

is positive semidefinite.

b) If there exists a vector $\mu^0 \in \mathbb{R}^m$ with (6.17) and if the matrix H given in (6.18) is positive definite, then x^0 is a solution of VP_{min}.

Remark: Note that the matrix

$$F''(x^0) - \sum_{i=1}^{m} \mu_i^0 c_i''(x^0) \in \mathbb{R}^{n\times n}$$

is the Hessian in x of the Lagrange function $L(x,\mu)$ evaluated at (x^0,μ^0).

6.4 Supplement

Lemma 6.1 *Let $Q \in \mathbb{R}^{n\times n}$ denote a symmetric matrix, $Q^T = Q$. The scalar function $F(x) = x^T Q x$ defined for $x \in \mathbb{R}^n$ has the gradient*

$$\nabla F(x) = \Big(D_1 F(x), \ldots, D_n F(x)\Big) = 2x^T Q\ .$$

First Proof: We have

$$\begin{aligned} F(x) &= \sum_i x_i (Qx)_i \\ &= \sum_i x_i \Big(\sum_j q_{ij} x_j \Big) \end{aligned}$$

Therefore, for $1 \le k \le n$,

$$\begin{aligned} D_k F(x) &= \sum_i \delta_{ik} (Qx)_i + \sum_i x_i \Big(\sum_j q_{ij} \delta_{jk} \Big) \\ &= (Qx)_k + \sum_i x_i q_{ik} \\ &= (Qx)_k + \sum_i q_{ki} x_i \\ &= 2(Qx)_k \end{aligned}$$

Written as a column vector,

$$(\nabla F(x))^T = 2Qx\ .$$

Written as a row vector,

$$\nabla F(x) = 2x^T Q\ .$$

Second Proof: For any $x, \xi \in \mathbb{R}^n$ and real $\varepsilon \neq 0$ we have

$$\begin{aligned} F(x+\varepsilon\xi) &= \langle x+\varepsilon\xi, Q(x+\varepsilon\xi)\rangle \\ &= \langle x, Qx\rangle + 2\varepsilon\langle Qx, \xi\rangle + \mathcal{O}(\varepsilon^2) \\ &= F(x) + 2\varepsilon\langle Qx, \xi\rangle + \mathcal{O}(\varepsilon^2) \end{aligned}$$

thus

$$\begin{aligned}\frac{1}{\varepsilon}(F(x+\varepsilon\xi)-F(x)) &= 2\langle Qx,\xi\rangle+\mathcal{O}(\varepsilon)\\ &= 2\sum_j (Qx)_j\xi_j+\mathcal{O}(\varepsilon)\end{aligned}$$

Therefore,

$$D_kF(x)=2(Qx)_k,\quad k=1,\dots,n\ .$$

6.5 The Implicit Function Theorem

Let $\mathbb{R}^n$ denote the state space and let $\mathbb{R}^m$ denote the parameter space. Let

$$\Phi:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^n$$

denote a C^1–function. Assume that $x_0\in\mathbb{R}^n$ and $\lambda_0\in\mathbb{R}^m$ satisfy

$$\Phi(x_0,\lambda_0)=0$$

and assume that the matrix

$$A=\Phi_x(x_0,\lambda_0)\in\mathbb{R}^{n\times n}$$

is nonsingular. Then there exist $\varepsilon_1>0$ and $\varepsilon_2>0$ so that for all $\lambda\in B_{\varepsilon_2}(\lambda_0)$ there exists a unique $x=x(\lambda)\in B_{\varepsilon_1}(x_0)$ with

$$\Phi(x(\lambda),\lambda)=0\quad\text{for}\quad\lambda\in B_{\varepsilon_2}(\lambda_0)\ .$$

The function $x(\lambda)$ is C^1 on $B_{\varepsilon_2}(\lambda_0)$ and $x(\lambda_0)=x_0$.

The fucntion $x(\lambda)$ is implicitly defined by the equation $\Phi(x(\lambda),\lambda)$.

7 Least Squares; Gram–Schmidt and QR–Factorization; Householder Reduction

7.1 Example of Data Fitting

Assume we are given m pairs of real numbers

$$(t_i, f_i) \in \mathbb{R}^2, \quad i = 1, 2, \ldots, m \ ,$$

and want to find a function $f(t)$ of the form

$$f(t) = x_1 + x_2 t + x_3 \sin t \tag{7.1}$$

which matches the data. The function $f(t)$ depends linearly on three parameters, x_1, x_2, x_3 and let us assume $m > 3$. How shall we choose the parameters x_1, x_2, x_3 to obtain the best fit

$$f(t_i) \sim f_i \quad \text{for} \quad i = 1, 2, \ldots, m \ ?$$

This is made precise in the following.

Let

$$A = \begin{pmatrix} 1 & t_1 & \sin t_1 \\ \vdots & \vdots & \vdots \\ 1 & t_m & \sin t_m \end{pmatrix}, \quad b = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} .$$

The requirement for the parameter vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

is

$$Ax \sim b \ .$$

In general, if $m > 3$, we do not expect that the system

$$Ax = b$$

is solvable, i.e., we do not expect to find a function $f(t)$ of the form (7.1) with

$$f(t_i) = f_i \quad \text{for} \quad i = 1, 2, \ldots, m \ .$$

Therefore, instead of trying to solve the system $Ax = b$, which probably has no solution, we will try to find a vector $x \in \mathbb{R}^3$ which minimizes the error

$$|Ax - b|^2 = \sum_{i=1}^m ((Ax - b)_i)^2 = \sum_{i=1}^m (f(t_i) - f_i)^2 \ .$$

The error consists of a sum of squares. Therefore, a vector $x^0 \in \mathbb{R}^3$ which minimizes the above expression, is called a least–squares solution of the system $Ax = b$.

Remarks: Why least squares? Good explanations, based on statistical concepts, are given by Meyer, pp. 446-448.

On Jan. 1, 1801, Giuseppe Piazzi observed Ceres, the largest dwarf planet between Mars and Jupiter. Ceres then came too close to the sun and first could not be rediscovered. Using the observed data, Carl Friedrich Gauss (1777–1855), calculated Ceres's orbit. Based on his computations, Ceres could then be found again. In his computations, Gauss invented and used the ideas of least squares. Gauss contributed to so many fields of mathematics, both pure and applied, that he is sometimes called "the Prince of Mathematics." He did extensive research on the Earth's magnetic field and in a system known as the Gaussian unit system, the unit of magnetic flux density is known as the gauss.

7.2 Least Squares Problems and the Normal Equations

Let $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$. In applications to least squares problems, one will typically have $m > n$, but it is not yet necessary to assume this.

We will say that a vector $x^0 \in \mathbb{R}^n$ is a least squares solution of the system

$$Ax = b$$

if

$$|Ax^0 - b| \leq |Ax - b| \quad \text{for all} \quad x \in \mathbb{R}^n \ . \tag{7.2}$$

The next lemma characterizes least squares solutions.

Lemma 7.1 *The vector $x^0 \in \mathbb{R}^n$ is a least squares solution of the system $Ax = b$ if and only if*

$$\langle Ax^0 - b, Ay \rangle = 0 \quad \textit{for all} \quad y \in \mathbb{R}^n \ .$$

Proof: Let $x, y \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}$ be arbitrary. Then we have

$$\begin{aligned} |A(x + \varepsilon y) - b|^2 &= \langle Ax + \varepsilon Ay - b, Ax + \varepsilon Ay - b \rangle \\ &= |Ax - b|^2 + \varepsilon^2 |Ay|^2 + 2\varepsilon \langle Ax - b, Ay \rangle \end{aligned}$$

From this we read off the following: If

$$\langle Ax - b, Ay \rangle = 0 \quad \text{for all} \quad y \in \mathbb{R}^n$$

then x is a least squares solution.

Conversely, assume that x is a least squares solution and consider the function

$$f(\varepsilon) = |Ax - b|^2 + \varepsilon^2 |Ay|^2 + 2\varepsilon \langle Ax - b, Ay \rangle$$

with

$$f'(0) = 2\langle Ax - b, Ay \rangle \ .$$

By assumption, x is a least squares solution, and we obtain that $f'(0) = 0$. $\diamond$

The lemma says that x^0 is a least squares solution of the system $Ax = b$ if and only if the error

$$Ax^0 - b$$

is orthogonal to $R(A)$. Since

$$R(A)^\perp = N(A^T)$$

we obtain that x^0 is a least squares solution of the system $Ax = b$ if and only if $Ax^0 - b$ lies in the nullspace of A^T, i.e.,

$$A^T(Ax^0 - b) = 0 \ .$$

We have proved the following result:

Theorem 7.1 *The vector $x^0 \in \mathbb{R}^n$ is a least squares solution of the system*

$$Ax = b$$

if and only if x^0 solves the so–called normal equations

$$A^T Ax = A^T b \ .$$

Warning: The matrix $A^T A$ is often ill–conditioned.

Lemma 7.2 *The normal equations are always solvable. The solution of the normal equations is unique if and only if the n columns of A are linearly independent.*

Proof: a) We first prove that

$$N(A) = N(A^T A) \ . \tag{7.3}$$

To show this, first note that $Ax = 0$ trivially implies $A^T Ax = 0$. Conversely, assume that $A^T Ax = 0$. Then we have

$$0 = \langle x, A^T Ax \rangle = \langle Ax, Ax \rangle = |Ax|^2 \ ,$$

thus $Ax = 0$. Therefore, $Ax = 0$ is equivalent to $A^T Ax = 0$, which yields (7.3).

We now use (7.3) to show that

$$R(A^T A) = R(A^T) \ .$$

Recall that $A \in \mathbb{R}^{m\times n}$. If $rank\, A = k$ then, using that $N(A) = N(A^T A)$:

$$dim\, N(A) = n - k = dim\, N(A^T A) \ .$$

Therefore,

$$dim\, R(A^T A) = k = dim\, R(A^T) \ .$$

Since the inclusion $R(A^T A) \subset R(A^T)$ is trivial, one obtains that

$$R(A^T A) = R(A^T) \ .$$

Since $A^T b \in R(A^T) = R(A^T A)$ the system

$$A^T A x = A^T b$$

is always solvable.

b) The solution of the normal equations is unique if and only if

$$\{0\} = N(A^T A) = N(A) \ .$$

The nullspace of A is trivial if and only if $Ax = 0$ implies $x = 0$. This implication holds if and only if the columns of A are linearly independent. $\diamond$

Summary: Let $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^n$ and assume that $m > n$. The system of equations $Ax = b$ has m equations for n unknowns $x_1, \ldots, x_n$. Typically, the system is not solvable if $n < m$. In the normal equations

$$A^T A x = A^T b$$

the matrix $A^T A$ is $n \times n$ and is typically nonsingular. However, $A^T A$ may be ill–conditioned, and if it is then one does not want to use Gaussian elimination to solve for x.

The QR–factorization of A and the Householder reduction give alternative methods to solve the normal equations.

7.3 The Gram–Schmidt Process and QR–Factorization

Let $A \in \mathbb{C}^{m\times n}$ have n linearly independent columns

$$a^1, \ldots, a^n \in \mathbb{C}^m \ .$$

We want to find orthonormal vectors

$$q^1, \ldots, q^n \in \mathbb{C}^m$$

so that

$$span\{q^1, \ldots, q^k\} = span\{a^1, \ldots, a^k\} \quad \text{for} \quad k = 1, \ldots, n \ .$$

The following process, called classical Gram–Schmidt process, constructs the vectors $q^1, \ldots, q^n$.

a) Set $q^1 = \frac{a^1}{|a^1|}$.

b) We wish to construct q^2 so that q^1, q^2 are orthonormal and

$$a^2 = \alpha q^1 + \beta q^2$$

for some scalars α, β. Suppose this holds. Then

$$\alpha = \langle q^1, a^2 \rangle$$

and $\beta \neq 0$. It follows that

$$q^2 = \frac{1}{\beta}\Big(a^2 - \langle q^1, a^2 \rangle q^1 \Big) \ .$$

Conversely, if we set

$$v^2 = a^2 - \langle q^1, a^2 \rangle q^1$$

and

$$q^2 = \frac{v^2}{|v^2|}$$

then q^1, q^2 are orthonormal and $span\,\{q^1, q^2\} = span\,\{a^1, a^2\}$.

c) Assume that $q^1, \ldots, q^{k-1}$ have been constructed. Proceeding as above, we find that q^k can be obtained as follows:

Set

$$v^k = a^k - \sum_{j=1}^{k-1} \langle q^j, a^k \rangle q^j$$

and

$$q^k = \frac{v^k}{|v^k|} \ .$$

We give a pseudo code for the classical Gram–Schmidt process:

Classical Gram–Schmidt: The linearly independent vectors $a^1, \ldots, a^n \in \mathbb{C}^m$ are given. The orthonormal vectors $q^1, \ldots, q^n \in \mathbb{C}^m$ and numbers r_{jk} for $1 \leq j \leq k \leq n$ are computed.

1) $r_{11} = |a^1|;\ q^1 = a^1/r_{11}$
2) for $k = 2, \ldots, n$:

for $j = 1, \ldots, k-1$
$r_{jk} = \langle q^j, a^k \rangle$
end j

$v^k = a^k - \sum_{j=1}^{k-1} r_{jk} q^j$
$r_{kk} = |v^k|$
$q^k = v^k / r_{kk}$
end k

The classical Gram–Schmidt process applied to linearly independent input vectors $a^1, \ldots, a^n \in \mathbb{C}^m$ computes orthonormal vectors $q^1, \ldots, q^n$ and numbers

$$r_{jk} = \langle q^j, a^k \rangle \quad \text{for} \quad 1 \leq j < k \leq n$$

and positive numbers

$$r_{kk} = |v^k| \quad \text{for} \quad k = 1, \ldots, n$$

so that

$$a^k = \sum_{j=1}^{k-1} r_{jk} q^j + r_{kk} q^k \ .$$

We set

$$A = (a^1, \ldots, a^n) \in \mathbb{C}^{m \times n}$$

and

$$Q = (q^1, \ldots, q^n) \in \mathbb{C}^{m \times n} \quad \text{and} \quad R = \begin{pmatrix} r_{11} & \ldots & r_{1n} \\ 0 & \ddots & \vdots \\ 0 & 0 & r_{nn} \end{pmatrix} \ .$$

Then we have

$$A = QR \quad \text{and} \quad Q^* Q = I_n \ .$$

The factorization $A = QR$ is called QR–factorization of A.

Theorem 7.2 *Let $A \in \mathbb{C}^{m \times n}$ have n linearly independent columns. There are unique matrices*

$$Q \in \mathbb{C}^{m \times n} \quad \textit{and} \quad R \in \mathbb{C}^{n \times n}$$

with the following properties:

$$A = QR \ ,$$

the columns of Q are orthonormal, the matrix R is upper triangular, and

$$r_{kk} > 0 \quad \textit{for} \quad k = 1, \ldots, n \ .$$

Proof: The classical Gram–Schmidt process produces the columns of Q and the matrix R. It remains to show uniqueness. To this end, assume that

$$A = QR = \tilde{Q}\tilde{R}$$

are two factorizations with the above properties. We then have

$$a^1 = r_{11} q^1 = \tilde{r}_{11} \tilde{q}^1 \ ,$$

which yields that

$$r_{11} = \tilde{r}_{11} \quad \text{and} \quad q^1 = \tilde{q}^1 \ .$$

Next, we have

$$a^2 = r_{12}q^1 + r_{22}q^2 = \tilde{r}_{12}q^1 + \tilde{r}_{22}\tilde{q}^2 \ .$$

Here

$$r_{12} = \langle q^1, a^2 \rangle = \tilde{r}_{12} \ .$$

etc. $\diamond$

In the classical Gram–Schmidt process the matrix R is constructed column–wise. It turns out to be numerically better to construct R row–wise to reduce round–off errors. The resulting process is called Modified Gram–Schmidt.

In the following, we assume $n = 3$, for simplicity. We give pseudo codes for Classical GS and Modified GS:

Classical GS

Column 1 of R and $\mathbf{q^1}$
$r_{11} = |a^1|$
$q^1 = a^1/r_{11}$

Column 2 of R and $\mathbf{q^2}$
$r_{12} = \langle q^1, a^2 \rangle$
$v^2 = a^2 - r_{12}q^1$
$r_{22} = |v^2|$
$q^2 = v^2/r_{22}$

Column 3 of R and $\mathbf{q^3}$
$r_{13} = \langle q^1, a^3 \rangle$
$r_{23} = \langle q^2, a^3 \rangle$
$v^3 = a^3 - r_{13}q^1 - r_{23}q^2$
$r_{33} = |v^3|$
$q^3 = v^3/r_{33}$

Modified GS

Row 1 of R and $\mathbf{q^1}$; updates of $\mathbf{a^2, a^3}$
$r_{11} = |a^1|$
$q^1 = a^1/r_{11}$
$r_{12} = \langle q^1, a^2 \rangle$
$r_{13} = \langle q^1, a^3 \rangle$
$\tilde{a}^2 = a^2 - r_{12}q^1$
$\tilde{a}^3 = a^3 - r_{13}q^1$

Row 2 of R and $\mathbf{q^2}$; update of $\mathbf{a^3}$
$v^2 = \tilde{a}^2$
$r_{22} = |v^2|$
$q^2 = v^2/r_{22}$
$r_{23} = \langle q^2, \tilde{a}^3 \rangle$
$\tilde{\tilde{a}}^3 = \tilde{a}^3 - r_{23}q^2$

Row 3 of R and $\mathbf{q^3}$

$v^3 = \tilde{\tilde{a}}^3$
$r_{33} = |v^3|$
$q^3 = v^3/r_{33}$

Difference between Classical and Modified GS: In Classical GS one computes

$$r_{23} = \langle q^2, a^3 \rangle \ .$$

In Modified GS one computes

$$r_{23} = \langle q^2, \tilde{a}^3 \rangle \quad \text{where} \quad \tilde{a}^3 = a^3 - r_{13} q^1 \ .$$

The two computed values for r_{23} agree, of course, in exact arithmetic since $\langle q^2, q^1 \rangle = 0$. Note that

$$r_{13} = \langle q^1, a^3 \rangle \ ,$$

thus

$$\tilde{a}^3 + \langle q^1, a^3 \rangle q^1 = a^3 \ . \tag{7.4}$$

It follows that

$$\langle q^1, \tilde{a}^3 \rangle = 0 \ .$$

Therefore, in equation (7.4) we have an orthogonal decomposition of a^3 and obtain that

$$|a^3|^2 = |\tilde{a}^3|^2 + |r_{13}|^2 \ ,$$

thus

$$|\tilde{a}^3| \leq |a^3| \ .$$

In general, the reduction process in Modified GS reduces the Euclidean norm of the vectors, which are used to computed inner products. This reduces the round–off errors.

7.4 Solution of the Normal Equations Using the QR–Factorization

Let $A \in \mathbb{C}^{m \times n}$ have n linearly independent columns and let $b \in \mathbb{C}^m$. The normal equations corresponding to the system

$$Ax = b \quad \text{where} \quad x \in \mathbb{C}^n$$

read

$$A^* A x = A^* b \ .$$

Here $A^* A \in \mathbb{C}^{n \times n}$ is nonsingular. If $A = QR$ is the QR factorization of A, then

$$A^* A = R^* R \quad \text{since} \quad Q^* Q = I_n \ .$$

The normal equations $R^*Rx = R^*Q^*b$ become

$$Rx = Q^*b \ ,$$

where R is upper–triangular.

When one compares the direct solution of the normal equations $A^*Ax = A^*b$ with the approach to use the QR–factorization, it is important to note that the factor R^* cancels in the equation $R^*Rx = R^*Q^*b$. This reduces the condition number of $A^*A = R^*R$ to the condition number of R when one solves the system $Rx = Q^*b$ instead of the normal equations. Roughly, one can expect that the condition number of R is about the square root of the condition number of A^*A.

7.5 Householder Reflectors

Another method to solve the normal equations $A^*Ax = A^*b$ is called Householder reduction. See the next section. In this section we introduce Householder reflectors.

We will use the following result about the eigenvalues of Hermitian and unitary matrices.

Lemma 7.3 *Let $H \in \mathbb{C}^{m\times m}$.*
a) If $H^* = H$ then all eigenvalues of H are real.
b) If $H^*H = I$ the all eigenvalues of H have the absolute value 1.

Proof: a) Let $Hx = \alpha x, x \neq 0$. We have

$$\begin{aligned} \bar{\alpha}|x|^2 &= \langle \alpha x, x\rangle \\ &= \langle Hx, x\rangle \\ &= \langle x, Hx\rangle \\ &= \langle x, \alpha x\rangle \\ &= \alpha|x|^2 \end{aligned}$$

It follows that $\bar{\alpha} = \alpha$, thus α is real.
b) Let $Hx = \alpha x, x \neq 0$. We have

$$\begin{aligned} |\alpha x|^2 &= \langle \alpha x, \alpha x\rangle \\ &= \langle Hx, Hx\rangle \\ &= \langle H^*Hx, x\rangle \\ &= |x|^2 \end{aligned}$$

If follows that $|\alpha| = 1$. $\diamond$

Let $u \in \mathbb{C}^m, |u| = 1$. The matrix

$$H = I - 2uu^* \in \mathbb{C}^{m\times m}$$

is called a **Householder reflector**. The mapping

$$x \to Hx = x - 2\langle u, x\rangle u$$

(from $\mathbb{C}^m$ onto $\mathbb{C}^m$) describes the reflection with respect to the hyperplane orthogonal to u. We have

$$\begin{aligned} H^2 &= I \\ H^* &= H \\ H^*H &= I \end{aligned}$$

Thus, H is unitary and Hermitian. Therefore, H has only the eigenvalues ± 1. It is clear that the hyperplane orthogonal to u is the eigenspace to the eigenvalue 1. Also, *span* $\{u\}$ is the eigenspace to the eigenvalue -1.

Lemma 7.4 *Let $a, b \in \mathbb{C}^m$ denote given vectors with*

$$|a| = |b| > 0, \quad a \neq b \ .$$

Set

$$H = I - 2uu^* \quad \textit{where} \quad u = \frac{a-b}{|a-b|} \ .$$

Then we have

$$Ha = b$$

if and only if $\langle a, b\rangle$ is real.

Proof: We have

$$Ha = a - \gamma(a-b) \quad \text{with} \quad \gamma = \frac{2}{|a-b|^2}\,\langle a-b, a\rangle \ .$$

Therefore,

$$Ha = (1-\gamma)a + \gamma b = b + (1-\gamma)(a-b) \ ,$$

and $Ha = b$ holds if and only if $\gamma = 1$. The condition $\gamma = 1$ is equivalent to

$$2\langle a-b, a\rangle = \langle a-b, a-b\rangle \ ,$$

i.e.,

$$\langle a-b, a+b\rangle = 0 \ ,$$

i.e.,

$$|a|^2 - |b|^2 + \langle a, b\rangle - \langle b, a\rangle = 0 \ .$$

This holds if and only if $\langle a, b\rangle$ is real. $\diamond$

Let $a \in \mathbb{C}^m$ be a given vector, $a \neq 0$. We want to find a vector

$$b = \alpha e_1 = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and a Householder reflector

$$H = I - 2uu^*$$

with $Ha = b$. Since $Ha = b$ implies

$$|\alpha| = |b| = |a|$$

we set

$$\alpha = |a|e^{i\omega}$$

where $\omega \in \mathbb{R}$ has to be determined. Write the first component of the vector a in the form

$$a_1 = |a_1|e^{i\phi} \quad \text{with} \quad \phi \in \mathbb{R} \ .$$

With these notations we have

$$\begin{aligned} \langle a, b \rangle &= \overline{a_1}|a|e^{i\omega} \\ &= |a_1|e^{-i\phi}\,|a|e^{i\omega} \\ &= |a_1||a|e^{i(\omega-\phi)} \end{aligned}$$

It is clear that $\langle a, b \rangle$ is real if we choose

$$\omega = \phi \quad \text{or} \quad \omega = \phi + \pi \ .$$

The choice

$$\omega = \phi + \pi$$

is better since possible cancellation errors are avoided when $a - b$ is formed. With $\omega = \phi + \pi$ we have

$$\begin{aligned} (a-b)_1 &= a_1 - \alpha \\ &= |a_1|e^{i\phi} - |a|e^{i\omega} \\ &= (|a_1| + |a|)e^{i\phi} \end{aligned}$$

The choice $\omega = \phi$ would lead to

$$\begin{aligned} (a-b)_1 &= a_1 - \alpha \\ &= |a_1|e^{i\phi} - |a|e^{i\omega} \\ &= (|a_1| - |a|)e^{i\phi} \end{aligned}$$

If $|a_1| \sim |a|$ then the choice $\omega = \phi$ leads to $b \sim a$ and cancellation errors occur when $a - b$ is formed.

We summarize the result in the following lemma, which will be used repeatedly in the Householder reduction process.

Lemma 7.5 *Let $a \in \mathbb{C}^m, a \neq 0, a_1 = |a_1|e^{i\phi}$. Set*

$$b = \alpha e_1 = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where

$$\alpha = -|a|e^{i\phi} \ .$$

Set

$$H = I - 2uu^* \quad \textit{where} \quad u = \frac{a-b}{|a-b|} \ .$$

Then we have

$$Ha = b = \alpha e_1 \ .$$

7.6 Householder Reduction

Let $A \in \mathbb{C}^{m \times n}$ have n linearly independent columns $a^1, \ldots, a^n \in \mathbb{C}^m$. We determine a number $\alpha = \alpha_1 \in \mathbb{C}$ with $|\alpha_1| = |a^1|$ and a Householder reflector $H_1 = I - 2uu^* \in \mathbb{C}^{m \times m}$ as in Lemma 7.5 and obtain

$$H_1 a^1 = \begin{pmatrix} \alpha_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{C}^m \ .$$

Define the matrix A_2 by

$$H_1 A = \begin{pmatrix} \alpha_1 & * & \ldots & * \\ 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{pmatrix} \quad \text{with} \quad A_2 \in \mathbb{C}^{(m-1)\times(n-1)} \ .$$

We now apply the same process to A_2 and construct a Householder reflector $\tilde{H}_2 \in \mathbb{C}^{(m-1)\times(m-1)}$ with

$$\tilde{H}_2 A_2 = \begin{pmatrix} \alpha_2 & * & \ldots & * \\ 0 & & & \\ \vdots & & A_3 & \\ 0 & & & \end{pmatrix} \quad \text{with} \quad A_3 \in \mathbb{C}^{(m-2)\times(n-2)} \ .$$

Note that $\tilde{H}_2$ has dimensions $(m-1)\times(m-1)$. To obtain an $m\times m$ matrix we supplement $\tilde{H}_2$ by a trivial border and set

$$H_2 = \begin{pmatrix} 1 & 0\ldots & 0 \\ 0 & & \\ \vdots & & \tilde{H}_2 \\ 0 & & \end{pmatrix} .$$

This yields

$$H_2H_1A = \begin{pmatrix} \alpha_1 & * & * & \ldots & * \\ 0 & \alpha_2 & * & \ldots & * \\ 0 & 0 & & & \\ \vdots & \vdots & & A_3 & \\ 0 & 0 & & & \end{pmatrix} .$$

The process can be continued. After n steps we have

$$H_n\cdots H_2H_1A = \begin{pmatrix} R \\ 0 \end{pmatrix} \in \mathbb{C}^{m\times n} \tag{7.5}$$

where

$$R = \begin{pmatrix} \alpha_1 & * & * & \ldots & * \\ 0 & \alpha_2 & * & \ldots & * \\ \vdots & \ddots & \ddots & * & * \\ \vdots & & \ddots & \ddots & * \\ 0 & \ldots & \ldots & 0 & \alpha_n \end{pmatrix} \in \mathbb{C}^{n\times n} .$$

Application to the Solution of the Normal Equations: Let $A\in\mathbb{C}^{m\times n}$ have n linearly independent columns and consider the normal equations

$$A^*Ax = A^*b \quad \text{for} \quad x\in\mathbb{C}^n .$$

Here $b\in\mathbb{C}^m$ is a given vector and $m\geq n$. Most often, one has $m>n$.

Let $H_1, H_2, \ldots, H_n \in \mathbb{C}^{m\times n}$ be constructed as above, thus

$$H_n\cdots H_2H_1A = \begin{pmatrix} R \\ 0 \end{pmatrix} \in \mathbb{C}^{m\times n} \tag{7.6}$$

where $R\in\mathbb{C}^{n\times n}$ is upper triangular.

Recall that $H_j^2 = I$ and $H_j = H_j^*$. Therefore, (7.6) yields that

$$\begin{aligned} A &= H_1\cdots H_n \begin{pmatrix} R \\ 0 \end{pmatrix} \\ A^* &= (R^*\ 0)H_n\cdots H_1 \\ A^*A &= (R^*\ 0)\begin{pmatrix} R \\ 0 \end{pmatrix} = R^*R \end{aligned}$$

The normal equations $A^*Ax = A^*b$ become

$$R^*Rx = R^*(Hb)^I \quad \text{with} \quad H = H_n \cdots H_1$$

where the vector $(Hb)^I$ contains the first n components of the vector $Hb \in \mathbb{C}^m$. It is interesting that the factor R^* cancels and one obtains the system

$$Rx = (Hb)^I$$

for the solution x of the normal equations $A^*Ax = A^*b$. Since R is upper triangular, the above system is easy to solve and the cancellation of R^* reduces the condition number.

8 The Singular Value Decomposition

8.1 Theoretical Construction of an SVD

Let $A \in \mathbb{C}^{m\times n}$ have $rank\, A = r$. Let

$$p = \min\{m,n\} \ .$$

We will show that one can factorize A in the form

$$A = U\Sigma V^* \tag{8.1}$$

where $U \in \mathbb{C}^{m\times m}$ is unitary, $V \in \mathbb{C}^{n\times n}$ is unitary and $\Sigma \in \mathbb{R}^{m\times n}$ is "almost" diagonal. With a suitable $p\times p$ diagonal matrix

$$\tilde{D} = diag(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0)$$

the matrix Σ is $\Sigma = \tilde{D}$ if $m = n$, the matrix Σ has the form

$$\Sigma = \begin{pmatrix} \tilde{D} \\ 0 \end{pmatrix}$$

if $m > n$ and the form

$$\Sigma = \begin{pmatrix} \tilde{D} & 0 \end{pmatrix}$$

if $m < n$. The values σ_j can be ordered as

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 \ .$$

Any factorization $A = U\Sigma V^*$ of A where the matrices U, V, Σ have the properties described above is called a singular value decomposition of A.

Theorem 8.1 *a) Any matrix $A \in \mathbb{C}^{m\times n}$ has an SVD.*

b) The values σ_j are unique. These numbers are called the (non–zero) singular values of A.

c) If A is real, then the matrices U and V can be chosen real as well.

Proof: The main difficulty of the proof is to show **existence** of an SVD. We first make a pretransformation from the equation (8.1) to the equation (8.2) below where B is a nonsingular square matrix of size $r\times r$. Recall that r denotes the rank of A. We will then prove existence of an SVD of B. Combined with the pretransformation, one obtains an SVD of A.

a) Pretransformation: Let $r = rank\, A$. Let $u^1, \ldots, u^r$ denote an ONB of $R(A)$ and let $u^{r+1}, \ldots, u^m$ denote an ONB of $N(A^*)$. Then $u^1, \ldots u^m$ is an ONB of $\mathbb{C}^m$. Let $U_0 \in \mathbb{C}^{m\times m}$ denote the matrix with columns u^j.

Let $v^1, \ldots, v^r$ denote an ONB of $R(A^*)$ and let $v^{r+1}, \ldots, v^n$ denote an ONB of $N(A)$. Then $v^1, \ldots, v^n$ is an ONB of $\mathbb{C}^n$. Let $V_0 \in \mathbb{C}^{n\times n}$ denote the matrix with columns v^j.

First consider Av^k for $1 \leq k \leq r$. We can write

$$Av^k = \sum_{j=1}^r b_{jk} u^j \ .$$

For $r+1 \le k \le n$ we have $Av^k = 0$. If $B \in \mathbb{C}^{r\times r}$ denotes the matrix with entries b_{jk} then one obtains that

$$AV_0 = U_0 \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \ . \tag{8.2}$$

Since, by assumption, the matrix A has rank r, the matrix $B \in \mathbb{C}^{r\times r}$ is nonsingular.

b) Existence of an SVD of B: Consider the quadratic form

$$F(\xi) = \xi^* B^* B \xi, \quad \xi \in \mathbb{C}^r \ ,$$

and maximize $F(\xi)$ over the sphere $|\xi| = 1$. Assume that the maximum of $F(\xi)$ is attained at $\xi = x$, where $x \in \mathbb{C}^r, |x| = 1$. We then know that x is an eigenvector of B^*B,

$$B^*Bx = \lambda_1 x, \quad 0 < \lambda_1 = \sigma_1^2 \quad \text{with} \quad \sigma_1 > 0 \ .$$

(See Section 6.2. The matrix B^*B is positive definite Hermitian and the result of Section 6.2 generalizes to the complex case.)

Here

$$\sigma_1^2 = \lambda_1 = x^* B^* B x = |Bx|^2 \ ,$$

thus

$$\sigma_1 = |B| \ .$$

Set

$$y = \frac{1}{\sigma_1} Bx \ .$$

Choose matrices $X, Y \in \mathbb{C}^{r\times(r-1)}$ so that the $r \times r$ matrices

$$R_x = (x|X), \quad R_y = (y|Y) \in \mathbb{C}^{r\times r}$$

are unitary. Note that

$$x^* X = 0 \quad \text{and} \quad y^* Y = 0 \ .$$

We will try to understand the structure of the $r \times r$ block matrix

$$R_y^* B R_x = \begin{pmatrix} y^* \\ Y^* \end{pmatrix} B \Big(x | X \Big) = \begin{pmatrix} y^*Bx & y^*BX \\ Y^*Bx & Y^*BX \end{pmatrix}$$

Note that y^*Bx is a scalar and Y^*BX has dimension $(r-1)\times(r-1)$. Also, y^*BX is a row vector with $r-1$ components and Y^*Bx is a column vector with $r-1$ components.

Since $y^* = \frac{1}{\sigma_1} x^* B^*$ we have

$$y^* B x = \frac{1}{\sigma_1} x^* B^* B x = \sigma_1 \; .$$

Also,

$$B^* B x = \lambda_1 x \; ,$$

thus

$$x^* B^* B = \lambda_1 x^* \; .$$

Therefore,

$$y^* B X = \frac{1}{\sigma_1} x^* B^* B X = \frac{\lambda_1}{\sigma_1} x^* X = 0 \; .$$

Furthermore,

$$Y^* B x = \sigma_1 Y^* y = 0 \; .$$

We obtain that

$$\begin{aligned} R_y^* B R_x &= \begin{pmatrix} y^* \\ Y^* \end{pmatrix} B \Big(x | X \Big) \\ &= \begin{pmatrix} y^* B x & y^* B X \\ Y^* B x & Y^* B X \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1 & 0 \\ 0 & B_2 \end{pmatrix} \; . \end{aligned}$$

with

$$B_2 = Y^* B X \in \mathbb{C}^{(r-1) \times (r-1)} \; .$$

In the equation

$$R_y^* B R_x = \begin{pmatrix} \sigma_1 & 0 \\ 0 & B_2 \end{pmatrix}$$

the matrices R_y and R_x are unitary and $\sigma_1 = |B|$. It follows that

$$\sigma_2 := |B_2| \le |B| = \sigma_1 \; .$$

Since B is non–singular, the matrix B_2 is also non–singular; thus $\sigma_2 > 0$. Applying the same process which we have applied to the $r \times r$ matrix B to the $(r-1) \times (r-1)$ matrix B_2 we obtain

$$R_{y(2)}^* B_2 R_{x(2)} = \begin{pmatrix} \sigma_2 & 0 \\ 0 & B_3 \end{pmatrix} \; .$$

We continue the process and obtain unitary matrices $P, Q \in \mathbb{C}^{r \times r}$ so that

$$P^*BQ = diag\,(\sigma_1, \ldots, \sigma_r) =: D\ .$$

This yields that

$$B = PDQ^*$$

is an SVD of B.

c) Application to A: Using equation (8.2) and $B = PDQ^*$ we obtain

$$A = U_0 \begin{pmatrix} PDQ^* & 0 \\ 0 & 0 \end{pmatrix} V_0^* = U_0 \begin{pmatrix} P & 0 \\ 0 & I_{m-r} \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q^* & 0 \\ 0 & I_{n-r} \end{pmatrix} V_0^*\ .$$

One obtains the singular value decomposition of A:

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

where

$$U = U_0 \begin{pmatrix} P & 0 \\ 0 & I_{m-r} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad V^* = \begin{pmatrix} Q^* & 0 \\ 0 & I_{n-r} \end{pmatrix} V_0^*\ .$$

Uniqueness of the Singular Values: Assume that

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

is an SVD of A. Then we have

$$A^*A = V diag(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0) V^*\ .$$

This shows that the numbers

$$\sigma_1^2, \ldots, \sigma_r^2$$

are the non–zero eigenvalues of A^*A. We will prove in the next chapter on determinants that the eigenvalues of any square matrix M are uniquely determined as the zeros of the characteristic polynomial $det\,(M - zI)$. This completes the proof of Theorem 8.1. $\diamond$

8.2 The SVD and the Four Fundamental Subspaces

Let $A \in \mathbb{C}^{m\times n}$ have $rank\,A = r$ and let $A = U\Sigma V^*$ denote an SVD of A with

$$U = (u^1, \ldots, u^m), \quad V = (v^1, \ldots, v^n)\ .$$

We will show that the columns of the matrices U and V give us bases of the four fundamental subspaces of A.

1) If $x \in \mathbb{C}^n$ then

$$Ax = \sum_{j=1}^{r} \sigma_j (v^{j*} x) u^j \ .$$

This shows that

$$R(A) \subset span\{u^1, \ldots, u^r\} \ .$$

Since $R(A)$ has dimension r, equality holds. Thus, the r vectors

$$u^1, \ldots, u^r$$

form an ONB of $R(A)$.

2) Recall that $N(A^*)$ is the orthogonal complement of $R(A)$. Therefore,

$$u^{r+1}, \ldots, u^m$$

is a basis of $N(A^*)$.

3) Note that

$$A^* = V \Sigma^T U^*$$

is an SVD of A^*. Therefore,

$$v^1, \ldots, v^r$$

is an ONB of $R(A^*)$ and

$$v^{r+1}, \ldots, v^n$$

is an ONB of $N(A)$.

In this way, the columns of the matrix U provide bases for $R(A)$ and $N(A^*)$. The columns of the matrix V provide bases for $R(A^*)$ and $N(A)$.

8.3 SVD and Least Squares

Consider the linear system

$$Ax = b \ .$$

As above, we assume that $A \in \mathbb{C}^{m \times n}$ and $b \in \mathbb{C}^m$ are given and that $A = U \Sigma V^*$ is an SVD of A.

The full rank case. First consider the case where $rank\, A = n \leq m$. In this case, there is a unique least squares solution, x_{ls}. The least squares solution is the unique solution of the normal equations

$$A^* A x = A^* b \ .$$

We have

$$A^*A = VD^2V^* \quad \text{where} \quad \Sigma = \begin{pmatrix} D \\ 0 \end{pmatrix}, \quad D = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}$$

and

$$A^*b = V\Sigma^T U^* b \ .$$

Set

$$U^*b = c = \begin{pmatrix} c^I \\ c^{II} \end{pmatrix} \quad \text{with} \quad c^I \in \mathbb{C}^n, \quad c^{II} \in \mathbb{C}^{m-n} \ .$$

Then the normal equations become

$$VD^2V^*x = VDc^I$$

or

$$DV^*x = c^I \ .$$

One obtains the least squares solution:

$$\begin{aligned} x_{ls} &= VD^{-1}c^I \\ &= \sum_{j=1}^n \frac{c_j}{\sigma_j}\, v^j \\ &= \sum_{j=1}^n \frac{u^{j*}b}{\sigma_j}\, v^j \\ &= \Big(\sum_{j=1}^n \frac{1}{\sigma_j} v^j u^{j*} \Big) b \end{aligned}$$

This formula for the least squares x_{ls} shows the following: Unless the right–hand side b of the system $Ax = b$ is special, the smallest singular values σ_j lead to the largest contribution in x_{ls}. This may be dangerous since the smallest σ_j may be contaminated by data errors.

It may be more reasonable to replace any small σ_j by zero and ignore the term

$$\frac{u^{j*}b}{\sigma_j}\, v^j \quad \text{if} \quad \sigma_j < tol$$

in the solution x_{ls}. The choice of tol depends on the application. If

$$\sigma_k \geq tol > \sigma_{k+1}$$

one may want to replace x_{ls} by

$$x_{ls}^{(k)} = \Big(\sum_{j=1}^{k} \frac{1}{\sigma_j} v^j u^{j*} \Big) b$$

The case of arbitrary rank. Let $A \in \mathbb{C}^{m\times n}$ have $rank\, A = r$. For a given $b \in \mathbb{C}^m$ the normal equations are

$$A^* A x = A^* b, \quad x \in \mathbb{C}^n \ .$$

Let us determine all solutions x of the normal equations.

We have

$$A^* = V \Sigma^T U^*$$

and

$$\begin{aligned} A^* b &= V \Sigma^T U^* b \\ &= V \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} U^* b \\ &= \sum_{j=1}^{r} \sigma_j \langle u^j, b \rangle v^j \ . \end{aligned}$$

If $x \in \mathbb{C}^n$ is arbitrary, then we can write

$$x = V y \quad \text{where} \quad y = V^* x \in \mathbb{C}^n \ .$$

Then we have

$$\begin{aligned} A^* A x &= V \Sigma^T \Sigma V^* x \\ &= V \Sigma^T \Sigma y \\ &= \sum_{j=1}^{r} y_j \sigma_j^2 v^j \end{aligned}$$

Comparing this expression with the expression for $A^* b$ we can establish that $x = V y$ solves the normal equations if and only if

$$y_j = \frac{1}{\sigma_j} \langle u^j, b \rangle \quad \text{for} \quad j = 1, \ldots, r \ .$$

This result tells us that $x \in \mathbb{C}^n$ solves the normal equations if and only if

$$x = \sum_{j=1}^{r} \frac{1}{\sigma_j} \langle u^j, b \rangle v^j + \sum_{j=r+1}^{n} y_j v^j$$

where

$$y_j \in \mathbb{C}$$

is arbitrary for $r+1 \le j \le n$. Clearly, the sum

$$\sum_{j=r+1}^{n} y_j v^j$$

is an arbitrary element of $N(A) = N(A^*A)$. The vector

$$x_{best} = \sum_{j=1}^{r} \frac{1}{\sigma_j} \langle u^j, b \rangle v^j$$

is the solution of the normal equations with the smallest Euclidean norm. The formula for x_{best} can also be written as

$$x_{best} = \Big(\sum_{j=1}^{r} \frac{1}{\sigma_j} v^j u^{j*} \Big) b \ .$$

This motivates to define the $n \times m$ matrix

$$A^\dagger = \sum_{j=1}^{r} \frac{1}{\sigma_j} v^j u^{j*}$$

which is called the **Moore–Penrose generalized inverse** of A. (The symbol $\dagger$ is called the dagger sign.)

If

$$A = U \Sigma V^*, \quad \Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad D = diag\,(\sigma_1, \ldots, \sigma_r) \ ,$$

then

$$A^\dagger = V \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^* \ .$$

Discussion of the Moore–Penrose Generalized Inverse:

Good properties: Every matrix $A \in \mathbb{C}^{m \times n}$ has a unique[5] Moore–Penrose generalized inverse $A^\dagger$. If A is a nonsingular square matrix, then $A^\dagger = A^{-1}$.

A bad property: $A^\dagger$ does not depend continuously on A. For example, let

$$A_\varepsilon = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} \ .$$

If $\varepsilon \neq 0$ then

$$A_\varepsilon^\dagger = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\varepsilon} \end{pmatrix} \ .$$

However, the matrix

[5] The uniqueness of $A^\dagger$ follows from the fact that for each $b \in \mathbb{C}^n$ the vector $x_{best} = A^\dagger b$ is the unique solution of $A^*Ax = A^*b$ which has the smallest Euclidean norm.

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

has the generalized inverse

$$A_0^\dagger = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} .$$

Thus,

$$|A_\varepsilon - A_0| \to 0 \quad \text{as} \quad \varepsilon \to 0 ,$$

but

$$|A_\varepsilon^\dagger - A_0^\dagger| \to \infty \quad \text{as} \quad \varepsilon \to 0 .$$

8.4 SVD and Rank

The rank of a matrix $A \in \mathbb{C}^{m\times n}$ does not depend continuously on A. The following is easy to show:

Lemma 8.1 *Let $A \in \mathbb{C}^{m\times n}$ be rank deficient, i.e., the rank of A is r where $r < min\{m,n\}$. If $\varepsilon > 0$ is arbitrary, then there exists $S \in \mathbb{C}^{m\times n}$ with $|S_\varepsilon| = \varepsilon$ so that the perturbed matrix $A+S$ has full rank, i.e.,*

$$rank\,(A+S) = \min\{m,n\} .$$

Proof: Let $A = U\Sigma V^*$ denote an SVD of A, where

$$\Sigma = D \quad \text{or} \quad \Sigma = \begin{pmatrix} D \\ 0 \end{pmatrix} \quad \text{or} \quad \Sigma = \begin{pmatrix} D & 0 \end{pmatrix} \tag{8.3}$$

with

$$D = diag(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) .$$

Replace D by

$$D_\varepsilon = diag(0, \ldots, 0, \varepsilon, \ldots, \varepsilon)$$

in formula (8.3) for Σ. Denote the result by Σ_ε:

$$\Sigma_\varepsilon = D_\varepsilon \quad \text{or} \quad \Sigma_\varepsilon = \begin{pmatrix} D_\varepsilon \\ 0 \end{pmatrix} \quad \text{or} \quad \Sigma_\varepsilon = \begin{pmatrix} D_\varepsilon & 0 \end{pmatrix}$$

and set

$$S_\varepsilon = U\Sigma_\varepsilon V^* .$$

Then we have $|S_\varepsilon| = \varepsilon$ and

$$A + S_\varepsilon = U(D + D_\varepsilon)V^*$$

has full rank for $\varepsilon > 0$. $\diamond$

Thus, through arbitrarily small perturbations, the rank can increase. However, as we will show, the rank cannot decrease through arbitrarily small perturbations. The singular values of A give the precise information formulated in the next theorem. Recall that $|S|$ denotes the matrix norm of a matrix $S \in \mathbb{C}^{m\times n}$ corresponding the Euclidean vector norms in $\mathbb{C}^n$ and $\mathbb{C}^m$.

Theorem 8.2 *Let $A \in \mathbb{C}^{m\times n}$ have $rank\, A = r \geq 1$ and let*

$$\sigma_1 \geq \ldots \geq \sigma_r > 0$$

denote the nonzero singular values of A. Let $0 \leq l < r$.

a) If $S \in \mathbb{C}^{m\times n}$ satisfies $|S| < \sigma_{l+1}$ then

$$rank\,(A+S) > l \ .$$

b) There exists $S \in \mathbb{C}^{m+n}$ with $|S| = \sigma_{l+1}$ so that

$$rank\,(A+S) = l \ .$$

The proof of this result will be given below.

The result of the above theorem can also be formulated using the distance of A from the set of matrices of rank l,

$$\mathcal{R}_l = \{B \in \mathbb{C}^{m\times n} \ : \ rank\, B = l\} \ .$$

We define

$$dist(A, \mathcal{R}_l) = inf\{|A - B| \ : \ B \in \mathcal{R}_l\} \ .$$

Theorem 8.3 *Let $A \in \mathbb{C}^{m\times n}$ have $rank\, A = r \geq 1$ and let $0 \leq l < r$. Then*

$$dist(A, \mathcal{R}_l) = \sigma_{l+1}$$

and there exists a matrix $B \in \mathcal{R}_l$ with

$$|A - B| = \sigma_{l+1} \ .$$

We will need the following simple result for the rank of a matrix product.

Lemma 8.2 *Let $A \in \mathbb{C}^{m\times n}, B \in \mathbb{C}^{n\times k}$, thus $AB \in \mathbb{C}^{m\times k}$. We have*

$$rank\,(AB) \leq rank\, B \ .$$

Proof: Let $rank\,(B) = r$ and let $v^1, \ldots, v^r$ denote a basis of $R(B)$. If $x \in R(AB)$ is arbitrary, then there exists $c \in \mathbb{C}^k$ with

$$x = ABc \ .$$

We then have $Bc \in R(B)$ and can write

$$Bc = \sum_{j=1}^{r} \alpha_j v^j \ .$$

Therefore,

$$x = ABc = \sum_{j=1}^{r} \alpha_j A v^j \ .$$

This shows that the r vectors $Av^1, \ldots, Av^r$ span $R(AB)$. The estimate follows.
$\diamond$

By considering

$$(AB)^* = B^* A^*$$

we also have

$$rank\,(AB) \le rank\,A \ .$$

For any finite matrix product:

$$rank(A_1 \ldots A_q) \le \min_j rank A_j \ .$$

The proof of Theorem 8.3 has two parts.

Part 1: We show that there exists a matrix $B \in \mathcal{R}_l$ with $|A - B| = \sigma_{l+1}$. Let

$$A = U \begin{pmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \ddots & & \\ & & & \sigma_r & \\ 0 & & & & 0 \end{pmatrix} V^*$$

and set

$$B = U \begin{pmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \sigma_l & & \\ & & & 0 & \\ 0 & & & & 0 \end{pmatrix} V^* \ .$$

Clearly, the matrix B has rank l. Since the application of U and of V^* does not change the length of any vector, it follows that

$$|A - B| = \sigma_{l+1} \ .$$

Part 2: Let $B \in \mathcal{R}_l$ be arbitrary. We will show that

$$|A - B| \ge \sigma_{l+1} \ .$$

We will construct a vector $x \in \mathbb{C}^n$ with $|x| = 1$ and $|(A - B)x| \geq \sigma_{l+1}$.

We start with a simple observation: Let $\alpha \in \mathbb{C}^m$ and set $y = U\alpha$. We have $|y| = |\alpha|$, i.e.,

$$|y|^2 = \sum_{j=1}^{m} |\alpha_j|^2 \quad \text{for} \quad y = \sum_{j=1}^{m} \alpha_j u^j \ . \tag{8.4}$$

Next, set

$$U_1 = (u^1, \ldots, u^{l+1}) \in \mathbb{C}^{m \times (l+1)}, \quad V_1 = (v^1, \ldots, v^{l+1}) \in \mathbb{C}^{n \times (l+1)} \ .$$

The matrix

$$U_1^* B V_1 \in \mathbb{C}^{(l+1) \times (l+1)}$$

is singular since B has rank l. There exists $c \in \mathbb{C}^{l+1}$ with $|c| = 1$ and

$$U_1^* B V_1 c = 0 \ .$$

Set $x = V_1 c \in \mathbb{C}^n$ and note that $|x| = 1$. We have

$$\begin{aligned} Ax &= U \Sigma V^* V_1 c \\ &= U \Sigma V^* (v^1, \ldots, v^{l+1}) c \\ &= U \Sigma \begin{pmatrix} I_{l+1} \\ 0 \end{pmatrix} c \\ &= \sum_{j=1}^{l+1} \sigma_j c_j u^j \end{aligned}$$

We write $Bx \in \mathbb{C}^m$ as $Bx = U\beta$ with $\beta = U^* Bx$ and obtain

$$\begin{aligned} \beta &= U^* B x \\ &= \begin{pmatrix} U_1^* \\ u^{(l+2)*} \\ \vdots \\ u^{m*} \end{pmatrix} B V_1 c \\ &= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \beta_{l+2} \\ \vdots \\ \beta_m \end{pmatrix} \end{aligned}$$

The last equation holds since

$$U_1^* B V_1 c = 0 \in \mathbb{C}^{l+1} \ .$$

It follows that

$$Bx = \sum_{j=l+2}^{m} \beta_j u^j \ .$$

We obtain that

$$Ax - Bx = \sum_{j=1}^{l+1} \sigma_j c_j u^j - \sum_{j=l+2}^{m} \beta_j u^j \ .$$

Using (8.4) we obtain that

$$\begin{aligned}
|(A-B)x|^2 &= \sum_{j=1}^{l+1} \sigma_j^2 |c_j|^2 + \sum_{j=l+2}^{m} |\beta_j|^2 \\
&\geq \sum_{j=1}^{l+1} \sigma_j^2 |c_j|^2 \\
&\geq \sigma_{l+1}^2 \sum_{j=1}^{l+1} |c_j|^2 \\
&= \sigma_{l+1}^2
\end{aligned}$$

This proves Theorem 8.3. $\diamond$

8.5 SVD and Filtering of Noisy Data

Let $A \in \mathbb{R}^{n \times n}$ denote a matrix whose n^2 entries a_{ij} are affected by noise. Assume that n is large (for example, $n = 10^3$), and we want to transmit A. Suppose we have

$$A = U \Sigma V^T = \sum_{j=1}^{n} \sigma_j u^j v^{jT} \ .$$

Let $tol > 0$ denote a tolerance and assume that

$$\sigma_1 \geq \ldots \geq \sigma_q \geq tol > \sigma_{q+1} \geq \ldots \geq \sigma_n \ .$$

If tol is a measure for the noise level, we may approximate A by

$$A_q = \sum_{j=1}^{q} \sigma_j u^j v^{jT} \ .$$

For example, if $n = 10^3$, but $q = 10$ then the transmission of A_q can be accomplished by transmitting the 20 vectors

$$u^1, \ldots, u^{10}, v^1, \ldots, v^{10} \in \mathbb{R}^{1,000}$$

and the ten numbers

$$\sigma_1 \geq \ldots \geq \sigma_{10} \geq tol \ .$$

These are

$$20 * 10^3 + 10$$

numbers. In contrast, the transmission of all entries of A would require to transmit 10^6 numbers. The cost of transmitting A_{10} is about 2% of the cost of transmitting A.

9 Determinants

We begin with an important geometric property of the determinant of a real $n \times n$ matrix A.

Let $A \in \mathbb{R}^{n\times n}$ denote a real $n \times n$ matrix with column vectors $a^1, \ldots, a^n \in \mathbb{R}^n$. The column vectors span the parallelepiped

$$P(a^1, \ldots, a^n) = \{x \in \mathbb{R}^n \ : \ x = \sum_{j=1}^n \alpha_j a^j, \quad 0 \le \alpha_j \le 1 \text{ for } j = 1, \ldots, n\} \ .$$

The determinant of A is the signed volume of this parallelepiped,

$$det(A) = vol(P(a^1, \ldots, a^n) \quad \text{or} \quad det(A) = -vol(P(a^1, \ldots, a^n) \ .$$

Note that $P(a^1, \ldots, a^n)$ is the image under A of the unit cube in $\mathbb{R}^n$. For $n = 2$ the unit cube is the unit square with corners

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

If two columns of the matrix $A \in \mathbb{R}^{n\times n}$ get exchanged, then the volume of the parallelepiped remains unchanged, but the determinant of the matrix gets multiplied by -1.

To study such sign–changes, we treat permutations and their signs in the next section.

9.1 Permutations and Their Signs

9.1.1 The Group S_n

Let n denote a positive integer. A bijective map σ from the set

$$\{1, 2, \ldots, n\}$$

onto itself is called a permutation of n elements. We represent a permutation σ of n elements by a $(2, n)$ matrix:

$$\sigma \simeq \begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma_1 & \sigma_2 & \ldots & \sigma_n \end{pmatrix}$$

where $\sigma_j = \sigma(j)$ for $j = 1, 2, \ldots, n$. Let S_n denote the set of all permutations of n elements. For $\sigma, \tau \in S_n$ one defines the product $\sigma\tau = \sigma \circ \tau$ by

$$(\sigma\tau)(j) = \sigma(\tau(j)) \quad \text{for} \quad j = 1, 2, \ldots, n \ .$$

Then S_n becomes the permutation group of n elements. Using induction, it is easy to show that the group S_n has $n!$ elements. For $n \ge 3$ the group S_n is non–commutative.

For example, if

$$\sigma \simeq \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad \tau \simeq \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$$

then

$$(\sigma\tau)(1) = 2 \quad \text{and} \quad (\tau\sigma)(1) = 3 \; .$$

The unit element of S_n is

$$id \simeq \begin{pmatrix} 1 & 2 & \ldots & n \\ 1 & 2 & \ldots & n \end{pmatrix} \; .$$

9.1.2 The Sign of a Permutation

For $\sigma \in S_n$ let $N = N(\sigma)$ denote the number of all pairs of integers (i, j) with

$$1 \le i < j \le n \quad \text{and} \quad \sigma_i > \sigma_j \; .$$

If we represent a permutation as

$$\sigma \simeq \begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma_1 & \sigma_2 & \ldots & \sigma_n \end{pmatrix}$$

then N is the number of all pairs (σ_i, σ_j) in the second row which are in wrong order, i.e., $i < j$ but $\sigma_i > \sigma_j$. For example, if

$$\sigma \simeq \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$$

then $N = 5$ since precisely the five pairs

$$(3, 2), \quad (3, 1), \quad (4, 2), \quad (4, 1), \quad (2, 1)$$

are in wrong order.

If $\sigma = id$ is the identity, then $N(\sigma) = N(id) = 0$.

Definition: Let $\sigma \in S_n$ denote a permutation of n elements an let $N = N(\sigma)$ denote the number of all pairs (i, j) with

$$1 \le i < j \le n \quad \text{but} \quad \sigma_i > \sigma_j \; .$$

Then the sign of σ is defined as

$$sgn(\sigma) = (-1)^{N(\sigma)} \; .$$

We will prove:

Theorem 9.1 *For $\sigma, \tau \in S_n$:*

$$sgn(\sigma\tau) = sgn(\sigma)\, sgn(\tau) \; .$$

Let us illustrate the sign of σ by a simple application. We will also use this application to prove the theorem.

Consider the polynomial $p(x)$ in n real variables defined by

$$\begin{array}{rcl} p(x) & = & p(x_1, x_2, \ldots, x_n) \\ & = & \displaystyle\prod_{1 \le i < j \le n} (x_i - x_j) \\ & = & (x_1 - x_2)(x_1 - x_3) \ldots (x_1 - x_n) \\ & & \qquad \cdot (x_2 - x_3) \ldots (x_2 - x_n) \\ & & \qquad\qquad \ldots \\ & & \qquad\qquad\qquad \cdot (x_{n-1} - x_n) \end{array}$$

Then we have, with $N = N(\sigma)$:

$$\begin{array}{rcl} p(x_{\sigma_1}, \ldots, x_{\sigma_n}) & = & \displaystyle\prod_{i<j} (x_{\sigma_i} - x_{\sigma_j}) \\ & = & (-1)^N \displaystyle\prod_{i<j} (x_i - x_j) \\ & = & sgn(\sigma)\, p(x_1, \ldots, x_n) \end{array}$$

The reason is simple: If the polynomial $p(x_{\sigma_1}, \ldots, x_{\sigma_n})$ has a factor $x_{\sigma_i} - x_{\sigma_j}$ with

$$i < j \quad \text{but} \quad \sigma_i > \sigma_j$$

then a sign change must be applied to the factor $x_{\sigma_i} - x_{\sigma_j}$ to obtain the corresponding factor in $p(x_1, \ldots, x_n)$.

Let us state the result:

Lemma 9.1 *Consider the polynomial*

$$p(x_1, x_2, \ldots, x_n) = \prod_{1 \le i < j \le n} (x_i - x_j)$$

and let $\sigma \in S_n$. Then we have

$$p(x_{\sigma_1}, x_{\sigma_2}, \ldots, x_{\sigma_n}) = sgn(\sigma)\, p(x_1, x_2, \ldots, x_n) \ . \tag{9.1}$$

We now prove Theorem 9.1. Let $\sigma, \tau \in S_n$ and let

$$x = (x_1, x_2, \ldots, x_n)$$

denote an argument of the polynomial $p(x)$. Set

$$y_j = x_{\sigma_j}, \quad j = 1, 2, \ldots, n \ ,$$

thus

$$\begin{aligned} y &= (y_1, y_2, \ldots, y_n) \\ &= (x_{\sigma_1}, x_{\sigma_2}, \ldots, x_{\sigma_n}) \end{aligned}$$

To prove the theorem, we compute

$$p(y_{\tau_1}, y_{\tau_2}, \ldots, y_{\tau_n})$$

in two different ways. First, since $y_j = x_{\sigma_j}$ we have

$$(y_{\tau_1}, y_{\tau_2}, \ldots, y_{\tau_n}) = (x_{(\sigma\tau)_1}, x_{(\sigma\tau)_2}, \ldots, x_{(\sigma\tau)_n})$$

and the previous Lemma yields that

$$\begin{aligned} p(x_{\sigma_1}, x_{\sigma_2}, \ldots, x_{\sigma_n}) &= sgn(\sigma)\, p(x_1, x_2, \ldots, x_n) \\ p(y_{\tau_1}, y_{\tau_2}, \ldots, y_{\tau_n}) &= sgn(\tau)\, p(y_1, y_2, \ldots, y_n) \\ &= sgn(\tau)\, p(x_{\sigma_1}, x_{\sigma_2}, \ldots, x_{\sigma_n}) \\ &= sgn(\tau)\, sgn(\sigma)\, p(x_1, x_2, \ldots, x_n) \end{aligned}$$

Second,

$$\begin{aligned} p(y_{\tau_1}, y_{\tau_2}, \ldots, y_{\tau_n}) &= p(x_{(\sigma\tau)_1}, x_{(\sigma\tau)_2}, \ldots, x_{(\sigma\tau)_n}) \\ &= sgn(\sigma\tau)\, p(x_1, x_2, \ldots, x_n) \end{aligned}$$

This shows that

$$sgn(\tau)\, sgn(\sigma) = sgn(\sigma\tau)\ .$$

Theorem 9.1 is proved. $\diamond$

9.1.3 Transpositions

A transposition is a permutation that exchanges precisely two elements of $\{1, 2, \ldots, n\}$ and leaves all other elements fixed. If i and j are two different elements in $\{1, 2, \ldots, n\}$ we write T_{ij} for the transposition that exchanges i and j. It is easy to see that every transposition has the sign -1:

$$sgn(T_{ij}) = -1\ . \tag{9.2}$$

To see this assume that $i < j$. Then

$$T_{ij} \simeq \begin{pmatrix} \ldots & i & \ldots & j & \ldots \\ \ldots & j & \ldots & i & \ldots \end{pmatrix}$$

where $\ldots$ stands for numbers $1 \le k \le n$ which remain fixed. The pairs in wrong order in the second row are:

$$(j, k) \quad \text{for} \quad k = i+1, \ldots, j-1 \quad \text{and} \quad k = i$$

and

$$(k,i) \quad \text{for} \quad k = i+1, \ldots, j-1 \ .$$

It follows that the number of pairs in wrong order is odd and (9.2) follows.

The next lemma will be shown by induction in n.

Lemma 9.2 *Let $n \geq 2$. Every $\sigma \in S_n$ can be written as a product of transpositions.*

Proof: For $n = 2$ the claim is clear. Let $n \geq 3$ and assume the claim holds for $n-1$. Let $\sigma \in S_n$. We may assume that $\sigma_n = k \neq n$ since otherwise we can consider σ as an element of S_{n-1}. Define

$$\tau = T_{kn}\,\sigma \ .$$

We then have

$$\tau(n) = T_{kn}(\sigma(n)) = T_{kn}(k) = n \ .$$

Thus, we can consider τ as an element of S_{n-1} and write τ as a product of transpositions. Then

$$\sigma = T_{kn}\,\tau$$

is also a product of transpositions. $\diamond$

Theorem 9.1 and the previous lemma have the following implication:

Lemma 9.3 *Let $\sigma \in S_n$ be any permutation. We have $sgn(\sigma) = 1$ if and only if one can write σ as an even number of transpositions. We have $sgn(\sigma) = -1$ if and only if one can write σ as an odd number of transpositions.*

Definition: The permutation σ is called even if $sgn(\sigma) = 1$. It is called odd if $sgn(\sigma) = -1$.

9.2 Volumes and Orientation: Intuitive Meaning of the Determinant

Let

$$A = (a^1, \ldots, a^n), \quad a^j \in \mathbb{R}^n \ ,$$

denote a real $n \times n$ matrix. The n columns a^j of A span the parallelepiped

$$P(a^1, \ldots, a^n) = \{x \in \mathbb{R}^n \ : \ x = \sum_{j=1}^n \alpha_j a^j, \quad 0 \leq \alpha_j \leq 1 \text{ for } j = 1, \ldots, n\} \ .$$

Geometrically, the determinant of A is the signed volume of the parallelepiped $P(a^1, \ldots, a^n)$:

$$det(A) = vol(P(a^1, \ldots, a^n)) \quad \text{or} \quad det(A) = -vol(P(a^1, \ldots, a^n)) \ .$$

Here the sign depends on the orientation of the n–tuple $(a^1, \ldots, a^n)$, which we discuss next.

Remarks on substitution in integrals: The fact that the determinant of a real matrix is related to volume is important for many results of analysis. For example, let Ω_1 and Ω_2 denote two open subsets of $\mathbb{R}^n$ and let $\phi : \Omega_1 \to \Omega_2$ denote a C^1–function which is $1-1$ and onto. Let $f : \Omega_2 \to \mathbb{R}$ be integrable. Then the following substitution rule holds:

$$\int_{\Omega_2} f(y)\, dy = \int_{\Omega_1} f(\phi(x)) |det\, \phi'(x)|\, dx \ .$$

To obtain this rule, it is important to related the determinant of the Jacobian $\phi'(x)$ to volume.

9.2.1 Orientation

Let $e^1, \ldots, e^n$ denote the standard basis of $\mathbb{R}^n$ and let $a^1, \ldots, a^n \in \mathbb{R}^n$ be arbitrary.

If the vectors $a^1, \ldots, a^n$ are linearly dependent, then the parallelepiped $P(a^1, \ldots, a^n)$ lies in a hyperplane of $\mathbb{R}^n$ and $P(a^1, \ldots, a^n)$ is called degenerate. Otherwise, if $a^1, \ldots, a^n$ are linearly independent, then $P(a^1, \ldots, a^n)$ is called non–degenerate. A non–degenerate parallelepiped does not fit into any hyperplane in $\mathbb{R}^n$.

Let $P(a^1, \ldots, a^n)$ be non–degenerate. If one can deform $P(a^1, \ldots, a^n)$ continuously into $P(e^1, \ldots, e^n)$ without passing through a degenerate state, then one says that the ordered n–tuple $(a^1, \ldots, a^n)$ is positively oriented. Otherwise, the n–tuple is called negatively oriented. We now express this more formally.

Definition: Let $(a^1, \ldots, a^n)$ denote an ordered n–tuple of linearly independent vectors $a^1, \ldots, a^n \in \mathbb{R}^n$. If there exist continuous functions

$$\alpha_j : [0,1] \to \mathbb{R}^n \quad \text{for} \quad j = 1, \ldots, n$$

with

$$\alpha_j(0) = a^j \quad \text{and} \quad \alpha_j(1) = e^j \quad \text{for} \quad j = 1, \ldots, n$$

so that the n vectors

$$\alpha_1(s), \ldots, \alpha_n(s) \in \mathbb{R}^n$$

are linearly independent for all $0 \le s \le 1$, then $(a^1, \ldots, a^n)$ is called positively oriented and we set

$$\mathcal{O}(a^1, \ldots, a^n) = 1 \ .$$

If such functions $\alpha_j(s)$ do not exist (but $a^1, \ldots, a^n$ are linearly independent), then the n–tuple $(a^1, \ldots, a^n)$ is called negatively oriented and we set

$$\mathcal{O}(a^1, \ldots, a^n) = -1 \; .$$

If $a^1, \ldots, a^n$ are linearly dependent then we set

$$\mathcal{O}(a^1, \ldots, a^n) = 0 \; .$$

An intuitive meaning of the determinant of a matrix

$$A = (a^1, \ldots, a^n) \in \mathbb{R}^{n \times n}$$

is

$$det(A) = \mathcal{O}(a^1, \ldots, a^n) \, vol(P(a^1, \ldots, a^n)) \; .$$

9.2.2 The Case $n = 2$

Consider the case $n = 2$. Parallelepipeds are parallelograms and $vol\,(P(a^1, a^2))$ is the area of the parallelogram spanned by a^1 and a^2.

For $\alpha > 0$ one obtains that

$$vol\,(P(\alpha a^1, a^2)) = \alpha \, vol\,(P(a^1, a^2)) \; .$$

For $\alpha < 0$,

$$vol\,(P(\alpha a^1, a^2) = |\alpha| \, vol\,(P(a^1, a^2)) \; .$$

If $\alpha < 0$ then multiplication of a^1 by α changes the orientation and one obtains that

$$det(\alpha a^1, a^2) = \alpha \, det(a^1, a^2) \; .$$

The rule

$$det(a^1 + \alpha a^2, a^2) = det(a^1, a^2)$$

is also geometrically plausible if we interpret $det(a, b)$ as the signed area of the parallelogram spanned by $a, b \in \mathbb{R}^2$.

Now consider the two parallelepipeds spanned by a^1, a^2 and b^1, a^2. We assume that $a^2 \neq 0$.

Let us first assume that the second component of a^2 is different from zero. Then there are scalars α, β, c_1, c_2 with

$$a^1 + \alpha a^2 = c_1 e^1$$

and

$$b^1 + \beta a^2 = c_2 e^1 \; .$$

We then obtain

$$\begin{aligned} det(a^1,a^2)+det(b^1,a^2) &= det(a^1+\alpha a^2,a^2)+det(b^1+\beta a^2,a^2) \\ &= det(c_1e^1,a^2)+det(c_2e^1,a^2) \\ &= det((c_1+c_2)e^1,a^2) \\ &= det(a^1+b^1+(\alpha+\beta)a^2,a^2) \\ &= det(a^1+b^1,a^2) \end{aligned}$$

Now consider the exceptional case where the second component of a^2 is zero. Then the first component of a^2 is different from zero and we have for suitable constant α,β,c_1,c_2:

$$a^1+\alpha a^2 = c_1e^2$$

and

$$b^1+\beta a^2 = c_2e^2\ .$$

The equation

$$det(a^1,a^2)+det(b^1,a^2) = det(a^1+b^1,a^2)$$

follows as above.

The rules

$$det(\alpha a^1,a^2) = \alpha\, det(a^1,a^2)$$

and

$$det(a^1,a^2)+det(b^1,a^2) = det(a^1+b^1,a^2)$$

say that the mapping

$$\begin{cases} \mathbb{R}^2 & \to & \mathbb{R} \\ x & \to & det(x,a^2) \end{cases}$$

is linear. Similarly, $x \to det(a^1,x)$ is linear. So far, we have assumed that

$$det(a,b)$$

is the signed area of the parallelogram spanned by $a,b\in\mathbb{R}^2$ and have used our intuition for *area* to derive these rules.

Generalizing from $n=2$ to general n, it is plausible that the determinant function

$$det(A) = \mathcal{O}(a^1,\ldots,a^n)\, vol(P(a^1,\ldots,a^n)) \quad \text{for} \quad A\in\mathbb{R}^{n\times n}$$

has the three properties that we introduce in the first theorem of the next section.

9.3 The Determinant as a Multilinear Function

In the following, let F denote a field. We will use the next theorem to define the determinant of a matrix $A \in F^{n\times n}$.

Theorem 9.2 *There exists a unique function*

$$d : F^n \times F^n \times \ldots \times F^n = (F^n)^n \to F$$

that has the following three properties:

(P1) For any fixed $j \in \{1, 2, \ldots, n\}$ and any fixed vectors $a^1, \ldots, a^{j-1}, a^{j+1}, \ldots, a^n \in F^n$ the map

$$b \to d(a^1, \ldots, a^{j-1}, b, a^{j+1}, \ldots, a^n)$$

from F^n to F is linear.

(P2) If $a^i = a^j$ for some $i \neq j$ then

$$d(a^1, \ldots, a^n) = 0\ .$$

(P3) The map d is normalized so that

$$d(e^1, e^2, \ldots, e^n) = 1\ .$$

Lemma 9.4 *If the map d has the properties $(P1)$ and $(P2)$ then d is alternating in the sense that the exchange of any two entries changes the sign:*

$$d(\ldots, a^i, \ldots, a^j, \ldots) = -d(\ldots, a^j, \ldots, a^i, \ldots)\ .$$

Proof: This follows from

$$\begin{aligned} d(a,b) &= d(a+b,b) \\ &= d(a+b, b-(a+b)) \\ &= d(a+b,-a) \\ &= d(b,-a) \\ &= -d(b,a) \end{aligned}$$

$\diamond$

Once we have proved the theorem, we define

$$det(A) = d(a^1, \ldots, a^n)$$

where $A = (a^1, \ldots, a^n)$. We will also prove the formula

$$det(A) = \sum_{\sigma\in S_n} \mathrm{sgn}\,(\sigma) \mathrm{a}_{\sigma_1 1} \ldots \mathrm{a}_{\sigma_n \mathrm{n}}\ .$$

Proof of Theorem 9.2: First assume that d is a function satisfying $(P1)$ and $(P2)$. We write the vectors $a^j \in F^n$ in terms of the vectors $e^1 = (1,0,\ldots,0)^T$ etc. We have

$$a^1 = \sum_{i_1=1}^n a_{i_1 1} e^{i_1}, \quad a^2 = \sum_{i_2=1}^n a_{i_2 2} e^{i_2}, \quad \text{etc.}$$

This yields

$$\begin{aligned} d(a^1, a^2, \ldots, a^n) &= d\Big(\sum_{i_1=1}^n a_{i_1 1} e^{i_1}, \ldots, \sum_{i_n=1}^n a_{i_n n} e^{i_n} \Big) \\ &= \sum_{i_1=1}^n \cdots \sum_{i_n=1}^n \; a_{i_1 1} \ldots \ldots a_{i_n n} \; d(e^{i_1}, \ldots, e^{i_n}) \end{aligned}$$

Using $(P2)$ it follows that $d(e^{i_1}, \ldots, e^{i_n}) = 0$ if two of the indices $i_1, \ldots, i_n$ are equal to each other. Therefore, in the above expression for $d(a^1, \ldots, a^n)$ we have to sum only over all permutations $\sigma \in S_n$ and obtain

$$d(a^1, a^2, \ldots, a^n) = \sum_{\sigma \in S_n} a_{\sigma_1 1} \ldots a_{\sigma_n n} \; d(e^{\sigma_1}, \ldots, e^{\sigma_n}) \; .$$

Using Lemma 9.4 one obtains:

$$d(e^{\sigma_1}, \ldots, e^{\sigma_n}) = sgn(\sigma) \, d(e^1, \ldots, e^n) \; .$$

Therefore,

$$d(a^1, a^2, \ldots, a^n) = \sum_{\sigma \in S_n} sgn(\sigma) \, a_{\sigma_1 1} \ldots a_{\sigma_n n} \, d(e^1, \ldots, e^n) \; . \tag{9.3}$$

In particular, we have shown uniqueness of any mapping d with properties $(P1), (P2), (P3)$.

If one defines

$$d(a^1, a^2, \ldots, a^n) = \sum_{\sigma \in S_n} sgn(\sigma) \, a_{\sigma_1 1} \ldots a_{\sigma_n n} \tag{9.4}$$

then the properties $(P1), (P2), (P3)$ are not difficult to prove.

Details:

(P1) Each term

$$a_{\sigma_1 1} a_{\sigma_2 2} \ldots a_{\sigma_n n}$$

depends linearly on each entry $a_{\sigma_j j}$. Therefore, (P1) holds.

(P2) Let $a^1 = a^2$, for example. Therefore,

$$a_{\sigma_1 1} = a_{\sigma_1 2} \quad \text{and} \quad a_{\sigma_2 1} = a_{\sigma_2 2} \; . \tag{9.5}$$

Consider the two terms

$$\begin{aligned} T_1 &= a_{\sigma_1 1} a_{\sigma_2 2} a_{\sigma_3 3} \ldots a_{\sigma_n n} \\ T_2 &= a_{\sigma_2 1} a_{\sigma_1 2} a_{\sigma_3 3} \ldots a_{\sigma_n n} \end{aligned}$$

We have $T_1 = T_2$ because of (9.5). If

$$\tau = T_{\sigma_1 \sigma_2} \sigma$$

then $\mathrm{sgn}\, \tau = -\mathrm{sgn}\, \sigma$ and

$$T_2 = a_{\tau_1 1} a_{\tau_2 2} a_{\tau_3 3} \ldots a_{\tau_n n} \; .$$

In the sum

$$d(a^1, a^2, \ldots, a^n) = \sum_{\sigma \in S_n} sgn(\sigma)\, a_{\sigma_1 1} \ldots a_{\sigma_n n}$$

the two terms T_1 and $-T_2$ cancel each other.

(P3) If $a^j = e^j$ for $1 \le j \le n$ then the only non–zero term in the sum (9.4) occurs for $\sigma = id$.

$\diamond$

For later reference, we note the next result, which follows from formula (9.3).

Lemma 9.5 *Let $d : F^n \times \ldots \times F^n \to F$ be a mapping that has the properties $(P1)$ and $(P2)$. Then we have*

$$d(b^1, \ldots, b^n) = det(b^1, \ldots, b^n)\, d(e^1, \ldots, e^n) \; . \tag{9.6}$$

9.4 Rules for Determinants

9.4.1 Product Formula

An important property of the determinant is the product formula.

Theorem 9.3 *For any $A, B \in F^{n \times n}$ we have*

$$det(AB) = det(A)\, det(B) \; .$$

Proof: Note that

$$AB = (Ab^1, \ldots, Ab^n) \; .$$

Define $d : F^n \times \ldots \times F^n \to F$ by

$$d(b^1, \ldots, b^n) = det(AB) = det(Ab^1, \ldots, Ab^n) \ .$$

It is easy to see that this function d has the properties $(P1)$ and $(P2)$. Therefore, using (9.6):

$$d(b^1, \ldots, b^n) = det(b^1, \ldots, b^n) \ d(e^1, \ldots, e^n) \ .$$

Since

$$Ae^j = a^j$$

we have

$$d(e^1, \ldots, e^n) = det(Ae^1, \ldots, Ae^n) = det(a^1, \ldots, a^n) = det(A) \ .$$

This proves the theorem. $\diamond$

9.4.2 The Cases $n = 1, 2, 3$

Recall the definition

$$det(A) = \sum_{\sigma \in S_n} sgn(\sigma) \, a_{\sigma_1 1} \ldots a_{\sigma_n n} \ .$$

For $n = 1$ this becomes

$$det(a_{11}) = a_{11} \ .$$

For $n = 2$:

$$det(A) = a_{11}a_{22} - a_{12}a_{21} \ .$$

For $n = 3$:

$$det(A) = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} \ .$$

9.4.3 Triangular Matrices

If A is lower triangular or upper triangular, then $det(A)$ is the product of the diagonal elements. In this case, the permutation $\sigma = id$ is the only permutation which can lead to a non–zero product

$$a_{\sigma_1 1} \ldots a_{\sigma_n n} \ . \tag{9.7}$$

The reason is simple: If $\sigma \neq id$, then there exist i and j with

$$\sigma_i < i \quad \text{and} \quad \sigma_j > j \ .$$

If A is upper triangular, for example, then

$$a_{\sigma_j j} = 0$$

and the product (9.7) equals zero. For $\sigma = id$ the product (9.7) equals $a_{11}a_{22}\dots a_{nn}$.

One obtains:

$$det \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & & \vdots \\ & & \ddots & \vdots \\ 0 & & & a_{nn} \end{pmatrix} = a_{11}a_{22}\dots a_{nn} \ .$$

9.4.4 Existence of A^{-1}

Let $A \in F^{n\times n}$ have an inverse, A^{-1}. From

$$AA^{-1} = I$$

we obtain that

$$det(A)\, det(A^{-1}) = det(I) = 1 \ .$$

In particular, we have shown that $det(A) \neq 0$ if A has an inverse.

If A has no inverse, then there is a vector $x \in F^n$ with

$$Ax = 0 \quad \text{and} \quad x \neq 0 \ .$$

For simplicity of notation, let $x_1 \neq 0$. From

$$x_1 a^1 + x_2 a^2 + \dots + x_n a^n = 0$$

we obtain

$$a^1 = \sum_{j=2}^n y_j a^j, \quad y_j = -x_j/x_1 \ .$$

We then have

$$\begin{aligned} det(a^1, a^2, \dots, a^n) &= det\Big(\sum_{j=2}^n y_j a^j, a^2, \dots a^n\Big) \\ &= \sum_{j=2}^n y_j det\Big(a^j, a^2, \dots a^n\Big) \\ &= 0 \end{aligned}$$

We have shown:

Theorem 9.4 *A matrix $A \in F^{n\times n}$ has an inverse A^{-1} if and only if $det(A) \neq 0$. If A has an inverse, then*

$$det(A^{-1}) = \frac{1}{det(A)} \ .$$

9.4.5 Transpose

Theorem 9.5 *For any $A \in F^{n \times n}$ we have*

$$det(A) = det(A^T) \ .$$

Proof: Set $B = A^T$, thus $b_{ij} = a_{ji}$. Abbreviate

$$p_\sigma = \prod_{j=1}^{n} a_{\sigma_j j}$$

and

$$q_\tau = \prod_{k=1}^{n} b_{\tau_k k} \ .$$

We then have, by the definition of the determinant,

$$det(A) = \sum_\sigma sgn(\sigma)\, p_\sigma$$

and

$$det(B) = \sum_\tau sgn(\tau)\, q_\tau \ .$$

Let $\tau = \sigma^{-1}$. We then have $\tau(\sigma_j) = j$, thus

$$\begin{aligned} p_\sigma &= \prod_{j=1}^{n} a_{\sigma_j j} \\ &= \prod_{j=1}^{n} a_{\sigma_j \tau(\sigma_j)} \\ &= \prod_{k=1}^{n} a_{k\tau(k)} \\ &= \prod_{k=1}^{n} b_{\tau_k k} \\ &= q_\tau \end{aligned}$$

Using Lemma 9.3, it is clear that $sgn(\sigma^{-1}) = sgn(\sigma)$. Therefore,

$$\begin{aligned} det(A) &= \sum_\sigma sgn(\sigma)\, p_\sigma \\ &= \sum_\sigma sgn(\sigma^{-1})\, q_{\sigma^{-1}} \\ &= \sum_\tau sgn(\tau)\, q_\tau \\ &= det(B) \\ &= det(A^T) \end{aligned}$$

We have also used that, if σ runs through all permutations in S_n, then so does σ^{-1}. $\diamond$

9.4.6 Block Matrices

Theorem 9.6 *Let $A \in F^{k\times k}, B \in F^{l\times l}, X \in F^{k\times l}$ and let*

$$C = \begin{pmatrix} A & X \\ 0 & B \end{pmatrix} \in F^{n\times n}, \quad n = k + l \ .$$

Then we have

$$det(C) = det(A)\, det(B) \ .$$

Proof: We have

$$\begin{aligned} det(A) &= \sum_{\sigma\in S_k} sgn(\sigma)\, a_{\sigma_1 1} \ldots a_{\sigma_k k} \\ det(B) &= \sum_{\tau\in S_l} sgn(\tau)\, b_{\tau_1 1} \ldots b_{\tau_k k} \end{aligned}$$

Now fix any $\sigma \in S_k$ and $\tau \in S_l$ and define $\phi \in S_n$ (with $n = k + l$) by

$$\phi \simeq \begin{pmatrix} 1 & \ldots & k & k+1 & \ldots & k+l \\ \sigma_1 & \ldots & \sigma_k & k+\tau_1 & \ldots & k+\tau_l \end{pmatrix} \ . \tag{9.8}$$

We have

$$sgn(\sigma) sgn(\tau) = sgn(\phi)$$

and

$$sgn(\sigma)\, a_{\sigma_1 1} \ldots a_{\sigma_k k}\, sgn(\tau)\, b_{\tau_1 1} \ldots b_{\tau_k k} = sgn(\phi) c_{\phi_1 1} \ldots c_{\phi_n n} \ .$$

Therefore,

$$det(A) det(B) = \sum_{\phi} sgn(\phi)\, c_{\phi_1 1} \ldots c_{\phi_n n} \tag{9.9}$$

where the sum is taken over all permutations $\phi \in S_n$ which have the form (9.8) for some $\sigma \in S_k, \tau \in S_l$.

However, if $\phi \in S_n$ does not have the form (9.8), then there exists an index $1 \le j \le k$ with $\phi_j > k$, thus

$$c_{\phi_j j} = 0 \ .$$

It follows that the sum in (9.9) equals

$$\sum_{\phi\in S_n} sgn(\phi)\, c_{\phi_1 1} \ldots c_{\phi_n n} = det(C) \ .$$

$\diamond$

9.4.7 Cramer's Rule

Theorem 9.7 *Let $A = (a^1, \ldots, a^n) \in F^{n\times n}$ be nonsingular. The solution x of the system $Ax = b$ is given by*

$$x_i = \frac{det(A_i)}{det(A)} \quad for \quad i = 1, \ldots, n$$

where the matrix $A_i \in F^{n\times n}$ is obtained from A by replacing the i–th column a^i of A by the right–hand side b of the system $Ax = b$.

Proof: We have

$$\begin{aligned} A_i &= A + (b - a^i)e^{iT} \\ &= A(I + A^{-1}(b - a^i)e^{iT}) \\ &= A(I + (x - e^i)e^{iT}) \\ &= AB \end{aligned}$$

where

$$B = I + (x - e^i)e^{iT} = \begin{pmatrix} I_{i-1} & * & 0 \\ 0 & x_i & 0 \\ 0 & * & I_{n-i} \end{pmatrix} .$$

(Note that the entries $I_{ii} = 1$ and $(e^i e^{iT})_{ii} = 1$ cancel each other.) It follows that

$$det(A_i) = det(A)\, det(B) = det(A) x_i \ .$$

$\diamond$

9.4.8 Determinant Formula for A^{-1}

If the 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is nonsingular, then its inverse is

$$A^{-1} = \frac{1}{detA} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} .$$

The following theorem generalizes this formula to $n \times n$ matrices.

Theorem 9.8 *Let $A \in F^{n\times n}$ be nonsingular and let $n > 1$. Let $A_{ji} \in F^{(n-1)\times(n-1)}$ be obtained from A by deleting row j and column i. Then the following formula holds for the elements of A^{-1}:*

$$(A^{-1})_{ij} = \frac{(-1)^{i+j}\, det(A_{ji})}{det(A)} .$$

Proof: Let x denote the first column of A^{-1}, thus $Ax = e^1$. By Cramer's rule we have

$$x_i = \frac{det(A_i)}{det(A)}$$

where

$$A_i = (a^1 \ldots a^{i-1} e^1 a^{i+1} \ldots a^n) \ .$$

It follows that

$$det(A_i) = (-1)^{i-1} \, det(e^1 a^1 \ldots a^{i-1} a^{i+1} \ldots a^n) = (-1)^{i+1} det(A_{1i})$$

and

$$(A^{-1})_{i1} = x_i = \frac{(-1)^{i+1} det(A_{1i})}{det(A)} \ .$$

The proof for the entries in the j–th column of A^{-1} is similar. $\diamond$

9.4.9 Column Expansion and Row Expansion

Theorem 9.9 *(Expansion with respect to column j) Let $A \in F^{n \times n}, n > 1$, and let $A_{ij} \in F^{(n-1)\times(n-1)}$ be obtained from A by deleting row i and column j. Then we have for each fixed j:*

$$det(A) = \sum_{i=1}^n (-1)^{i+j} a_{ij} \, det(A_{ij}) \ .$$

Proof: Let $j = 1$ for simplicity of notation. Write the first column of A as

$$a^1 = \sum_{i=1}^n a_{i1} e^i \ .$$

We have

$$\begin{aligned} det(A) &= det\Big(\sum_i a_{i1} e^i, a^2 \ldots a^n \Big) \\ &= \sum_i a_{i1} \, det(e^i a^2 \ldots a^n) \end{aligned}$$

Here we have with entry 1 in row i:

$$(e^i a^2 \ldots a^n) = \begin{pmatrix} 0 & \vdots & & \vdots \\ 0 & \vdots & & \vdots \\ 1 & a^2 & \ldots & a^n \\ 0 & \vdots & & \vdots \\ 0 & \vdots & & \vdots \end{pmatrix}$$

and, therefore,

$$det(e^i a^2 \ldots a^n) = det \begin{pmatrix} 0 & * & \ldots & * \\ 0 & * & \ldots & * \\ 1 & 0 & \ldots & 0 \\ 0 & * & \ldots & * \\ 0 & * & \ldots & * \end{pmatrix}$$

where $*$ stands for the matrix entries $a_{\alpha\beta}$. Exchanging two rows $i-1$ times one obtains that

$$det(e^i a^2 \ldots a^n) = (-1)^{i+1} det \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & & & \\ \vdots & & A_{i1} & \\ 0 & & & \end{pmatrix} = (-1)^{i+1} det(A_{i1}) \ .$$

This proves the formula. $\diamond$

Using that $det(A) = det(A^T)$ one can turn the column expansion formula into row expansion.

Theorem 9.10 *(Expansion with respect to row i) Let $A \in F^{n\times n}, n > 1$, and let $A_{ij} \in F^{(n-1)\times(n-1)}$ be obtained from A by deleting row i and column j. Then we have for each fixed i:*

$$det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij}\, det(A_{ij}) \ .$$

9.5 Remarks on Computing Determinants

If n is large, the formula

$$det(A) = \sum_{\sigma \in S_n} sgn(\sigma)\, a_{\sigma_1 1} \ldots a_{\sigma_n n}$$

is not useful for evaluating $det(A)$. For example, let $n = 100$. To compute any product in the sum takes 100 operations. The sum has

$$N = 100! \sim 9.33 * 10^{157}$$

terms. Thus one needs about

$$Q_A \sim 10^{160}$$

operations. A teraflop machine performs 10^{12} operations per second. The age of the universe is, very roughly,

$$T \sim 2 * 10^{10} \text{ years } \sim 2 * 10^{10} * 3 * 10^7 \text{ sec} \ .$$

(The number of seconds in a year is $365 * 86,400 \sim 3 * 10^2 * 10^5$.) Thus, a teraflop machine starting at the time of the big bang, has performed about

$$Q_B \sim 6 * 10^{29}$$

operations. Thus, the number Q_B is off from Q_A by a factor $\sim 10^{130}$.

However, we can perform the LU–factorization (with partial pivoting) of A in $\sim 10^6$ operations. (Here we assume that we can carry out exact arithmetic in the field F.) If the factorization breaks down, then $det(A) = 0$. If it does not break down, it yields the determinant in $\sim 10^6$ operations, which takes about $10^{-6} sec$ on a teraflop machine.

9.6 The Permanent of a Matrix

The permanent of $A \in F^{n \times n}$ is defined as

$$per(A) = \sum_{\sigma \in S_n} a_{\sigma_1 1} \ldots a_{\sigma_n n} \; .$$

The definition agrees with that of $det(A)$, except that the factors $sgn(\sigma)$ multiplying $a_{\sigma_1 1} \ldots a_{\sigma_n n}$ are missing in the definition of $per(A)$.

Can you obtain an algorithm which computes $per(A)$ in a number of operations $Q(n)$ with polynomial bound in n? Thus, one would like to have an algorithm with

$$Q(n) \le C q^n$$

for all large n, where C and q do not depend on n.

For the computation of $det(A)$ the algorithm based on LU–factorization yields $Q_{det}(n) \le C n^3$.

For the computation of $per(A)$ no algorithm with polynomial bound is known. In fact, if such an algorithm can be shown to exist, then $P = NP$. (The computation of $per(A)$ is an NP–complete problem.) The question if $P = NP$ or $P \ne NP$ is the most important open problem of theoretical computer science.

The computation of $per(A)$ comes up in the so-called marriage problem. Given a set of n women, $W_1, \ldots, W_n$, and a set of n men, $M_1, \ldots, M_n$. For $1 \le i, j \le n$ define

$$a_{ij} = 1 \quad \text{if} \quad W_i \text{ can marry } M_j$$

and

$$a_{ij} = 0 \quad \text{if} \quad W_i \text{ cannot marry } M_j \; .$$

Clearly, this leads to an $n \times n$ matrix $A = (a_{ij})$ with entries a_{ij} equal zero or one. If $\sigma \in S_n$ and

$$a_{\sigma_1 1} \ldots a_{\sigma_n n} = 1$$

then woman W_{σ_j} can marry man M_j for every $1 \le j \le n$. Such a permutation is called a solution of the marriage problem encoded in A. Then

$$per(A)$$

is the number of solutions of the marriage problem encoded in A.

9.7 The Characteristic Polynomial

Let $A \in \mathbb{C}^{n\times n}$ and let $z \in \mathbb{C}$. Define the characteristic polynomial of A by

$$\begin{aligned} p_A(z) &= det(A - zI) \\ &= \sum_{\sigma \in S_n} sgn(\sigma)\,(a_{\sigma_1 1} - \delta_{\sigma_1 1} z)\ldots(a_{\sigma_n n} - \delta_{\sigma_n n} z) \end{aligned}$$

It is clear that $p_A(z)$ is a polynomial in z of degree n.

In fact, if $\sigma = id$ then

$$\begin{aligned} (a_{\sigma_1 1} - \delta_{\sigma_1 1} z)\ldots(a_{\sigma_n n} - \delta_{\sigma_n n} z) &= (a_{11} - z)\ldots(a_{nn} - z) \\ &= (-1)^n z^n + (-1)^{n-1}(a_{11} + \ldots + a_{nn})z^{n-1} + q(z) \end{aligned}$$

where

$$\partial q(z) \le n - 2 \ .$$

If $\sigma \neq id$ then there are at least two indices j with

$$\sigma_j \neq j, \quad \text{thus} \quad \delta_{\sigma_j j} = 0 \ .$$

This yields that the degree of

$$(a_{\sigma_1 1} - \delta_{\sigma_1 1} z)\ldots(a_{\sigma_n n} - \delta_{\sigma_n n} z)$$

is $\le n-2$. It follows that the characteristic polynomial $p_A(z)$ has degree n and has the form

$$p_A(z) = (-1)^n z^n + (-1)^{n-1}\, tr(A) z^{n-1} + \alpha_{n-2} z^{n-2} + \ldots + \alpha_1 z + \alpha_0$$

with

$$\begin{aligned} tr(A) &= a_{11} + \ldots + a_{nn} \\ det(A) &= p_A(0) = \alpha_0 \end{aligned}$$

9.8 Vandermond Determinants

A square matrix of the form

$$A_n(x_1, x_2, \dots, x_n) = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{pmatrix} \tag{9.10}$$

is called a Vandermond matrix.

For example,

$$A_2(\alpha, \beta) = \begin{pmatrix} 1 & \alpha \\ 1 & \beta \end{pmatrix}, \quad A_3(\alpha, \beta, \gamma) = \begin{pmatrix} 1 & \alpha & \alpha^2 \\ 1 & \beta & \beta^2 \\ 1 & \gamma & \gamma^2 \end{pmatrix} .$$

Let

$$V_n(x_1, x_2, \dots, x_n) = det\, A_n(x_1, x_2, \dots, x_n)$$

denote the determinant of the Vandermond matrix (9.10). We claim that the following product formula holds:

$$V_n(x_1, x_2, \dots, x_n) = \Pi_{1 \le i < j \le n}(x_j - x_i) .$$

The formula holds for $n = 2$:

$$det \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} = x_2 - x_1 .$$

We prove the general formula by induction in n. The formula clearly holds if we have $x_j = x_i$ for some $j \neq i$. Therefore, we may assume that the numbers $x_1, x_2, \dots, x_n$ are all distinct.

We fix $x_1, \dots, x_{n-1}$ and consider the polynomial

$$p(x) = V_n(x_1, x_2, \dots, x_{n-1}, x) = det \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \dots & x_{n-1}^{n-1} \\ 1 & x & x^2 & \dots & x^{n-1} \end{pmatrix} \tag{9.11}$$

It is clear that $p(x)$ is a polynomial of degree $\le n-1$ with zeros at $x_1, x_2, \dots, x_{n-1}$. Therefore,

$$p(x) = \alpha(x - x_1)(x - x_2)\dots(x - x_{n-1}) \tag{9.12}$$

where α is independent of x, but depends on $x_1, x_2, \dots, x_{n-1}$. The polynomial $p(x)$ has the form

$$p(x) = \alpha x^{n-1} + q(x) \quad \text{where} \quad \partial q(x) \le n - 2 .$$

Expand the determinant in (9.11) with respect to the last row to obtain that the coefficient α of x^{n-1} equals

$$\alpha = V_{n-1}(x_1, \ldots, x_{n-1}) \ .$$

By the induction hypothesis,

$$\alpha = V_{n-1}(x_1, \ldots, x_{n-1}) = \Pi_{1 \le i < j \le n-1}(x_j - x_i)$$

and (9.12) yields that

$$\begin{aligned} V_n(x_1, x_2, \ldots, x_{n-1}, x_n) &= p(x_n) \\ &= \alpha \, \Pi_{1 \le i \le n-1}(x_n - x_i) \\ &= V_{n-1}(x_1, \ldots, x_{n-1}) \, \Pi_{1 \le i \le n-1}(x_n - x_i) \\ &= \Pi_{1 \le i < j \le n}(x_j - x_i) \ . \end{aligned}$$

This completes the induction. $\diamond$

10 Eigenvalues, Eigenvectors, and Transformation to Block–Diagonal Form

Eigenvalues and eigenvectors of matrices play a fundamental role in many applications. For example, properties of solutions of linear systems of ODEs

$$x'(t) = Ax(t) \quad \text{and} \quad Mu''(t) + Ku(t) = 0$$

depend on eigenvalues and eigenvectors.

Consider a first order ODE system $x' = Ax$, for example, where

$$A \in \mathbb{C}^{n\times n} \quad \text{and} \quad x = x(t) \in \mathbb{C}^n \; .$$

Let $T \in \mathbb{C}^{n\times n}$ denote a non–singular matrix and introduce new variables $y = y(t) \in \mathbb{C}^n$ by the transformation

$$x(t) = Ty(t) \; .$$

The system $x' = Ax$ transforms to

$$y' = By \quad \text{where} \quad B = T^{-1}AT \; .$$

A transformation from A to $T^{-1}AT$ is called a similarity transformation. If the transformed matrix $B = T^{-1}AT$ is simple, in some sense, then it may be easy to analyze the system $y' = By$ and use the transformation $x = Ty$ to understand the original system $x' = Ax$. In the simplest case, the matrix $B = T^{-1}AT$ is diagonal, but this diagonal form cannot always be achieved.

In this chapter we will discuss how a matrix $A \in \mathbb{C}^{n\times n}$ can be transformed to **block–diagonal** form,

$$T^{-1}AT = \begin{pmatrix} M_1 & 0 & \dots & 0 \\ 0 & M_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & M_s \end{pmatrix} \; .$$

The columns of the transformation matrix T will be eigenvectors and generalized eigenvectors of A. Each block matrix $M_j \in \mathbb{C}^{m_j\times m_j}$ has only one eigenvalue λ_j, and $\lambda_1, \lambda_2, \dots, \lambda_s$ are the distinct eigenvalues of A. The above transformation of A to block–diagonal form will be established in two steps: By Schur's transformation to upper triangular form and by decoupling transformations.

In Chapter 11 we will show how to further transform the blocks M_j to Jordan canonical form.

10.1 Eigenvalues Are Zeros of the Characteristic Polynomial

Definition: *Let $A \in \mathbb{C}^{n\times n}$. A number $\lambda \in \mathbb{C}$ is called an eigenvalue of A if there exists a vector $x \in \mathbb{C}^n, x \neq 0$, with $Ax = \lambda x$. If λ is an eigenvalue of A, then*

$$E_\lambda = \{x \in \mathbb{C}^n \ : \ Ax = \lambda x\} = N(A - \lambda I)$$

is called the geometric eigenspace (or simply the eigenspace) of A to the eigenvalue λ. Any vector $x \in E_\lambda \setminus \{0\}$ is called an eigenvector of A to the eigenvalue λ.

We know from Theorem 9.4 that a matrix $B \in \mathbb{C}^{n\times n}$ has no inverse if and only if $det(B) = 0$. Therefore, $\lambda \in \mathbb{C}$ is an eigenvalue of A if and only if

$$det(A - \lambda I) = 0 \ .$$

Lemma 10.1 *Let $A \in \mathbb{C}^{n\times n}$ and let*

$$p_A(z) = det(A - zI), \quad z \in \mathbb{C} \ ,$$

denote the characteristic polynomial of A. A number $\lambda \in \mathbb{C}$ is an eigenvalue of A if and only if

$$p_A(\lambda) = 0 \ .$$

10.2 The Geometric and Algebraic Multiplicities of Eigenvalues

By the fundamental theorem of algebra, there are uniquely determined distinct numbers

$$\lambda_1, \ldots, \lambda_s \in \mathbb{C}$$

and integers

$$m_1, \ldots, m_s \in \{1, 2, \ldots, n\}$$

so that

$$p_A(z) = (\lambda_1 - z)^{m_1} \cdots (\lambda_s - z)^{m_s}, \quad \sum m_j = n \ .$$

The numbers λ_j are the distinct eigenvalues of A. The integer m_j is called the algebraic multiplicity of the eigenvalue λ_j. In general, the number m_j is different from the geometric multiplicity d_j of λ_j, which is defined as the dimension of the geometric eigenspace,

$$d_j = dim\, E_{\lambda_j} \ .$$

We will see later that

$$1 \le d_j = dim\, E_{\lambda_j} \le m_j, \quad j = 1, \ldots, s \ .$$

In words, the geometric multiplicity of any eigenvalue λ_j never exceeds the algebraic multiplicity.

Example: The matrix

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

has the characteristic polynomial

$$p_A(z) = z^2 \ .$$

The only eigenvalue of A is $\lambda_1 = 0$. The algebraic multiplicity of $\lambda_1 = 0$ is $m_1 = 2$. The geometric eigenspace is

$$E_0 = N(A) = span\{e^1\} \ .$$

We see that the geometric multiplicity d_1 of the eigenvalue $\lambda_1 = 0$ equals $d_1 = 1$.

Summary: Every matrix $A \in \mathbb{C}^{n\times n}$ has a non–empty set of eigenvalues,

$$\sigma(A) = \{\lambda_1, \ldots, \lambda_s\} \ .$$

The set $\sigma(A)$ of eigenvalues of A is called the spectrum of A. The eigenvalues of A are the zeros of the characteristic polynomial $p_A(z) = det(A - zI)$.

Remark: It is not true that every linear operator L has an eigenvalue. For example, let $U = C[0,1]$ and let $L : U \to U$ be the integral operator defined by

$$(Lu)(t) = \int_0^t u(s)\, ds, \quad 0 \le t \le 1 \ .$$

Assume that $Lu = \lambda u$, i.e.,

$$\int_0^t u(s)\, ds = \lambda u(t) \quad \text{for} \quad 0 \le t \le 1 \ .$$

First assume that $\lambda \neq 0$. Differentiation yields

$$u(t) = \lambda u'(t) \ ,$$

thus

$$u'(t) = \frac{1}{\lambda} u(t) \ ,$$

thus

$$u(t) = u(0)e^{t/\lambda} \ .$$

But we have $u(0) = 0$, thus $u \equiv 0$. Second, if $\lambda = 0$, then

$$\int_0^t u(s)\, ds = 0 \quad \text{for} \quad 0 \le t \le 1 \ .$$

Again, differentiation yields that $u(t) = 0$ for $0 \le t \le 1$.

10.3 Similarity Transformations

Definition: *A matrix $A \in \mathbb{C}^{n\times n}$ is called similar to a matrix $B \in \mathbb{C}^{n\times n}$ if there exists a non–singular matrix $T \in \mathbb{C}^{n\times n}$ with*

$$T^{-1}AT = B\ .$$

The transformation from A to $T^{-1}AT$ is called a similarity transformation.

The following is easy to prove:

Lemma 10.2 *1. A is similar to A.*
2. If A is similar to B, then B is similar to A.
3. If A is similar to B and B is similar to C, then A is similar to C.

The lemma says that *similarity of matrices* is an equivalence relation in the set $\mathbb{C}^{n\times n}$. Therefore, the set $\mathbb{C}^{n\times n}$ decomposes into disjoint similarity classes. An aim, that we will address later, is to determine in each similarity class a matrix that is *as simple as possible*. This problem leads to Jordan's normal form.

Lemma 10.3 *If A and B are similar, then $p_A(z) = p_B(z)$. Consequently, similar matrices A and B have the same spectrum, $\sigma(A) = \sigma(B)$. Also, if λ_j is an eigenvalue of A with algebraic multiplicity m_j and geometric multiplicity d_j, then λ_j is an eigenvalue of B with the same multiplicities.*

Proof: We have

$$\begin{aligned} p_B(z) &= det(B - zI) \\ &= det(T^{-1}AT - zI) \\ &= det(T^{-1}(A - zI)T) \\ &= det(T^{-1})\, det(A - zI)\, det(T) \\ &= p_A(z) \end{aligned}$$

This yields that $\sigma(A) = \sigma(B)$ and also implies the agreement of the algebraic multiplicities. Further, if $Ax = \lambda x$ and $x = Ty$, then $ATy = \lambda Ty$, thus $By = \lambda y$. Thus, if $x \in E_\lambda(A)$, then $T^{-1}x \in E_\lambda(B)$. The converse also holds and the equality

$$T(E_\lambda(B)) = E_\lambda(A)$$

follows. Since T is nonsingular, the above equality implies that the eigenspaces $E_\lambda(A)$ and $E_\lambda(B)$ have the same dimension. $\diamond$

It is reasonable to ask if the previous lemma has a converse. More precisely, assume that $A, B \in \mathbb{C}^{n\times n}$ are matrices that have the same spectrum, $\sigma(A) = \sigma(B)$, and assume that for each $\lambda_j \in \sigma(A)$ we have

$$m_j(A) = m_j(B), \quad d_j(A) = d_j(B)\ .$$

Can we conclude that A and B are similar? The answer is no, in general.

Example: Consider the matrices

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} .$$

It is easy to see that $\lambda_1 = 0$ is the only eigenvalue and

$$m_1(A) = m_1(B) = 4 .$$

Also, since $rank(A) = rank(B) = 2$, we have

$$d_1(A) = d_1(B) = 2 .$$

However, $A^2 = 0$ and $B^2 \neq 0$. Therefore, A and B are not similar to each other.

The first part of the following theorem is not difficult to prove. The second part will be shown later.

Theorem 10.1 *Let $A, B \in \mathbb{C}^{n\times n}$.*

1. If A is similar to B then $\sigma(A) = \sigma(B)$ and, for every $\lambda_j \in \sigma(A)$, we have

$$rank((A - \lambda_j I)^r) = rank((B - \lambda_j I)^r) \quad \textit{for} \quad r = 1, 2, \ldots, n . \qquad (10.1)$$

2. Conversely, if $\sigma(A) = \sigma(B)$ and if (10.1) holds for every $\lambda_j \in \sigma(A)$, then A is similar to B.

10.4 Schur's Transformation to Upper Triangular Form

Similarity transformations do not change the eigenstructure[6] of a matrix A. To better understand the eigenstructure of A, one applies similarity transformations to A that lead to simpler matrices. A first and important step is the transformation of A to upper triangular form by a similarity transformation with a unitary matrix.

Theorem 10.2 *(Schur) Let $A \in \mathbb{C}^{n\times n}$ have the characteristic polynomial*

$$p_A(z) = (\mu_1 - z)(\mu_2 - z)\ldots(\mu_n - z) .$$

Here $\mu_1, \ldots, \mu_n$ are the not necessarily distinct eigenvalues of A, listed in any order. Each eigenvalue is listed according to its algebraic multiplicity. There exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ so that U^*AU is upper–triangular,

$$U^*AU = R = \begin{pmatrix} \mu_1 & \ldots & r_{1n} \\ & \ddots & \vdots \\ 0 & & \mu_n \end{pmatrix} .$$

The eigenvalues of A appear on the diagonal of R in any desired order.

[6] If $B = T^{-1}AT$ then A and B have the same eigenvalues λ_j and for any exponent $r = 1, 2, \ldots$ the nullspace of $(A - \lambda_j I)^r$ has the same dimension as the nullspace of $(B - \lambda_j I)^r$. In particular, the eigenspaces $E_{\lambda_j}(A)$ and $E_{\lambda_j}(B)$ have the same dimensions.

Proof: We use induction in n, the case $n = 1$ being trivial.

Let $\mu_1 \in \sigma(A)$. There exists a vector $u^1 \in \mathbb{C}^n$ with $|u^1| = 1$ and $Au^1 = \mu_1 u^1$. Choose $u^2, \ldots, u^n$ so that the matrix

$$U_1 = (u^1, \ldots, u^n)$$

is unitary. We then have

$$AU_1 = (\mu_1 u^1, Au^2, \ldots, Au^n)$$

and

$$U_1^* A U_1 = \begin{pmatrix} \mu_1 & * & \ldots & * \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix}$$

where $B \in \mathbb{C}^{(n-1)\times(n-1)}$. The matrix B has the characteristic polynomial

$$p_B(z) = (\mu_2 - z) \ldots (\mu_n - z) \ .$$

By the induction hypothesis, there exists a unitary matrix $V \in \mathbb{C}^{(n-1)\times(n-1)}$ with

$$V^* B V = \begin{pmatrix} \mu_2 & \ldots & * \\ & \ddots & * \\ 0 & & \mu_n \end{pmatrix} \ .$$

Setting

$$U_2 = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & & & \\ \vdots & & V & \\ 0 & & & \end{pmatrix}$$

one obtains that

$$U_2^* U_1^* A U_1 U_2 = \begin{pmatrix} \mu_1 & * & \ldots & * \\ 0 & \mu_2 & * & * \\ \vdots & & \ddots & * \\ 0 & \ldots & 0 & \mu_n \end{pmatrix}$$

The matrices U_1 and U_2 are unitary; the product $U = U_1 U_2$ is unitary. $\diamond$

Remark: The transformation by a unitary matrix U is always well–conditioned since

$$|U| = |U^{-1}| = 1 \ .$$

Schur's theorem implies that one can always transform to *upper triangular* using well–conditioned transformations. On the other hand, the transformation of a matrix A to *diagonal* form (if possible) may lead to a transformation matrix T for which

$$|T||T^{-1}|$$

is very large. This happens frequently if A has eigenvalues that are not well–separated. Consider the example

$$A = \begin{pmatrix} \varepsilon & 1 \\ 0 & 0 \end{pmatrix}, \quad 0 < \varepsilon << 1 \ .$$

We have

$$At^1 = \varepsilon t^1, \quad At^2 = 0 \ ,$$

with

$$t^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad t^2 = \begin{pmatrix} 1 \\ -\varepsilon \end{pmatrix} \ .$$

Setting $T = (t^1, t^2)$ one obtains

$$AT = T \begin{pmatrix} \varepsilon & 0 \\ 0 & 0 \end{pmatrix} \ ,$$

thus $T^{-1}AT$ is diagonal. In this case

$$|T||T^{-1}| = \mathcal{O}(1/\varepsilon) \ .$$

10.5 Transformation of Normal Matrices to Diagonal Form

A matrix $A \in \mathbb{C}^{n \times n}$ is called normal if $AA^* = A^*A$. We recall: If $A^* = A$ then A is called Hermitian; if $A^* = -A$ then A is called skew Hermitian; if $A^*A = I$ then A is unitary.

It is not difficult to show that Hermitian matrices, skew Hermitian matrices, unitary matrices, and diagonal matrices are all normal.

Lemma 10.4 *If A is normal and U is unitary, then U^*AU is also normal.*

Proof: This follows from

$$\begin{aligned} (U^*AU)(U^*AU)^* &= U^*AA^*U \\ (U^*AU)^*(U^*AU) &= U^*A^*AU \end{aligned}$$

$\diamond$

Lemma 10.5 *Let $A \in \mathbb{C}^{n \times n}$. Then A is normal if and only if*

$$|Ax| = |A^*x| \quad \textit{for all} \quad x \in \mathbb{C}^n \ .$$

Proof: a) First assume A to be normal. We have

$$\begin{aligned} |Ax|^2 &= \langle Ax, Ax \rangle \\ &= \langle x, A^*Ax \rangle \\ &= \langle x, AA^*x \rangle \\ &= \langle A^*x, A^*x \rangle \\ &= |A^*x|^2 \end{aligned}$$

We will prove the converse below. $\diamond$

Lemma 10.6 *If B is normal and upper triangular, then B is diagonal.*

Proof: First let $n = 2$, for simplicity. Let

$$B = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix}, \quad B^* = \begin{pmatrix} \bar{a} & 0 \\ \bar{b} & \bar{c} \end{pmatrix} .$$

We have

$$Be^1 = ae^1, \quad |Be^1| = |a|$$

and

$$B^*e^1 = \begin{pmatrix} \bar{a} \\ \bar{b} \end{pmatrix}, \quad |B^*e^1|^2 = |a|^2 + |b|^2 .$$

By the previous lemma, it follows that $b = 0$. For general n we also consider Be^1 and B^*e^1 and obtain that the first column of B^*, except for the diagonal entry, is zero. We then consider Be^2 and B^*e^2 etc. $\diamond$

Together with Schur's Theorem, we obtain the following important result:

Theorem 10.3 *If A is normal, then there exists a unitary matrix U so that U^*AU is diagonal. The converse also holds, i.e., if there is a unitary matrix U so that U^*AU is diagonal, then A is normal.*

One can also express this result as follows:

Theorem 10.4 *A matrix $A \in \mathbb{C}^{n \times n}$ is normal if and only if the vector space $\mathbb{C}^n$ has an orthonormal basis consisting of eigenvectors of A.*

We now complete the proof of Lemma 10.5.

Assume that $|Ax| = |A^*x|$ for all $x \in \mathbb{C}^n$. The matrices

$$H_1 = A^*A \quad \text{and} \quad H_2 = AA^*$$

are Hermitian and satisfy

$$\langle H_1x, x \rangle = \langle H_2x, x \rangle \quad \text{for all} \quad x \in \mathbb{C}^n .$$

We set $H = H_1 - H_2$ and obtain

$$\langle Hx, x \rangle = 0 \quad \text{for all} \quad x \in \mathbb{C}^n \ .$$

Clearly, the Hermitian matrix H is normal. There exists a unitary matrix U so that $U^* H U = \Lambda$ is diagonal. Setting $U^* x = y$ we obtain that

$$0 = \langle Hx, x \rangle = \langle U \Lambda U^* x, x \rangle = \langle \Lambda y, y \rangle \ .$$

Since

$$0 = \langle \Lambda y, y \rangle \quad \text{for all} \quad y \in \mathbb{C}^n$$

it follows that $\Lambda = 0$, thus $H = 0$ and $H_1 = H_2$. $\diamond$

10.6 Special Classes of Matrices

Theorem 10.5 *Let $A \in \mathbb{C}^{n \times n}$. We have:*
1. *If $A = A^*$ then all eigenvalues of A are real.*
2. *If $A = -A^*$ then all eigenvalues of A are purely imaginary.*
3. *If $A^* A = I$ then all eigenvalues of A have absolute value one.*

Proof: 1. Let $Ax = \lambda x, |x| = 1$. We have

$$\begin{aligned}
\lambda &= \lambda |x|^2 \\
&= \langle x, \lambda x \rangle \\
&= \langle x, Ax \rangle \\
&= \langle Ax, x \rangle \\
&= \langle \lambda x, x \rangle \\
&= \bar{\lambda}
\end{aligned}$$

which shows that λ is real. The proofs of 2. and 3. are similar. $\diamond$

Theorem 10.6 *Let $A \in \mathbb{R}^{n \times n}, A = A^T$. Then there is a real orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ so that $Q^T A Q$ is real, diagonal.*

Proof: The eigenvalues of A are real. By Schur's theorem, there is a unitary matrix U so that $U^* A U$ is upper triangular. The proof of Schur's theorem shows that one can choose U real if A and its eigenvalues are real. The proof of Theorem 10.3 shows that $U^T A U$ is diagonal. $\diamond$

10.7 Applications to ODEs

1) Recall that the scalar ODE

$$m u''(t) + k u(t) = 0$$

with $m > 0, k > 0$ has the general solution

$$u(t) = c_1 \cos(\omega t) + c_2 \sin(\omega t)$$

where $\omega = \sqrt{k/m}$.

Consider the system of ODEs

$$Mu''(t) + Ku(t) = 0 \tag{10.2}$$

where M and K are positive definite Hermitian matrices in $\mathbb{C}^{n\times n}$ and $u(t) \in \mathbb{C}^n$. There exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ with

$$U^*MU = D^2, \quad D = diag(d_1, \ldots, d_n), \quad d_j > 0 \ .$$

Write

$$M = UD^2U^* = UDU^*UDU^* = V^2$$

with

$$V = UDU^*, \quad V = V^* > 0 \ .$$

Using the new variable

$$v(t) = Vu(t)$$

the system (10.2) becomes $MV^{-1}v'' + KV^{-1}v = 0$, thus

$$v''(t) + V^{-1}KV^{-1}v(t) = 0 \ .$$

Here

$$K_1 := V^{-1}KV^{-1}$$

is positive definite Hermitian. There exists a unitary matrix U_1 with

$$U_1^*K_1U_1 = D_1^2, \quad D_1 = diag(\alpha_1, \ldots, \alpha_n), \quad \alpha_j > 0 \ .$$

The system $v'' + K_1 v = 0$ becomes

$$v''(t) + U_1 D_1^2 U_1^* v(t) = 0 \ .$$

Using the variable

$$q(t) = U_1^* v(t)$$

one obtains the diagonal system

$$q''(t) + D_1^2 q(t) = 0$$

or

$$q_j''(t) + \alpha_j^2 q_j(t) = 0, \quad j = 1, 2, \ldots, n$$

with general solution

$$q_j(t) = c_1 \cos(\alpha_j t) + c_2 \sin(\alpha_j t) \ .$$

It follows that all solutions $u(t)$ of the second order system (10.2) are oscillatory.

2) Consider the ODE system

$$u''(t) = Au(t)$$

where $A = A^T \in \mathbb{R}^{n\times n}$. There exists an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ so that

$$Q^{-1}AQ = \Lambda = diag(\lambda_1, \ldots, \lambda_n), \quad \lambda_j \in \mathbb{R} \ .$$

Introduce a new vector variable $v(t)$ by the transformation

$$u(t) = Qv(t)$$

and obtain

$$v''(t) = \Lambda v(t) \ ,$$

i.e.,

$$v_j''(t) = \lambda_j v_j(t), \quad j = 1, \ldots, n \ .$$

a) Let $\lambda_j < 0$. Write $\lambda_j = -\kappa_j^2, \kappa_j > 0$. The general solution of the equation

$$v_j'' + \kappa_j^2 v_j = 0$$

is an oscillation

$$v_j(t) = \alpha \sin(\kappa_j t) + \beta \cos(\kappa_j t) \ .$$

b) Let $\lambda_j > 0$. Write $\lambda_j = \kappa_j^2, \kappa_j > 0$. The general solution of the equation

$$v_j'' = \kappa_j^2 v_j$$

is

$$v_j(t) = \alpha e^{\kappa_j t} + \beta e^{\kappa_j t} \ .$$

The exponentially growing term (if present) makes the stationary solution $u \equiv 0$ of the system $u'' = Au$ unstable.

c) Let $\lambda_j = 0$. The general solution of the equation

$$v_j'' = 0$$

is

$$v_j(t) = \alpha t + \beta \ .$$

The growing term αt (if present) makes the stationary solution $u \equiv 0$ of the system $u'' = Au$ unstable.

One obtains that the solution $u \equiv 0$ of the system $u'' = Au$ is stable if and only if all eigenvalues λ_j of $A = A^T$ are negative.

10.8 Hadamard's Inequality

The following theorem is easy to prove.

Theorem 10.7 *Let $P \in \mathbb{C}^{n\times n}$ be Hermitian and let*

$$\langle x, Px\rangle > 0 \quad \textit{for all} \quad x \in \mathbb{C}^n, \quad x \neq 0 \ .$$

(One calls P a positive definite Hermitian matrix.) Then all eigenvalues λ_j of P are real and positive.

Proof: Let $Px = \lambda x, x \in \mathbb{C}^n, x \neq 0$. We have

$$\begin{aligned}
\lambda |x|^2 &= \langle x, \lambda x\rangle \\
&= \langle x, Px\rangle \\
&= \langle Px, x\rangle \\
&= \langle \lambda x, x\rangle \\
&= \bar{\lambda} |x|^2
\end{aligned}$$

It follows that λ is real and $\lambda > 0$ since $\langle x, Px\rangle > 0$. $\diamond$

Theorem 10.8 *(Hadamard's Inequality) Let*

$$A = (a^1, \ldots, a^n) \in \mathbb{C}^{n\times n} \ ,$$

i.e., $a^j \in \mathbb{C}^n$ is the j–th column of A. Then the estimate

$$|det(A)| \leq |a^1||a^2| \cdots |a^n|$$

holds. Here $|a^j|$ denotes the Euclidean vector norm of a^j.

We first prove the important geometric–arithmetic mean inequality.

Theorem 10.9 *Let $x_1, \ldots, x_n$ denote n positive real numbers. Then the inequality*

$$\Big(x_1 x_2 \cdots x_n\Big)^{1/n} \leq \frac{1}{n}\,(x_1 + x_2 + \cdots + x_n)$$

holds.

Proof: Define the real function

$$f(x) = e^{x-1} - x, \quad x \in \mathbb{R} \ .$$

We have

$$f'(x) = e^{x-1} - 1, \quad f''(x) = e^{x-1} > 0 \ .$$

Since

$$f(1) = 0, \quad f'(1) = 0 \quad \text{and} \quad f''(x) > 0 \quad \text{for all} \quad x \in \mathbb{R} \ ,$$

one obtains that

$$f(x) \geq 0 \quad \text{for all} \quad x \in \mathbb{R} \ ,$$

thus

$$x \leq e^{x-1} \quad \text{for all} \quad x \in \mathbb{R} \ .$$

Set

$$\alpha = \frac{1}{n}\,(x_1 + x_2 + \cdots + x_n) \ .$$

Then we have

$$\frac{x_j}{\alpha} \leq \exp\Big(\frac{x_j}{\alpha} - 1\Big) \ .$$

Therefore,

$$\begin{aligned}
\frac{x_1}{\alpha}\cdot\frac{x_2}{\alpha}\cdots\frac{x_n}{\alpha} &\leq \exp\Big(\frac{x_1}{\alpha}-1\Big)\cdot\exp\Big(\frac{x_2}{\alpha}-1\Big)\cdots\exp\Big(\frac{x_n}{\alpha}-1\Big) \\
&= \exp\Big(\frac{x_1+x_2+\ldots+x_n}{\alpha} - n\Big) \\
&= \exp\Big(\frac{\alpha n}{\alpha} - n\Big) \\
&= e^0 \\
&= 1
\end{aligned}$$

This proves that

$$x_1 x_2 \cdots x_n \leq \alpha^n \ ,$$

i.e.,

$$\Big(x_1 x_2 \cdots x_n\Big)^{1/n} \leq \alpha \ .$$

$\diamond$

Proof of Hadamard's Inequality: We may assume that

$$\alpha_j := |a^j| > 0 \quad \text{for} \quad j = 1, \ldots, n$$

and that $det(A) \neq 0$. Set

$$m^j = \frac{1}{\alpha_j}\, a^j, \quad M = (m^1, \ldots, m^n) \ .$$

We then have

$$\begin{aligned} det(A) &= det(\alpha_1 m^1, \ldots, \alpha_n m^n) \\ &= \alpha_1 \cdots \alpha_n \, det(M) \end{aligned}$$

and must prove that

$$|det(M)| \leq 1 \ .$$

Set

$$P = M^* M \ .$$

Then P is positive definite, Hermitian. Also,

$$p_{jj} = m^{j*} m^j = 1 \quad \text{for} \quad j = 1, \ldots, n \ .$$

This shows that $tr(P) = n$.

Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalue of P, thus $\lambda_j > 0$. We have $\sum_j \lambda_j = n$ and

$$\begin{aligned} |det(M)|^2 &= det(P) \\ &= \lambda_1 \cdots \lambda_n \\ &\leq \Big(\frac{1}{n} \sum_j \lambda_j \Big)^n \\ &= 1^n \\ &= 1 \end{aligned}$$

This proves the bound $|det(M)| \leq 1$. $\diamond$

Another Proof of Hadamard's Inequality: We can write

$$A = QR \quad \text{where} \quad Q^* Q = I \quad \text{and} \quad R \ \text{ is upper triangular} \ .$$

If

$$A = (a^1, \ldots, a^n), \quad Q = (q^1, \ldots, q^n)$$

then

$$A = QR = (q^1, \ldots, q^n) \begin{pmatrix} r_{11} & \ldots & r_{1j} & \ldots & r_{1n} \\ & \ddots & \vdots & & \vdots \\ & & r_{jj} & & \vdots \\ & & & \ddots & \vdots \\ 0 & & & & r_{nn} \end{pmatrix}$$

implies that

$$a^j = (QR)^j = \sum_{i=1}^j r_{ij} q^i \ .$$

(Here $(QR)^j$ denotes the j–th column of QR.) Thus

$$|a^j|^2 = \sum_{i=1}^j |r_{ij}|^2 \geq |r_{jj}^2| \ .$$

Using the estimate $|r_{jj}| \leq |a^j|$ we obtain that

$$|det(A)| = |det(R)| = \Pi_j |r_{jj}| \leq \Pi_j |a^j| \ .$$

10.9 Diagonalizable Matrices

A class of matrices, larger than the class of normal matrices, are the diagonalizable matrices.

Definition: *A matrix $A \in \mathbb{C}^{n \times n}$ is called diagonalizable if there exists a non–singular matrix T so that $T^{-1}AT$ is diagonal.*

Assume

$$T^{-1}AT = \Lambda = diag(\lambda_1, \ldots, \lambda_n), \quad T = (t^1, \ldots, t^n) \ .$$

This yields that $AT = T\Lambda$, thus

$$At^j = \lambda_j t^j, \quad j = 1, \ldots, n \ .$$

In other words, if $T^{-1}AT$ is diagonal, then the columns of T contain the eigenvectors of A. The converse also holds. One obtains:

Theorem 10.10 *A matrix $A \in \mathbb{C}^{n \times n}$ is diagonalizable if and only if the vector space $\mathbb{C}^n$ has a basis of eigenvectors of A.*

The following result holds in finite and infinite dimensions.

Theorem 10.11 *Let U be a vector space and let $L : U \to U$ denote any linear operator. If $u_1, \ldots, u_k$ are eigenvectors of L to distinct eigenvalues, then $u_1, \ldots, u_k$ are linearly independent.*

Proof: Use induction in k. The claim is obvious for $k = 1$. Suppose any $k-1$ eigenvectors to $k-1$ distinct eigenvalues are linearly independent.

Let $Lu_j = \lambda_j u_j, j = 1, \ldots, k$. Assume

$$c_1 u_1 + \ldots + c_k u_k = 0 \ . \tag{10.3}$$

Multiplication by λ_k yields

$$c_1 \lambda_k u_1 + \ldots + c_k \lambda_k u_k = 0 \ . \tag{10.4}$$

Applying L to (10.3) one obtains

$$c_1\lambda_1 u_1 + \ldots + c_k\lambda_k u_k = 0 \; . \tag{10.5}$$

Subtracting (10.5) from (10.4) we have

$$c_1(\lambda_1 - \lambda_k)u_1 + \ldots + c_{k-1}(\lambda_{k-1} - \lambda_k)u_{k-1} = 0 \; . \tag{10.6}$$

By the induction assumption, $u_1, \ldots, u_{k-1}$ are linearly independent. Therefore, the coefficients are zero. Since the λ_j are distinct, it follows that $c_1 = \ldots = c_{k-1} = 0$. $\diamond$

Theorem 10.12 *Assume that the matrix $A \in \mathbb{C}^{n\times n}$ has n distinct eigenvalues. Then A is diagonalizable.*

Theorem 10.13 *The set $\mathcal{D}_n$ of all diagonalizable matrices in $\mathbb{C}^{n\times n}$ is dense in $\mathbb{C}^{n\times n}$.*

Proof: Let $A \in \mathbb{C}^{n\times n}$ be arbitrary and let $U^*AU = \Lambda + R$ where U is unitary, Λ is diagonal and R is strictly upper triangular. Let $\varepsilon_0 > 0$ be given and let

$$D_\varepsilon = diag(\varepsilon_1, \ldots, \varepsilon_n), \quad |\varepsilon_j| \le \varepsilon_0 \quad \text{for} \quad j = 1, \ldots, n \; .$$

Consider the matrix

$$A_\varepsilon = U(\Lambda + D_\varepsilon + R)U^* \; .$$

Then A_ε has the eigenvalues $\lambda_j + \varepsilon_j$ and we can choose the ε_j so that the eigenvalues of A_ε are distinct. Therefore, A_ε is diagonalizable. Also,

$$|A - A_\varepsilon| = |UD_\varepsilon U^*| = |D_\varepsilon| \le \varepsilon_0 \; .$$

This proves that, given any $\varepsilon_0 > 0$, there is a diagonalizable matrix A_ε with $|A - A_\varepsilon| \le \varepsilon_0$. $\diamond$

Though the last result is of some theoretical interest, it is not useful in practice since the transformation matrix T_ε of A_ε to diagonal form may have a very large condition number. Put differently, the study of the spectral properties of matrices that are not diagonalizable deserves some special attention. This is addressed with the transformation to Jordan form.

10.10 Transformation to Block–Diagonal Form

10.10.1 Two Auxiliary Results

Let $A = D + R$ denote an upper triangular matrix where D is diagonal and R is strictly upper triangular. The following lemma formulates a technical tool that allows one to *scale down* the strictly upper triangular part R by applying a similarity transformation.

Lemma 10.7 *Let $A = D + R \in \mathbb{C}^{n\times n}$ where D is diagonal and R is strictly upper triangular. For $0 < \varepsilon \le 1$ consider the diagonal matrix*

$$T_\varepsilon = diag(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1}) \ .$$

Then

$$T_\varepsilon^{-1} A T_\varepsilon = D + R_\varepsilon$$

where

$$|R_\varepsilon|_\infty \le \varepsilon |R|_\infty \ .$$

Proof: Let r_{ij} for $1 \le i < j \le n$ denote the strictly upper triangular elements of R. It is easy to check that

$$(R_\varepsilon)_{ij} = r_{ij}\varepsilon^{j-i}, \quad 1 \le i < j \le n \ .$$

Thus, every nonzero entry of R gets multiplied by a positive factor less than or equal to ε. It then follows that

$$|(R_\varepsilon)_{ij}| \le \varepsilon |r_{ij}| \quad \text{for all} \quad i,j \quad \text{and} \quad 0 < \varepsilon \le 1 \ .$$

Also,

$$|R|_\infty = \max_i \sum_j |r_{ij}| \ ,$$

thus

$$|R_\varepsilon|_\infty \le \varepsilon |R|_\infty \ .$$

$\diamond$

Lemma 10.8 *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. If $B \in \mathbb{C}^{n\times n}$ satisfies the bound*

$$|B| < \frac{1}{|A^{-1}|}$$

then $A + B$ is also nonsingular.

Proof: Let $(A + B)x = 0$, thus $Ax = -Bx$ and

$$x = -A^{-1}Bx \ .$$

It follows that

$$|x| \le |A^{-1}||B||x| \ .$$

By assumption,

$$|A^{-1}||B| < 1 \ ,$$

thus $x = 0$. This yields that $A + B$ is nonsingular. $\diamond$

10.10.2 The Blocking Lemma

Consider a block matrix of the form

$$A = \begin{pmatrix} M_1 & M_{12} \\ 0 & M_2 \end{pmatrix}$$

where

$$M_1 \in \mathbb{C}^{k\times k}, \quad M_2 \in \mathbb{C}^{l\times l}, \quad M_{12} \in \mathbb{C}^{k\times l} \ .$$

The matrix M_{12} describes a coupling between the blocks. We want to eliminate the coupling by a similarity transformation, $T^{-1}AT$. We claim that this can be done if

$$\sigma(M_1) \cap \sigma(M_2) = \emptyset \ , \tag{10.7}$$

i.e., if M_1 and M_2 have no common eigenvalue. To this end, consider a matrix T of the form

$$T = \begin{pmatrix} I_k & S \\ 0 & I_l \end{pmatrix}, \quad S \in \mathbb{C}^{k\times l} \ .$$

It is easy to check that

$$T^{-1} = \begin{pmatrix} I_k & -S \\ 0 & I_l \end{pmatrix}$$

and

$$T^{-1}AT = \begin{pmatrix} I_k & -S \\ 0 & I_l \end{pmatrix} \begin{pmatrix} M_1 & M_{12} \\ 0 & M_2 \end{pmatrix} \begin{pmatrix} I_k & S \\ 0 & I_l \end{pmatrix} = \begin{pmatrix} M_1 & X \\ 0 & M_2 \end{pmatrix}$$

with

$$X = M_{12} + M_1 S - S M_2 \ .$$

In order to achieve a decoupling, we must find $S \in \mathbb{C}^{k\times l}$ so that $X = 0$. Note that the condition

$$M_1 S - S M_2 = -M_{12} \tag{10.8}$$

consists of kl linear equations for kl unknowns s_{ij}. Therefore, the equation (10.8) has a unique solution S if we can prove that $M_1 S - S M_2 = 0$ implies $S = 0$.

Lemma 10.9 *Let $M_1 \in \mathbb{C}^{k\times k}$ and $M_2 \in \mathbb{C}^{l\times l}$ have disjoint spectra, i.e., assume (10.7). If $S \in \mathbb{C}^{k\times l}$ satisfies $M_1 S - S M_2 = 0$ then $S = 0$.*

Proof: Case 1: M_1 and M_2 are diagonal,

$$M_1 = diag(d_1, \ldots, d_k), \quad M_2 = diag(e_1, \ldots, e_l) \ .$$

The matrix equation $M_1 S - S M_2 = 0$ becomes

$$d_i s_{ij} - s_{ij} e_j = 0 \quad \text{for} \quad i = 1, \ldots, k \quad \text{and} \quad j = 1, \ldots, l \ .$$

Since $d_i \neq e_j$ we conclude that $s_{ij} = 0$.

Case 2: The matrices M_1 and M_2 are upper triangular,

$$M_1 = D_1 + R_1, \quad M_2 = D_2 + R_2 \ ,$$

where D_1, D_2 are diagonal and R_1, R_2 are strictly upper triangular. For $0 < \varepsilon << 1$ define the diagonal scaling matrices

$$T_1 = diag(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{k-1}), \quad T_2 = diag(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{l-1}) \ .$$

Then we have

$$T_1^{-1} M_1 T_1 = D_1 + P_1(\varepsilon) \quad \text{where} \quad |P_1(\varepsilon)| \leq C\varepsilon \ .$$

Similarly,

$$T_2^{-1} M_2 T_2 = D_2 + P_2(\varepsilon) \quad \text{where} \quad |P_2(\varepsilon)| \leq C\varepsilon \ .$$

From

$$M_1 = T_1(D_1 + P_1(\varepsilon))T_1^{-1}, \quad M_2 = T_2(D_2 + P_2(\varepsilon))T_2^{-1}$$

we obtain that

$$T_1(D_1 + P_1(\varepsilon))T_1^{-1} S - S T_2(D_2 + P_2(\varepsilon))T_2^{-1} = 0 \ ,$$

thus

$$(D_1 + P_1(\varepsilon))(T_1^{-1} S T_2) - (T_1^{-1} S T_2)(D_2 + P_2(\varepsilon)) = 0 \ .$$

By choosing $\varepsilon > 0$ small, the perturbation terms $P_j(\varepsilon)$ can be made arbitrarily small. Since the limit system, obtained for $\varepsilon = 0$, is nonsingular (by Case 1), it follows that

$$T_1^{-1} S T_2 = 0 \ .$$

This yields that $S = 0$.

Case 3: Let $M_1 \in \mathbb{C}^{k \times k}$ and $M_2 \in \mathbb{C}^{l \times l}$ be arbitrary, satisfying (10.7). With unitary matrices U_1, U_2 and upper triangular matrices N_1, N_2 we have

$$U_1^* M_1 U_1 = N_1, \quad U_2^* M_2 U_2 = N_2 \ .$$

We write the equation $M_1 S - S M_2 = 0$ in the form

$$U_1 N_1 U_1^* S - S U_2 N_2 U_2^* = 0 \ ,$$

thus

$$N_1(U_1^* S U_2) - (U_1^* S U_2) N_2 = 0 \ .$$

Using the result of Case 2, we conclude that $U_1^* S U_2 = 0$, i.e., $S = 0$. $\diamond$

This leads to the following blocking lemma:

Lemma 10.10 *(Blocking Lemma) Consider a block matrix*

$$A = \begin{pmatrix} M_1 & M_{12} \\ 0 & M_2 \end{pmatrix}$$

where

$$M_1 \in \mathbb{C}^{k\times k}, \quad M_2 \in \mathbb{C}^{l\times l}, \quad M_{12} \in \mathbb{C}^{k\times l} \ .$$

If M_1 and M_2 have no common eigenvalue then there exists a unique transformation matrix T of the form

$$T = \begin{pmatrix} I_k & S \\ 0 & I_l \end{pmatrix}, \quad S \in \mathbb{C}^{k\times l} \ ,$$

so that

$$T^{-1}AT = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \ .$$

Example: Let

$$A = \begin{pmatrix} \lambda_1 & b \\ 0 & \lambda_2 \end{pmatrix}, \quad \lambda_1 \neq \lambda_2$$

and

$$T = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} 1 & -s \\ 0 & 1 \end{pmatrix} \ .$$

Then we have

$$T^{-1}AT = \begin{pmatrix} \lambda_1 & x \\ 0 & \lambda_2 \end{pmatrix}$$

with

$$x = b + s(\lambda_1 - \lambda_2) \ .$$

We see that $T^{-1}AT$ is diagonal if and only if

$$s = \frac{b}{\lambda_2 - \lambda_1} \ .$$

This shows that the condition number of the transformation is

$$|T||T^{-1}| \sim \Big(1 + \frac{|b|}{|\lambda_2 - \lambda_1|}\Big)^2 \ .$$

We see that the condition number of T can be very large if the eigenvalues λ_1 and λ_2 of A are close to each other.

10.10.3 Repeated Blocking

By applying Schur's theorem and then, repeatedly, the Blocking Lemma, one obtains the following result:

Theorem 10.14 *Let $A \in \mathbb{C}^{n\times n}$ have the characteristic polynomial*

$$p_A(z) = det(A - zI) = (\lambda_1 - z)^{m_1} \dots (\lambda_s - z)^{m_s}$$

with distinct zeros $\lambda_1, \dots, \lambda_s$. Then there exists a nonsingular transformation matrix T so that $T^{-1}AT$ has the block form

$$T^{-1}AT = \begin{pmatrix} M_1 & 0 & \dots & 0 \\ 0 & M_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & M_s \end{pmatrix}$$

where each M_j has size $m_j \times m_j$ and

$$M_j = \lambda_j I_{m_j} + R_j$$

where R_j is strictly upper triangular.

To transform A further to Jordan canonical form, it suffices to consider each of the block matrices M_j separately. Thus we are lead to consider a matrix of the form

$$M = \lambda I + R$$

where R is strictly upper triangular. To understand the eigenvectors and generalized eigenvectors of such a matrix M is the core difficulty of the transformation to Jordan canonical form. We will treat this in Chapter 12.

First, we want to reinterpret Theorem 10.14 in terms of generalized eigenspaces.

10.11 Generalized Eigenspaces

Let $A \in \mathbb{C}^{n\times n}$ and let λ be an eigenvalue of A. Recall that the space

$$E_\lambda = N(A - \lambda I)$$

is called the eigenspace or geometric eigenspace of A to the eigenvalue λ. The space

$$gE_\lambda = \{x \in \mathbb{C}^n \ : \ (A - \lambda I)^j x = 0 \text{ for some } j \in \mathbb{N}\}$$

is called the generalized eigenspace of A to the eigenvalue λ.

Example: Let

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

In this case $A^2=0$. We have

$$E_0 = span\{e^1\}, \quad gE_0 = \mathbb{C}^2 \ .$$

Assume that $T^{-1}AT$ is a similarity transformation of A to block form as in Theorem 10.14. We claim that the matrix T contains in its columns bases of all the generalized eigenspaces of A. For simplicity of notation, assume that there are only two blocks,

$$T^{-1}AT = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}, \quad M_j = \lambda_j I_{m_j} + R_j \ , \quad m_1 + m_2 = n \ ,$$

and the matrices R_j are strictly upper triangular. Note that

$$R_1^{m_1} = 0, \quad R_2^{m_2} = 0 \ .$$

Lemma 10.11 *Under the above assumptions, the first m_1 columns of T form a basis of gE_{λ_1} and the last m_2 columns of T form a basis of gE_{λ_2}.*

Proof: We have

$$A - \lambda_1 I = T \begin{pmatrix} R_1 & 0 \\ 0 & X \end{pmatrix} T^{-1}, \quad X = (\lambda_2 - \lambda_1)I + R_2, \quad det(X) \neq 0 \ .$$

Therefore,

$$(A - \lambda_1 I)^j = T \begin{pmatrix} R_1^j & 0 \\ 0 & X^j \end{pmatrix} T^{-1}, \quad j = 1, 2, \ldots$$

Since $R_1^{m_1} = 0$ we have

$$(A - \lambda_1 I)^j = T \begin{pmatrix} 0 & 0 \\ 0 & X^j \end{pmatrix} T^{-1}, \quad j \geq m_1 \ .$$

Since X^j is nonsingular, one obtains that

$$rank((A - \lambda_1 I)^j) = n - m_1, \quad j \geq m_1 \ ,$$

thus

$$dim(gE_{\lambda_1}) = m_1 \ .$$

Partition T as

$$T = (T^I, T^{II})$$

where T^I contains the first m_1 columns of T. For $j \geq m_1$ we have

$$(A - \lambda_1 I)^j T = T \begin{pmatrix} 0 & 0 \\ 0 & X^j \end{pmatrix} = (0, T^{II} X^j) \ .$$

This shows that

$$(A - \lambda_1 I)^j T^I = 0 \ .$$

Therefore, the columns t^k with $1 \le k \le m_1$ lie in the nullspace of $(A - \lambda_1 I)^j = gE_{\lambda_1}$. Since the dimension of gE_{λ_1} equals m_1, the m_1 columns of T^I form a basis of gE_{λ_1}. The proof for gE_{λ_2} is similar. $\diamond$

In the general case, one obtains:

Theorem 10.15 *Let $A \in \mathbb{C}^{n \times n}$ and let*

$$p_A(z) = det(A - zI) = (\lambda_1 - z)^{m_1} \ldots (\lambda_s - z)^{m_s}$$

denote the characteristic polynomial of A with distinct eigenvalues $\lambda_1, \ldots, \lambda_s$. The generalized eigenspace of A to the eigenvalue λ_j is

$$gE_{\lambda_j} = N((A - \lambda_j I)^{m_j}) \ .$$

Its dimension is m_j, the algebraic multiplicity of λ_j. If $T^{-1}AT$ has the block form described in Theorem 10.14, then the first m_1 columns of T form a basis of gE_{λ_1}, the next m_2 columns of T form a basis of gE_{λ_2} etc.

One calls the elements of

$$gE_{\lambda_j} \setminus \{0\}$$

generalized eigenvectors of A to the eigenvalue λ_j. The previous theorem implies that $\mathbb{C}^n$ always has a basis consisting of generalized eigenvectors of A.

In the Jordan form theorem, one constructs particular bases in the generalized eigenspaces.

10.12 The Direct Sum of the Generalized Eigenspaces

Let W denote a vector space and let $U_1, U_2, \ldots, U_s$ denote subspaces of W. By definition, the set

$$U := U_1 + U_2 + \ldots + U_s \tag{10.9}$$

consists of all vectors of the form

$$u_1 + u_2 + \ldots + u_s \quad \text{where} \quad u_j \in U_j \quad \text{for} \quad j = 1, 2, \ldots, s \ .$$

It is not difficult to show that the set $U = U_1 + U_2 + \ldots + U_s$ defined in this way is a subspace of W.

Assume that

$$W = U_1 + U_2 + \ldots + U_s \ .$$

One says that W is the *direct sum* of the subspaces $U_1, U_2, \ldots, U_s$ if for every $w \in W$ for $j = 1, 2, \ldots, s$ there exist **unique** vectors $u_j \in U_j$ with

$$w = u_1 + u_2 + \ldots + u_s \; .$$

If W is the direct sum of $U_1, U_2, \ldots, U_s$ then one writes

$$W = U_1 \oplus U_2 \oplus \ldots \oplus U_s \; .$$

We will show:

Theorem 10.16 *Let $\lambda_1, \ldots, \lambda_s$ denote the distinct eigenvalues of the matrix $A \in \mathbb{C}^{n \times n}$ and let gE_{λ_j} denote the generalized eigenspace to the eigenvalue λ_j. Then we have*

$$\mathbb{C}^n = gE_{\lambda_1} \oplus \ldots \oplus gE_{\lambda_s} \; ,$$

i.e., the vector space $\mathbb{C}^n$ is the direct sum of the generalized eigenspaces of A.

Proof: Let $w \in \mathbb{C}^n$ be given. We have to show that for $j = 1, \ldots, s$ there exists a unique $u_j \in gE_{\lambda_j}$ so that

$$w = \sum_{j=1}^{s} u_j \; .$$

Existence of u_j: Let $T^{-1}AT$ denote a transformation of A to block-diagonal form. See Theorem 10.14. Let

$$T = (T^{(1)} \ldots T^{(s)})$$

where the columns of $T^{(j)}$ form a basis of gE_{λ_j}. See Theorem 10.15. Set $x := T^{-1}w$, thus $w = Tx$. Write

$$x = \begin{pmatrix} x^{(1)} \\ \vdots \\ x^{(s)} \end{pmatrix} \in \mathbb{C}^n$$

where $x^{(1)}$ contains the first m_1 components of x, etc. We have

$$w = Tx = \sum_{j=1}^{s} T^{(j)} x^{(j)} = \sum_{j=1}^{s} u_j$$

where $u_j = T^{(j)} x^{(j)} \in gE_{\lambda_j}$.

Uniqueness of u_j: Assume that $w = \sum_{j=1}^{s} v_j$ where $v_j \in gE_{\lambda_j}$. We have $v_j = T^{(j)} y^{(j)}$ for a vector $y^{(j)} \in \mathbb{C}^{m_j}$. Set

$$y = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(s)} \end{pmatrix} \in \mathbb{C}^n$$

Then we have

$$w = \sum_{j=1}^{s} v_j = \sum_{j=1}^{s} T^{(j)} y^{(j)} = Ty$$

and the equation $w = Tx$ implies that $y = x$, thus $v_j = u_j$ follows from

$$v_j = T^{(j)} y^{(j)} = T^{(j)} x^{(j)} = u_j \quad \text{for} \quad j = 1, \ldots, s \ .$$

$\diamond$

10.13 Summary

Let $A \in \mathbb{C}^{n \times n}$ have the distinct eigenvalues $\lambda_1, \ldots, \lambda_s$ and the characteristic polynomial

$$p_A(z) = det(A - zI) = (\lambda_1 - z)^{m_1} \ldots (\lambda_s - z)^{m_s} \ .$$

Let

$$gE_{\lambda_j} = N((A - \lambda_j I)^{m_j})$$

denote the generalized eigenspace to the eigenvalue λ_j. The algebraic multiplicity of the eigenvalue λ_j is

$$\begin{aligned} m_j &= \text{multiplicity of } \lambda_j \text{ as a zero of } p_A(z) \\ &= dim(gE_{\lambda_j}) \end{aligned}$$

The space $\mathbb{C}^n$ is the direct sum of the generalized eigenspaces of A:

$$\mathbb{C}^n = gE_{\lambda_1} \oplus \ldots \oplus gE_{\lambda_s} \ .$$

Every space gE_{λ_j} is invariant under A:

$$A(gE_{\lambda_j}) \subset gE_{\lambda_j} \ .$$

Denote by

$$B_j = (A - \lambda_j I)\Big|_{gE_{\lambda_j}}$$

the restriction of the operator $A - \lambda_j I$ to the generalized eigenspace gE_{λ_j}. Then the operator

$$B_j : gE_{\lambda_j} \to gE_{\lambda_j}$$

is nilpotent, i.e., there is an exponent $r \in \{1, 2, \ldots\}$ with $B_j^r = 0$.

Definition: *The exponent $i = i_j$ with*

$$B_j^{i-1} \neq 0, \quad B_j^i = 0$$

is called the **Riesz index** *of the eigenvalue λ_j of A.*

We will learn in Chapter 12 that the Riesz index $i = i_j$ equals the dimension of the largest Jordan block to the eigenvalue λ_j.

Example: Let

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

Then we have

$$A^2 \neq 0, \quad A^3 = 0 .$$

The Riesz index of the eigenvalue $\lambda_1 = 0$ is $i = 3$.

Theorem 10.17 *Let $A \in \mathbb{C}^{n\times n}$ denote a matrix with eigenvalues λ_j as above. Let*

$$T = (t^1, t^2, \ldots, t^n) \in \mathbb{C}^{n\times n} .$$

The following two conditions are equivalent:

1. The matrix T is nonsingular and $T^{-1}AT$ has block–diagonal form

$$T^{-1}AT = \begin{pmatrix} M_1 & & & 0 \\ & M_2 & & \\ & & \ddots & \\ 0 & & & M_s \end{pmatrix}$$

where the block M_j has dimensions $m_j \times m_j$ and has λ_j as its only eigenvalue.

2. The first m_1 columns of T form a basis of gE_{λ_1}, the next m_2 columns of T form a basis of gE_{λ_2} etc.

Using Schur's Theorem and the Blocking Lemma we have obtained existence of a transformation matrix T satisfying the conditions of the theorem.

10.14 The Cayley–Hamilton Theorem

If

$$p(z) = \sum_{j=0}^{k} a_j z^j$$

is a polynomial with $a_j \in \mathbb{C}$ and if $A \in \mathbb{C}^{n\times n}$ then one defines the matrix

$$p(A) = \sum_{j=0}^{k} a_j A^j$$

where $A^0 = I$.

The Cayley–Hamilton Theorem says that every matrix A satisfies its own characteristic equation, i.e., $p_A(A) = 0$ where $p_A(z) = det(A - zI)$ denotes the characteristic polynomial of A. We will use Theorem 10.14 to prove this.

The following result is also used:

Lemma 10.12 *Let $p(z)$ and $q(z)$ be polynomials with product $r(z) = p(z)q(z)$. Then, for every $n \times n$ matrix A,*

$$r(A) = p(A)q(A) \ .$$

Theorem 10.18 *(Cayley–Hamilton) Let*

$$p_A(z) = det(A - zI) = (\lambda_1 - z)^{m_1} \dots (\lambda_s - z)^{m_s}$$

denote the characteristic polynomial of A. Then

$$p_A(A) = 0 \ .$$

Proof: Let $T^{-1}AT = M$ denote the block matrix of Theorem 10.14. The matrix

$$R_j = M_j - \lambda_j I_{m_j} \in \mathbb{C}^{m_j \times m_j}$$

is strictly upper triangular. Therefore,

$$(\lambda_j I_{m_j} - M_j)^{m_j} = 0, \quad j = 1, \dots, s \ .$$

Obtain:

$$(\lambda_1 I - M)^{m_1} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & X_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & X_s \end{pmatrix}, \quad (\lambda_2 I - M)^{m_2} = \begin{pmatrix} Y_1 & 0 & \dots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & Y_3 & 0 \\ 0 & \dots & 0 & Y_s \end{pmatrix},$$

etc. Taking the product of these matrices, one obtains that

$$p_A(M) = (\lambda_1 I - M)^{m_1} \dots (\lambda_s I - M)^{m_s} = 0 \ .$$

Finally, $A = TMT^{-1}$, thus

$$p_A(A) = T p_A(M) T^{-1} = 0 \ .$$

$\diamond$

11 Similarity Transformations and Systems of ODEs

Our goal is to explain how similarity transformations, $T^{-1}AT = B$, can be used to simplify a given system of ODEs, $u' = Au + f(t)$.

Consider an initial value problem

$$u'(t) = Au(t) + f(t), \quad u(0) = u^{(0)} . \tag{11.1}$$

Here $u(t), f(t), u^{(0)} \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$. The unknown vector function $u(t)$ is often called the state variable. The forcing function $f(t)$ and the initial value $u^{(0)}$ are given and the system matrix A is assumed to be constant. If $f : [0,\infty) \to \mathbb{C}^n$ is a continuous function, then the initial value problem has a unique solution $u \in C^1\Big([0,\infty), \mathbb{C}^n\Big)$. In other words, the system (11.1) determines how the state variable $u(t)$ evolves in the state space $\mathbb{C}^n$.

11.1 The Scalar Case

The simplest case occurs for $n = 1$ and $f \equiv 0$. The initial–value problem becomes

$$u'(t) = \lambda u(t), \quad u(0) = u^{(0)} ,$$

with solution

$$u(t) = e^{\lambda t} u^{(0)} .$$

If $\lambda = \alpha + i\beta$ with real α, β, then

$$u(t) = e^{\alpha t}\Big(\cos(\beta t) + i \sin(\beta t)\Big) u^{(0)} .$$

The solution grows in magnitude if $\alpha > 0$; it decays if $\alpha < 0$. If $\alpha = 0$ then $|u(t)|$ is constant.

The forced equation

$$u'(t) = \lambda u(t) + f(t), \quad u(0) = u^{(0)} ,$$

has the solution

$$u(t) = e^{\lambda t} u^{(0)} + \int_0^t e^{\lambda(t-s)} f(s)\, ds .$$

Here, for every fixed s, the function

$$q(t) = e^{\lambda(t-s)} f(s)$$

satisfies

$$q'(t) = \lambda q(t), \quad q(s) = f(s) .$$

Thus, also for the forced problem, the sign of $\operatorname{Re}\lambda$ is important.

11.2 Introduction of New Variables

Consider the system (11.1). To better understand the system, one often introduces new variables $v(t)$ by

$$u(t) = Tv(t)$$

where $T \in \mathbb{C}^{n\times n}$ is a nonsingular transformation matrix that must be determined. One obtains

$$Tv'(t) = ATv(t) + f(t)$$

or

$$v'(t) = T^{-1}ATv(t) + T^{-1}f(t) \ .$$

Thus one obtains the transformed initial value problem

$$v'(t) = Bv(t) + g(t), \quad v(0) = v^{(0)} \ , \tag{11.2}$$

with

$$B = T^{-1}AT, \quad g(t) = T^{-1}f(t), \quad v^{(0)} = T^{-1}u^{(0)} \ .$$

The aim of the transformation is to obtain a new system (11.2) that is easier to understand than the given system (11.2).

If $B = T^{-1}AT = \Lambda$ is diagonal, then the system (11.2) decouples into n scalar equations that we know how to solve. For each component $v_j(t)$ of $v(t)$ we have

$$v_j(t) = e^{\lambda_j t}v_j^{(0)} + \int_0^t e^{\lambda_j(t-s)}\, g_j(s)\, ds \ .$$

Then the transformation $u(t) = Tv(t)$ gives us the solution of the original system (11.1).

Warning: It is possible that the condition number of the transformation,

$$\kappa(T) = |T||T^{-1}| \ ,$$

is very large. This should be avoided because, otherwise, the relations

$$u(t) = Tv(t), \quad g(t) = T^{-1}f(t) \ ,$$

may be very distorting.

A rule of thumb: The simpler the matrix $B = T^{-1}AT$, the larger the condition number of T. Thus, in applications, one has to find the right compromise between the simplicity of B and an acceptable size of the condition number $\kappa(T) = |T||T^{-1}|$.

Upper Triangular Form: It is always possible to transform to upper triangular form using a unitary transformation, $U^*AU = \Lambda + R$. The upper triangular system becomes

$$v'(t) = (\Lambda + R)v(t) + g(t)$$

where Λ is diagonal and R is strictly upper triangular. One obtains $v_n' = \lambda_n v_n + g_n$, which is a scalar equation for $v_n(t)$. Once $v_n(t)$ is known, one obtains a scalar equation for $v_{n-1}(t)$, etc. Thus, in principle, one has to solve only forced scalar equations. However, the equations are not decoupled; $v_n(t)$ influences all other variables, $v_{n-1}(t)$ influences the variables $v_j(t)$ for $1 \le j \le n-2$, etc.

Separation of Modes: Sometimes one wishes to separate growing and decaying modes. Suppose Schur's transformation leads to a blocked system (assuming $f \equiv 0$):

$$u'(t) = \begin{pmatrix} M_1 & M_{12} \\ 0 & M_2 \end{pmatrix} u(t) \ .$$

Suppose that all eigenvalues λ_j of M_1 satisfy

$$\mathrm{Re}\,\lambda_j \le -\delta < 0$$

and that all eigenvalues λ_j of M_2 satisfy

$$\mathrm{Re}\,\lambda_j \ge \delta > 0 \ .$$

One can eliminate the coupling through M_{12}. There is a transformation

$$u(t) = Tv(t), \quad T = \begin{pmatrix} I & S \\ 0 & I \end{pmatrix} ,$$

so that

$$T^{-1} \begin{pmatrix} M_1 & M_{12} \\ 0 & M_2 \end{pmatrix} T = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} .$$

In the v–variables, the system becomes

$$v'(t) = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} v(t) \ .$$

If one partitions

$$v = \begin{pmatrix} v^I \\ v^{II} \end{pmatrix}$$

then one obtains the two decoupled systems

$$\frac{dv^I}{dt} = M_1 v^I, \quad \frac{dv^{II}}{dt} = M_2 v^{II} \ .$$

Here v^I contains the decaying modes and v^{II} contains the growing modes.

12 The Jordan Normal Form

Every matrix $A \in \mathbb{C}^{n\times n}$ can be transformed by a similarity transformation to Jordan normal form, $T^{-1}AT = \Lambda + J$. Here Λ is diagonal and J is a nilpotent Jordan matrix. Existence of the transformation can be shown in three steps. Step 1 is Schur's transformation to upper triangular form, $U^*AU = R$; see Theorem 10.2. Step 2 is repeated blocking, based on the Blocking Lemma 10.10; see Theorem 10.14. In Step 3 it suffices to study the transformation of a single block, $M = \lambda I + R$, where R is strictly upper triangular, thus nilpotent. In this chapter we carry out the details of Step 3.

The Jordan normal form is named after Camille Jordan, 1838-1922, a French mathematician.

12.1 Preliminaries and Examples

Definition: *Let U be a vector space and let $L : U \to U$ be a linear operator. Then L is called nilpotent if $L^j = 0$ for some $j \in \{1, 2, \ldots\}$.*

Example: Let $R \in \mathbb{C}^{n\times n}$ be strictly upper triangular. Then $R^n = 0$. Thus R (or, more precisely, the linear operator defined by R) is nilpotent.

Lemma 12.1 *Let $L : U \to U$ be linear and nilpotent. If λ is an eigenvalue of L, then $\lambda = 0$.*

Proof: Let $Lx = \lambda x, x \neq 0$. One obtains that

$$L^2x = \lambda Lx = \lambda^2 x$$

etc. Therefore, if $L^j = 0$, then

$$0 = L^j x = \lambda^j x \ ,$$

thus $\lambda = 0$. $\diamond$

Lemma 12.2 *Let $A \in \mathbb{C}^{n\times n}$ be nilpotent. Then $A^n = 0$.*

Proof: We transform A to upper triangular form: $U^*AU = R$. Since all eigenvalues of A are zero, the matrix R is *strictly* upper triangular. Therefore, $R^n = 0$ and $A^n = UR^nU^* = 0$. $\diamond$

We will describe below the Jordan form of a general nilpotent matrix. Let us first consider the cases $n = 2$ and $n = 3$.

Example: $n = 2$. Let $A \in \mathbb{C}^{2\times 2}$ be nilpotent, thus $A^2 = 0$. There are two cases:

Case 1: $A = 0$. In this case the Jordan normal form of A is $J = 0$.

Case 2: $A \neq 0$, but $A^2 = 0$. We claim that A is similar to

$$J = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \ .$$

Let $T = (t^1, t^2)$ be a nonsingular matrix. We want to determine T so that

$$AT = TJ \; .$$

Since

$$AT = (At^1, At^2) \quad \text{and} \quad TJ = (t^1, t^2) \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} (0, t^1)$$

this requires that

$$At^1 = 0 \quad \text{and} \quad At^2 = t^1 \; .$$

By assumption, $A \neq 0$. Choose any t^2 so that

$$t^1 := At^2 \neq 0 \; .$$

Then $At^1 = A^2 t^2 = 0$. We must show that the two vectors $t^1 = At^2$ and t^2 are linearly independent. Let

$$\alpha A t^2 + \beta t^2 = 0 \; .$$

Applying A, we see that $\beta = 0$. Then $\alpha = 0$ follows. This shows that every matrix $A \in \mathbb{C}^{2\times 2}$ with

$$A \neq 0, \quad A^2 = 0 \; ,$$

is similar to

$$J = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \; .$$

Example: $n = 3$. Let $A \in \mathbb{C}^{3\times 3}$ be nilpotent, thus $A^3 = 0$. There are three cases:

Case 1: $A = 0$**.** In this case the Jordan normal form of A is $J = 0$.

Case 2: $A \neq 0$**, but** $A^2 = 0$**.** We claim that A is similar to

$$J_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \; .$$

Case 3: $A \neq 0$**,** $A^2 \neq 0$**, but** $A^3 = 0$**.** We claim that A is similar to

$$J_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \; .$$

Case 1 is obvious. Let us consider **Case 3** first. We want to determine a nonsingular matrix $T = (t^1, t^2, t^3)$ with

$$AT = TJ_3$$

where

$$TJ_3 = (t^1, t^2, t^3) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = (0, t^1, t^2)$$

The equation $AT = TJ_3$ requires that

$$(At^1, At^2, At^3) = (0, t^1, t^2) \ .$$

In other words, we want to have

$$At^1 = 0, \quad At^2 = t^1, \quad At^3 = t^2 \ .$$

Clearly, if t^3 is chosen, then t^2 and t^1 are determined. Since $A^2 \neq 0$ we can choose t^3 so that $A^2 t^3 \neq 0$. Then define

$$t^2 := At^3, \quad t^1 := At^2 \ .$$

Since $A^3 = 0$ we have

$$At^1 = A^2 t^2 = A^3 t^3 = 0 \ .$$

It remains to show that the three vectors

$$t^1 = A^2 t^3, \quad t^2 = At^3, \quad t^3$$

are linearly independent. Let

$$\alpha A^2 t^3 + \beta A t^3 + \gamma t^3 = 0 \ .$$

Apply first A^2 to obtain that $\gamma = 0$. Then apply A to obtain $\beta = 0$, etc.

Next consider **Case 2** where $A \neq 0$, but $A^2 = 0$. This case is more complicated than Case 3.

The condition

$$AT = TJ_2$$

where

$$TJ_2 = (t^1, t^2, t^3) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = (0, t^1, 0)$$

requires that

$$(At^1, At^2, At^3) = (0, t^1, 0) \ ,$$

thus that equation $AT = TJ_2$ requires that

$$At^1 = At^3 = 0, \quad At^2 = t^1 \ .$$

We must determine two vectors, t^1 and t^3, in $N(A)$ so that the system

$$At^2 = t^1$$

is solvable for t^2 and so that the three vectors t^1, t^2, t^3 are linearly independent. The vector t^1 must lie in

$$N(A) \cap R(A) \ .$$

Then t^2 must be chosen with $At^2 = t^1$ and the vector t^3 must be chosen in $N(A)$ so that the three vectors t^1, t^2, t^3 are linearly independent. It is not obvious that this can always be done.

To show that the vectors t^1, t^2, t^3 can always be constructed, we first apply a Schur transformation (see Theorem 10.2) to A. We may then assume that

$$A = \begin{pmatrix} 0 & a & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix} \ .$$

It is easy to check that

$$A^2 = \begin{pmatrix} 0 & 0 & ac \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \ .$$

The assumption $A^2 = 0$ yields that $a = 0$ or $c = 0$.

Case 2a: Let $a \neq 0$, but $c = 0$. Thus,

$$A = \begin{pmatrix} 0 & a & b \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \ .$$

We see that

$$N(A) = span\{e^1, \begin{pmatrix} 0 \\ b \\ -a \end{pmatrix}\} \ .$$

Also,

$$R(A) = span\{e^1\} \ .$$

Therefore, we can choose

$$t^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad t^3 = \begin{pmatrix} 0 \\ b \\ -a \end{pmatrix} \ .$$

The system

$$At^2 = t^1$$

is solved by

$$t^2 = \begin{pmatrix} 0 \\ 1/a \\ 0 \end{pmatrix} \ .$$

This leads to the nonsingular transformation matrix

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/a & b \\ 0 & 0 & -a \end{pmatrix} .$$

One easily checks that

$$AT = TJ_2 .$$

Case 2b: Let $a = c = 0$, but $b \neq 0$. We have

$$A = \begin{pmatrix} 0 & 0 & b \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} .$$

We see that

$$N(A) = span\{e^1, e^2\}, \quad R(A) = span\{e^1\} .$$

We can take

$$t^1 = e^1, \quad t^3 = e^2, \quad t^2 = \begin{pmatrix} 0 \\ 0 \\ 1/b \end{pmatrix} .$$

The transformation matrix is

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1/b & 0 \end{pmatrix} .$$

One easily checks that $AT = TJ_2$.

Case 2c: Let $a = 0$, but $c \neq 0$. We have

$$A = \begin{pmatrix} 0 & 0 & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{pmatrix}, \quad c \neq 0 .$$

We see that

$$N(A) = span\{e^1, e^2\}, \quad R(A) = span\{\begin{pmatrix} b \\ c \\ 0 \end{pmatrix} .$$

Therefore, we can choose

$$t^1 = \begin{pmatrix} b \\ c \\ 0 \end{pmatrix}, \quad t^3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad t^2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} .$$

The transformation matrix is

$$T = \begin{pmatrix} b & 0 & 1 \\ c & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} .$$

One easily checks that $AT = TJ_2$.

Remark: In the case $n = 3, A \neq 0, A^2 = 0$, our considerations have shown that

$$dim\, N(A) = 2, \quad dim\, R(A) = 1 \ ,$$

and

$$R(A) \subset N(A) \ .$$

The vector t^1 must be chosen in $R(A) \cap N(A)$. We have seen that this choice is possible. Also, we can find t^2 with $At^2 = t^1$ and $t^3 \in N(A)$ so that the three vectors t^1, t^2, t^3 are linearly independent. At this point, it is not obvious that a similarity transformation to a normal form J can always be made for any nilpotent matrix $A \in \mathbb{C}^{n \times n}$ if $n > 3$.

12.2 The Rank of a Matrix Product

Let $A \in \mathbb{C}^{m \times n}, B \in \mathbb{C}^{n \times p}$, thus $AB \in \mathbb{C}^{m \times p}$. The matrices A and B determine linear maps between the vector spaces $\mathbb{C}^p, \mathbb{C}^n$, and $\mathbb{C}^m$:

$$\mathbb{C}^m \xleftarrow{A} \mathbb{C}^n \xleftarrow{B} \mathbb{C}^p \ .$$

The matrix product AB determines a map from $\mathbb{C}^p$ to $\mathbb{C}^m$. Since

$$rank(AB) = dim\, R(AB) \quad \text{and} \quad rank(B) = dim\, R(B)$$

and since $dim\, R(AB)$ cannot be larger than $dim R(B)$ it is clear that

$$rank(AB) \leq rank(B) \ .$$

The following theorem gives a more precise formula for the rank of a matrix product. Note that the space $N(A) \cap R(B)$ is a subspace of $\mathbb{C}^n$.

Theorem 12.1 *Let $A \in \mathbb{C}^{m \times n}, B \in \mathbb{C}^{n \times p}$, thus $AB \in \mathbb{C}^{m \times p}$. For the rank of $AB \in \mathbb{C}^{m \times p}$ we have*

$$rank(AB) = rank(B) - dim\Big(N(A) \cap R(B)\Big) \ .$$

Proof: Let $s := dim\Big(N(A) \cap R(B)\Big)$ and $s + t := dim(R(B))$. Let

$$x^1, \ldots, x^s$$

be a basis of $N(A) \cap R(B)$ and let

$$x^1, \ldots, x^s, z^1, \ldots, z^t$$

be a basis of $R(B)$. We claim that the t vectors

$$Az^1, \ldots, Az^t$$

form a basis of $R(AB)$. If this is shown then

$$\begin{aligned} rank(AB) &= dim(R(AB)) \\ &= t \\ &= (s+t) - s \\ &= rank(B) - dim\Big(N(A) \cap R(B)\Big) \ , \end{aligned}$$

and the theorem is proved.

We continue to prove that the t vectors $Az^1, \ldots, Az^t$ form a basis of $R(AB)$. First, since $z^j \in R(B)$ we can write $z^j = B\xi^j$, thus $Az^j = AB\xi^j \in R(AB)$. Thus we have shown that

$$span\{Az^1, \ldots, Az^t\} \subset R(AB) \ .$$

Second, let $b \in R(AB)$ be arbitrary. We can write $b = AB\xi$. Here $B\xi \in R(B)$, thus

$$B\xi = a_1 x^1 + \ldots + a_s x^s + b_1 z^1 + \ldots + b_t z^t \ .$$

Since $x^j \in N(A)$ we obtain

$$b = AB\xi = b_1 Az^1 + \ldots + b_t Az^t \ .$$

So far, we have shown that

$$span\{Az^1, \ldots, Az^t\} = R(AB) \ .$$

Third, it remains to prove that the vectors $Az^1, \ldots, Az^t$ are linearly independent. Let

$$b_1 Az^1 + \ldots + b_t Az^t = 0 \ .$$

This implies that

$$A(b_1 z^1 + \ldots + b_t z^t) = 0 \ .$$

thus $b_1 z^1 + \ldots + b_t z^t \in N(A)$. Since $z^j \in R(B)$ we have shown that

$$b_1 z^1 + \ldots + b_t z^t \in N(A) \cap R(B)$$

and can write

$$b_1 z^1 + \ldots + b_t z^t = a_1 x^1 + \ldots + a_s x^s \ .$$

Since the $s+t$ vectors

$$x^1, \ldots, x^s, z^1, \ldots, z^t$$

are linearly independent, we conclude that all coefficients a_i, b_j are zero, proving the linear independence of $Az^1, \ldots, Az^t$. $\diamond$

12.3 Nilpotent Jordan Matrices

The following matrices J_k of size $k \times k$ are called elementary Jordan blocks:

$$J_1 = (0), \quad J_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad J_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} .$$

In general, let

$$J_k = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \\ \vdots & & & 0 & 1 \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix} \in \mathbb{R}^{k\times k} .$$

It is easy to check that

$$rank(J_k^j) = k - j \quad \text{for} \quad 0 \le j \le k \quad \text{and} \quad J_k^j)0 \quad \text{for} \quad j \ge k . \tag{12.1}$$

Any block matrix of the form

$$J = \begin{pmatrix} J_{k_1} & 0 & \dots & 0 \\ 0 & J_{k_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & J_{k_q} \end{pmatrix} \tag{12.2}$$

is called a nilpotent Jordan matrix. The numbers $k_1, \dots, k_q$ are called the block sizes of the nilpotent Jordan matrix J.

Assume that J is a nilpotent Jordan matrix (12.2) of size $n \times n$. We will show how the ranks of the powers of J,

$$r_j := rank(J^j), \quad j = 0, 1, \dots$$

determine the number q of blocks of J and the block sizes, $k_1, \dots, k_q$.

Let

$$m := \max\{k_1, \dots, k_q\}$$

denote the maximal block size and let

$$b_j := \text{number of blocks of size } j \quad \text{for} \quad j = 1, \dots, m .$$

We have

$$q = b_1 + b_2 + \dots + b_m$$

since the number of all blocks is q.

We want to discuss how the ranks of the powers J^j ($j = 0, \dots, m+1$) and the number of block sizes b_j ($j = 1, \dots, m$) are related to each other. We have

$$r_0 = rank(J^0) = rank(I) = n, \quad r_m = r_{m+1} = 0 \ .$$

Furthermore,

$$r_1 = rank(J) = n - q = r_0 - (b_1 + b_2 + \ldots + b_m) \ .$$

This holds since $rank(J_k) = k - 1$. Also,

$$\begin{aligned} r_2 &= rank(J^2) \\ &= n - 2q + b_1 \\ &= r_1 - (b_2 + \ldots + b_m) \end{aligned}$$

Here $b_2 + \ldots + b_m$ is the number of blocks of size ≥ 2. In general,

$$r_j = r_{j-1} - (b_j + \ldots + b_m) \quad \text{for} \quad j = 1, \ldots, m \ . \tag{12.3}$$

The reason is the following: If one compares J^j with J^{j-1} then one observes a rank drop of one for each block $J_{k_\nu}^{j-1}$ that is nonzero in J^{j-1}. In other words, if one goes from J^{j-1} to J^j, then the rank drops by the number of blocks of size $\geq j$, i.e., by $b_j + \ldots + b_m$.

Using (12.3) for $j+1$ instead of j we have

$$r_{j+1} = r_j - (b_{j+1} + \ldots + b_m) \ . \tag{12.4}$$

Therefore,

$$\begin{aligned} b_j + b_{j+1} + \ldots b_m &= r_{j-1} - r_j \\ b_{j+1} + \ldots b_m &= r_j - r_{j+1} \end{aligned}$$

Subtraction yields that

$$b_j = r_{j-1} - 2r_j + r_{j+1} \ .$$

We summarize:

Lemma 12.3 *Let J denote the nilpotent Jordan matrix (12.2). Let $r_j = rank(J^j)$ and let b_j denote the number of elementary Jordan blocks of size j in J. Then we have*

$$b_j = r_{j-1} - 2r_j + r_{j+1}, \quad j = 1, 2, \ldots, m \ .$$

Here m is the maximal block size of the blocks J_k in J, thus $A^m = 0$.

12.4 The Jordan Form of a Nilpotent Matrix

The main theorem is the following:

Theorem 12.2 *Let $A \in \mathbb{C}^{n\times n}$ be an arbitrary nilpotent matrix. There exists a nonsingular matrix $T \in \mathbb{C}^{n\times n}$ and a nilpotent Jordan matrix $J \in \mathbb{C}^{n\times n}$ of the form (12.2) so that $T^{-1}AT = J$. The block sizes k_ν of J are uniquely determined by A. That is, the nilpotent Jordan matrix J is uniquely determined except for the order of the blocks, which is arbitrary.*

The theorem is proved in several steps.

The uniqueness statement follows from Lemma 12.3: If $T^{-1}AT = J$ then

$$rank(A^j) = rank(J^j)$$

for every $j = 0, 1, \ldots$ These rank numbers determine the block sizes.

To motivate the construction of T, assume first that

$$T^{-1}AT = J = \begin{pmatrix} J_k & 0 \\ 0 & J_l \end{pmatrix} , \tag{12.5}$$

i.e., J consists of two blocks of sizes k and l, respectively. Let

$$T = \Big(x^k, x^{k-1}, \ldots, x^1, y^l, y^{l-1}, \ldots, y^1\Big) .$$

Then the equation $AT = TJ$ reads

$$\Big(Ax^k, Ax^{k-1}, \ldots, Ax^1, Ay^l, Ay^{l-1}, \ldots, Ay^1\Big) = \Big(0, x^k, \ldots, x^2, 0, y^l, \ldots, y^2\Big) .$$

Thus, the equation $AT = TJ$ holds if and only if

$$Ax^k = 0, \quad Ax^{k-1} = x^k, \ldots, Ax^1 = x^2$$

and

$$Ay^l = 0, \quad Ay^{l-1} = y^l, \ldots, Ay^1 = y^2 .$$

Thus, the string of vectors

$$x^k, x^{k-1}, \ldots, x^2, x^1$$

(these are the first k column vectors of T) has the form

$$A^{k-1}x^1, A^{k-2}x^1, \ldots, Ax^1, x^1$$

with $A^k x^1 = 0$ if (12.5) holds. A similar form holds for the y–string.

Definition: *A string of vectors*

$$A^{k-1}x, A^{k-2}x, \ldots, Ax, x$$

is called a Jordan string (or a Jordan chain) of length k for the matrix A (to the eigenvalue zero) if

$$A^{k-1}x \neq 0, \quad A^k x = 0 \ .$$

We see that $T^{-1}AT$ has the form (12.5) if and only if the first k columns of T form a Jordan string of length k and the last l columns of T form a Jordan string of length l. In addition, to make T nonsingular, all the vectors in both strings must be linearly independent.

Lemma 12.4 *Let*

$$A^{k-1}x, A^{k-2}x, \ldots, Ax, x$$

denote a Jordan string for A. These vectors are linearly independent.

Proof: Assume that

$$a_{k-1}A^{k-1}x + \ldots + a_1 Ax + a_0 x = 0 \ . \tag{12.6}$$

Apply A^{k-1} to the equation and note that $A^k x = 0$. One obtains that $a_0 A^{k-1}x = 0$. Since $A^{k-1}x \neq 0$ we conclude that $a_0 = 0$. Then apply A^{k-2} to (12.6) to obtain $a_1 = 0$, etc. $\diamond$

Lemma 12.5 *Let*

$$A^{k-1}x, A^{k-2}x, \ldots, Ax, x$$

and

$$A^{l-1}y, A^{l-2}y, \ldots, Ay, y$$

denote two Jordan strings for A. If the two vectors at the beginning of the strings,

$$A^{k-1}x \quad \textit{and} \quad A^{l-1}y \ ,$$

are linearly independent, then the $k+l$ vectors in both strings are linearly independent.

Proof: Assume that

$$a_{k-1}A^{k-1}x + \ldots + a_1 Ax + a_0 x + b_{l-1}A^{l-1}y + \ldots + b_1 Ay + b_0 y = 0 \ . \tag{12.7}$$

First assume that $k = l$. Apply A^{k-1} to the equation (12.7). Note that $A^k x = A^l y = 0$. One obtains that

$$a_0 A^{k-1}x + b_0 A^{l-1}y = 0$$

and $a_0 = b_0 = 0$ follows. Applying A^{k-2} to (12.7) one obtains $a_1 = b_1 = 0$ etc.

Second, let $k > l$. Apply A^{k-1} to the equation. Note that $A^k x = 0$ and $A^{k-1}y = 0$ since $k > l$. One obtains $a_0 A^{k-1}x = 0$. Since $A^{k-1}x \neq 0$ we conclude that $a_0 = 0$. Then apply A^{k-2} to (12.7), etc. $\diamond$

It is not difficult to generalize the lemma and its proof to any finite number of Jordan strings:

Lemma 12.6 *Consider a set of q Jordan strings for A:*

$$A^{k_\alpha-1}x^\alpha, A^{k_\alpha-2}x^\alpha, \ldots, Ax^\alpha, x^\alpha, \quad \alpha = 1, \ldots, q \ .$$

Assume that the q vectors

$$z^\alpha := A^{k_\alpha-1}x^\alpha, \quad \alpha = 1, \ldots, q \ ,$$

at the beginning of the strings are linearly independent. Then all the vectors in the q strings are linearly independent.

Note that the vectors $z^\alpha = A^{k_\alpha-1}x^\alpha$ at the beginning of the strings lie in $N(A)$ and in the range spaces $R(A^{k_\alpha-1})$. These vectors must be linearly independent if we want to use the corresponding strings as columns in the transformation matrix T.

Proving Theorem 12.2, then, amounts to showing that $\mathbb{C}^n$ has a basis consisting of Jordan strings of A. We have to understand the intersections of the range spaces $R(A^j)$ with $N(A)$ in order to obtain the vectors $z^\alpha = A^{k_\alpha-1}x^\alpha$ at the beginning of the strings.

A key result, showing that one gets enough vectors to get a basis of $\mathbb{C}^n$, is the following.

Lemma 12.7 *Let $A \in \mathbb{C}^{n\times n}$ be nilpotent. Define the spaces*

$$\mathcal{M}_j = R(A^j) \cap N(A) \quad for \quad j = 0, 1, \ldots, n$$

and let

$$d_j := dim(\mathcal{M}_j), \quad j = 0, 1, \ldots, n \ .$$

Let m be the smallest number with $\mathcal{M}_m = \{0\}$, i.e.

$$d_m = 0 < d_{m-1} \le d_{m-2} \le \ldots \le d_1 \le d_0 = dim(N(A)) \ .$$

Then we have

$$d_0 + d_1 + \ldots + d_{m-1} = n \ .$$

Proof: a) Set $r_j = rank(A^j)$. We write the rank formula of Theorem 12.1 in the form

$$dim\Big(N(A) \cap R(B)\Big) = rank(B) - rank(AB) \ .$$

Applying the formula with $B = A^j$ we obtain that

$$dim\Big(N(A) \cap R(A^j)\Big) = rank(A^j) - rank(A^{j+1})$$

thus

$$d_j = r_j - r_{j+1}, \quad j = 0, \ldots, m-1 \ .$$

Therefore,

$$\begin{aligned} d_0 + d_1 + \ldots + d_{m-1} &= (r_0 - r_1) + (r_1 - r_2) + \ldots + (r_{m-1} - r_m) \\ &= r_0 - r_m \\ &= n - r_m \ . \end{aligned}$$

b) It remains to show that $r_m = 0$, i.e., $A^m = 0$. By the definition of m we have

$$R(A^m) \cap N(A) = \{0\} \ . \tag{12.8}$$

Suppose $A^m \neq 0$. Then $A^m x \neq 0$ for some $x \in \mathbb{C}^n$. There exists $i \geq 0$ so that

$$y := A^{m+i} x \neq 0, \quad Ay = A^{m+i+1} x = 0 \ .$$

Then we have

$$0 \neq y = A^m(A^i x) \in R(A^m) \cap N(A) \ ,$$

contradicting (12.8). $\diamond$

We can now complete the proof of Theorem 12.2 as follows: Choose d_{m-1} linearly independent vectors

$$z^\alpha \in R(A^{m-1}) \cap N(A) = \mathcal{M}_{m-1} \ .$$

These vectors have the form

$$z^\alpha = A^{m-1} x^\alpha$$

and each z^α gives us a Jordan string of length m,

$$z^\alpha = A^{m-1} x^\alpha, \ldots, A x^\alpha, x^\alpha \ . \tag{12.9}$$

In total, this gives us d_{m-1} times m vectors; here we use that there are m vectors in the string (12.9). Then supplement the d_{m-1} basis vectors z^α of $\mathcal{M}_{m-1}$ by $d_{m-2} - d_{m-1}$ vectors z^β to obtain a basis of $\mathcal{M}_{m-2}$. Each vector z^β gives us a Jordan string of length $m-1$. Thus we obtain an additional $d_{m-2} - d_{m-1}$ times $m-1$ vectors, etc.

In total, the number of all the vectors in all the constructed Jordan strings is

$$N = d_{m-1} m + (d_{m-2} - d_{m-1})(m-1) + \ldots + (d_1 - d_2) 2 + (d_0 - d_1) 1 \ .$$

It is easy to see that

$$N = d_{m-1} + d_{m-2} + \ldots + d_1 + d_0 = n \ ,$$

where the last equation has been shown in Lemma 12.7.

Finally, by Lemma 12.6, the total set of constructed vectors is linearly independent. Using the n vectors as columns of T, we have $AT = TJ$, thus $T^{-1}AT = J$. This completes the proof of Theorem 12.2.

Remarks: We have shown that the number q of elementary Jordan blocks in $J = T^{-1}AT$ (see (12.2)) equals $dim(N(A))$. At the beginning of each Jordan string in T is a vector in $N(A)$. Roughly speaking, one chooses these vectors in $N(A)$ as deeply as possible in the iterated range spaces $R(A^j)$.

If m is the maximal block size of all the elementary Jordan blocks in J then $A^m = 0$, but $A^{m-1} \neq 0$. The maximal block size m agrees with the number m introduced in Lemma 12.7.

The number b_m of blocks of size m equals the dimension of $R(A^{m-1}) \cap N(A)$. In general, if b_j is the number of blocks of size j and $r_j = rank(A^j)$, then, using Lemma 12.3,

$$\begin{aligned} b_j &= (r_{j-1} - r_j) - (r_j - r_{j+1}) \\ &= d_{j-1} - d_j \\ &= dim\Big(R(A^{j-1}) \cap N(A)\Big) - dim\Big(R(A^j) \cap N(A)\Big) \ . \end{aligned}$$

This confirms that the number b_j of blocks of size j equals the number of constructed Jordan strings of length j.

12.5 The Jordan Form of a General Matrix

Let $A \in \mathbb{C}^{n \times n}$ be an arbitrary matrix. Its characteristic polynomial has the form

$$p_A(z) = det(A - zI) = (\lambda_1 - z)^{m_1} \cdots (\lambda_s - z)^{m_s}$$

where $\lambda_1, \ldots, \lambda_s$ are the distinct eigenvalues of A and

$$\sum_j m_j = n \ .$$

Using Schur's theorem (Theorem 10.2) and the Blocking Lemma (see Lemma 10.10 and Theorem 10.14), we know that there exists a transformation matrix S so that

$$S^{-1}AS = \begin{pmatrix} M_1 & 0 & \ldots & 0 \\ 0 & M_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & M_s \end{pmatrix}$$

where

$$M_j = \lambda_j I_{m_j} + R_j, \quad R_j \quad \text{is nilpotent}, \quad j = 1, \ldots, s \ .$$

Since $R_j \in \mathbb{C}^{m_j \times m_i}$ is nilpotent, there exists a transformation matrix

$$\Phi_j \in \mathbb{C}^{m_j \times m_j}$$

so that

$$\Phi_j^{-1} R_j \Phi_j = J^{(j)}$$

is a nilpotent Jordan matrix. Note that

$$\Phi_j^{-1} M_j \Phi_j = \lambda_j I_{m_j} + J^{(j)} \; .$$

Let Φ denote the block diagonal matrix

$$\Phi = \begin{pmatrix} \Phi_1 & 0 & \dots & 0 \\ 0 & \Phi_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \Phi_s \end{pmatrix}$$

and let $T = S\Phi$. We then have

$$\begin{aligned} T^{-1}AT &= \Phi^{-1}S^{-1}AS\Phi \\ &= \Phi^{-1}M\Phi \\ &= \begin{pmatrix} \lambda_1 I_{m_1} + J^{(1)} & & 0 \\ & \ddots & \\ 0 & & \lambda_s I_{m_s} + J^{(s)} \end{pmatrix} \; . \end{aligned}$$

This is a transformation of A to Jordan normal form.

Note that

$$T^{-1}AT = \Lambda + J$$

where J is a nilpotent Jordan matrix and Λ is diagonal,

$$\Lambda = \begin{pmatrix} \lambda_1 I_{m_1} & & \\ & \ddots & \\ & & \lambda_s I_{m_s} \end{pmatrix} \; .$$

12.6 Application: The Matrix Exponential

Let $A \in \mathbb{C}^{n \times n}$ and consider the initial value problem

$$u'(t) = Au(t), \quad u(0) = u^{(0)} \; .$$

The solution is

$$u(t) = e^{tA}u^{(0)}$$

where the matrix exponential is defined by the exponential series,

$$e^{tA} = \sum_{j=0}^{\infty} \frac{1}{j!} (tA)^j \ .$$

It is often difficult to understand e^{tA} directly in terms of this series.

If we introduce a new variable $v(t)$ by

$$u(t) = Tv(t)$$

then the initial value problem for $u(t)$ transforms to

$$v'(t) = Bv(t) \quad \text{where} \quad B = T^{-1}AT$$

and

$$v(0) = T^{-1}u(0) = T^{-1}u^{(0)} \ .$$

We then have

$$v(t) = e^{tB}T^{-1}u^{(0)}$$

and

$$u(t) = Te^{tB}T^{-1}u^{(0)} \ .$$

To make e^{tB} as simple as possible, we let B denote the Jordan normal form of A. We then have to understand the exponential of each block

$$\lambda_j I_k + J_k =: \lambda I + J$$

where $\lambda_j \in \sigma(A)$. We have

$$\begin{aligned} e^{t(\lambda I + J)} &= e^{t\lambda I} e^{tJ} \\ &= e^{t\lambda} e^{tJ} \end{aligned}$$

Here, for $J = J_k$,

$$e^{tJ_k} = I_k + \frac{1}{1!}\, tJ_k + \frac{1}{2!}\, t^2 J_k^2 + \ldots + \frac{1}{(k-1)!}\, t^{k-1} J_k^{k-1} \ .$$

One obtains

$$e^{tJ_k} = \begin{pmatrix} 1 & t & t^2/2 & \ldots & t^{k-1}/(k-1)! \\ 0 & 1 & t & \ldots & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & 1 & t \\ 0 & & & 0 & 1 \end{pmatrix} \ .$$

The following Theorem is rather easy to prove. We recall that an eigenvalue λ of A is called semi–simple if the generalized eigenspace gE_λ equals the geometric eigenspace E_λ. This holds if and only if the Jordan matrix J corresponding to λ is the zero matrix.

Theorem 12.3 *Let $A \in \mathbb{C}^{n\times n}$.*
(a) The limit relation

$$e^{tA} \to 0 \quad as \quad t \to \infty$$

holds if and only if

$$\mathrm{Re}\,\lambda < 0 \quad for\ all \quad \lambda \in \sigma(A)\ .$$

(b) There exists a constant $C > 0$ with

$$|e^{tA}| \le C \quad for\ all \quad t \ge 0$$

if and only if for each $\lambda \in \sigma(A)$ we have
Case 1: $\mathrm{Re}\,\lambda < 0$ *;*
or
Case 2: $\mathrm{Re}\,\lambda = 0$ *and λ is semi–simple.*

12.7 Application: Powers of Matrices

Theorem 12.4 *Let $A \in \mathbb{C}^{n\times n}$.*
(a) The limit relation

$$A^j \to 0 \quad as \quad j \to \infty$$

holds if and only if

$$|\lambda| < 1 \quad for\ all \quad \lambda \in \sigma(A)\ .$$

(b) There exists a constant $C > 0$ with

$$|A^j| \le C \quad for\ all \quad j = 1, 2, \ldots$$

if and only if for each $\lambda \in \sigma(A)$ we have
Case 1: $|\lambda| < 1$; or
Case 2: $|\lambda| = 1$ and λ is semi–simple, i.e, the algebraic multiplicity of λ equals its geometric multiplicity.

Proof: Let $T^{-1}AT = B$ denote a transformation of A to Jordan normal form. Clearly, $A^j \to 0$ as $j \to \infty$ holds if and only if $B^j \to 0$ as $j \to \infty$. Similarly, $|A^j|$ is a bounded sequence if and only if $|B^j|$ is bounded.

Consider a Jordan block

$$Q = \lambda I + J, \quad J = J_k = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \\ \vdots & & & 0 & 1 \\ 0 & \ldots & \ldots & \ldots & 0 \end{pmatrix} \in \mathbb{R}^{k\times k}\ .$$

First let $|\lambda| < 1$. Let

$$T_\varepsilon = diag(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{k-1}) \ .$$

We have

$$T_\varepsilon^{-1} Q T_\varepsilon = \lambda I + \varepsilon J \ .$$

Therefore, if $\varepsilon > 0$ is small enough, then $|\lambda I + \varepsilon J| < 1$. It follows that

$$Q^j \to 0 \quad \text{as} \quad j \to \infty$$

if $|\lambda| < 1$. This proves (a). Also, if Case 1 or Case 2 holds for every eigenvalue λ of A, then the boundedness of $|A^j|$ follows.

Assume $|\lambda| > 1$ for some eigenvalue λ of A. We have

$$|(\lambda I + J)^j| \geq |\lambda|^j \to \infty \quad \text{as} \quad j \to \infty \ .$$

Now let $|\lambda| = 1$ and assume that λ is not semi–simple, thus $J = J_k$ with $k \geq 2$. We have

$$\begin{aligned} (\lambda I + J)^j &= \sum_{l=0}^{k-1} \begin{pmatrix} j \\ l \end{pmatrix} \lambda^{j-l} J^l \\ &= \lambda^j I + j\lambda^{j-1} J + \ldots \end{aligned}$$

The vector

$$(\lambda I + J)^j e^2$$

has the entry $j\lambda^{j-1}$ in component 1. Therefore,

$$|(\lambda I + J)^j| \geq j \ .$$

It follows that $|A^j| \to \infty$ as $j \to \infty$ if A has an eigenvalue λ with $|\lambda| = 1$ which is not semi–simple.

13 Complementary Subspaces and Projectors

Let W be a vector space and let $L : W \to W$ be a linear operator. A subspace U of W is called invariant under L if $L(U) \subset U$. To analyze an operator $L : W \to W$ one can try to write the space W as a sum of subspaces which are invariant under L. One can then investigate L separately on each invariant subspace: *divide and conquer*.

We will explain how a direct sum decomposition $W = U \oplus V$ of a space W is related to projectors. An important result of this chapter is the *Spectral Representation Theorem*: Any diagonalizable matrix A can be written in the form

$$A = \lambda_1 P_1 + \cdots + \lambda_s P_s \ ,$$

where $\lambda_1, \ldots, \lambda_s$ are the distinct eigenvalues of A and P_j is the projector onto the eigenspace $N(A - \lambda_j I)$ along the range space $R(A - \lambda_j I)$. The spaces $N(A - \lambda_j I)$ and $R(A - \lambda_j I)$ are complementary subspaces of $\mathbb{C}^n$ which are both invariant under A.

13.1 Complementary Subspaces

Definition: *Let U and V be two subspaces of the vector space W. The spaces U, V are called complementary if every $w \in W$ can be written as*

$$w = u + v \quad \textit{with} \quad u \in U, \quad v \in V \ , \tag{13.1}$$

and the decomposition (13.1) is unique. I.e., if (13.1) holds and if

$$w = u_1 + v_1 \quad \textit{with} \quad u_1 \in U, \quad v_1 \in V \ , \tag{13.2}$$

then $u = u_1$ and $v = v_1$. If U and V are complementary subspaces of W then one writes

$$W = U \oplus V$$

and calls W the direct sum of U and V.

Lemma 13.1 *Let U and V be subspaces of $W = \mathbb{C}^n$. Then $\mathbb{C}^n = U \oplus V$ if and only if*

$$dim\, U + dim\, V = n$$

and

$$U \cap V = \{0\} \ .$$

Proof: Let $u_1, \ldots, u_k$ be a basis of U and let $v_1, \ldots, v_l$ be a basis of V.

First assume that $\mathbb{C}^n = U \oplus V$. We claim that (a) the $k+l$ vectors $u_1, \ldots, v_l$ are linearly independent. Assuming that

$$\sum_i \alpha_i u_i + \sum_j \beta_j v_j = 0$$

we conclude, using uniqueness of the decomposition $0+0=0$, that

$$\sum_i \alpha_i u_i = \sum_j \beta_j v_j = 0 \; .$$

This implies $\alpha_i = \beta_j = 0$. Next we show that (b) the $k+l$ vectors $u_1, \ldots, v_l$ generate $\mathbb{C}^n$. If $w \in \mathbb{C}^n$ is given, we can write $w = u + v = \sum_i \alpha_i u_i + \sum_j \beta_j v_j$. From (a) and (b) we conclude that the $k+l$ vectors $u_1, \ldots, v_l$ form a basis of $\mathbb{C}^n$. Therefore, $k+l=n$. Next let $w \in U \cap V$. Since $w - w = 0 = 0 + 0$ with $w \in U, -w \in V$, we conclude that $w = 0$, proving that $U \cap V = \{0\}$.

Second, assume $dim\, U + dim\, V = n$ and $U \cap V = \{0\}$. Similarly as before, it follows that the n vectors $u_1, \ldots, v_l$ form a basis of $\mathbb{C}^n$. One then concludes that $\mathbb{C}^n = U \oplus V$. $\diamond$

13.2 Projectors

Definition: *Let W be a vector space. A linear operator $P : W \to W$ is called a projector if $P^2 = P$.*

Lemma 13.2 *Let U and V be complementary subspaces of the vector space W, i.e., $W = U \oplus V$. Given $w \in W$, let $u \in U$ and $v \in V$ be determined so that $w = u + v$. (By assumption, u and v are unique.) The mapping*

$$P : W \to W, \quad w \to Pw = u \; ,$$

is linear and is a projector, i.e., $P^2 = P$.

Proof: Let $w_1, w_2 \in W$ and let

$$w_1 = u_1 + v_1, \quad w_2 = u_2 + v_2, \quad u_j \in U, \quad v_j \in V \; .$$

We then have

$$\alpha w_1 + \beta w_2 = (\alpha u_1 + \beta u_2) + (\alpha v_1 + \beta v_2) \; .$$

This implies that

$$P(\alpha w_1 + \beta w_2) = \alpha u_1 + \beta u_2 = \alpha P w_1 + \beta P w_2 \; ,$$

showing linearity of P.

If $u \in U$, then the equation

$$u = u + 0, \quad u \in U, \quad 0 \in V \; ,$$

yields that $Pu = u$. Therefore, for any $w \in W$,

$$P^2 w = Pw$$

since $Pw \in U$. $\diamond$.

The projector $P : W \to W$ determined in the previous lemma is called the projector onto U along V. It is easy to see that $Q = I - P$ is the projector onto V along U.

We have seen that any pair U, V of complementary subspaces of W determines a projector $P : W \to W$, the projector onto U along V. Conversely, let $P : W \to W$ be any projector. It is easy to see that the subspaces

$$R(P) \quad \text{and} \quad N(P)$$

are complementary and that P is the projector onto $R(P)$ along $N(P)$. To summarize, any pair of complementary subspaces of a vector space determines a projector and, conversely, any projector determines a pair of complementary subspaces.

13.3 Matrix Representations of a Projector

Let U and V denote complementary subspaces of $W = \mathbb{C}^n$,

$$\mathbb{C}^n = U \oplus V \ ,$$

and let $P : \mathbb{C}^n \to \mathbb{C}^n$ denote the projector onto U along V. Let $u_1, \ldots, u_k$ be a basis of U and let $v_1, \ldots, v_l$ be a basis of V where $k + l = n$. The $n \times n$ matrix

$$T = (u_1, \ldots u_k, v_1, \ldots, v_l) =: (T^I, T^{II})$$

is non–singular. Here T^I has k columns and T^{II} has l columns.

We want to determine the matrix representation of P. To this end, let $w \in \mathbb{C}^n$ be given and write

$$w = \sum_{i=1}^k \alpha_i u_i + \sum_{j=1}^l \beta_j v_j = T \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \ .$$

Then we have

$$Pw = u = \sum_{i=1}^k \alpha_i u_i \ .$$

This leads to the matrix form of P as follows: We have

$$\begin{aligned} Pw &= T^I \alpha \\ &= (T^I, T^{II}) \begin{pmatrix} \alpha \\ 0 \end{pmatrix} \\ &= T \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ &= T \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} T^{-1} w \ . \end{aligned}$$

We have derived the following matrix representation of P:

$$P = T \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} T^{-1} \ . \tag{13.3}$$

Another way to write the projector P is as follows: Let $S = (T^{-1})^T$ and partition T and S as

$$T = (T^I, T^{II}), \quad S = (S^I, S^{II})$$

where T^I and S^I have k columns. Then we have

$$T^{-1} = S^T = \begin{pmatrix} (S^I)^T \\ (S^{II})^T \end{pmatrix} \ .$$

Using (13.3), it is not difficult to show that

$$P = T^I (S^I)^T \ . \tag{13.4}$$

Here, typically, one leaves P in factorized form.

Example: Let $n = 2$ and

$$U = span\{\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad V = span\{\begin{pmatrix} 1 \\ 1 \end{pmatrix}\} \ .$$

Then we have

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \quad S = (T^{-1})^T = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \ .$$

The projector P onto U along V reads, according to (13.3),

$$P = T \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} T^{-1} = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} \ .$$

The alternative representation (13.4) is

$$P = \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1 \ , \ \ -1) \ .$$

The factorized form clearly shows that Pw is a multiple of e^1.

Remark: The matrix representation (13.3) shows that $tr(P) = k = dim\, R(P)$ for any projector $P \in \mathbb{C}^{n \times n}$. (Recall that the trace of a matrix is a coefficient of its characteristic polynomial and, therefore, remains unchanged under similarity transformations.)

13.4 Orthogonal Complementary Subspaces

The decomposition

$$\mathbb{C}^n = U \oplus V$$

is particularly useful if U and V are orthogonal subspaces, i.e., if $V = U^\perp$. We will show that this occurs if and only if the corresponding projector P onto U along V is Hermitian.

Theorem 13.1 *Let $\mathbb{C}^n = U \oplus V$ and let P denote the projector onto U along V. Then U is orthogonal to V if and only if $P = P^*$.*

Proof: First assume that $P = P^*$. Let $u \in U = R(P)$ and let $v \in V = N(P)$ be arbitrary. Write $u = Px$ and obtain

$$\begin{aligned} \langle u, v \rangle &= \langle Px, v \rangle \\ &= \langle x, Pv \rangle \\ &= 0 . \end{aligned}$$

This shows that U and V are orthogonal.

Conversely, let $V = U^\perp$. If $u_1, \ldots, u_k$ is an ONB of U and $v_1, \ldots, v_l$ is an ONB of V then $k + l = n$ and $u_1, \ldots, v_l$ is an ONB of $\mathbb{C}^n$. Form the matrix $T = (u_1, \ldots, v_l)$ as in the previous section. This matrix is unitary, thus

$$T^{-1} = T^* .$$

The formula

$$P = T \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} T^*$$

shows that $P^* = P$. $\diamond$

Remark: A projector P with $P^* = P$ is sometimes called an orthogonal projector since $R(P)$ and $N(P)$ are orthogonal. However, P is not an orthogonal matrix unless $P = I$.

13.5 Invariant Complementary Subspaces and Transformation to Block Form

Let $A \in \mathbb{C}^{n \times n}$. Assume that

$$\mathbb{C}^n = U \oplus V$$

and

$$A(U) \subset U, \quad A(V) \subset V . \tag{13.5}$$

In other words, the complementary subspaces U and V of $\mathbb{C}^n$ are both invariant under A. As above, let $u_1, \ldots, u_k$ be a basis of U and let $v_1, \ldots, v_l$ be a basis of V where $k + l = n$. We form the $n \times n$ matrix

$$T = (u_1, \ldots u_k, v_1, \ldots, v_l) =: (T^I, T^{II}) ,$$

which is non–singular, and consider the similarity transform

$$T^{-1}AT =: B .$$

We claim that the invariance (13.5) implies that B is a block matrix

$$B = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \tag{13.6}$$

where B_1 is $k \times k$ and B_2 is $l \times l$.

Indeed,

$$Au_j = \sum_{i=1}^{k} \alpha_{ij} u_i, \quad j = 1, \ldots, k ,$$

and

$$Av_j = \sum_{i=1}^{l} \beta_{ij} v_i, \quad j = 1, \ldots, l .$$

It then is not difficult to see that

$$\begin{aligned} AT &= (Au_1, \ldots, Au_k, Av_1, \ldots, Av_l) \\ &= \Big(\sum_{i=1}^{k} \alpha_{i1} u_i, \ldots, \sum_{i=1}^{l} \beta_{i1} v_i, \ldots \Big) \\ &= TB \end{aligned}$$

if B has the block form (13.6) and

$$B_1 = (\alpha_{ij}), \quad B_2 = (\beta_{ij}) .$$

13.6 The Range–Nullspace Decomposition

Let $B \in \mathbb{C}^{n \times n}$ denote a singular matrix of $rank\, B = r$. Since

$$N = N(B) \quad \text{and} \quad R = R(B)$$

have ranks

$$rank\, N = n - r \quad \text{and} \quad rank\, R = r$$

the spaces N and R are complementary subspaces of $\mathbb{C}^n$ if and only if

$$N \cap R = \{0\} .$$

The example

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

with

$$N = R = span\{e^1\}$$

shows that the spaces N and R are not always complementary.

Lemma 13.3 *Let $B \in \mathbb{C}^{n\times n}$ denote a singular matrix. The subspaces $N = N(B)$ and $R = R(B)$ are complementary subspaces of $\mathbb{C}^n$ if and only if the eigenvalue 0 of B is semi–simple, i.e., the geometric and algebraic multiplicities of the eigenvalue 0 are the same.*

Proof: We know that 0 is semi–simple if and only if[7]

$$N(B) = N(B^2) \ .$$

First assume 0 to be semi–simple and let

$$w \in N \cap R \ .$$

Then there exists a vector $x \in \mathbb{C}^n$ with $w = Bx$ and $0 = Bw = B^2x$. It follows that $x \in N(B^2) = N(B)$, thus $w = Bx = 0$. We have shown that $N \cap R = \{0\}$ if 0 is a semi–simple eigenvalue of B.

Second, assume $N \cap R = \{0\}$. We want to prove that $N(B) = N(B^2)$. To this end, let $x \in N(B^2)$ be arbitrary and set $y = Bx$. We then have $By = B^2x = 0$, thus $y \in N \cap R$. The assumption $N \cap R = \{0\}$ yields that $y = 0$, thus $Bx = 0$, thus $x \in N(B)$. This argument shows that $N(B^2) \subset N(B)$. Since the inclusion $N(B) \subset N(B^2)$ is trivial, we have shown that $N(B) = N(B^2)$, i.e., 0 is semi–simple. $\diamond$

To summarize, if $B \in \mathbb{C}^{n\times b}$ is a singular matrix with semi–simple eigenvalue 0, then we have

$$\mathbb{C}^n = N(B) \oplus R(B) \ . \tag{13.7}$$

This is called the range–nullspace decomposition of $\mathbb{C}^n$ determined by B. It is clear that both spaces, $N(B)$ and $R(B)$, are invariant under B.

Let $A \in \mathbb{C}^{n\times n}$ denote any matrix and let λ_1 be a semi–simple eigenvalue of A. Setting $B = A - \lambda_1 I$ and applying the above result, one obtains the range–nullspace decomposition

$$\mathbb{C}^n = N(A - \lambda_1 I) \oplus R(A - \lambda_1 I) \ . \tag{13.8}$$

Here

$$N(A - \lambda_1 I) = E(\lambda_1)$$

is the eigenspace corresponding to λ_1.

[7] The inclusion $N(B) \subset N(B^2)$ always holds. If $N(B) \neq N(B^2)$ then there exists x with $B^2x = 0, Bx \neq 0$ and the algebraic multiplicity of the eigenvalue 0 exceeds its geometric multiplicity. Thus, if $N(B) \neq N(B^2)$ then $\lambda = 0$ is not semi–simple.

13.7 The Spectral Theorem for Diagonalizable Matrices

Let $A \in \mathbb{C}^{n\times n}$ denote a diagonalizable matrix with characteristic polynomial

$$p_A(z) = det(A - zI) = (\lambda_1 - z)^{m_1} \dots (\lambda_s - z)^{m_s}$$

where $\lambda_1, \dots, \lambda_s$ are the distinct eigenvalues of A. The assumption that A is diagonalizable is equivalent to saying that all eigenvalues of A are semi–simple. If

$$U_j := E(\lambda_j) = N(A - \lambda_j I)$$

denotes the eigenspace to λ_j then

$$dim\, U_j = m_j \ .$$

We also set

$$V_j = R(A - \lambda_j I) \ .$$

Taking $j = 1$, for instance, we have the range–nullspace decomposition

$$\mathbb{C}^n = U_1 \oplus V_1 \ .$$

Here

$$dim\, U_1 = m_1, \quad dim V_1 = n - m_1 = m_2 + \dots + m_s \ .$$

Lemma 13.4 *Under the above assumptions we have*

$$V_1 = U_2 \oplus \dots \oplus U_s \ .$$

Thus, if A is a diagonalizable matrix with distinct eigenvalues $\lambda_1, \dots, \lambda_s$, then the range $V_1 = R(A - \lambda_1 I)$ is the direct sum of the eigenspaces $U_j = N(A - \lambda_j I)$ corresponding to all the eigenvalues different from λ_1.

Proof: Let us show first that the eigenspace U_2 is a subspace of the range space V_1. To this end, let $x \in U_2$, thus

$$Ax = \lambda_2 x \ ,$$

thus

$$(A - \lambda_1 I)x = (\lambda_2 - \lambda_1)x \ .$$

Dividing by $\lambda_2 - \lambda_1$ we obtain that $x \in V_1$. The same arguments apply to U_3 etc. One obtains that

$$U_2 + \dots + U_s \subset V_1 \ .$$

It is not difficult to show that the sum $U_2 + \dots + U_s$ is direct and has dimension $m_2 + \dots + m_s$. Then the claim follows. $\diamond$

For each $j = 1, \ldots, s$ we have the range–nullspace decomposition

$$\mathbb{C}^n = U_j \oplus V_j, \quad U_j = N(A - \lambda_j I), \quad V_j = R(A - \lambda_j I) \; .$$

Let P_j denote the projector onto U_j along V_j.

The spectral theorem for diagonalizable matrices is the following:

Theorem 13.2 *Let $A \in \mathbb{C}^{n \times n}$ denote a diagonalizable matrix with distinct eigenvalues $\lambda_1, \ldots, \lambda_s$ and let P_j denote the projector onto the eigenspace $U_j = N(A - \lambda_j i)$ along the range space $V_j = R(A - \lambda_j I)$. Then we have*

$$\begin{aligned} A &= \lambda_1 P_1 + \ldots + \lambda_s P_s \\ I &= P_1 + \ldots + P_s \\ P_i P_j &= \delta_{ij} P_i = \delta_{ij} P_j \quad \text{for} \quad 1 \le i, j \le s \; . \end{aligned}$$

The representation $A = \sum \lambda_j P_j$ is called the spectral representation of A.

Proof: Consider a transformation matrix

$$T = (T^{(1)}, \ldots, T^{(s)})$$

where the columns of $T^{(j)}$ form a basis of U_j. Then we have

$$T^{-1} A T = \Lambda = \begin{pmatrix} \lambda_1 I_{m_1} & & \\ & \ddots & \\ & & \lambda_s I_{m_s} \end{pmatrix} \; .$$

This gives the representation

$$A = T \begin{pmatrix} \lambda_1 I_{m_1} & & \\ & \ddots & \\ & & \lambda_s I_{m_s} \end{pmatrix} T^{-1} \; .$$

The projector P_1 is

$$P_1 = T \begin{pmatrix} I_{m_1} & & \\ & 0 & \\ & & 0 \end{pmatrix} T^{-1} \; .$$

A similar formula holds for P_2 etc. The claims of the theorem are then obvious. $\diamond$

Example: Let

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \; .$$

The eigenvalues of A are

$$\lambda_1 = 3, \quad \lambda_2 = 1$$

with algebraic multiplicities $m_1 = m_2 = 1$. We have

$$A\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 3\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad A\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

Therefore, the eigenspaces are

$$U_1 = E(3) = span\{\begin{pmatrix} 1 \\ -1 \end{pmatrix}\}, \quad U_2 = E(1) = span\{\begin{pmatrix} 1 \\ 1 \end{pmatrix}\} .$$

The transformation matrix T has the eigenvectors as columns,

$$T = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad T^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} .$$

The transformation of A to diagonal form is

$$T^{-1}AT = \Lambda = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

leading to the representation

$$A = T\begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}T^{-1} .$$

The projectors are

$$P_1 = T\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}T^{-1}, \quad P_2 = T\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}T^{-1} .$$

The spectral representation of A is

$$A = 3P_1 + P_2 .$$

Using the matrix

$$S = (T^{-1})^T = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

and the projector representation (13.4), we have

$$P_1 = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}(1\ ,\ -1), \quad P_2 = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}(1\ ,\ 1) .$$

In this way, we can write Aw as

$$Aw = \frac{3}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}(1\ ,\ -1)w + \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}(1\ ,\ 1)w .$$

Evaluating the inner products, we have

$$Aw = \frac{3}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}(w_1 - w_2) + \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}(w_1 + w_2) .$$

The point is that Aw is written directly as a linear combination of the eigenvectors of A,

$$Aw = \alpha(w) \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \beta(w) \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

The coefficients, $\alpha(w)$ and $\beta(w)$, are linear functionals of w which are also directly displayed.

13.8 Functions of A

Let A be a diagonalizable matrix with spectral representation

$$A = \sum \lambda_j P_j .$$

It is then easy to obtain functions of A in terms of its spectral representation. For example,

$$\begin{aligned} A^2 &= \sum \lambda_j^2 P_j \\ A^3 &= \sum \lambda_j^3 P_j \\ e^A &= \sum e^{\lambda_j} P_j \end{aligned}$$

A matrix B with $B^2 = A$ is called a square root of A, often written as $B = A^{1/2}$. One should note, however, that square roots a typically not unique. A square root of A can be obtained as

$$A^{1/2} = \sum \lambda_j^{1/2} P_j .$$

Similarly, if A is non–singular, a logarithm of A can be obtained as

$$\log A = \sum (\log \lambda_j) P_j .$$

Here $\log \lambda_j$ is any complex logarithm of λ_j. Since $e^{\log \lambda_j} = \lambda_j$ it follows that

$$e^{\log A} = A .$$

Remark: A non–diagonalizable matrix may not have a square root. For example, the matrix

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

does not have a root. If $B^2 = J$ then $B^4 = J^2 = 0$, which implies that B is nilpotent. Since B is 2×2 we conclude that $B^2 = 0$, a contradiction.

13.9 Spectral Projectors in Terms of Right and Left Eigenvectors

Definition: *Let λ be an eigenvalue of $A \in \mathbb{C}^{n \times n}$. A column vector $x \in \mathbb{C}^n$ is called a right eigenvector of A to the eigenvalue λ if $Ax = \lambda x$ and $x \neq 0$. A* *row vector y^T is called a left eigenvector of A to the eigenvalue λ if $y^T A = \lambda y^T$ and $y \neq 0$.*

Let A be a diagonalizable matrix, as above. Recall the relation

$$T^{-1}AT = \Lambda \quad \text{or} \quad AT = T\Lambda$$

introduced above. Here the columns of T are right eigenvectors of A. If $(T^{-1})^T =: S$ we can also write the equation $T^{-1}AT = \Lambda$ as

$$S^T A = \Lambda S^T \ .$$

This relation says that the rows of S are left eigenvectors of A. If we partition column-wise,

$$T = \Big(T^{(1)}, \ldots, T^{(s)}\Big), \quad S = \Big(S^{(1)}, \ldots, S^{(s)}\Big) \ ,$$

where $T^{(j)}$ and $S^{(j)}$ contain m_j columns, then $T^{(j)}$ contains right eigenvectors to λ_j in its columns and $(S^{(j)})^T$ contains left eigenvectors to λ_j in its rows.

As above, let P_j denote the projector onto the eigenspace $U_j = E(\lambda_j)$ along the sum of the other eigenspaces. Then, corresponding to (13.4), we have the product representation

$$P_j = T^{(j)}(S^{(j)})^T \ . \tag{13.9}$$

Let us consider the special case where the eigenvalues of A are all distinct, i.e., all eigenspaces have dimension one. Then, for each eigenvalue λ_j of A, there are non–zero vectors x_j and y_j with

$$Ax_j = \lambda_j x_j, \quad y_j^T A = \lambda_j y_j^T \ .$$

The vectors x_j and y_j are uniquely determined, up to scalar factors. The representation (13.9) becomes

$$P_j = \alpha_j x_j y_j^T$$

where α_j is a scalar which we will determine below.

Lemma 13.5 *Let $x_1, \ldots, x_n$ and $y_1^T, \ldots, y_n^T$ be right and left eigenvectors of A to the distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. Then we have*

$$y_j^T x_i = 0 \quad \textit{for} \quad i \neq 0, \quad y_j^T x_j \neq 0 \ .$$

Proof: For $i \neq j$ we have

$$\lambda_j y_j^T x_i = y_j^T A x_i = \lambda_i y_j^T x_i \ .$$

This yields $y_j^T x_i = 0$ since $\lambda_i \neq \lambda_j$. If, in addition, the equality $y_j^T x_j = 0$ would also hold, then y_j would be orthogonal to a basis of $\mathbb{C}^n$ and the equation $y_j = 0$ would follow. $\diamond$

Since $y_j^T x_j \neq 0$ we may assume, after scaling, that

$$y_j^T x_j = 1 \ .$$

Then we have

$$\Big(x_j y_j^T\Big)^2 = x_j y_j^T x_j y_j^T = x_j y_j^T \ .$$

It follows that the projector P_j is given by

$$P_j = x_j y_j^T \ .$$

We have shown:

Theorem 13.3 *Let $A \in \mathbb{C}^{n \times n}$ have n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. There exist non–zero vectors x_j and y_j with*

$$A x_j = \lambda_j x_j, \quad y_j^T A = \lambda_j y_j^T, \quad y_j^T x_j = 1 \ .$$

In terms of these vectors, the spectral projectors are the rank 1 matrices

$$P_j = x_j y_j^T \ .$$

The matrix A has the spectral representation

$$A = \sum_{j=1}^n \lambda_j x_j y_j^T \ .$$

14 The Resolvent and Projectors

In this chapter $A \in \mathbb{C}^{n \times n}$ denotes a general complex matrix, diagonalizable or not. How do the eigenvalues of A change if A gets perturbed? One can show that they change continuously. (See Theorem 14.4.) However, a multiple eigenvalue generally breaks up non–smoothly under perturbations. A simple example is

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

Its only eigenvalue $\lambda_1 = 0$ is algebraically double, but geometrically simple. The perturbed matrix

$$A_\varepsilon = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}, \quad \varepsilon > 0 ,$$

has the distinct eigenvalues

$$\lambda_1(\varepsilon) = \sqrt{\varepsilon}, \quad \lambda_2(\varepsilon) = -\sqrt{\varepsilon}$$

and the function $\sqrt{\varepsilon}$ is not differentiable at $\varepsilon = 0$.

An aim of this chapter is to show that projectors onto sums of generalized eigenspaces behave better under perturbations of A than the perturbed eigenvalues. The projectors are analytic functions of the perturbation. The resolvent of A, introduced in the next section, is used to obtain an integral representation of projectors.

The subject connects linear algebra with complex variables.

14.1 The Resolvent of a Matrix

Let $A \in \mathbb{C}^{n \times n}$. With $\lambda_1, \ldots, \lambda_s$ we denote the distinct eigenvalues of A, thus the set

$$\sigma(A) = \{\lambda_1, \ldots, \lambda_s\}$$

is the spectrum of A. The matrix valued function

$$z \to R(z) = (zI - A)^{-1}, \quad z \in \mathbb{C} \setminus \sigma(A) ,$$

is called the resolvent of A.

The next theorem states that the resolvent $R(z) = (zI - A)^{-1}$ is an analytic function defined on the open set $\mathbb{C} \setminus \sigma(A)$ with a pole at each eigenvalue of A. We recall that the index i_j of the eigenvalue λ_j of A is the index of nilpotency of the operator

$$(\lambda_j I - A)\Big|_{gE_{\lambda_j}} , \tag{14.1}$$

and i_j equals the size of the largest Jordan block to λ_j. In formula (14.1), the space

$$gE_{\lambda_j} = \{x \in \mathbb{C}^n \ : \ (\lambda_j I - A)^n x = 0\}$$

denotes the generalized eigenspace of A to the eigenvalue λ_j and the operator in (14.1) is the restriction of the operator $\lambda_j I - A$ to the generalized eigenspace gE_{λ_j}. The operator (14.1) maps the space gE_{λ_j} into itself.

Theorem 14.1 *1. The resolvent*

$$(zI - A)^{-1} = R(z) = (r_{jk}(z))_{1 \leq j,k \leq n}$$

depends analytically on $z \in \mathbb{C} \setminus \sigma(A)$. In fact, every entry $r_{jk}(z)$ is a rational function of z.

2. At every eigenvalue λ_j of A the resolvent has a pole. The order of the pole at λ_j equals the index of the eigenvalue λ_j.

Proof: 1. The formula for the inverse of a matrix in terms of determinants shows that $r_{jk}(z)$ is a rational function of z; see Theorem 9.8.

Remarks: Under suitable assumptions, the resolvent generalizes to linear operators A in Banach spaces. The following proof of the analyticity of $R(z)$ can be generalized to certain operators on Banach spaces.

Let $z_0 \in \mathbb{C} \setminus \sigma(A)$ and let $|z - z_0| < \varepsilon$ where

$$\varepsilon = \frac{1}{|R(z_0)|} \ .$$

We write

$$\begin{aligned} zI - A &= (z_0 I - A) - (z_0 - z)I \\ &= (z_0 I - A)\Big(I - (z_0 - z)R(z_0)\Big) \end{aligned}$$

where

$$|z_0 - z||R(z_0)| < \varepsilon |R(z_0)| = 1 \ .$$

The geometric sum formula applies to the inverse of $I - (z_0 - z)R(z_0)$ and one obtains that

$$R(z) = \sum_{j=0}^{\infty} (z_0 - z)^j \Big(R(z_0)\Big)^{j+1} \quad \text{for} \quad |z - z_0| < \frac{1}{|R(z_0)|} \ .$$

This shows that $R(z)$ is a power series as a function of z in a neighborhood of any point z_0 outside of the spectrum of A.

2. Let

$$J_k = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \\ \vdots & & & 0 & 1 \\ 0 & \ldots & \ldots & \ldots & 0 \end{pmatrix}$$

denote an elementary Jordan block of dimension $k \times k$ and consider $(zI - J_k)^{-1}$ for $z \neq 0$. We have $zI - J_k = z\Big(I - \frac{1}{z}\, J_k\Big)$, thus

$$(zI - J_k)^{-1} = \frac{1}{z}\,\Big(I + \frac{1}{z}\,J_k + \ldots + \frac{1}{z^{k-1}}\,J_k^{k-1}\Big)$$

where $J_k^{k-1} \neq 0$ and $J_k^k = 0$. This shows that $(zI - J_k)^{-1}$ has a pole of order k at $z = 0$.

We have shown in Section 12.5 that there exists a transformation matrix T so that $T^{-1}AT$ has Jordan form,

$$T^{-1}AT = \begin{pmatrix} B_1 & & 0 \\ & \ddots & \\ 0 & & B_s \end{pmatrix} =: B$$

with

$$B_j = \lambda_j I_{m_j} + J^{(j)}$$

where $J^{(j)}$ is a nilpotent Jordan matrix. I.e., $J^{(j)}$ is a block diagonal matrix whose diagonal blocks are elementary Jordan blocks. From $A = TBT^{-1}$ one obtains that

$$zI - A = T(zI - B)T^{-1}$$

and

$$(zI - A)^{-1} = T(zI - B)^{-1}T^{-1}\ .$$

Here $(zI - B)^{-1}$ is a block diagonal matrix with diagonal blocks

$$(zI_{m_j} - B_j)^{-1} = \Big((z - \lambda_j)I_{m_j} - J^{(j)}\Big)^{-1}\ .$$

Such a block has a pole of order i_j at $z = \lambda_j$. $\diamond$

14.2 Integral Representation of Projectors

The next theorem states that the residue of the resolvent $R(z) = (zI - A)^{-1}$ at the eigenvalue $z = \lambda_j$ is the projector P_j onto the generalized eigenspace gE_{λ_j} along the sum of the other generalized eigenvalues.

We recall that $\mathbb{C}^n$ is the direct sum of the generalized eigenspaces of A:

$$\mathbb{C}^n = gE_{\lambda_1} \oplus gE_{\lambda_2} \oplus \ldots \oplus gE_{\lambda_s}\ .$$

See Theorem 10.16.

In the following, let

$$\Gamma_{\lambda_j r} = \partial D(\lambda_j, r)$$

denote the positively oriented circle of radius r centered at λ_j. This circle has the parameterization

$$z(\phi) = \lambda_j + re^{i\phi}, \quad 0 \le \phi \le 2\pi \; .$$

Theorem 14.2 *Let $A \in \mathbb{C}^{n\times n}$ and let $\lambda_1, \ldots, \lambda_s$ denote the distinct eigenvalues of A. Assume that $r > 0$ is so small that λ_j is the only eigenvalue of A in the closed disk*

$$\bar{D}(\lambda_j, r) = \{z \; : \; |z - \lambda_j| \le r\} \; .$$

Then we have

$$\frac{1}{2\pi i}\int_{\Gamma_{\lambda_j r}} (zI - A)^{-1}\, dz = P_j$$

where P_j is the projector onto the generalized eigenspace gE_{λ_j} along the sum of the other generalized eigenspaces.

We will prove this below.

Together with Cauchy's integral theorem, one obtains the following result from Theorem 14.2.

Theorem 14.3 *Let Γ denote any simply closed positively oriented curve in $\mathbb{C} \setminus \sigma(A)$ and let Ω denote the region surrounded by Γ. Consider the following two sums of generalized eigenspaces:*

$$\begin{aligned} U &= \sum_{\lambda_j \in \Omega} gE_{\lambda_j} \\ V &= \sum_{\lambda_j \notin \bar{\Omega}} gE_{\lambda_j} \end{aligned}$$

Then

$$P = \frac{1}{2\pi i}\int_{\Gamma} (zI - A)^{-1}\, dz \tag{14.2}$$

is the projector onto U along V.

If the matrix A has non–simple eigenvalues, then the spectrum of A generally behaves badly (continuously, but not smoothly) under perturbations of A. We will prove in the last section that an appropriate sum of eigenprojectors behaves much better under perturbations of A, however. This is a consequence of the representation (14.2). Eigenprojectors and their sums are better behaved mathematical objects than the spectrum itself.

14.3 Proof of Theorem 14.2

First consider an elementary Jordan block of dimension $k \times k$

$$J_k = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} .$$

The generalized eigenspace to the eigenvalue $\lambda_1 = 0$ is

$$gE_0 = \mathbb{C}^k =: U$$

and the direct sum of all other generalized eigenspaces is

$$\{0\} =: V \ .$$

The projector onto U along V is $P = I_k$.

The resolvent of J_k is

$$\begin{aligned} (zI_k - J_k)^{-1} &= z^{-1}(I_k - \frac{1}{z}J_k)^{-1} \\ &= \frac{1}{z}\Big(I_k + \frac{1}{z}J_k + \ldots + \frac{1}{z^{k-1}}J_k^{k-1}\Big) \quad \text{for} \quad z \in \mathbb{C} \setminus \{0\} \ . \end{aligned}$$

If $\Gamma_r = \partial D(0,r)$ denotes the positively oriented circle of radius r, centered at $z = 0$, then we know from complex variables that

$$\frac{1}{2\pi i}\int_{\Gamma_r} \frac{dz}{z^j} = \begin{cases} 1 & \text{for} \quad j = 1 \\ 0 & \text{for} \quad j = 2, 3, \ldots \end{cases}$$

Therefore, the above formula for the resolvent $(zI_k - J_k)^{-1}$ yields that

$$\frac{1}{2\pi i}\int_{\Gamma_r} (zI_k - J_k)^{-1}\, dz = I_k \ .$$

The next lemma follows easily.

Lemma 14.1 *Let Γ_r denote the boundary curve of $D(0,r)$ and let $\Gamma_{\lambda r}$ denote the boundary curve of $D(\lambda, r)$.*

1. If

$$J = \begin{pmatrix} J_{k_1} & & \\ & \ddots & \\ & & J_{k_q} \end{pmatrix} \in \mathbb{C}^{m\times m}$$

is any nilpotent Jordan matrix, then

$$\frac{1}{2\pi i}\int_{\Gamma_r} (zI_m - J)^{-1}\, dz = I_m \ .$$

2. If $B = \lambda I_m + J$ then

$$\frac{1}{2\pi i}\int_{\Gamma_{\lambda r}} (zI_m - B)^{-1}\, dz = I_m \ .$$

To prove Theorem 14.2 we assume that $j = 1$ for simplicity of notation. As above, let $T^{-1}AT = B$ denote the Jordan form of A, thus $A = TBT^{-1}$.

We have

$$\begin{aligned} \frac{1}{2\pi i}\int_{\Gamma_{\lambda_1 r}} (zI-A)^{-1}\,dz &= \frac{1}{2\pi i} T\Big(\int_{\Gamma_{\lambda_1 r}} (zI-B)^{-1}\,dz\Big)T^{-1} \\ &= T\begin{pmatrix} I_{m_1} & 0 \\ 0 & 0 \end{pmatrix} T^{-1} \end{aligned}$$

According to Section 13.3, this is the projector onto $U = gE_{\lambda_1}$ along

$$V = gE_{\lambda_2} \oplus \ldots \oplus gE_{\lambda_s} \; .$$

This proves the theorem. $\diamond$

14.4 Application: Sums of Eigenprojectors under Perturbations

We first consider a simple example which shows that multiple eigenvalues generally behave badly under perturbations of the matrix.

Example: Let

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \; , \quad Q = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \; .$$

Thus A has the eigenvalue $\lambda_1 = 0$ of algebraic multiplicity 3 and geometric multiplicity 1. The perturbed matrix

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

has the characteristic polynomial

$$det(A+\varepsilon Q - zI) = det\begin{pmatrix} -z & 1 & 0 \\ 0 & -z & 1 \\ \varepsilon & 0 & -z \end{pmatrix} = -z^3 + \varepsilon \; .$$

If $\varepsilon \neq 0$ is a complex number,

$$\varepsilon = |\varepsilon| e^{i\theta} \; ,$$

then the three eigenvalues of $A + \varepsilon Q$ are

$$\lambda_j(\varepsilon) = |\varepsilon|^{1/3} e^{i(\theta + 2\pi j)/3}, \quad j = 1,2,3 \; .$$

We note that the above expression depends continuously on ε but is not differentiable at $\varepsilon = 0$. Of course, the expression $A + \varepsilon Q$ depends analytically on ε. The example shows that analytic dependence of a perturbation of A on a parameter ε does *not* imply analytic dependence of the eigenvalues of the perturbed matrix on the parameter.

Next we show precisely in which sense the eigenvalues of a matrix depend continuously on the matrix entries.

Theorem 14.4 *Let $A \in \mathbb{C}^{n\times n}$ have the characteristic polynomial*

$$p_A(z) = det(A - zI) = (\lambda_1 - z)^{m_1} \ldots (\lambda_s - z)^{m_s}$$

with distinct numbers $\lambda_1, \ldots, \lambda_s$. Let $r > 0$ be chosen so small that the s disks

$$\bar{D}(\lambda_j, r), \quad j = 1, \ldots, s \ ,$$

are disjoint. Then there exists $\delta > 0$ with the following property: If $Q \in \mathbb{C}^{n\times n}$ satisfies $|Q| < \delta$ then, for $j = 1, \ldots, s$, the matrix $A + Q$ has precisely m_j eigenvalues in $D(\lambda_j, r)$, where each eigenvalue is counted according to its algebraic multiplicity.

Proof: Fix any $1 \le j \le s$ and consider the circle

$$\Gamma = \partial D(\lambda_j, r) \ .$$

We have

$$\min\{|p_A(z)| \ : \ z \in \Gamma\} > 0 \ ,$$

and, if $\delta > 0$ is small enough, then (by continuity as a function of Q)

$$\min\{|p_{A+Q}(z)| \ : \ z \in \Gamma\} > 0$$

for all $Q \in \mathbb{C}^{n\times n}$ with $|Q| < \delta$. By the residue theorem, the integer

$$\frac{1}{2\pi i} \int_\Gamma \frac{p'_{A+Q}(z)}{p_{A+Q}(z)}\, dz =: m(A+Q)$$

equals the number of zeros of $p_{A+Q}(z)$ in $D(\lambda_j, r)$. For $Q = 0$ we have $m(A) = m_j$. Since $m(A+Q)$ depends continuously on Q and is integer valued, it follows that $m(A+Q) = m_j$ for all Q with $|Q| < \delta$. $\diamond$

In the following we use the same notation as in the previous theorem and its proof. Fix $Q \in \mathbb{C}^{n\times n}$ and consider the matrices

$$A + \varepsilon Q, \quad |\varepsilon| < \varepsilon_0 \ ,$$

where $\varepsilon \in \mathbb{C}$ is a small parameter. Fix $1 \le j \le s$. If $|\varepsilon||Q| < \delta$ then, by the previous theorem, the matrix $A + \varepsilon Q$ has m_j eigenvalues in $D(\lambda_j, r)$. These eigenvalues, which depend on ε, form the so–called λ_j–group.

Let $U(\varepsilon)$ denote the sum of all generalized eigenspaces to the eigenvalues of the λ_j–group and let $V(\varepsilon)$ denote the sum of all other generalized eigenspaces. By Theorem 14.3 the projector

$$P(\varepsilon) = \frac{1}{2\pi i} \int_\Gamma \Big(zI - (A + \varepsilon Q) \Big)^{-1} \, dz$$

is the projector onto $U(\varepsilon)$ along $V(\varepsilon)$ if $|\varepsilon|$ is small enough.

We now show that the projector $P(\varepsilon)$ depends analytically on ε. Set $R(z) = (zI - A)^{-1}$ and assume that $|\varepsilon|$ is so small that

$$\max_{z \in \Gamma} |\varepsilon||R(z)||Q| < 1 \ .$$

We have

$$zI - (A + \varepsilon Q) = (zI - A)\Big(I - \varepsilon R(z) Q \Big) \ ,$$

thus

$$\Big(zI - (A + \varepsilon Q) \Big)^{-1} = \sum_{j=0}^\infty \varepsilon^j (R(z)Q)^j R(z)$$

for $z \in \Gamma$. The convergence of the series is uniform for $z \in \Gamma$. Therefore,

$$P(\varepsilon) = \frac{1}{2\pi i} \sum_{j=0}^\infty \varepsilon^j \int_\Gamma (R(z)Q)^j R(z) \, dz \ .$$

This proves that $P(\varepsilon)$ depends analytically on ε for $|\varepsilon| < \varepsilon_0$ if ε_0 is sufficiently small.

15 Approximate Solution of Large Linear Systems: GMRES

Suppose you want to solve a linear system

$$Ax = b$$

where $A \in \mathbb{R}^{n \times n}$ is a given nonsingular matrix and $b \in \mathbb{R}^n$ is a given vector. Assume that A is a full matrix, i.e., you cannot take advantage of sparsity patterns of A. Gaussian elimination needs about

$$N \sim \frac{2}{3} n^3$$

operations. Take an extreme case where

$$n = 10^8 \ .$$

The number of operations needed is

$$N \sim \frac{2}{3} 10^{24} \ .$$

Suppose you have a petaflop machine which performs about 10^{15} floating point operations per sec.[8]

The computation would take about

$$T \sim \frac{2}{3} 10^9 \, sec \ .$$

Since 1 year $\sim 3 * 10^7$ sec you have

$$T \sim 22 \, years \ .$$

Clearly, you must settle for an approximate solution of $Ax = b$ that can be computed faster.

GMRES is an acronym for *generalized minimal residual* algorithm. If $x_0 \in \mathbb{R}^n$ is any given vector, then $b - Ax_0$ is called the residual of x_0 for the system $Ax = b$ and

$$\phi(x_0) = |b - Ax_0|$$

is the Euclidean norm of the residual of x_0.

The idea of GMRES is to generate an ONB (ortho–normal basis) in a so–called Krylov subspace K_m of $\mathbb{R}^n$ and to determine the vector $\tilde{x} \in x_0 + K_m$ which minimizes the residual over the affine subspace

$$x_0 + K_m \ .$$

In other words, the vector $\tilde{z} \in K_m$ will be determined so that

[8] We ignore the difficulty that A may not fit into memory.

$$|b - A(x_0 + \tilde{z})| < |b - A(x_0 + z)| \quad \text{for all} \quad z \in K_m, \quad z \neq \tilde{z} \ . \qquad (15.1)$$

Then $\tilde{x} = x_0 + \tilde{z}$ will be the computed approximation to the exact solution $x_{ex} = A^{-1}b$. It is generally difficult to analyze how close $\tilde{x}$ is to $x_{ex} = A^{-1}b$. However, one can compute the size of the residual,

$$|b - A\tilde{x}| \ .$$

If this residual is sufficiently small, one accepts $\tilde{x}$ as a good approximation to $x = A^{-1}b$.

Note that

$$Ax_{ex} - A\tilde{x} = b - A\tilde{x} \ ,$$

thus

$$x_{ex} - \tilde{x} = A^{-1}(b - A\tilde{x}) \ ,$$

thus

$$|x_{ex} - \tilde{x}| \leq |A^{-1}||b - A\tilde{x}| \ .$$

15.1 GMRES

Let A be a nonsingular real $n \times n$ matrix and let $b \in \mathbb{R}^n$. We want to obtain an approximate solution $\tilde{x}$ of the system $Ax = b$. Regarding the matrix A, we assume that we can compute Av for any given vector v, but we will not manipulate the matrix elements of A.

15.1.1 The Arnoldi Process

Let x_0 denote an initial approximation for $x_{ex} = A^{-1}b$. For example, we can take $x_0 = 0$ if nothing better is known. Let

$$r_0 = b - Ax_0$$

denote its residual. With

$$K_m = K_m(r_0) = span\{r_0, Ar_0, \ldots, A^{m-1}r_0\}$$

we denote the m–th Krylov subspace for r_0. We assume that K_m has dimension m.

We want to compute the vector $z \in K_m$ which minimizes the Euclidean vector norm of the residual,

$$|b - A(x_0 + z)| \ ,$$

over K_m. Precisely, we want to determine $\tilde{z} \in K_m$ with

$$|b - A(x_0 + \tilde{z})| < |b - A(x_0 + z)| \quad \text{for all} \quad z \in K_m, \quad z \neq \tilde{z} \; . \tag{15.2}$$

In applications, m is much less than n; for instance, we can have $m = 20$ and $n = 10^8$. To perform the above minimization over K_m we compute an ONB $v_1, \ldots, v_m, v_{m+1}$ of K_{m+1}. The following is a pseudo–code for the so–called **Arnoldi process**:

```
      v_1 = r_0/|r_0|
      for j = 1 to m
C      By induction hypothesis, v_1,...,v_j form an ONB of K_j.
      v = A v_j
        for i = 1 to j
        h_ij = <v_i, v>
        v = v - h_ij v_i
        end i
      h_{j+1,j} = |v|
      v_{j+1} = v/h_{j+1,j}
C      The vectors v_1,...,v_j, v_{j+1} form an ONB of K_{j+1}.
      end j
```

Remark: Under the assumption that A is a full $n \times n$ matrix and $m << n$, the main computational work is the evaluation of the matrix times vector products $v = Av_j$ for $j = 1, \ldots, m$. This costs about $2\,m\,n^2$ operations. The remaining work is $\mathcal{O}(m^2 n)$, which is negligible. If $n = 10^8$ and $m = 20$ then the number of operations is about

$$2mn^2 = 4 \cdot 10^{17} \; .$$

If we can perform 10^{15} operations per second, the execution time is

$$T_M \sim 400 sec \; .$$

Upon completion, the Arnoldi process has produced vectors

$$v_1, v_2, \ldots, v_{m+1} \in \mathbb{R}^n$$

and numbers h_{ij} for $1 \leq j \leq m$ and $1 \leq i \leq j+1$. We collect the h_{ij} in a matrix:

$$H_m = \begin{pmatrix} h_{11} & & \ldots & h_{1m} \\ h_{21} & h_{22} & \ldots & h_{2m} \\ & \ddots & \ddots & \vdots \\ & & h_{m,m-1} & h_{mm} \\ 0 & & & h_{m+1,m} \end{pmatrix} \in \mathbb{R}^{(m+1)\times m} \; .$$

The matrix H_m has upper Hessenberg form.

Lemma 15.1 *Assume that the Krylov subspace K_{m+1} has dimension $m+1$. Then the following holds for the vectors $v_1, \ldots, v_{m+1}$ and the matrix H_m computed in the Arnoldi process:*

a) The vectors $v_1, \ldots, v_{m+1}$ are orthonormal (ON).

b) For $1 \le j \le m+1$, the vectors

$$v_1, \ldots, v_j$$

span K_j.

c) If we set

$$V_m = (v_1, \ldots, v_m), \quad V_{m+1} = (v_1, \ldots, v_m, v_{m+1}) \ ,$$

then

$$AV_m = V_{m+1} H_m \ .$$

In fact,

$$Av_j = \sum_{i=1}^{j+1} h_{ij} v_i \quad for \quad 1 \le j \le m \ .$$

Proof: 1) Using induction, let us assume that the vectors $v_1, \ldots, v_{j-1}$ form an ONB of K_{j-1} and that $v_1, \ldots, v_j$ form an ONB of K_j. Set $v = Av_j$ and define

$$\tilde{K}_{j+1} = span\,\{v_1, \ldots, v_j, v\} \ .$$

We claim that

$$\tilde{K}_{j+1} = K_{j+1} \ .$$

We must show that $v \in K_{j+1}$ and $A^j r_0 \in \tilde{K}_{j+1}$. First, v_j has the form

$$v_j = \sum_{i=0}^{j-1} \alpha_i A^i r_0 \ .$$

Therefore, $v = Av_j \in A(K_j) \subset K_{j+1}$.

Second, setting $y = A^{j-1} r_0$, we have $A^j r_0 = Ay$. Here y has the form

$$y = \sum_{i=1}^{j-1} \beta_i v_i + \beta_j v_j \ .$$

It follows that

$$Ay \in A(K_{j-1}) + \beta_j A v_j \subset K_j + \beta_j A v_j \subset \tilde{K}_{j+1} \ .$$

2) In the second part of the proof, we consider the following part of the Arnoldi process:

$v = Av_j$
  for $i = 1$ to j

$h_{ij} = \langle v_i, v \rangle$
$v = v - h_{ij} v_i$
end i

For $j = 1$ one computes

$$v = Av_1 \quad \text{and} \quad h_{11} = \langle v_1, v \rangle$$

and

$$v^{(1)} = v - \langle v_1, v \rangle v_1 \ .$$

Note that $v^{(1)}$ is orthogonal to v_1 and that

$$span\,\{v, v_1\} = span\,\{v^{(1)}, v_1\} \ .$$

In the next step, one computes

$$v^{(2)} = v^{(1)} - \langle v_2, v^{(1)} \rangle v_2 \ .$$

The vector $v^{(2)}$ is orthogonal to v_1 and v_2. Also,

$$span\,\{v, v_1, v_2\} = span\,\{v^{(2)}, v_1, v_2\} \ .$$

The arguments can be continued. One obtains that, after normalization of the last vector computed in the loop, the vectors

$$v_1, \ldots, v_j, v_{j+1}$$

form an ONB of K_{j+1}.

3) We have

$$h_{j+1,j} v_{j+1} = v^{(j)} = Av_j - \sum_{i=1}^{j} h_{ij} v_i \ ,$$

thus

$$Av_j = \sum_{i=1}^{j+1} h_{ij} v_i \ .$$

For example,

$$\begin{aligned} Av_1 &= h_{11} v_1 + h_{21} v_2 \\ Av_2 &= h_{12} v_1 + h_{22} v_2 + h_{32} v_3 \end{aligned}$$

These relations then imply that

$$AV_m = V_{m+1} H_m$$

where V_m, V_{m+1}, and H_m are defined above. $\diamond$

15.1.2 Application to Minimization over K_m

Recall that K_m denotes the Krylov subspace,

$$K_m = span\{r_0, Ar_0, \ldots, A^{m-1} r_0\} \subset \mathbb{R}^n \ ,$$

and $v_1, \ldots, v_m$ denotes the computed ONB of K_m. The matrix

$$V_m = (v_1, \ldots, v_m) \in \mathbb{R}^{n \times m}$$

has the j–th column v_j and the equation

$$AV_m = V_{m+1} H_m$$

is been proved in the previous lemma. An arbitrary vector $z \in K_m$ can be written in the form

$$z = V_m y, \quad y \in \mathbb{R}^m \ .$$

We have

$$\begin{aligned} b - A(x_0 + z) &= b - Ax_0 - AV_m y \\ &= r_0 - AV_m y \\ &= r_0 - V_{m+1} H_m y \ . \end{aligned}$$

Set $\beta = |r_0|$; then we have

$$r_0 = \beta v_1 = \beta V_{m+1} e^1 \quad \text{where} \quad e^1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{m+1} \ .$$

Therefore,

$$r_0 - V_{m+1} H_m y = V_{m+1} (\beta e^1 - H_m y) \ .$$

Since the columns of V_{m+1} are orthonormal, we obtain

$$\begin{aligned} |b - A(x_0 + z)| &= |r_0 - V_{m+1} H_m y| \\ &= |\beta e^1 - H_m y| \quad \text{where} \quad z = V_m y \quad \text{and} \quad \beta = |r_0| \ . \end{aligned}$$

In other words, minimizing the norm

$$|b - A(x_0 + z)|, \quad z \in K_m \ , \tag{15.3}$$

is equivalent to minimizing

$$|\beta e^1 - H_m y|, \quad y \in \mathbb{R}^m \ . \tag{15.4}$$

If $\tilde{y}$ minimizes (15.4) then $\tilde{z} = V_m \tilde{y}$ minimizes (15.3). We summarize:

Lemma 15.2 *If $\tilde{y} \in \mathbb{R}^m$ is the least squares solution of the system*

$$H_m y = \beta e^1$$

then $\tilde{z} = V_m \tilde{y} \in K_m$ solves the minimization problem (15.1).

One obtains the vector

$$x_{app} = x_0 + \tilde{z} \in \mathbb{R}^n$$

as approximate solution of the system $Ax = b$. The residual of x_{app} is

$$b - A(x_0 + \tilde{z}) = b - Ax_0 - A\tilde{z} = r_0 - A\tilde{z} \ .$$

If the norm of this residual is small enough, one can accept x_{app} as approximation to x_{ex}. If the norm of the residual is too large, one can either increase m or restart GMRES with x_0 replaced by $x_{app} = x_0 + \tilde{z}$.

Remark 1: The system $H_m y = \beta e^1$ has $m+1$ equations for the unknown vector $y \in \mathbb{R}^m$. The matrix $H_m \in \mathbb{R}^{(m+1)\times m}$ has upper Hessenberg form. One can obtain the least squares solution $\tilde{y} \in \mathbb{R}^m$ of the system $H_m y = \beta e^1$ by solving the normal equations

$$H_m^T H_m y = \beta H_m^T e^1 \ .$$

If the matrix $H_m^T H_m$ is ill–conditioned (which is often the case), one can use the QR–factorization of H_m or one can apply Householder reflectors. Special algorithms have been developed which take into account that H_m has upper Hessenberg form. However, if $m << n$, the numerical work to compute the least squares solution of the system $H_m y = \beta e^1$ is often negligible compared with the work to compute the vectors Av_j in the Arnoldi process.

Remark 2: The following will be used in the next section. Let

$$K_{m,\mathbb{C}} = span_{\mathbb{C}}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}$$

denote the span of the vectors $r_0, Ar_0, \ldots, A^{m-1}r_0$ in $\mathbb{C}^m$. Thus $K_{m,\mathbb{C}}$ consists of all vectors in $\mathbb{C}^n$ of the form

$$z = \alpha_1 r_0 + \ldots + \alpha_m A^{m-1} r_0 \quad \text{where} \quad \alpha_j \in \mathbb{C} \ .$$

Let $\tilde{z} \in K_{m,\mathbb{C}}$ satisfy

$$|b - A(x_0 + \tilde{z})| < |b - A(x_0 + z)| \quad \text{for all} \quad z \in K_{m,\mathbb{C}}, \quad z \neq \tilde{z} \ .$$

For $z \in K_{m,\mathbb{C}}$ we have $z = z_{Re} + iz_{Im}$ with

$$\begin{aligned} z_{Re} &= \beta_1 r_0 + \ldots + \beta_m A^{m-1} r_0 \\ z_{Im} &= \gamma_1 r_0 + \ldots + \gamma_m A^{m-1} r_0 \end{aligned}$$

where $\beta_j, \gamma_j \in \mathbb{R}$. Therefore,

$$|b - A(x_0 + z)|^2 = |b - A(x_0 + z_{Re})|^2 + Az_{im}|^2 \ .$$

This shows that taking the minimum of

$$|b - A(x_0 + z)|$$

over $z \in K_{m,\mathbb{C}}$ equals the minimum of $|b - A(x_0 + z)|$ over $z \in K_m$. We will use this in the proof of Theorem 15.1 below.

15.2 Error Estimates

A simple pretransformation. Consider a system $Ax = b$ where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ are given and $x_{ex} \in \mathbb{R}^n$ is the unknown exact solution. Let x_0 denote a known approximation of x_{ex}, which we use as a starting point in GMRES. The GMRES process then computes an approximate solution $\tilde{z} = \tilde{z}(x_0, m)$, which depends on x_0 and on the number m of steps in GMRES. To simplify matters, let us first eliminate the dependency of $\tilde{z}$ on x_0. We introduce a new unknown $\tilde{x}$ by writing the unknown x in the system $Ax = b$ as

$$x = x_0 + \tilde{x} \ .$$

The system $Ax = b$ becomes

$$Ax_0 + A\tilde{x} = b \quad \text{or} \quad A\tilde{x} = \tilde{b} \quad \text{with} \quad \tilde{b} = b - Ax_0 \ .$$

Dropping the tilde (˜) notation in the system $A\tilde{x} = \tilde{b}$, one obtains the system

$$Ax = b$$

where $x_0 = 0$ is now the best know approximate solution.

A bound for the residual. GMRES applied to $Ax = b$ with $x_0 = 0$ computes the vector

$$\tilde{z} \in K_{m,\mathbb{C}} = span_{\mathbb{C}}\{b, Ab, \ldots, A^{m-1}b\}$$

which minimizes

$$|b - Az| \quad \text{for} \quad z \in K_{m,\mathbb{C}} \ .$$

We want to derive a bound for the residual error $|b - A\tilde{z}|$.

Let P_j denote the vector space of all complex polynomials $p(\lambda)$ of degree $\leq j$. Since

$$K_{m,\mathbb{C}} = span_{\mathbb{C}}\{b, Ab, \ldots, A^{m-1}b\}$$

the vectors $z \in K_{m,\mathbb{C}}$ can be written as

$$z = p(A)b, \quad p \in P_{m-1} \ ,$$

and the norm of their residual is

$$|b - Ap(A)b|, \quad p \in P_{m-1} \ .$$

We can write

$$b - Ap(A)b = q(A)b \quad \text{where} \quad q \in P_m, \quad q(0) = 1 \ .$$

One then obtains that the vector $\tilde{z} = \tilde{p}(A)b$, which is computed by GMRES after m steps, satisfies

$$|b - A\tilde{z}| = \min\{|q(A)b| \ : \ q \in P_m, \ q(0) = 1\} \ . \tag{15.5}$$

In other words, if one considers the expression

$$|q(A)b| \ ,$$

where q varies over all complex polynomials of degree less than or equal to m satisfying $q(0) = 1$, then the minimal value of the expression $|q(A)b|$ equals the norm of the residual of the GMRES approximation $\tilde{z}$.

To better understand the expression on the right–hand side of (15.5), assume that A is diagonalizable and

$$T^{-1}AT = \Lambda \ .$$

We have $A = T\Lambda T^{-1}$ and $q(A) = Tq(\Lambda)T^{-1}$. Therefore,

$$|q(A)b| \leq |T||T^{-1}||b||q(\Lambda)| \ .$$

Here

$$|q(\Lambda)| = \max_{\lambda \in \sigma(A)} |q(\lambda)| \ .$$

This yields the estimate

$$|b - A\tilde{z}| \leq |T||T^{-1}||b| \min_{q \in P_m, q(0)=1} \max_{\lambda \in \sigma(A)} |q(\lambda)| \tag{15.6}$$

for the residual of $\tilde{z}$.

A simple implication is the following:

Theorem 15.1 *Assume that the nonsingular matrix $A \in \mathbb{R}^{n \times n}$ is diagonalizable and has only m distinct eigenvalues. The vector $\tilde{z} \in K_m$ computed by GMRES in m steps solves the system,*

$$A\tilde{z} = b \ .$$

Proof: Let $\lambda_1, \ldots, \lambda_m$ denote the m distinct eigenvalues of A and let

$$q(\lambda) = \frac{(\lambda_1 - \lambda) \cdots (\lambda_m - \lambda)}{\lambda_1 \cdots \lambda_m} \ . \tag{15.7}$$

Then $q \in P_m, q(0) = 1$, and

$$\max_{\lambda\in\sigma(A)} |q(\lambda)| = 0 \ .$$

The equation $A\tilde{z} = b$ follows from (15.6). $\diamond$

The following is plausible, but not precise: If A has only m clusters of eigenvalues, then the vector $\tilde{z} \in K_m$ computed by GMRES in m steps will be a good approximation to the exact solution of the system $Ax = b$. The reason is that the number

$$\min_{q\in P_m, q(0)=1} \max_{\lambda\in\sigma(A)} |q(\lambda)|$$

will be small.

Somewhat more precisely, assume that there are m complex numbers $\lambda_1, \ldots, \lambda_m$ and $\varepsilon > 0$ so that all eigenvalues λ of A lie in the union of disks

$$\cup_{j=1}^m D(\lambda_j, \varepsilon) \ .$$

If all numbers λ_j are $\mathcal{O}(1)$ and are bounded away from the origin, then the polynomial $q(\lambda)$ defined in (15.7) satisfies

$$\max_{\lambda\in\sigma(A)} |q(\lambda)| = \mathcal{O}(\varepsilon) \ .$$

A further assumption is that the condition number $|T||T^{-1}|$ in the estimate (15.6) is not very large.

15.3 Research Project: GMRES and Preconditioning

Idea: Use GMRES with preconditioning. The preconditioning should lead to a matrix P_1AP_2 with a few clusters of eigenvalues.

Preconditioning 1: Choose a simple invertible matrix P_1 and replace the system $Ax = b$ by

$$P_1Ax = P_1b \ .$$

Preconditioning 2: Choose a simple invertible matrix P_2 and replace the system $P_1Ax = P_1b$ by

$$P_1AP_2y = P_1b \quad \text{where} \quad x = P_2y \ .$$

Apply GMRES to the system $P_1AP_2y = P_1b$ and obtain an approximate solution $\tilde{y}$. Then set $\tilde{x} = P_2\tilde{y}$.

Difficulty: One wants to choose the preconditioners P_1 and P_2 so that the eigenvalues of P_1AP_2 get clustered into a few clusters. But it is not well understood how the choices of P_1 and P_2 achieve such clustering.

16 The Courant–Fischer Min–Max Theorem

The eigenvalues of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ can be characterized by extremal properties of the quadratic form x^*Ax on the unit sphere in $\mathbb{C}^n$. This characterization is useful if one wants to understand how eigenvalues change if A is perturbed by a Hermitian matrix.

16.1 The Min–Max Theorem

Let $A \in \mathbb{C}^{n\times n}$ denote a Hermitian matrix. We know that $\mathbb{C}^n$ has an ONB $u_1,\ldots,u_n$ of eigenvectors of A and that the eigenvalues λ_j are real. We may assume the λ_j to be ordered:

$$Au_j = \lambda_j u_j, \quad \lambda_n \le \ldots \le \lambda_1 \ .$$

The eigenvalues are not necessarily distinct. Each eigenvalue is listed by its multiplicity. (Recall that the algebraic and geometric multiplicities are the same since A is similar to a diagonal matrix.)

Let

$$S = \{x \ : \ x \in \mathbb{C}^n, \ |x| = 1\}$$

denote the unit sphere in $\mathbb{C}^n$. We wish to characterize the eigenvalues λ_j of A in terms of the quadratic form

$$\phi(x) = x^*Ax, \quad x \in S \ .$$

Since

$$(x^*Ax)^* = x^*Ax$$

we first note that $\phi(x)$ is real valued.

Let

$$U = (u_1,\ldots,u_n), \quad \Lambda = diag(\lambda_1,\ldots,\lambda_n) \ .$$

Then U is unitary and

$$U^*AU = \Lambda, \quad A = U\Lambda U^* \ .$$

This follows from

$$AU = (\lambda_1 u_1,\ldots,\lambda_n u_n) = U\Lambda \ .$$

If $x \in S$ then $y := U^*x \in S$ and

$$\begin{aligned}\phi(x) &= x^*U\Lambda U^*x \\ &= y^*\Lambda y \\ &= \sum_{j=1}^n \lambda_j |y_j|^2 \ .\end{aligned}$$

Note that

$$Ue^j = u_j, \quad e^j = U^* u_j \ .$$

Thus, if $x = u_j$ then $y = U^* x = U^* u_j = e^j$. In particular,

$$\phi(u_j) = \lambda_j \ . \tag{16.1}$$

We obtain:

Lemma 16.1 *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian and let $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$ denote the eigenvalues of A. Let $Au_j = \lambda_j u_j$ where the vectors $u_1, \ldots, u_n$ form an ONB of $\mathbb{C}^n$. Under these assumptions, the quadratic form $\phi(x) = x^* Ax$ satisfies:*

$$\lambda_n \leq \phi(x) \leq \lambda_1, \quad x \in S \ .$$

Furthermore,

$$\lambda_1 = \max_{x\in S} \phi(x) = \phi(u_1) \ .$$

and

$$\lambda_n = \min_{x\in S} \phi(x) = \phi(u_n) \ .$$

We now wish to characterize the other eigenvalues by extremal properties of ϕ. We carry this out for λ_2 (assuming $n \geq 2$).

Let $V \subset \mathbb{C}^n$ denote a subspace of dimension $dim\, V = n - 1$. We first claim that

$$\max_{x\in V\cap S} \phi(x) \geq \lambda_2 \ . \tag{16.2}$$

To show this, set

$$Y_2 = span\{e^1, e^2\} \ .$$

Since $dim\, U^*(V) = n - 1$ and $dim\, Y_2 = 2$ the intersection $Y_2 \cap U^*(V)$ contains a non–zero vector y, and we may assume that $|y| = 1$. Setting $x = Uy$ we have $x \in V \cap S$ and

$$\begin{aligned} \phi(x) &= \sum_{j=1}^n \lambda_j |y_j|^2 \\ &= \lambda_1 |y_1|^2 + \lambda_2 |y_2|^2 \quad \text{(since } y \in Y_2) \\ &\geq \lambda_2 \ . \end{aligned}$$

This proves (16.2).

We now show that there exists a subspace V of dimension $n-1$ so that equality holds in (16.2). To this end, set $V_2 = span\,\{u_1\}^\perp$. If $x \in V_2 \cap S$ then

$$x = \sum_{j=2}^{n} y_j u_j, \quad y = (0, y_2, \ldots, y_n)^T = U^* x \ .$$

We have

$$\phi(x) = \sum_{j=2}^{n} \lambda_j |y_j|^2 \le \lambda_2 \ .$$

Since $x \in V_2 \cap S$ was arbitrary, this shows that

$$\max_{x \in V_2 \cap S} \phi(x) \le \lambda_2 \ .$$

Furthermore, $u_2 \in V_2 \cap S$ (since u_2 is orthogonal to u_1) and $\phi(u_2) = \lambda_2$. See (16.1).

Therefore,

$$\max_{x \in V_2 \cap S} \phi(x) = \phi(u_2) = \lambda_2 \ .$$

Thus, we have shown the following min–max formula for λ_2:

Lemma 16.2 *Under the assumptions of Lemma 16.1 the quadratic form $\phi(x) = x^* A x$ satisfies:*

$$\min_{dim V = n-1} \ \max_{x \in V \cap S} \phi(x) = \lambda_2 \ .$$

Here the minimum is taken over all subspaces V of $\mathbb{C}^n$ that have dimension $n-1$.

The min–max is attained for

$$V = V_2 = span\,\{u_1\}^\perp, \quad x = u_2 \ .$$

We now prove a corresponding max–min characterization of λ_2. Let V denote a subspace of $\mathbb{C}^n$ of dimension 2. If

$$\tilde{Y}_2 = span\,\{e^1\}^\perp$$

then $\tilde{Y}_2$ has dimension $n-1$. Therefore, there exists a non–zero vector $y \in \tilde{Y}_2 \cap U^*(V)$, and we may assume that $y \in S$. If $x = Uy$ then $x \in V \cap S$ and

$$\begin{aligned} \phi(x) &= \sum_{j=2}^{n} \lambda_j |y_j|^2 \\ &\le \lambda_2 \end{aligned}$$

This shows that

$$\min_{x \in V \cap S} \phi(x) \le \lambda_2 \tag{16.3}$$

whenever V is a subspace of $\mathbb{C}^n$ of dimension 2. Next, consider

$$\tilde{V}_2 = span\{u_1, u_2\} \ .$$

If $x \in \tilde{V}_2 \cap S$ then

$$x = y_1 u_1 + y_2 u_2 \quad \text{(where } y_j \in \mathbb{C}\text{)}$$

and

$$\begin{aligned} \phi(x) &= \lambda_1 |y_1|^2 + \lambda_2 |y_2|^2 \\ &\geq \lambda_2 \ . \end{aligned}$$

Thus, since $x \in \tilde{V}_2 \cap S$ was arbitrary,

$$\min_{x \in \tilde{V}_2 \cap S} \phi(x) \geq \lambda_2 \ .$$

Setting $x = u_2$ we see that

$$\min_{x \in \tilde{V}_2 \cap S} \phi(x) = \lambda_2$$

where the minimum is attained at $x = u_2$. Together with (16.3) we have shown the following max–min formula for λ_2:

Lemma 16.3 *Under the assumptions of Lemma 16.1 the quadratic form* $\phi(x) = x^* Ax$ *satisfies:*

$$\max_{dim V = 2} \quad \min_{x \in V \cap S} \phi(x) = \lambda_2 \ .$$

Here the maximum is taken over all subspaces V *of* $\mathbb{C}^n$ *which have dimension 2.*

The max–min is attained for

$$V = \tilde{V}_2 = span\,\{u_1, u_2\}, \quad x = u_2 \ .$$

It is not difficult to generalize the results of Lemma 16.2 and Lemma 16.3 and obtain the following characterizations of λ_j:

Theorem 16.1 *Let* $A \in \mathbb{C}^{n \times n}$ *be a Hermitian matrix with eigenvalues* $\lambda_1 \geq \ldots \geq \lambda_n$ *and orthonormal eigenvectors* u_j, $Au_j = \lambda_j u_j$. *Let* $\phi(x) = x^* Ax, |x| = 1$.

We have

$$\lambda_j = \min_{dim V = n+1-j} \quad \max_{x \in V \cap S} \phi(x)$$

and

$$\lambda_j = \max_{dim V = j} \quad \min_{x \in V \cap S} \phi(x) \ .$$

16.2 Eigenvalues of Perturbed Hermitian Matrices

Let $A \in \mathbb{C}^{n\times n}$ denote a Hermitian matrix with eigenvalues

$$\lambda_1 \geq \ldots \geq \lambda_n, \quad Au_j = \lambda_j u_j$$

where $u_1, \ldots, u_n \in \mathbb{C}^n$ are ON. Let $E \in \mathbb{C}^{n\times n}$ denote any Hermitian matrix with

$$|E| \leq \varepsilon$$

and consider the perturbed matrix

$$B = A + E$$

with eigenvalues

$$\beta_1 \geq \ldots \geq \beta_n \ .$$

We claim that

$$\lambda_j - \varepsilon \leq \beta_j \leq \lambda_j + \varepsilon, \quad j = 1, \ldots, n \ . \tag{16.4}$$

First note that

$$x^*Bx = x^*Ax + x^*Ex \quad \text{and} \quad |x^*Ex| \leq \varepsilon \quad \text{for} \quad x \in S \ ,$$

thus

$$x^*Ax - \varepsilon \leq x^*Bx \leq x^*Ax + \varepsilon \quad \text{for all} \quad x \in S \ .$$

To prove (16.4), we take $j = 2$ for simplicity.

As in the previous section, let

$$V_2 = span\,\{u_1\}^\perp \quad \text{and} \quad \tilde{V}_2 = span\,\{u_1, u_2\} \ .$$

We have

$$\begin{aligned}
\beta_2 &= \min_{dim V = n-1} \ \max_{x \in V \cap S} x^*Bx \\
&\leq \max_{x \in V_2 \cap S} x^*Bx \\
&\leq \max_{x \in V_2 \cap S} \Big(x^*Ax + \varepsilon \Big) \\
&= \lambda_2 + \varepsilon \ .
\end{aligned}$$

Similarly,

$$
\begin{aligned}
\beta_2 &= \max_{dimV=2} \min_{x \in V \cap S} x^* B x \\
&\geq \min_{x \in \tilde{V}_2 \cap S} x^* B x \\
&\geq \min_{x \in \tilde{V}_2 \cap S} \Big(x^* A x - \varepsilon \Big) \\
&= \lambda_2 - \varepsilon \ .
\end{aligned}
$$

Remark: An inclusion like (16.4) does not hold if a general matrix A with real eigenvalues is perturbed. For example, let

$$A = \begin{pmatrix} 1 & 10^{10} \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 10^{10} \\ \varepsilon & 2 \end{pmatrix} \ .$$

The eigenvalues of B are the zeros of the polynomial

$$p_B(z) = z^2 - 3z + 2 - \varepsilon * 10^{10} \ ,$$

i.e.,

$$\beta_{1,2} = \frac{3}{2} \pm \sqrt{\frac{1}{4} + \varepsilon * 10^{10}} \ .$$

For $\varepsilon = 0$ we obtain the eigenvalues of A,

$$\lambda_1 = 2, \quad \lambda_2 = 1 \ .$$

If

$$10^{-10} << \varepsilon << 1$$

then the eigenvalues of B are

$$\beta_{1,2} \sim \frac{3}{2} \pm \sqrt{\varepsilon} * 10^5 \ .$$

We see that (16.4) does not hold at all.

16.3 Eigenvalues of Submatrices

Let $B \in \mathbb{C}^{(n+1)\times(n+1)}$ denote a Hermitian matrix with eigenvalues

$$\beta_1 \geq \ldots \geq \beta_{n+1} \ .$$

Partition B as

$$B = \begin{pmatrix} A & c \\ c^* & \alpha \end{pmatrix} \quad \text{where} \quad A \in \mathbb{C}^{n \times n} \ . \tag{16.5}$$

The matrix A is called the leading principal submatrix of order n of B. Let

$$\lambda_1 \geq \ldots \geq \lambda_n$$

denote the eigenvalues of A. We claim that these are interlaced with those of B, i.e.,

$$\beta_1 \geq \lambda_1 \geq \beta_2 \geq \ldots \geq \lambda_n \geq \beta_{n+1} \; . \tag{16.6}$$

To prove this, let $U \in \mathbb{C}^{n \times n}$ be unitary with

$$U^* A U = \Lambda = diag(\lambda_1, \ldots, \lambda_n) \; .$$

Set

$$\tilde{U} = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\tilde{U}^* B \tilde{U} = \tilde{B} \; .$$

Then, since $\tilde{U}$ is unitary, the matrix $\tilde{B}$ also has the eigenvalues β_j.

To show that $\beta_j \geq \lambda_j$ we apply Theorem 16.1 to the matrix $\tilde{B}$. Let

$$V_j = span\,\{e^1, \ldots, e^j\} \subset \mathbb{C}^{n+1} \quad \text{where} \quad 1 \leq j \leq n \; .$$

If $x \in V_j \cap S$ then

$$x^* \tilde{B} x = \sum_{i=1}^{j} \lambda_i |x_i|^2 \geq \lambda_j \; .$$

This shows that

$$\min_{x \in V_j \cap S} x^* \tilde{B} x \geq \lambda_j \; .$$

Also, by Theorem 16.1,

$$\begin{aligned} \beta_j &= \max_{dim V = j} \; \min_{x \in V \cap S} x^* \tilde{B} x \\ &\geq \min_{x \in V_j \cap S} x^* \tilde{B} x \\ &\geq \lambda_j \end{aligned}$$

Next we will prove that $\beta_j \leq \lambda_{j-1}$. Let

$$W_j = span\,\{e^{j-1}, e^j, \ldots, e^n\} \subset \mathbb{C}^{n+1} \quad \text{where} \quad 2 \leq j \leq n+1 \; .$$

Note that W_j has dimension $n + 2 - j$. If $x \in W_j \cap S$ then

$$x^* \tilde{B} x = \sum_{i=j-1}^{n} \lambda_i |x_i|^2 \leq \lambda_{j-1} \; ,$$

showing that

$$\max_{x \in W_j \cap S} x^* \tilde{B} x \le \lambda_{j-1} \; .$$

Also, by Theorem 16.1,

$$\begin{aligned} \beta_j &= \min_{dim V = n+2-j} \; \max_{x \in V \cap S} x^* \tilde{B} x \\ &\le \max_{x \in W_j \cap S} x^* \tilde{B} x \\ &\le \lambda_{j-1} \end{aligned}$$

We have shown the following result.

Lemma 16.4 *Let B denote an $(n+1) \times (n+1)$ Hermitian matrix of the form (16.5) with eigenvalues $\beta_1 \ge \ldots \ge \beta_{n+1}$. Let A denote the leading principal submatrix of order n of B with eigenvalues $\lambda_1 \ge \ldots \ge \lambda_n$. Then the λ_j are interlaced with the eigenvalues of B as in (16.6).*

An $n \times n$ matrix A is called a principal submatrix of order n of $B \in \mathbb{C}^{(n+1)\times(n+1)}$ if A is obtained from B by deleting row j and column j of B, for some $j \in \{1, \ldots, n+1\}$. We claim that the eigenvalues of A interlace with those of B, as stated in (16.6).

To show this, consider the permutation $\sigma \in S_{n+1}$ which interchanges j and $n+1$ and leaves all other elements of $\{1, 2, \ldots, n+1\}$ fixed. If P is the corresponding permutation matrix, then $P^{-1} = P^T = P$ and

$$P^T B P$$

has the same eigenvalues as B. In addition, the matrix A is the leading principal submatrix of order n of $P^T B P$. The claim follows from the previous lemma.

Example: Let

$$B = \begin{pmatrix} a & c \\ \bar{c} & b \end{pmatrix} \quad \text{where} \quad a, b \in \mathbb{R} \; .$$

Lemma 16.4 says that

$$\beta_1 \ge a \ge \beta_2$$

if $\beta_1 \ge \beta_2$ are the eigenvalues of B. It is easy to check this directly. We have

$$det(B - \beta I) = (a - \beta)(b - \beta) - |c|^2 \; .$$

The eigenvalues β_j of B are the solutions of

$$(\beta - a)(\beta - b) = |c|^2 \; .$$

The inequalities

$$\beta_2 \le a, b \le \beta_1$$

follow since the parabola $(\beta - a)(\beta - b)$ intersects the line $\beta \equiv |c|^2$ at β–values outside the interval between a and b.

17 Introduction to Control Theory

Let $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}$ and consider the initial value problem

$$x'(t) = Ax(t) + Bu(t), \quad x(0) = 0 \ , \tag{17.1}$$

for $x(t) \in \mathbb{R}^n$. We think of $x(t)$ as the state vector of a system and of $u(t) \in \mathbb{R}^m$ as a control function, which we can choose to control the evolution of the state $x(t)$.

A first question of control theory is this: Given any time $t_1 > 0$ and any state $x^{(1)} \in \mathbb{R}^n$, under what assumptions on the pair of matrices A and B does there exist a control function $u(t)$ so that the solution $x(t)$ of the IVP (17.1) satisfies $x(t_1) = x^{(1)}$? In other words, under what assumptions can we control the system so that the state $x(t)$ moves from $x(0) = 0$ to any given state $x^{(1)}$? This question leads to the concept of controllability.

It turns out that the assumption $x(0) = 0$ is not restrictive and the length of the time interval $t_1 > 0$ is also unimportant.

17.1 Controllability

Definition: *Fix any $t_1 > 0$. The system (17.1) is called controllable in the interval $0 \le t \le t_1$ if for any $x^{(1)} \in \mathbb{R}^n$ there exists a control function*

$$u : [0, t_1] \to \mathbb{R}^m, \quad u \in C \ ,$$

so that the solution $x(t)$ of the system (17.1) satisfies $x(t_1) = x^{(1)}$.

If $B = 0$ then, obviously, the system (17.1) is not controllable. Therefore, in the following, we always assume $B \neq 0$.

Define the matrix

$$M_n = (B, AB, A^2B, \dots, A^{n-1}B) \in \mathbb{R}^{n\times(mn)} \ .$$

Note that every part

$$A^jB$$

has the same dimensions which B has, i.e., A^jB has m columns and n rows.

The following theorem implies that the controllability of the system (17.1) does not depend on the time interval $0 \le t \le t_1$.

Theorem 17.1 *The system (17.1) is controllable in $0 \le t \le t_1$ if and only if $\operatorname{rank} M_n = n$.*

Example 1: Let

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

We have

$$AB = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} .$$

By Theorem 17.1, the corresponding system (17.1) is controllable.

Example 2: Let

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

We have

$$AB = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} .$$

By Theorem 17.1, the corresponding system (17.1) is not controllable.

Before proving Theorem 17.1, we first prove the following lemma.

Lemma 17.1 *For $k = 1, 2, \ldots$ set*

$$M_k = (B, AB, A^2B, \ldots, A^{k-1}B) \in \mathbb{R}^{n \times (mk)}$$

and let

$$r_k = rank\, M_k \ .$$

If $r_k = r_{k+1}$ then $r_{k+1} = r_{k+2}$.

Proof: First recall that

$$r_k = rank\, M_k = dim\, range\, M_k \quad \text{for} \quad k = 1, 2, \ldots$$

We have

$$M_{k+1} = (M_k, A^k B)$$

and the assumption $r_{k+1} = r_k$ implies that every column of $A^k B$ lies in the range of M_k. If $(A^k B)_j$ denotes the j–th column of $A^k B$ then there exists a vector $c_j \in \mathbb{R}^{mk}$ with

$$(A^k B)_j = M_k c_j \quad \text{for} \quad j = 1, \ldots, m \ .$$

Set

$$C = (c_1, c_2, \ldots, c_m) \in \mathbb{R}^{(mk) \times m}$$

and obtain that

$$
\begin{aligned}
A^k B &= M_k C \\
A^{k+1} B &= A M_k C \\
&= (AB, A^2B, \ldots, A^kB)C \\
&= (B, AB, A^2B, \ldots, A^kB)\tilde{C} \\
&= M_{k+1}\tilde{C}
\end{aligned}
$$

with

$$
\tilde{C} = \begin{pmatrix} 0 \\ C \end{pmatrix} \quad \text{where} \quad 0 \in \mathbb{R}^{m\times m} \ .
$$

Thus the columns of $A^{k+1}B$ lie in

$$
range\, M_{k+1} = range\, M_k \ .
$$

Therefore,

$$
r_{k+2} = r_{k+1} = r_k \ .
$$

$\diamond$

Proof of Theorem 17.1: For M_k and $r_k = rank M_k$ we use the notation of the previous lemma.

1) Assume first that $rank\, M_n = n$. In the following, we will construct a control function $u(t)$ for which the solution $x(t)$ of the IVP (17.1) satisfies $x(t_1) = x^{(1)}$. We will use that the solution of (17.1) is given by

$$
x(t) = \int_0^t e^{A(t-s)}\, Bu(s)\, ds \ . \tag{17.2}
$$

Define the matrix

$$
K = \int_0^{t_1} e^{-At} BB^T e^{-A^T t}\, dt \in \mathbb{R}^{n\times n} \ .
$$

It is clear that $K = K^T$. We will show that K is positive definite. Set

$$
C(t) = B^T e^{-A^T t} \in \mathbb{R}^{m\times n} \ .
$$

If $a \in \mathbb{R}^n$ is arbitrary, we have

$$
\begin{aligned}
a^T K a &= \int_0^{t_1} a^T C(t)^T C(t) a\, dt \\
&= \int_0^{t_1} |C(t)a|^2\, dt
\end{aligned}
$$

This shows that $a^T K a \geq 0$ and if

$$
a^T K a = 0
$$

then

$$a^T C(t)^T = a^T e^{-At} B = 0 \quad \text{for} \quad 0 \le t \le t_1 \; .$$

As above, let $a \in \mathbb{R}^n$ be arbitrary and define the vector function

$$\phi(t) = a^T e^{-At} B \quad \text{for} \quad 0 \le t \le t_1 \; .$$

Note that $\phi(t)$ is a row vector of dimension m. If one assumes that $a^T K a = 0$ then

$$\phi(t) = a^T e^{-At} B = 0 \quad \text{for} \quad 0 \le t \le t_1 \; .$$

Therefore,

$$\begin{aligned} \phi(0) &= a^T B = 0 \\ \phi'(t) &= -a^T e^{-At} AB = 0 \\ \phi'(0) &= -a^T AB = 0 \\ \phi''(t) &= a^T e^{-At} A^2 B = 0 \end{aligned}$$

etc.

Setting $t = 0$ one obtains that $a^T K a = 0$ implies that

$$\begin{aligned} a^T B &= 0 \\ a^T AB &= 0 \\ a^T A^2 B &= 0 \end{aligned}$$

etc. Therefore,

$$a^T M_n = a^T (B, AB, A^2 B, \ldots, A^{n-1} B) = 0 \; .$$

Since, by assumption, M_n has n linearly independent columns it follows that $a = 0$. Thus we have shown that $a^T K a > 0$ if $a \neq 0$, i.e.,

$$K = K^T > 0 \; .$$

Set

$$u(t) = B^T e^{-A^T t} K^{-1} e^{-At_1} x^{(1)} \in \mathbb{R}^m \; . \tag{17.3}$$

Then the solution of the IVP (17.1) satisfies

$$\begin{aligned} x(t_1) &= \int_0^{t_1} e^{A(t_1 - s)} B u(s) \, ds \\ &= e^{At_1} \Big(\int_0^{t_1} e^{-As} B B^T e^{-A^T s} ds \Big) K^{-1} e^{-At_1} x^{(1)} \\ &= e^{At_1} K K^{-1} e^{-At_1} x^{(1)} \\ &= x^{(1)} \end{aligned}$$

This proves that the control function $u(t)$ given in (17.3) leads to a solution $x(t)$ of the IVP (17.1) with $x(t_1) = x^{(1)}$.

2) Assume that $rank\, M_n < n$, i.e., $r_n < n$. Since

$$1 \le r_1 \le r_2 \le \ldots \le r_n < n$$

there exists $k \in \{1, \ldots, n-1\}$ with

$$r_k = r_{k+1} < n \; .$$

Using the above Lemma we conclude that M_n is a **strict** subspace of $\mathbb{R}^n$ and

$$range\, M_j = range M_n \neq \mathbb{R}^n \quad \text{for all} \quad j \ge n \; . \tag{17.4}$$

For the solution $x(t)$ of the IVP (17.1) we have

$$\begin{aligned} x(t_1) &= \int_0^{t_1} e^{A(t_1-t)} Bu(t)\, dt \\ &= \int_0^{t_1} \sum_{j=0}^{\infty} \frac{1}{j!} A^j B (t_1 - t)^j u(t)\, dt \\ &= \sum_{j=0}^{\infty} A^j B \alpha_j \end{aligned}$$

with

$$\alpha_j = \frac{1}{j!} \int_0^{t_1} (t_1 - t)^j u(t)\, dt \in \mathbb{R}^m \; .$$

Because of (17.4) we have

$$\sum_{j=0}^{J} A^j B \alpha_j \in range M_n \quad \text{for all} \quad J \; ,$$

and, since M_n is a closed subspace of $\mathbb{R}^n$, we obtain that

$$x(t_1) = \lim_{J \to \infty} \sum_{j=0}^{J} A^j B \alpha_j \in M_n$$

for every control function $u(t)$. This proves: If $range\, M_n$ is a strict subspace of $\mathbb{R}^n$, then the system $x' = Ax + Bu$ is not controllable in $0 \le t \le t_1$. $\diamond$

17.2 General Initial Data

Consider the IVP

$$x'(t) = Ax(t) + Bu(t), \quad x(0) = x^{(0)} \; , \tag{17.5}$$

where $x^{(0)} \in \mathbb{R}^n$ is given. Also, let $x^{(1)} \in \mathbb{R}^n$ be given and let $t_1 > 0$ be fixed. In the following theorem we show that the assumption $x(0) = 0$ in (17.1) is not restrictive.

Theorem 17.2 *Assume that the system (17.1) is controllable, i.e., rank* $M_n = n$*. (See the previous theorem.) Then there exists a control function* $u(t)$ *so that the solution of (17.5) satisfies* $x(t_1) = x^{(1)}$*.*

Proof: By the previous theorem, there exists a control function $u(t)$ so that the solution $y(t)$ of the IVP

$$y'(t) = Ay(t) + Bu(t), \quad y(0) = 0 \ ,$$

satisfies

$$y(t_1) = x^{(1)} - e^{At_1}x^{(0)} \ .$$

Set

$$x(t) = e^{At}x^{(0)} + y(t) \ .$$

Then we have

$$x(0) = x^{(0)} \quad \text{and} \quad x(t_1) = x^{(1)}$$

and

$$\begin{aligned} x'(t) &= Ae^{At}x^{(0)} + y'(t) \\ &= Ae^{At}x^{(0)} + Ay(t) + Bu(t) \\ &= A\Big(e^{At}x^{(0)} + y(t)\Big) + Bu(t) \\ &= Ax(t) + Bu(t) \end{aligned}$$

Therefore, $x(t)$ satisfies the differential equation $x' = Ax + Bu$, the initial condition $x(0) = x^{(0)}$ and the end condition $x(t_1) = x^{(1)}$. $\diamond$

17.3 Control of the Inverted Pendulum

The **standard pendulum equation** is

$$ml\phi'' = -mg\sin\phi \ ,$$

thus

$$\phi'' + \omega^2 \sin\phi = 0 \quad \text{with} \quad \omega^2 = \frac{g}{l} \ .$$

For small ϕ one replaces $\sin\phi$ by ϕ and obtains the linear equation

$$\phi'' + \omega^2\phi = 0$$

with general solution

$$\phi(t) = a\cos(\omega t) + b\sin(\omega t) \ .$$

The **inverted pendulum equation** is

$$ml\phi'' = mg \sin \phi \ .$$

Here ϕ is the angle between the pendulum and the upper vertical line. One obtains the equation

$$\phi'' - \omega^2 \sin \phi = 0 \quad \text{with} \quad \omega^2 = \frac{g}{l} \ .$$

Replacing $\sin \phi$ by ϕ yields the linearized equation

$$\phi'' - \omega^2 \phi = 0$$

with general solution

$$\phi(t) = ae^{\omega t} + be^{-\omega t} \ .$$

The exponentially growing term $e^{\omega t}$ makes it clear that the state $\phi = 0$ is unstable, which is physically obvious.

The **controlled inverted pendulum equation** is

$$ml\phi'' = mg \sin \phi - mu'' \cos \phi \ .$$

Here $u = u(t)$ is the position of the base point on the x–axis.

One obtains

$$\phi'' = \omega^2 \sin \phi - \frac{1}{l} u'' \cos \phi \ .$$

Linearization about $\phi = 0$ yields

$$\phi'' = \omega^2 \phi - \frac{1}{l} u'' \ .$$

As a first order system:

$$\begin{pmatrix} \phi \\ \phi' \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ \omega^2 & 0 \end{pmatrix} \begin{pmatrix} \phi \\ \phi' \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} (-\frac{u''}{l})$$

We can apply the general theory with

$$A = \begin{pmatrix} 0 & 1 \\ \omega^2 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \ .$$

One obtains that

$$AB = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

and

$$M_2 = (B|AB) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \ .$$

The linearized system is controllable.

17.4 Derivation of the Controlled Reversed Pendulum Equation via Lagrange

Let

$$\begin{aligned} x(t) &= u(t) + l \sin\phi(t) \\ y(t) &= l \cos\phi(t) \end{aligned}$$

denote the coordinates of the mass point. One obtains

$$\begin{aligned} x' &= u' + l\phi' \cos\phi \\ y' &= -l\phi' \sin\phi \end{aligned}$$

The kinetic energy of the point mass m at $\Big(x(t), y(t)\Big)$ is

$$\begin{aligned} E_{kin} &= \frac{m}{2}\Big(x'^2 + y'^2\Big) \\ &= \frac{m}{2}\Big((u' + l\phi'\cos\phi)^2 + (l\phi'\sin\phi)^2\Big) \\ &= \frac{m}{2}u'^2 + mlu'\phi'\cos\phi + \frac{m}{2}\, l^2\phi'^2 \end{aligned}$$

The potential energy is

$$E_{pot} = mgy = mgl\cos\phi\ .$$

The Lagrange function is

$$\begin{aligned} L(\phi, \phi') &= E_{kin} - E_{pot} \\ &= \frac{m}{2}u'^2 + mlu'\phi'\cos\phi + \frac{m}{2}\, l^2\phi'^2 - mgl\cos\phi \end{aligned}$$

Lagrange's equation is

$$\frac{\partial L}{\partial \phi} - \frac{d}{dt}\frac{\partial L}{\partial \phi'} = 0\ .$$

We have

$$\begin{aligned} \frac{\partial L}{\partial \phi} &= -mlu'\phi'\sin\phi + mgl\sin\phi \\ \frac{\partial L}{\partial \phi'} &= mlu'\cos\phi + ml^2\phi' \\ \frac{d}{dt}\frac{\partial L}{\partial \phi'} &= mlu''\cos\phi - mlu'\phi'\sin\phi + ml^2\phi'' \end{aligned}$$

The Lagrange equation

$$\frac{\partial L}{\partial \phi} = \frac{d}{dt}\frac{\partial L}{\partial \phi'}$$

yields

$$mgl \sin\phi = mlu'' \cos\phi + ml^2\phi'' \ .$$

Dividing by ml^2 yields

$$\phi'' = \frac{g}{l}\sin\phi - \frac{1}{l}u''\cos\phi \ .$$

17.5 The Inverted Double Pendulum

We derive the equations of motion using the Lagrange function

$$L = L(t, \phi_1, \phi_1', \phi_2, \phi_2') \ .$$

We have

$$\begin{aligned}
x_1 &= u + l_1 \sin\phi_1 \\
y_1 &= l_1 \cos\phi_1 \\
x_2 &= x_1 + l_2 \sin\phi_2 \\
y_2 &= y_1 + l_2 \cos\phi_2
\end{aligned}$$

with time derivatives

$$\begin{aligned}
x_1' &= u' + l_1 i\phi_1' \cos\phi_1 \\
y_1' &= -l_1\phi_1' \sin\phi_1 \\
x_2' &= x_1' + l_2\phi_2' \cos\phi_2 \\
y_2' &= y_1' - l_2\phi_2' \sin\phi_2
\end{aligned}$$

The kinetic energy is

$$E_{kin} = \frac{m_1}{2}\left(x_1'^2 + y_1'^2\right) + \frac{m_2}{2}\left(x_2'^2 + y_2'^2\right) \ .$$

The potential energy is

$$E_{pot} = m_1 g y_1 + m_2 g y_2 \ .$$

For the Lagrange function one obtains

$$L = \frac{m_1}{2}\left(x_1'^2 + y_1'^2\right) + \frac{m_2}{2}\left(x_2'^2 + y_2'^2\right) - m_1 g y_1 - m_2 g y_2 \ .$$

The dynamic equations are

$$\begin{aligned}
\frac{d}{dt}\frac{\partial L}{\partial \phi_1'} &= \frac{\partial L}{\partial \phi_1} \\
\frac{d}{dt}\frac{\partial L}{\partial \phi_2'} &= \frac{\partial L}{\partial \phi_2}
\end{aligned}$$

17.6 Optimal Control

Consider the initial–value problem

$$x' = f(x, u), \quad 0 \le t \le t_1, \quad x(0) = x_0 \tag{17.6}$$

where $x(t) \in \mathbb{R}^n$ is the state vector and $u(t) \in \mathbb{R}^m$ is the control function. Here

$$f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$$

is a given smooth function. If the control function $u(t)$ is chosen, then the IVP (17.6) determines the evolution of the state vector $x(t)$ uniquely. We ignore the possibility that $x(t)$ may not exist for $0 \le t \le t_1$.

Let

$$J(u) = \psi(x(t_1)) + \int_0^{t_1} l(x(t), u(t))\, dt$$

denote the so–called objective function. Here

$$\psi : \mathbb{R}^n \to \mathbb{R} \quad \text{and} \quad l : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$$

are given smooth functions. In optimal control theory, one tries to determine a control function $u(t)$ which maximizes the objective function $J(\cdot)$.

Choose any smooth function $\lambda : [0, t_1] \to \mathbb{R}^n$ and define the modified objective function

$$\tilde{J}(u, \lambda) = J(u) - \int_0^{t_1} \lambda(t)^T \Big(x'(t) - f(x(t), u(t)) \Big)\, dt\ .$$

It is clear that

$$\tilde{J}(u, \lambda) = J(u)$$

since $x(t)$ is always assumed to solve $x' = f(x, u)$. A smart choice for $\lambda(t)$ will be made below.

Define the Hamiltonian

$$H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$$

by

$$H(\lambda, x, u) = \lambda^T f(x, u) + l(x, u)$$

and note that

$$H(\lambda(t), x(t), u(t)) - \lambda(t)^T x'(t) = -\lambda(t)^T \Big(x'(t) - f(x(t), u(t)) \Big) + l(x(t), u(t))\ .$$

Therefore,

$$\tilde{J}(u, \lambda) = \psi(x(t_1)) + \int_0^{t_1} \Big(H(\lambda(t), x(t), u(t)) - \lambda(t)^T x'(t) \Big)\, dt\ .$$

Let $u(t)$ be a local maximum of the objective function $J(u)$. With $x(t)$ we always denote the solution of the IVP (17.6).

Let $v(t)$ denote a control function which is a small change of $u(t)$. More precisely, assume that

$$\int_0^{t_1} |u_i(t) - v_i(t)|\,dt \le \varepsilon \quad \text{for} \quad i = 1, 2, \ldots, m \ .$$

Let $x(t) + \delta(t)$ denote the solution of the IVP

$$x' + \delta' = f(x + \delta, v), \quad x(0) = x_0, \quad \delta(0) = 0 \ ,$$

i.e., the change of the control from $u(t)$ to $v(t)$ changes the state function from $x(t)$ to $x(t) + \delta(t)$. Under reasonable assumptions one can show that

$$\max_{0 \le t \le t_1} |\delta(t)| \le C\varepsilon \ .$$

Here $|\cdot|$ denotes the Euclidean vector norm on $\mathbb{R}^n$.

We have

$$\tilde{J}(x+\delta, v) = \psi(x(t_1)+\delta(t_1)) + \int_0^{t_1} H(\lambda(t), x(t)+\delta(t), v(t))\,dt - \int_0^{t_1} \lambda(t)^T (x'(t)+\delta'(t))\,dt \ .$$

The change in the objective function is

$$\begin{aligned}
\Delta\tilde{J} &= \tilde{J}(u+\delta,\lambda) - \tilde{J}(u,\lambda) \\
&= \psi(x(t_1)+\delta(t_1)) - \psi(x(t_1)) + \int_0^{t_1} \Big(H(\lambda, x+\delta, v) - H(\lambda, x, u) \Big)\,dt - \int_0^{t_1} \lambda(t)^T \delta'(t)\,dt
\end{aligned}$$

Here

$$-\int_0^{t_1} \lambda(t)^T \delta'(t)\,dt = -\lambda(t_1)^T \delta(t_1) + \int_0^{t_1} \lambda'(t)^T \delta(t)\,dt \ .$$

Furthermore, note that

$$\begin{aligned}
H(\lambda, x+\delta, v) - H(\lambda, x, u) &= H(\lambda, x+\delta, v) - H(\lambda, x, v) + H(\lambda, x, v) - H(\lambda, x, u) \\
&= H_x(\lambda, x, v)\delta + \mathcal{O}(\varepsilon^2) + H(\lambda, x, v) - H(\lambda, x, u)
\end{aligned}$$

Also,

$$\int_0^{t_1} H_x(\lambda, x, v)\delta\,dt = \int_0^{t_1} H_x(\lambda, x, u)\delta\,dt + \mathcal{O}(\varepsilon^2) \ .$$

Neglecting terms of order $\mathcal{O}(\varepsilon^2)$ one obtains that

$$
\begin{aligned}
\Delta \tilde{J} &= \tilde{J}(u+\delta,\lambda) - \tilde{J}(u,\lambda) \\
&= \Big(\psi_x(x(t_1)) - \lambda(t_1)^T\Big)\delta(t_1) + \int_0^{t_1} \Big(H_x(\lambda,x,u) + \lambda'(t)^T\Big)\delta(t)\,dt \\
&\quad + \int_0^{t_1} \Big(H(\lambda,x,v) - H(\lambda,x,u)\Big)\,dt
\end{aligned}
$$

Choose $\lambda(t)$ as the solution of the following IVP:

$$
\begin{aligned}
\lambda'(t)^T &= -H_x(\lambda,x,u) \\
\lambda(t_1)^T &= \psi_x(x(t_1))
\end{aligned}
$$

With this choice of $\lambda(t)$ one obtains that

$$
\Delta \tilde{J} = \int_0^{t_1} \Big(H(\lambda(t),x(t),v(t)) - H(\lambda(t),x(t),u(t))\Big)\,dt
$$

The assumption that the control function $u(t)$ locally maximizes $J(\cdot)$ yields that

$$
\Delta \tilde{J} \le 0 \ .
$$

This implies that for every $0 \le t \le t_1$ we have

$$
H(\lambda(t),x(t),v) \le H(\lambda(t),x(t),u(t)) \quad \text{for all} \quad v \in \mathbb{R}^m \ .
$$

This result says the following: If $u(t)$ is an optimal control function then, for every fixed time $t \in [0,t_1]$, the vector $u(t) \in \mathbb{R}^m$ maximizes the function

$$
v \to H(\lambda(t),x(t),v), \quad v \in \mathbb{R}^m \ .
$$

This result is called the **Pontryagin Maximum Principle**. One obtains that

$$
H_u(\lambda(t),x(t),u(t)) = 0 \quad \text{for} \quad 0 \le t \le t_1 \ .
$$

Using that

$$
H(\lambda,x,u) = \lambda^T f(x,u) + l(x,u)
$$

one obtains

$$
\lambda(t)^T f_u(x(t),u(t)) + l_u(x(t),u(t)) = 0 \quad \text{for} \quad 0 \le t \le t_1 \ .
$$

To summarize, the optimal control function $u(t) \in \mathbb{R}^m$, the state vector $x(t) \in \mathbb{R}^n$ and the vector function $\lambda(t) \in \mathbb{R}^n$ satisfy the following differential–algebraic system

$$
\begin{aligned}
x' &= f(x,u) && (17.7) \\
-\lambda'^T &= \lambda^T f_x(x,u) + l_x(x,u) && (17.8) \\
\lambda^T f_u(x,u) + l_u(x,u) &= 0 && (17.9)
\end{aligned}
$$

and the boundary conditions

$$x(0) = x_0, \quad \lambda(t_1) = \psi_x(x(t_1))^T \ . \tag{17.10}$$

The differential–algebraic system for the vector function

$$\begin{pmatrix} x(t) \\ \lambda(t) \\ u(t) \end{pmatrix} \in \mathbb{R}^{2n+m}$$

consists of $2n$ first–order differential equations and m algebraic equations. One expects $2n$ free constants. Typically, the $2n$ free constants are determined by the $2n$ boundary conditions (17.10).

17.7 Linear Systems with Quadratic Cost Function

Assume that the IVP has the form

$$x' = Ax + Bu, \quad x(0) = x_0$$

where $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$. Also, assume that the objective function has the form

$$J(u) = \int_0^{t_1} l(x(t), u(t))\, dt$$

with

$$l(x,u) = -\frac{1}{2}\Big(x^T Q x + u^T R u\Big), \quad \psi(x) \equiv 0 \ .$$

Here $Q \in \mathbb{R}^{n \times n}, R \in \mathbb{R}^{m \times m}$ and it is assumed that

$$Q = Q^T \geq 0, \quad R = R^T > 0 \ .$$

Since

$$f_x = A, \quad f_u = B, \quad l_x = -x^T Q, \quad l_u = -u^T R$$

the differential–algebraic system (17.7), (17.8), (17.9) for x, λ, u becomes

$$\begin{aligned} x' &= Ax + Bu && (17.11) \\ -\lambda' &= A^T \lambda - Qx && (17.12) \\ \lambda^T B - u^T R &= 0 && (17.13) \end{aligned}$$

with boundary conditions

$$x(0) = x_0, \quad \lambda(t_1) = 0 \ . \tag{17.14}$$

First assume that the functions $x(t), \lambda(t), u(t)$ satisfy the differential–algebraic system (17.11), (17.12), (17.13) and the boundary conditions (17.14). The algebraic equation $\lambda^T B - u^T R = 0$ yields that

$$u = R^{-1} B^T \lambda \ .$$

Eliminating $u(t)$ from (17.11), one then obtains the following differential system for x, λ:

$$\begin{aligned} x' &= Ax + BR^{-1}B^T\lambda \\ \lambda' &= Qx - A^T\lambda \end{aligned}$$

with boundary conditions

$$x(0) = x_0, \quad \lambda(t_1) = 0 \ .$$

In the following, assume that x, λ solve the above BVP and also assume that

$$\lambda(t) = -P(t)x(t), \quad P(t_1) = 0 \ ,$$

where $P : [0, t_1] \to \mathbb{R}^{n \times n}$ is a smooth matrix valued function. Then we have

$$x' = (A - BR^{-1}B^T P)x \tag{17.15}$$

and

$$-Px' - P'x = (Q + A^T P)x \ . \tag{17.16}$$

Multiply (17.15) by P and add (17.16) to obtain that

$$-P'x = \Big(PA + A^T P - PBR^{-1}B^T P + Q\Big)x \ .$$

This motivates to determine $P(t)$ as the solution of the IVP

$$-P' = PA + A^T P - PBR^{-1}B^T P + Q, \quad P(t_1) = 0 \tag{17.17}$$

and then to determine $x(t)$ as the solution of

$$x' = \Big(A - BR^{-1}B^T P\Big)x, \quad x(0) = x_0 \ . \tag{17.18}$$

We now prove the following converse.

Lemma 17.2 *Let $P(t)$ and $x(t)$ denote the solutions of the IVPs (17.17) and (17.18). Then $P(t)$ is a symmetric matrix. Set*

$$\lambda = -Px, \quad u = R^{-1}B^T\lambda \ .$$

Then x, λ, u solve the differential–algebraic system (17.11), (17.12), (17.13) and the boundary conditions (17.14).

Proof: Set $\tilde{R} = BR^{-1}B^T$. Since $R = R^T > 0$ the matrix R^{-1} is symmetric and $\tilde{R}$ is also symmetric. The matrix function $P(t)$ satisfies

$$-P' = PA + A^TP - P\tilde{R}P + Q, \quad P(t_1) = 0 \ .$$

Taking the transpose of the differential equation, one obtains that $P^T(t)$ satisfies the same differential equation. Using uniqueness of solutions of initial–value problems, one obtains that $P(t) \equiv P^T(t)$.

We have

$$\begin{aligned}
\lambda &= -Px \\
u &= R^{-1}B^T\lambda \\
&= -R^{-1}B^TPx \\
x' &= Ax - BR^{-1}B^TPx \\
&= Ax + Bu
\end{aligned}$$

Therefore, (17.11) holds. Second,

$$\begin{aligned}
-\lambda' &= (Px)' \\
&= P'x + Px' \\
&= -(PA + A^TP - P\tilde{R}P + Q)x + P(Ax + Bu) \\
&= -A^TPx - Qx + P\tilde{R}Px + PBu \\
&= A^T\lambda - Qx + P\tilde{R}Px - PBR^{-1}B^TPx \\
&= A^T\lambda - Qx
\end{aligned}$$

This shows that (17.12) holds. Third,

$$\lambda^TB - u^TR = \lambda^TB - \lambda^TBR^{-1}R = 0 \ ,$$

which proves that (17.13) holds. The boundary condition $\lambda(t_1) =$ is satisfied since

$$\lambda(t_1) = -P(t_1)x(t_1)$$

and $P(t_1) = 0$. $\diamond$

18 The Discrete Fourier Transform

We first recall Fourier expansion. Replacing integrals by sums will motivate the Discrete Fourier Transform.

18.1 Fourier Expansion

Let $u(t)$ and $v(t)$ denote 1–periodic functions from $\mathbb{R}$ to $\mathbb{C}$, which are sufficiently regular. (For example, $u, v \in C(\mathbb{R}, \mathbb{C})$.)

One defines their L_2–inner product by

$$(u, v) = \int_0^1 \bar{u}(t) v(t)\, dt \ .$$

An important observation is the following: For $k \in \mathbb{Z}$ let

$$u_k(t) = e^{2\pi i k t} \ ,$$

i.e., $u_k(t)$ is a 1–periodic function with $|k|$ waves in the interval $0 \le t \le 1$. Then we have

$$(u_k, u_j) = \int_0^1 e^{2\pi i (j-k)}\, dt = \delta_{jk} \quad \text{for} \quad j, k \in \mathbb{Z} \ .$$

If $u(t)$ is a sufficiently regular 1–periodic function, then one can write $u(t)$ in the form

$$u(t) = \sum_{j=-\infty}^{\infty} c_j e^{2\pi i j t} \ .$$

Taking the inner product with $u_k(t)$ and formally exchanging integration and summation, one obtains that

$$c_k = (u_k, u) = \int_0^1 e^{-2\pi i k t} u(t)\, dt \ .$$

One can prove the following:

Theorem 18.1 *Let $u \in L_2(0, 1)$ and set*

$$\hat{u}(j) = (u_j, u) = \int_0^1 e^{-2\pi i j t} u(t)\, dt \ .$$

Then the function $u(t)$ is given by the Fourier series

$$u(t) = \sum_{j=-\infty}^{\infty} \hat{u}(j) e^{2\pi i j t} \ .$$

The Fourier series converges to $u(t)$ in the L_2–norm. If $u \in C^1[0, 1]$ and $u(0) = u(1)$ then the Fourier series converges in maximum norm to $u(t)$.

The numbers

$$\hat{u}(j) = \int_0^1 e^{-2\pi i j t} u(t)\, dt, \quad j \in \mathbb{Z} ,$$

are called the Fourier coefficients of the function $u(t)$.

18.2 Discretization

Recall the trapezoidal rule

$$\int_a^b g(t)\, dt \sim \frac{b-a}{2} \Big(g(a) + g(b) \Big) .$$

Let $n \in \mathbb{N}$ and let $h = \frac{1}{n}$ denote a step–size. The $n+1$ points

$$t_k = kh, \quad k = 0, 1, \ldots, n ,$$

form an equidistant grid in the interval $0 \le t \le 1$. If $g : [0,1] \to \mathbb{C}$ is a continuous function with $g(0) = g(1)$, the trapezoidal approximation to $\hat{g}(j)$ with step size h is:

$$\begin{aligned} \hat{g}(j) &= \int_0^1 g(t)\, dt \\ &\sim \sum_{k=0}^{n-1} \frac{h}{2} \Big(g(t_k) + g(t_{k+1}) \Big) \\ &= h \sum_{k=0}^{n-1} g(t_k) \end{aligned}$$

Let's apply this formula to the integral

$$\hat{u}(j) = \int_0^1 e^{-2\pi i j t} u(t)\, dt .$$

One obtains the approximation

$$\begin{aligned} \hat{u}(j) &\sim h \sum_{k=0}^{n-1} u(t_k) e^{-2\pi i jk/n} \\ &= h \sum_{k=0}^{n-1} u(t_k) \xi^{jk} \end{aligned}$$

with

$$\xi = \xi_n = e^{-2\pi i/n} .$$

We now replace the grid function

$$\Big(u(t_0), u(t_1), \ldots, u(t_{n-1}\Big)$$

by a column vector

$$u = \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} \in \mathbb{C}^n$$

and define

$$v_j = \sum_{k=0}^{n-1} u_k \xi^{jk} \quad \text{for} \quad j = 0, 1, \ldots, n-1 \quad \text{where} \quad \xi = e^{-2\pi i/n} \ . \tag{18.1}$$

The vector $v = DFT(u) \in \mathbb{C}^n$ is called the Discrete Fourier Transform of the vector $u \in \mathbb{C}^n$.

For a smooth, 1–periodic function $u(t)$ the formula

$$u(t) = \sum_{j=-\infty}^{\infty} \hat{u}(j) e^{2\pi i j t} \tag{18.2}$$

holds. It expresses the function $u(t)$ in terms of its Fourier coefficients. We therefore expect that we can also use the discrete Fourier transform $v = DFT(u)$ of a vector $u \in \mathbb{C}^n$ to express u in terms of v. In fact, we will prove that

$$u_k = \frac{1}{n} \sum_{j=0}^{n-1} v_j \omega^{jk} \quad \text{for} \quad k = 0, 1, \ldots, n-1 \quad \text{where} \quad \omega = e^{2\pi i/n} = \bar{\xi} = 1/\xi \ . \tag{18.3}$$

This is the discrete analogue of the formula (18.2).

18.3 DFT as a Linear Transformation

Let $n \in \mathbb{N}$ be fixed and set

$$\xi = e^{-2\pi i/n}, \quad \omega = \bar{\xi} = e^{2\pi i/n} \ .$$

Define the following complex symmetric $n \times n$ matrices:

$$F = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & \xi & \xi^2 & \ldots & \xi^{n-1} \\ 1 & \xi^2 & \xi^4 & \ldots & \xi^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \xi^{n-1} & \xi^{(n-1)2} & \ldots & \xi^{(n-1)(n-1)} \end{pmatrix} \tag{18.4}$$

$$G = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega & \omega^2 & \dots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \dots & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega^{n-1} & \omega^{(n-1)2} & \dots & \omega^{(n-1)(n-1)} \end{pmatrix} \tag{18.5}$$

Thus

$$F = \left(\xi^{jk}\right)_{0 \le j,k \le n-1} \quad \text{and} \quad G = \left(\omega^{jk}\right)_{0 \le j,k \le n-1} .$$

The mapping

$$DFT_n : \begin{cases} \mathbb{C}^n & \to \mathbb{C}^n \\ x & \to Fx \end{cases}$$

is called the discrete Fourier transform of order n.

Thus, if

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} \in \mathbb{C}^n \quad \text{and} \quad y = DFT_n(x) = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{pmatrix}$$

then

$$y_j = \sum_{k=0}^{n-1} \xi^{jk} x_k = \sum_{k=0}^{n-1} e^{-2\pi ijk/n} x_k \quad \text{for} \quad 0 \le j \le n-1 .$$

This is the same as formula (18.1) with u replaced by x and v replaced by y.

Lemma 18.1 *The matrices*

$$\frac{1}{\sqrt{n}} F \quad \textit{and} \quad \frac{1}{\sqrt{n}} G$$

are unitary and

$$F^{-1} = \frac{1}{n} G .$$

Proof: Consider two different columns of F:

$$f^j = \begin{pmatrix} 1 \\ \xi^j \\ \xi^{2j} \\ \vdots \\ \xi^{(n-1)j} \end{pmatrix} \quad \text{and} \quad f^l = \begin{pmatrix} 1 \\ \xi^l \\ \xi^{2l} \\ \vdots \\ \xi^{(n-1)l} \end{pmatrix} \quad \text{where} \quad 0 \le j,l \le n-1 \quad \text{and} \quad j \ne l .$$

Their inner product is

$$\begin{aligned}
\langle f^j, f^l \rangle &= \sum_{k=0}^{n-1} \xi^{-jk} \xi^{lk} \\
&= \sum_{k=0}^{n-1} \xi^{(l-j)k} \\
&= \sum_{k=0}^{n-1} q^k \quad (\text{with } q = \xi^{l-k} \neq 1) \\
&= \frac{q^n - 1}{q - 1} \\
&= 0
\end{aligned}$$

since $q^n = 1$. (This follows from $\xi^n = 1$.)

Thus, the columns of F are orthogonal. It is also clear that each column of F has Euclidean length $|f^j| = \sqrt{n}$ since

$$|f^j|^2 = \sum_{k=0}^{n-1} 1 = n \ .$$

Therefore, the matrix $\frac{1}{\sqrt{n}} F$ is unitary. The same arguments show that $\frac{1}{\sqrt{n}} G$ is unitary.

The inverse of the unitary matrix $\frac{1}{\sqrt{n}} F$ is

$$\Big(\frac{1}{\sqrt{n}} F \Big)^{-1} = \frac{1}{\sqrt{n}} F^* = \frac{1}{\sqrt{n}} G \ .$$

This yields that

$$F^{-1} = \frac{1}{n} G \ .$$

$\diamond$

If $y = Fx = DFT(x)$ then

$$x = F^{-1} y = \frac{1}{n} Gy \ ,$$

thus

$$x_k = \frac{1}{n} \sum_{j=0}^{n-1} \omega^{jk} y_j \quad \text{for} \quad k = 0, 1, \ldots, n-1 \quad \text{where} \quad \omega = e^{2\pi i/n} \ .$$

This proves formula (18.3), the inversion of the discrete Fourier transform.

18.4 Fourier Series and DFT

Let $u(t)$ denote a smooth, 1–periodic function from $\mathbb{R}$ to $\mathbb{C}$, thus

$$u(t) = \sum_{j=-\infty}^{\infty} \hat{u}(j) e^{2\pi i j t}$$

with

$$\hat{u}(j) = \int_0^1 e^{-2\pi i j t} u(t)\, dt \; . \tag{18.6}$$

Let $n \in \mathbb{N}$ and let $h = \frac{1}{n}$ denote a step–size. Let

$$u^h = \begin{pmatrix} u(0) \\ u(h) \\ \vdots \\ u((n-1)h) \end{pmatrix} \in \mathbb{C}^n$$

denote the restriction of $u(t)$ to the h–grid. How is the vector

$$v^h = DFT(u^h)$$

related to the Fourier coefficients $\hat{u}(j)$?

Note that the components of v^h are the n numbers

$$v_j^h = \sum_{k=0}^{n-1} u(kh) e^{-2\pi i j k h}, \quad j = 0, 1, \ldots, n-1 \; . \tag{18.7}$$

Therefore, the number hv_j^h is the approximation of $\hat{u}(j)$ if one replaces the integral in (18.6) using the trapezoidal rule with step size h.

At first, it is confusing that v_j^h is only defined for integers j with $0 \le j \le n-1$, but $\hat{u}(j)$ is also defined for negative j. Note, however, that formula (18.7) can also be used to define v_j^h for all integers j, and one then obtains values with

$$v_j^h = v_{j+n}^h \quad \text{for all} \quad j \in \mathbb{Z} \; .$$

In words, the function

$$\left\{ \begin{array}{ccc} \mathbb{Z} & \to & \mathbb{C} \\ j & \to & v_j^h \end{array} \right. \tag{18.8}$$

has period n. In particular,

$$v_{n-1}^h = v_{-1}^h, \quad v_{n-2}^h = v_{-2}^h$$

etc.

It is therefore reasonable to expect that

$$\hat{u}(j) \sim \left\{ \begin{array}{lll} hv_j^h & \text{for} & 0 \le j < \frac{n}{2} \\ hv_{j+n}^h & \text{for} & -\frac{n}{2} < j < 0 \end{array} \right.$$

where $v^h = DFT(u^h)$.

Example: Consider the 1–periodic function

$$u(t) = 2\cos(2\pi * 40t) + 6\cos(2\pi * 7t)$$

with frequencies $2\pi * 40$ and $2\pi * 7$.

The functions $2\cos(2\pi * 40t)$ and $6\cos(2\pi * 7t)$ as well as their sum $u(t)$ are shown in Figures 1, 2, 3.

Since

$$\cos(2\pi jt) = \frac{1}{2}\Big(e^{2\pi ijt} + e^{-2\pi ijt}\Big)$$

one obtains for the Fourier coefficients of $u(t)$:

$$\hat{u}(j) = \begin{cases} 1 & \text{for} \quad j = \pm 40 \\ 3 & \text{for} \quad j = \pm 7 \\ 0 & \text{for} \quad j \in \mathbb{Z} \setminus \{-40, -7, 7, 40\} \end{cases}$$

We now choose the step–size

$$h = \frac{1}{n} \quad \text{with} \quad n = 512 = 2^9$$

and let

$$u^h = \begin{pmatrix} u(0) \\ u(h) \\ \vdots \\ u(511\,h) \end{pmatrix} \in \mathbb{R}^{512}$$

denote the restriction of $u(t)$ to the h–grid. For

$$v^h = DFT(u^h)$$

one obtains almost exactly:

$$hv^h_j = \begin{cases} 1 & \text{for} \quad j = 40 \ \text{ and } \ j = 472 \\ 3 & \text{for} \quad j = 7 \ \text{ and } \ j = 505 \\ 0 & \text{otherwise} \end{cases}$$

Here we have used the n–periodicity of the extended function (18.8) and

$$512 - 40 = 472 \quad \text{and} \quad 512 - 7 = 505 \ .$$

The grid function

$$hv^h_j \quad \text{for} \quad j = 0, 1, \ldots, 511$$

is shown in Figure 6.

We will now perturb the signal $u(t)$ and then consider the DFT of the perturbed signal. As above, let

$$n = 512, \quad h = \frac{1}{n}, \quad t_j = jh \quad \text{for} \quad j = 0, 1, \ldots, n-1 \ .$$

We determine the noise function (with maximal amplitude 5) by

$$noise(j) = 10 * (rand - 0.5) \quad \text{for} \quad j = 1, 2, \ldots, n$$

where *rand* is MATLAB's random number, which is uniformly distributed in the interval from 0 to 1. A typical noise function

$$f(t_j) = noise(j+1), \quad j = 0, 1, \ldots, n-1 \ ,$$

is shown in Figure 4 and the perturbed signal $u(t_j) + f(t_j)$ is shown in Figure 5

In Figure 3, the low frequency part $6\cos(2\pi * 7t)$ (see Figure 2) shows up rather clearly. The high frequency part $2\cos(2\pi * 40t)$ (see Figure 1) is more difficult to detect, but can still be recognized.

After the noise is added, one obtains the grid function $u(t_j) + f(t_j)$ shown in Figure 5. The low frequency part $6\cos(2\pi * 7t)$, consisting of seven bumps, still shows up rather clearly, but the high frequency part $2\cos(2\pi * 40t)$ is not visible at all.

Figure 7 shows the real part of the DFT of the perturbed signal $u(t_j)+f(t_j)$, multiplied by h. It is interesting that the high frequency part $2\cos(2\pi * 40t)$ shows up clearly on the Fourier side, with peaks near $j = 40$ and $j = 512 - 40$.

Discrete Fourier transformation is a useful tool to detect periodic structures in signals. And there are many other applications.

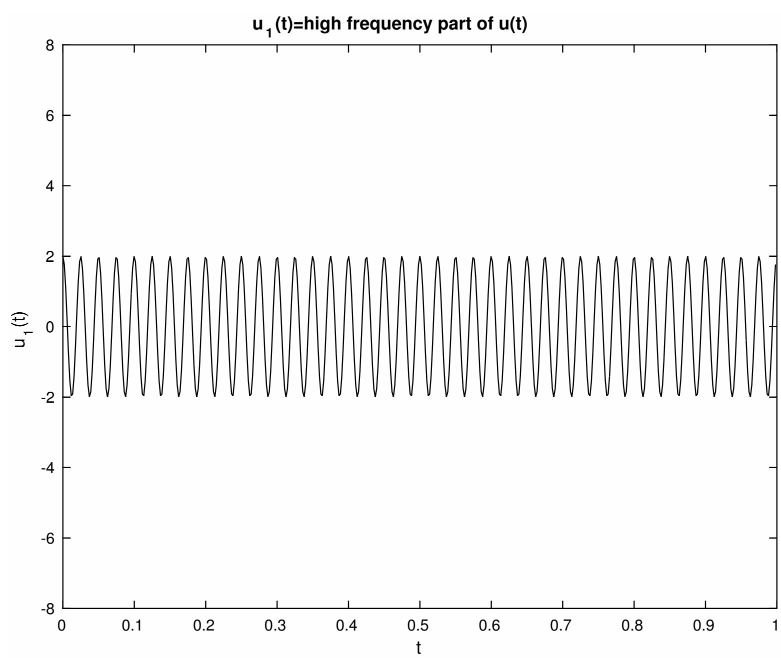

Figure 1: $u_1(t) = 2\cos(2\pi * 40t)$

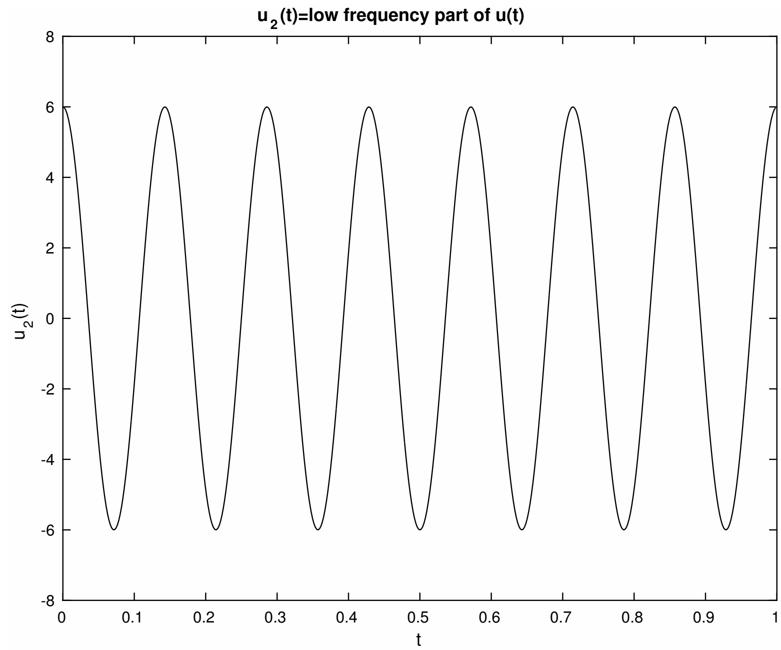

Figure 2: $u_2(t) = 6\cos(2\pi * 7t)$

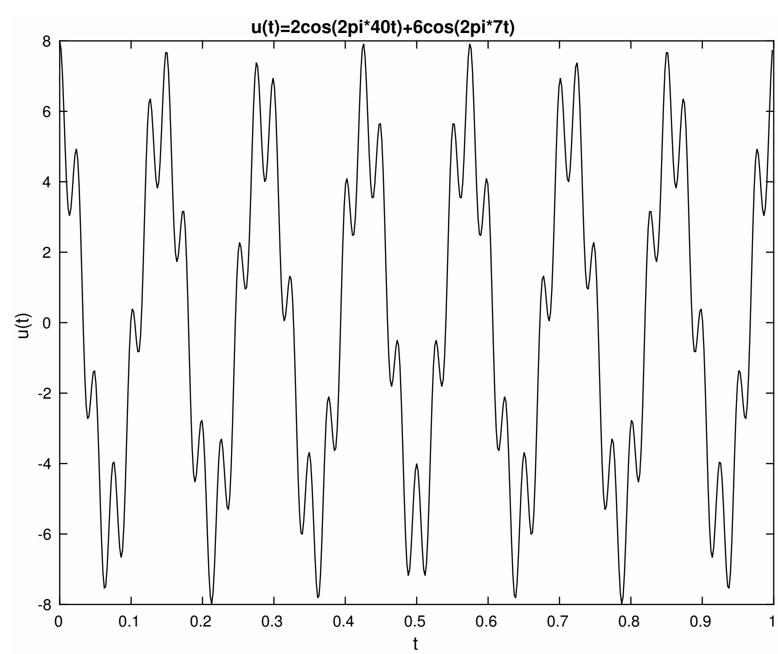

Figure 3: $u(t) = 2\cos(2\pi * 40t) + 6\cos(2\pi * 7t)$

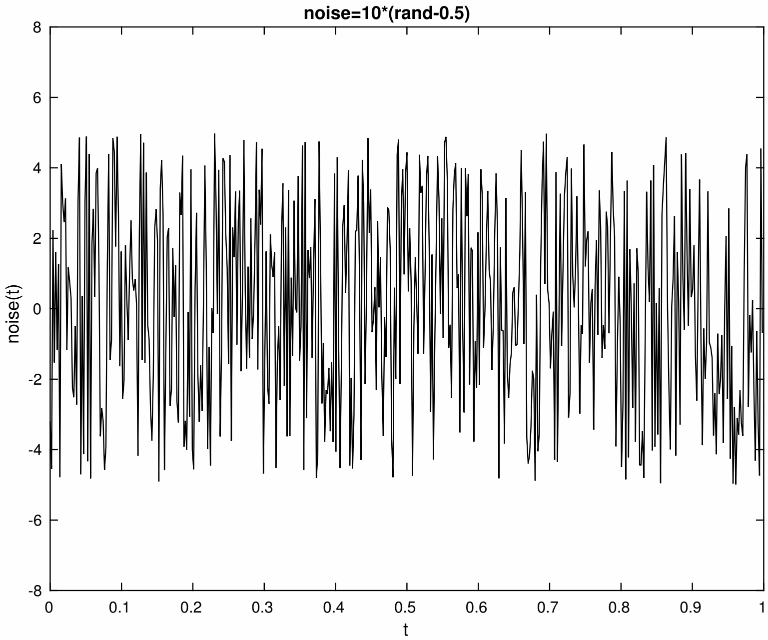

Figure 4: Noise generated with rand

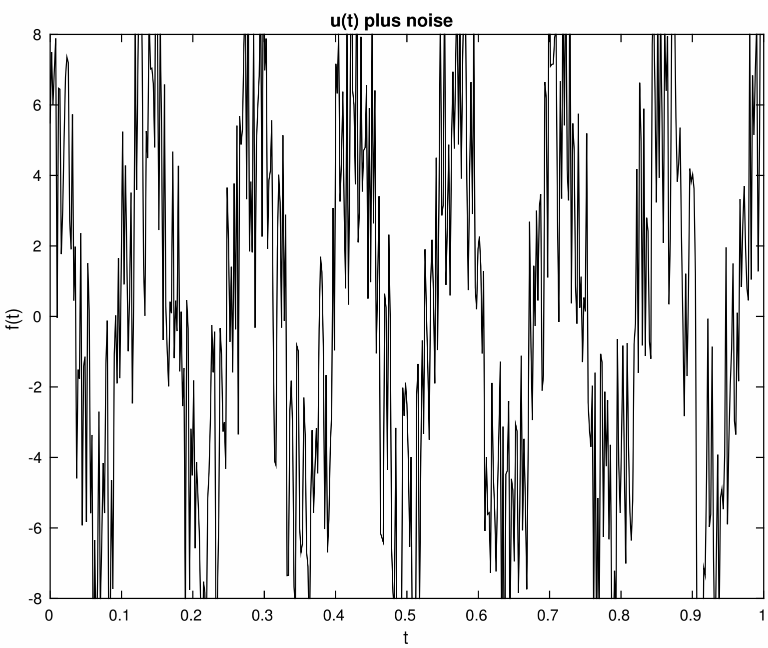

Figure 5: The signal $u(t)$ plus noise

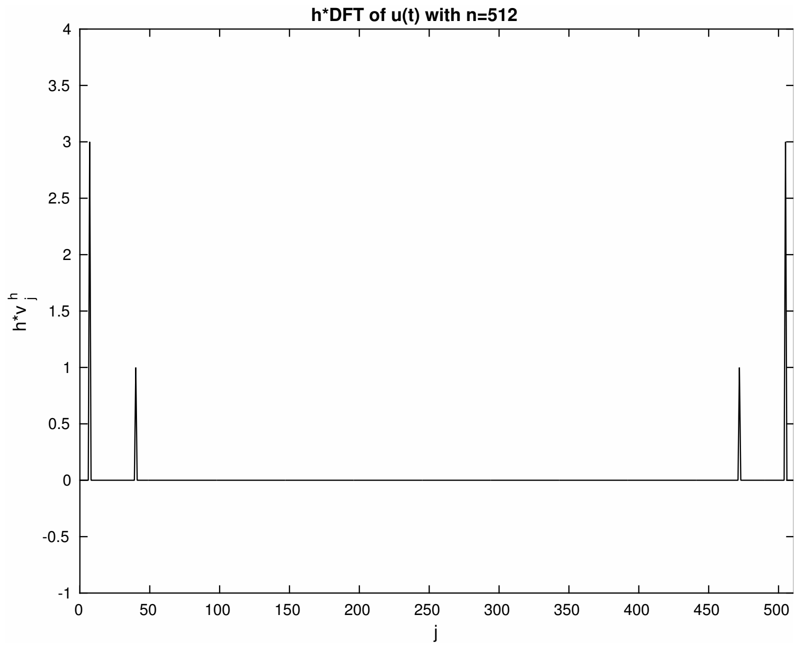

Figure 6: The discrete Fourier transform of $u(t)$ (multiplied by h)

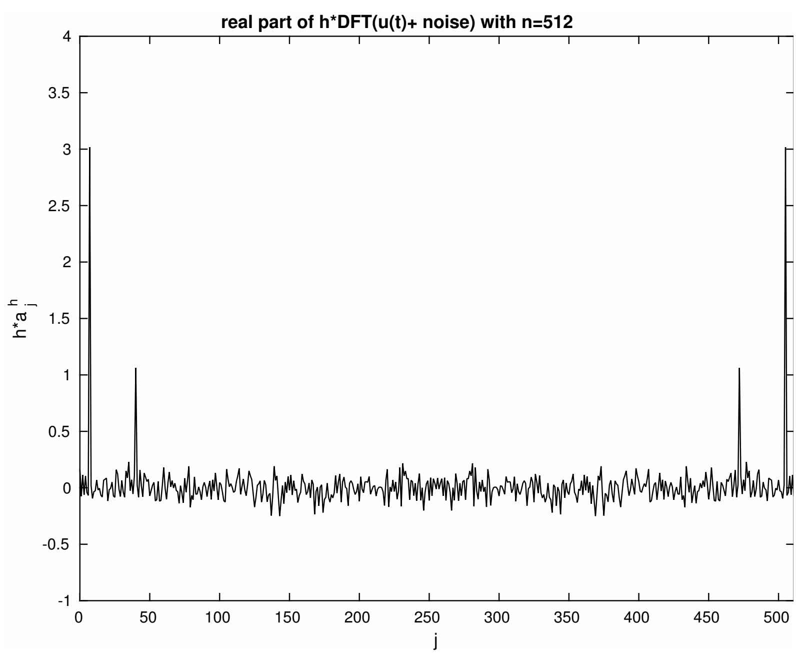

Figure 7: The real part of the discrete Fourier transform of $u(t)$ plus noise (multiplied by h)

19 Fast Fourier Transform

Let $N = 2n$ and let

$$x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} \in \mathbb{C}^N, \quad y = DFT_N(x) \in \mathbb{C}^N \ .$$

Using the matrix multiplication $y = F_N x$ to compute y requires $N^2 = 4n^2$ multiplications.

We now describe how y can be computed by computing DFTs of two vectors of dimension $n = N/2$.

Let

$$x^{(1)} = x^{(even)} = \begin{pmatrix} x_0 \\ x_2 \\ \vdots \\ x_{N-2} \end{pmatrix} \in \mathbb{C}^n, \quad x^{(2)} = x^{(odd)} = \begin{pmatrix} x_1 \\ x_3 \\ \vdots \\ x_{N-1} \end{pmatrix} \in \mathbb{C}^n \ .$$

Let

$$y^{(1)} = DFT_n\Big(x^{(1)}\Big) \quad \text{and} \quad y^{(2)} = DFT_n\Big(x^{(2)}\Big) \ .$$

We have

$$\begin{aligned} y_k &= \sum_{j=0}^{N-1} e^{-2\pi ijk/N} x_j \quad \text{for} \quad 0 \le k \le N-1 \\ y_k^{(1)} &= \sum_{l=0}^{n-1} e^{-2\pi ilk/n} x_{2l} \quad \text{for} \quad 0 \le k \le n-1 \\ y_k^{(2)} &= \sum_{l=0}^{n-1} e^{-2\pi ilk/n} x_{2l+1} \quad \text{for} \quad 0 \le k \le n-1 \end{aligned}$$

For $0 \le k \le n-1$ we have

$$\begin{aligned} y_k &= \sum_{j=0,\ j \text{ even}}^{N-1} e^{-2\pi ijk/2n} x_j + \sum_{j=0,\ j \text{ odd}}^{N-1} e^{-2\pi ijk/2n} x_j \\ &= \sum_{l=0}^{n-1} e^{-2\pi ilk/n} x_{2l} + e^{-2\pi ik/N} \sum_{l=0}^{n-1} e^{-2\pi ilk/n} x_{2l+1} \\ &= y_k^{(1)} + e^{-2\pi ik/N} y_k^{(2)} \end{aligned}$$

To obtain the second line, set $j = 2l$ in the first sum of the first line and set $j = 2l + 1$ in the second sum of the first line.

We also have for $0 \le k \le n-1$:

$$\begin{aligned}
y_{n+k} &= \sum_{j=0}^{N-1} e^{2\pi i(n+k)/2n} \\
&= \sum_{j=0, j \text{ even}}^{N-1} e^{-2\pi ij(n+k)/2n} x_j + \sum_{j=0, j \text{ odd}}^{N-1} e^{-2\pi ij(n+k)/2n} x_j \\
&= \sum_{l=0}^{n-1} e^{-2\pi ilk/n} x_{2l} - e^{-2\pi ik/N} \sum_{l=0}^{n-1} e^{-2\pi ilk/n} x_{2l+1} \\
&= y_k^{(1)} - e^{-2\pi ik/N} y_k^{(2)}
\end{aligned}$$

Here we have used that

$$\begin{aligned}
e^{-2\pi i(2l+1)(n+k)/2n} &= e^{-2\pi i(n+k)/2n} e^{-2\pi ilk/n} \\
&= -e^{-2\pi ik/N} e^{-2\pi ilk/n}
\end{aligned}$$

In the following, we will only count the number of complex *multiplications* and will ignore the number of additions and subtractions. We will assume that $N = 2n$. Suppose we need M_n multiplications to compute DFT_n of a vector of dimension n. Then, to compute $y^{(1)}$ and $y^{(2)}$ we need $2M_n$ multiplications. To compute $y = DFT_N x$ we need an additional n multiplications if we use the formulas

$$y_k = y_k^{(1)} + e^{-2\pi ik/N} y_k^{(2)} \quad \text{for} \quad 0 \le k \le n-1$$

and

$$y_{n+k} = y_k^{(1)} - e^{-2\pi ik/N} y_k^{(2)} \quad \text{for} \quad 0 \le k \le n-1 \; .$$

One then obtains for the number M_N of multiplications to compute $y = DFT_N x$:

$$M_N = 2M_n + n \; .$$

Suppose that

$$M_n = \frac{n}{2} \log_2 n \; .$$

Then obtain for $N = 2n$:

$$\begin{aligned}
M_N &= 2M_n + n \\
&= n \log_2 n + n \\
&= \frac{N}{2} \log_2(N/2) + n \\
&= \frac{N}{2} \Big((\log_2 N) - 1 \Big) + n \\
&= \frac{N}{2} \log_2 N
\end{aligned}$$

For short, if $N = 2n$ then $M_n = \frac{n}{2} \log_2 n$ implies that $M_N = \frac{N}{2} \log_2 N$.

If N is a power of 2, then the above idea can be repeated. For $n = 2$ we have

$$F_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and may set

$$M_2 = 1 \ .$$

One obtains that for $N = 2^k$ where $k \in \mathbb{N}$:

$$M_N = \frac{N}{2} \log_2 N = \frac{N}{2} k \ .$$

The reduction of numerical work from $\sim N^2$ to $\sim N \log_2 N$ is significant.

The algorithm, where the above idea is carried out repeatedly, is called Fast–Fourier Transform, FFT.

FFT was published in 1965 by James Cooley and John Tukey. However, it is reported that Gauss had already used the idea in hand–computations around 1805.

20 Eigenvalues Under Perturbations

Let $A \in \mathbb{C}^{n\times n}$ denote a matrix with n distinct eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. We will study how the eigenvalues and corresponding eigenvectors change if A is replaced by a disturbed matrix $A + \varepsilon B$.

A useful result, which we prove first, says that A has right and left eigenvectors, which are biorthogonal. This result will help us to study the perturbation problem.

Recall our notation for the inner product in $\mathbb{C}^n$: If $a, b \in \mathbb{C}^n$ are column vectors with components a_j and b_j then

$$\langle a, b\rangle = a^*b = \sum_{j=1}^n \bar{a}_j b_j \ .$$

20.1 Right and Left Eigenvectors

Let $A \in \mathbb{C}^{n\times n}$ denote a matrix with n distinct eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Let $u_1, \ldots, u_n \in \mathbb{C}^n$ denote corresponding right eigenvectors,

$$Au_j = \lambda_j u_j, \quad u_j \neq 0 \ .$$

Theorem 20.1 *a) Under the above assumption, the matrix A^* has the n distinct eigenvalues $\bar{\lambda}_1, \ldots, \bar{\lambda}_n$.*

b) If $A^* v_k = \bar{\lambda}_k v_k$ and if v_k is properly scaled, then

$$\langle v_k, u_j\rangle = \delta_{jk} \quad \textit{for} \quad j,k = 1, 2, \ldots, n \ .$$

Proof: a) The characteristic polynomial of A^* is

$$\begin{aligned} det(A^* - \lambda I) &= det(\bar{A} - \lambda I) \\ &= \overline{det}(A - \bar{\lambda} I) \end{aligned}$$

and the zeros are $\bar{\lambda}_1, \ldots, \bar{\lambda}_n$.

b) For $j \neq k$ we have

$$\begin{aligned} \lambda_j \langle v_k, u_j\rangle &= \langle v_k, Au_j\rangle \\ &= \langle A^* v_k, u_j\rangle \\ &= \langle \bar{\lambda}_k v_k, u_j\rangle \\ &= \lambda_k \langle v_k, u_j\rangle \end{aligned}$$

It follows that $\langle v_k, u_j\rangle = 0$ since $\lambda_j \neq \lambda_k$.

c) Let $A^* v_k = \bar{\lambda}_k v_k$. We have obtained that $\langle v_k, u_j\rangle = 0$ for all j which are different from k. If $\langle v_k, u_k\rangle = 0$ then v_k is orthogonal to a basis of $\mathbb{C}^n$, thus $v_k = 0$. Therefore, if $v_k \neq 0$ then $\langle v_k, u_k\rangle \neq 0$. The claim follows. $\diamond$

Terminology: The equation

$$A^* v_k = \bar{\lambda}_k v_k$$

can also be written as

$$v_k^* A = \lambda_k v_k^* \ .$$

Therefore, one calls v_k^* a left eigenvector of A to the eigenvalue λ_k. Theorem 20.1 can also be stated as follows:

Theorem 20.2 *Let $A \in \mathbb{C}^{n\times n}$ have n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. There exists a basis $u_1, \ldots, u_n$ of right eigenvectors of A,*

$$Au_j = \lambda_j u_j, \quad j = 1, \ldots, n \ ,$$

and a basis $v_1, \ldots, v_n$ of left eigenvectors of A,

$$v_k^* A = \lambda_k v_k^*, \quad k = 1, \ldots, n \ .$$

The two bases are biorthogonal:

$$\langle v_k, u_j \rangle \begin{cases} = 0 & \textit{for} \quad j \neq k \\ \neq 0 & \textit{for} \quad j = k \end{cases}$$

After proper scaling of the eigenvectors one obtains that

$$\langle v_k, u_j \rangle = \delta_{jk} \quad \textit{for} \quad j, k = 1, 2, \ldots, n \ .$$

20.2 Perturbations of A

We use the same notation as in the previous section and assume that the matrix $A \in \mathbb{C}^{n\times n}$ has the n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. We also assume that the right eigenvectors $u_1, \ldots, u_n$ and the left eigenvectors $v_1, \ldots, v_n$ are chosen so that

$$Au_j = \lambda_j u_j, \quad v_k^* A = \lambda_k v_k^*, \quad \langle v_k, u_j \rangle = \delta_{jk} \quad \text{for} \quad j, k = 1, \ldots, n \ .$$

Consider the perturbed matrix

$$A + \varepsilon B \ ,$$

where $B \in \mathbb{C}^{n\times n}$ and $\varepsilon \in \mathbb{C}$. We want to understand how the eigenvalues and eigenvectors change if $|\varepsilon|$ is small. For simplicity of notation, we study the perturbation of λ_1 and u_1.

We first prove the following auxiliary result.

Lemma 20.1 *Under the above assumptions, we have*

$$range(A - \lambda_1 I) = \{ b \in \mathbb{C}^n \ : \ \langle v_1, b \rangle = 0 \} \ . \tag{20.1}$$

Proof: Let $x \in \mathbb{C}^n$ be arbitrary and write

$$x = \sum_{j=1}^{n} \alpha_j u_j \ .$$

Then $b = (A - \lambda_1 I)x$ is the general vector in $range(A - \lambda_1 I)$ and

$$b = \sum_{j=2}^{n} \alpha_j (\lambda_j - \lambda_1) u_j \ .$$

Therefore, $\langle v_1, b \rangle = 0$. This proves that

$$range(A - \lambda_1 I) \subset \{ b \in \mathbb{C}^n \ : \ \langle v_1, b \rangle = 0 \} \ .$$

Since both spaces in the above formula have the same dimension $n-1$, the equality (20.1) follows. $\diamond$

To study the eigenvalue problem for $A + \varepsilon B$ near λ_1 and u_1 we first proceed formally and make the ansatz

$$\lambda(\varepsilon) = \lambda_1 + \varepsilon \mu, \quad u(\varepsilon) = u_1 + \varepsilon q$$

for an eigenvalue and a corresponding eigenvector of $A + \varepsilon B$. The condition for $\mu \in \mathbb{C}$ and $q \in \mathbb{C}^n$ is

$$(A + \varepsilon B)(u_1 + \varepsilon q) = (\lambda_1 + \varepsilon \mu)(u_1 + \varepsilon q) \ .$$

Multiplying out and using that $Au_1 = \lambda_1 u_1$ yields the condition

$$\varepsilon(Aq + Bu_1) + \varepsilon^2 Bq = \varepsilon(\mu u_1 + \lambda_1 q) + \varepsilon^2 \mu q \ . \tag{20.2}$$

Proceeding formally, we neglect the ε^2–terms and obtain the condition

$$Aq + Bu_1 = \mu u_1 + \lambda_1 q$$

for the number $\mu \in \mathbb{C}$ and the vector $q \in \mathbb{C}^n$. Rewriting the above condition yields the equation

$$(A - \lambda_1 I)q = \mu u_1 - Bu_1 \ . \tag{20.3}$$

Here A, B, λ_1 and u_1 are known and $\mu \in \mathbb{C}$ as well as $q \in \mathbb{C}^n$ need to be determined. It is clear that a solution q of the system (20.3) exists if and only if the right–hand side lies in $range(A - \lambda_1 I)$. Using the previous lemma, this yields the condition

$$0 = \langle v_1, \mu u_1 - Bu_1 \rangle \ ,$$

i.e.,

$$\mu = \langle v_1, Bu_1 \rangle \ .$$

With this choice of μ, let us solve the system (20.3) for q and write

$$q = \sum_{j=1}^{n} \alpha_j u_j \ . \tag{20.4}$$

We obtain that

$$(A - \lambda_1 I) q = \sum_{j=2}^{n} \alpha_j (\lambda_j - \lambda_1) u_j \ .$$

Assuming that $\mu = \langle v_1, B u_1 \rangle$, equation (20.3) holds if and only if

$$\langle v_k, (A - \lambda_1 I) q \rangle = \langle v_k, \mu u_1 - B u_1 \rangle \quad \text{for} \quad k = 2, \ldots, n \ .$$

Using formula (20.4) for q one obtains the equivalent conditions

$$\alpha_k (\lambda_k - \lambda_1) = \langle v_k, \mu u_1 - B u_1 \rangle \quad \text{for} \quad k = 2, \ldots, n \ .$$

i.e.,

$$\alpha_k = \frac{\langle v_k, B u_1 \rangle}{\lambda_1 - \lambda_k} \quad \text{for} \quad k = 2, \ldots, n \ .$$

The value of α_1 in the sum (20.4) is arbitrary. The choice $\alpha_1 = 0$ is equivalent to the condition $\langle v_1, q \rangle = 0$. One obtains the following result:

Lemma 20.2 *Under the above assumptions, consider the system*

$$(A - \lambda_1 I) q = \mu u_1 - B u_1, \quad \langle v_1, q \rangle = 0 \tag{20.5}$$

where A, B, λ_1, u_1 and v_1 are known and $\mu \in \mathbb{C}$ as well as $q \in \mathbb{C}^n$ have to be determined.

The system is uniquely solvable for μ, q. The solution is given by

$$\mu = \langle v_1, B u_1 \rangle, \quad q = \sum_{k=2}^{n} \alpha_k u_j \quad \textit{with} \quad \alpha_k = \frac{\langle v_k, B u_1 \rangle}{\lambda_1 - \lambda_k} \ .$$

Our formal computations suggest that the matrix

$$A + \varepsilon B$$

has the eigenvalue

$$\lambda(\varepsilon) = \lambda_1 + \varepsilon \mu + \mathcal{O}(\varepsilon^2) \tag{20.6}$$

and the corresponding eigenvector

$$u(\varepsilon) = u_1 + \varepsilon q + \mathcal{O}(\varepsilon^2) \tag{20.7}$$

if $|\varepsilon|$ is small. Here μ and q are determined as described in the previous lemma.

It is not a matter of linear algebra, however, to prove the formulas (20.6) and (20.7). Note that we neglected ε^2–terms in (20.2) and claim that they lead to $\mathcal{O}(\varepsilon^2)$–terms in the formulas (20.6) and (20.7). One can prove the validity of (20.6) and (20.7) using the **Implicit Function Theorem** of nonlinear analysis.

When applying the implicit function theorem, it is crucial to note that the linear system (20.5) for μ, q is nonsingular. In fact, the system (20.5) for

$$\begin{pmatrix} q \\ \mu \end{pmatrix} \in \mathbb{C}^{n+1}$$

reads in matrix form

$$\begin{pmatrix} A - \lambda_1 I & -u_1 \\ v_1^* & 0 \end{pmatrix} \begin{pmatrix} q \\ \mu \end{pmatrix} = \begin{pmatrix} -Bu_1 \\ 0 \end{pmatrix} .$$

If $B = 0$ then the uniqueness statement in the previous lemma implies that $q = 0, \mu = 0$. Therefore, the $(n+1) \times (n+1)$ matrix of the above system is nonsingular.

21 Perron–Frobenius Theory

In this chapter a real matrix $A = (a_{ij})$ is called positive if all its entries are strictly positive, $a_{ij} > 0$. The matrix A is called non–negative if all entries are non–negative, $a_{ij} \geq 0$.

In 1907, Oskar Perron published results on spectral properties of positive matrices. His results were extended to non–negative matrices by Georg Frobenius in 1912.

Positive and non–negative matrices play a role in probability theory since probabilities cannot be negative. We consider an application to Markov processes in Section 21.3.

Notations: For $A \in \mathbb{C}^{n \times n}$ let

$$\rho(A) = \max\{|\lambda| \ : \ \lambda \in \sigma(A)\}$$

denote its spectral radius.

If $A \in \mathbb{R}^{n \times n}$ then $A > 0$ means that

$$a_{ij} > 0$$

for all matrix elements of A. Similarly, $A \geq 0$ means that

$$a_{ij} \geq 0$$

for all matrix elements of A. If $x, y \in \mathbb{R}^n$ then

$$x > y \quad \text{means that} \quad x_j > y_j \quad \text{for} \quad j = 1, \ldots, n \ .$$

Similarly,

$$x \geq y \quad \text{means that} \quad x_j \geq y_j \quad \text{for} \quad j = 1, \ldots, n \ .$$

If $x \in \mathbb{C}^n$ then let

$$|x|_{ab} = (|x_1|, |x_2|, \ldots, |x_n|)^T \ .$$

Remarks on History: Oskar Perron (1880–1975) proved spectral properties for positive matrices, $A > 0$. Perron was a professor at the universities of Heidelberg and Munich. Perron's paradox illustrates the danger of simply assuming that a solution of an optimization problem exists: Let N be the largest integer. If $N > 1$ then $N^2 > N$, contradicting the definition on N. Hence $N = 1$.

Georg Frobenius (1849–1917) extended some of Perron's results to certain non–negative matrices, $A \geq 0$. Frobenius taught at ETH Zurich and the University of Berlin. In this case, the older mathematician expanded on the work of the younger.

21.1 Perron's Theory

Theorem 21.1 *(Perron) Let $A \in \mathbb{R}^{n\times n}, A > 0$. The following holds:*

1) The spectral radius of A is positive, $r := \rho(A) > 0$.

2) The spectral radius $r = \rho(A)$ is an algebraically simple eigenvalue of A.

3) There exists an eigenvector $\xi \in \mathbb{R}^n$ with

$$A\xi = r\xi, \quad \xi > 0 \ .$$

The vector ξ is called Perron's eigenvector of A.

4) If $\lambda \in \sigma(A)$ and $\lambda \neq \rho(A)$ then

$$|\lambda| < \rho(A) \ .$$

5) If $y \in \mathbb{R}^n, y \geq 0$, is an eigenvector of A,

$$Ay = \lambda y \ ,$$

then $\lambda = r = \rho(A)$ and y is a multiple of ξ.

6) There exist vectors $\xi > 0$ and $\eta > 0$ with

$$A\xi = r\xi \quad \textit{and} \quad A^T\eta = r\eta \ .$$

If $y \in \mathbb{R}^n$ is arbitrary, then the convergence

$$\frac{1}{r^j}\, A^j y \to c\xi \quad \textit{as} \quad j \to \infty \quad \textit{where} \quad c = \frac{\langle \eta, y\rangle}{\langle \eta, \xi\rangle}$$

holds.

7) Let the vectors ξ and η be as in 6) and assume the scaling

$$\langle \eta, \xi \rangle = 1 \ .$$

It holds that

$$\frac{1}{r^j}\, A^j \to \xi\eta^T \quad \textit{as} \quad j \to \infty \ .$$

The limit matrix $\xi\eta^T$ is a projector of rank one.

Proof: a) If $\rho(A) = 0$ then $\sigma(A) = \{0\}$ and A is nilpotent, $A^n = 0$. This is not possible if $A > 0$.

b) In the following, we will assume that $r = \rho(A) = 1$. This is no restriction since we can replace A by $A_0 = \frac{1}{\rho(A)}\, A$.

There exists $\lambda \in \sigma(A)$ with $|\lambda| = 1$ and there exists $x \in \mathbb{C}^n$ with

$$Ax = \lambda x, \quad x \neq 0 \ .$$

Set $\xi = |x|_{ab}$; then

$$\xi \in \mathbb{R}^n, \quad \xi \geq 0, \quad \xi \neq 0 \ .$$

We have

$$
\begin{aligned}
\xi &= |x|_{ab} \\
&= |\lambda x|_{ab} \\
&= |Ax|_{ab} \\
&\le A|x|_{ab} \\
&= A\xi
\end{aligned}
$$

Thus, the vector $\xi \in \mathbb{R}^n$ satisfies

$$A\xi \ge \xi \ge 0, \quad \xi \ne 0 \; .$$

We claim that $A\xi = \xi$.

Suppose this does not hold. Then set

$$A\xi - \xi =: y \ge 0, \quad y \ne 0 \; .$$

We obtain that

$$A^2\xi - A\xi = Ay > 0 \; .$$

Let $z := A\xi$. Then we have

$$Az > z > 0 \; .$$

There exists $\varepsilon > 0$ so that

$$\frac{1}{1+\varepsilon}\, Az \ge z > 0 \; .$$

If we set

$$B = \frac{1}{1+\varepsilon}\, A$$

then $Bz \ge z > 0$, thus

$$B^j z \ge z > 0 \quad \text{for all} \quad j = 1, 2, \ldots$$

However, since $\rho(B) < 1$ we have $B^j \to 0$ as $j \to \infty$. This contradiction proves that $A\xi = \xi$.

We have proved that $r = \rho(A)$ is an eigenvalue of A and we have proved 3).

To continue the proof of Perron's Theorem, we will use the following:

Lemma 21.1 *For $j = 1, 2, \ldots, n$ let $z_j = r_j e^{i\alpha_j}$ denote complex numbers with*

$$|z_j| = r_j > 0, \quad \alpha_j \in \mathbb{R} \; .$$

Then the equation

$$|z_1 + z_2 + \ldots + z_n| = r_1 + r_2 + \ldots + r_n \tag{21.1}$$

holds if and only if

$$e^{i\alpha_1} = e^{i\alpha_2} = \ldots = e^{i\alpha_n} \ . \tag{21.2}$$

Proof: First let $n = 2$ and set

$$\phi = \alpha_2 - \alpha_1, \quad c = \cos\phi, \quad s = \sin\phi \ .$$

We have

$$\begin{aligned}
|z_1 + z_2|^2 &= |r_1 + r_2 e^{i\phi}|^2 \\
&= |r_1 + r_2 c + i r_2 s|^2 \\
&= (r_1 + r_2 c)^2 + (r_2 s)^2 \\
&= r_1^2 + 2 r_1 r_2 c + r_2^2
\end{aligned}$$

This equals $(r_1 + r_2)^2$ if and only if

$$1 = c = \cos\phi \ .$$

This is equivalent to $\phi = 2\pi j$ for some integer j, i.e., $|z_1 + z_2| = |z_1| + |z_2|$ holds if and only if

$$e^{i\alpha_1} = e^{i\alpha_2} \ .$$

For general n it is clear that (21.2) implies (21.1). Also, if

$$e^{i\alpha_1} \neq e^{i\alpha_2} \ ,$$

for example, then

$$|z_1 + z_2| < r_1 + r_2 \ .$$

Therefore,

$$\begin{aligned}
|z_1 + z_2 + \ldots + z_n| &\leq |z_1 + z_2| + r_3 + \ldots + r_n \\
&< r_1 + r_2 + r_3 + \ldots + r_n
\end{aligned}$$

This proves the lemma. $\diamond$

c) To continue the proof of Perron's Theorem, we assume again that $r = \rho(A) = 1$.

Let $\lambda \in \sigma(A), |\lambda| = 1$, and let $x \in \mathbb{C}^n$,

$$Ax = \lambda x, \quad x \neq 0 \ .$$

We have shown above that the vector

$$h := |x|_{ab}$$

satisfies

$$Ah = h > 0 \ .$$

(In the arguments above the vector $|x|_{ab}$ was called ξ.)

We now claim that

$$A\xi = \xi > 0 \quad \text{and} \quad Ah = h > 0$$

implies that h is a multiple ξ. To prove this set

$$M = \max_j \frac{h_j}{\xi_j} \ .$$

If h is not a multiple of ξ then

$$h \le M\xi, \quad h \ne M\xi \ , \tag{21.3}$$

and there exists j with

$$h_j = M\xi_j \ . \tag{21.4}$$

Applying A to the estimate (21.3) we obtain

$$h = Ah < MA\xi = M\xi \ ,$$

which contradicts (21.4).

So far, we have shown that

$$Ax = \lambda x, \quad x \ne 0, \quad |\lambda| = \rho(A) = 1$$

implies that

$$|x|_{ab} = M\xi$$

for some number $M > 0$ where

$$A\xi = \xi > 0 \ .$$

By scaling x, we may assume that $|x|_{ab} = \xi$, i.e.,

$$x_j = \xi_j e^{i\alpha_j}, \quad j = 1, 2, \ldots, n \ ,$$

with real α_j.

We have

$$|Ax|_{ab} = |\lambda x|_{ab} = |x|_{ab} = A|x|_{ab} \ .$$

Consider the first component of this equation, for example. It says that

$$|\sum_{j=1}^n a_{1j} x_j| = \sum_{j=1}^n a_{1j} |x_j| \ .$$

By the previous lemma, we obtain that

$$e^{i\alpha_1} = e^{i\alpha_2} = \ldots = e^{i\alpha_n} \ .$$

Thus, x is a multiple of ξ. The argument proves that A has no eigenvalue λ with $|\lambda| = \rho(A)$ except $\lambda = \rho(A)$. It also proves that the eigenvalue $\rho(A)$ is geometrically simple.

d) It remains to prove that the eigenvalue $\rho(A)$ is not just geometrically simple, but also algebraically simple. We may assume again that $\rho(A) = 1$. If this eigenvalue is not algebraically simple then there exists $z \in \mathbb{C}^n$ with

$$Az - z = \xi \quad \text{where} \quad A\xi = \xi > 0 \ .$$

Details: The matrix $B = A - I$ has the eigenvalue 0, which is geometrically simple. Suppose its algebraic multiplicity is $k \geq 2$. There exists a Jordan chain of k vectors,

$$B^{k-1}x_0, B^{k-2}x_0, \ldots, Bx_0, x_0 \ .$$

Here $\xi := B^{k-1}x_0$ satisfies $B\xi = 0$, thus $A\xi = \xi$. The vector $z := B^{k-2}x_0$ satisfies $Bz = \xi$, thus $Az - z = \xi$.
End of Details.

Write

$$z = a + ib, \quad a, b \in \mathbb{R}^n \ .$$

Obtain that

$$Aa - a + i(Ab - b) = \xi \ .$$

Since $\xi \in \mathbb{R}^n$ obtain that $Ab = b$ and

$$Aa - a = \xi, \quad a \in \mathbb{R}^n \ .$$

Choose a real number α so large that

$$y := a + \alpha\xi > 0 \ .$$

Since $Aa - a = \xi$ and $A\xi - \xi = 0$ we have

$$\begin{aligned} Ay - y &= A(a + \alpha\xi) - a - \alpha\xi \\ &= \xi + \alpha(A\xi - \xi) \\ &= \xi \ , \end{aligned}$$

thus

$$Ay > y > 0 \ .$$

There exists $\varepsilon > 0$ with

$$Ay \geq (1+\varepsilon)y \ ,$$

which yields that

$$A^j y \geq (1+\varepsilon)^j y \ .$$

Therefore, $|A^j y| \to \infty$ as $j \to \infty$.

On the other hand, for some number $M > 0$ we have

$$y \leq M\xi \ ,$$

which yields that

$$A^j y \leq M A^j \xi = M\xi \ .$$

The contradiction implies that a vector z with

$$Az - z = \xi$$

does not exist. The eigenvalue $\rho(A)$ is algebraically simple.

To prove 5), let $\eta > 0$ denote Perron's eigenvector for A^T,

$$A^T \eta = r\eta \ .$$

Let $y \in \mathbb{R}^n, y \geq 0, y \neq 0, Ay = \lambda y$. We have (note that λ is real)

$$\begin{aligned} \lambda\langle \eta, y\rangle &= \langle \eta, \lambda y\rangle \\ &= \langle \eta, Ay\rangle \\ &= \langle A^T\eta, y\rangle \\ &= r\langle \eta, y\rangle \end{aligned}$$

The equation $\lambda = r$ follows.

6) First assume that A has a complete set of eigenvectors, $\xi, v_2, \ldots, v_n$:

$$A\xi = r\xi, \quad Av_k = \lambda_k v_k, \quad |\lambda_k| < r \ .$$

Write

$$y = c\xi + \sum_{k=2}^n c_k v_k$$

and obtain that

$$\frac{1}{r^j} A^j y = c\xi + \sum_{k=2}^n c_k \Big(\frac{\lambda_k}{r}\Big)^j v_k \ .$$

It follows that

$$\frac{1}{r^j} A^j y \to c\xi \ . \tag{21.5}$$

Since $A^T\eta = r\eta$ it follows that

$$\langle \eta, \frac{1}{r^j}\, A^j \rangle = \langle \eta, y \rangle$$

is independent of j. In the above equation, use the limit relation (21.5) to obtain that

$$\langle \eta, c\xi \rangle = \langle \eta, y \rangle \ ,$$

i.e., $c = \langle \eta, y \rangle / \langle \eta, \xi \rangle$.

In the general case, there exists $T \in \mathbb{C}^{n\times n}$ with

$$T^{-1}AT = \begin{pmatrix} r & 0 \\ 0 & B \end{pmatrix}, \quad \rho(B) < r \ ,$$

and ξ is the first column of T. One obtains that

$$\frac{1}{r}\, A = T \begin{pmatrix} 1 & 0 \\ 0 & \tilde{B} \end{pmatrix} T^{-1}, \quad \rho(\tilde{B}) < 1 \ ,$$

and

$$\lim_{j\to\infty} \frac{1}{r^j}\, A^j y = T \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} T^{-1} y \ .$$

Therefore,

$$\lim_{j\to\infty} \frac{1}{r^j}\, A^j y = T \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix} = c\xi$$

The equation

$$c = \frac{\langle \eta, y \rangle}{\langle \eta, \xi \rangle} > 0$$

follows as above.

7) Assume that ξ and η are scaled so that $\langle \eta, \xi \rangle = 1$. Then, by 6), we have for every $y \in \mathbb{R}^n$:

$$\lim_{j\to\infty} \frac{1}{r^j}\, A^j y = \langle \eta, y \rangle \xi \ .$$

Here

$$\langle \eta, y \rangle \xi = (\eta^T y)\xi = \xi \eta^T y \ .$$

Since the convergence holds for every $y \in \mathbb{R}^n$ it follows that

$$\frac{1}{r^j}\, A^j \to \xi \eta^T \ .$$

It is clear that

$$(\xi\eta^T)^2 = \xi\eta^T\xi\eta^T = \xi\eta^T \ .$$

Thus, the limit matrix is a projector.

This completes the proof of Perron's Theorem. $\diamond$

21.2 Frobenius's Theory

In this section let $n \geq 2, A \in \mathbb{R}^{n\times n}, A \geq 0$. The matrix

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

has the spectral radius $\rho(A) = 0$. Clearly, property 1) of Perron's Theorem 21.1 does not hold for all non–negative matrices.

However, for every non–negative matrix A the spectral radius is an eigenvalue corresponding to a non–negative eigenvector:

Theorem 21.2 *Let $A \in \mathbb{R}^{n\times n}, A \geq 0$. Set $r = \rho(A)$. There exists $\xi \in \mathbb{R}^n, \xi \geq 0, \xi \neq 0$, with*

$$A\xi = r\xi \ .$$

Proof: Let $E = (e_{ij}) \in \mathbb{R}^{n\times n}$ denote the matrix with entries $e_{ij} = 1$ for all i, j. Set

$$A_k = A + \frac{1}{k}E \quad \text{for} \quad k = 1, 2, \ldots$$

Clearly, $A_k > 0$, and Perron's Theorem applies to A_k. Set $r_k = \rho(A_k)$, thus $r_k > 0$. There exists

$$\xi^{(k)} \in \mathbb{R}^n, \quad \xi^{(k)} > 0, \quad |\xi^{(k)}|_\infty = 1$$

with

$$(A + \frac{1}{k}E)\xi^{(k)} = r_k\xi^{(k)} \ .$$

Since the sequences $\xi^{(k)} \in \mathbb{R}^n$ and $r_k \in \mathbb{R}$ are bounded there exists a subsequence $k \in \mathbb{N}_1$ with

$$\xi^{(k)} \to \xi, \quad r_k \to r^* \quad \text{as} \quad k \to \infty, \quad k \in \mathbb{N}_1 \ .$$

It follows that

$$\xi \geq 0, \quad |\xi|_\infty = 1, \quad A\xi = r^*\xi \ .$$

It remains to prove that $r^* = r$, i.e., that r^* is the spectral radius of A. Since r^* is an eigenvalue of A we have $r^* \leq r$.

Suppose that $r^* < r$, thus $r - r^* = \delta > 0$. The matrix A has an eigenvalue λ with $|\lambda| = r$. For large k the matrix A_k has an eigenvalue λ_k near λ, i.e., with

$$|\lambda - \lambda_k| < \frac{\delta}{2} \; .$$

It then follows that

$$r_k \geq r^* + \frac{\delta}{2}$$

for all large k. This contradicts the convergence $r_k \to r^*$ for $k \in \mathbb{N}_1$. $\diamond$

In the proof of the next theorem we will use two simple results on eigenvalues:

Lemma 21.2 *1) Let $\lambda \in \mathbb{C}$ denote an eigenvalue of the matrix $A \in \mathbb{C}^{n\times n}$ with algebraic multiplicity k. Then the eigenvalue $1+\lambda$ of $I + A$ also has algebraic multiplicity k.*

2) Let $\lambda \in \mathbb{C}$ denote an eigenvalue of the matrix $B \in \mathbb{C}^{n\times n}$ and let $m \in \mathbb{N}, m \geq 2$. Then λ^m is an eigenvalue of B^m, and if λ^m is an algebraically simple eigenvalue of B^m then the eigenvalue λ of B is also algebraically simple.

Proof: 1) By Schur's theorem there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ so that

$$U^*AU = \Lambda + R$$

where R is strictly upper–triangular and

$$\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

is diagonal. If λ is an eigenvalue of A of algebraic multiplicity k then λ comes up k times in the string $\lambda_1, \lambda_2, \ldots, \lambda_n$. The eigenvalue $1+\lambda$ of

$$I + A = U^*(I + \Lambda)U$$

comes up k times in the string

$$1+\lambda_1, 1+\lambda_2, \ldots, 1+\lambda_n$$

2) As above, let U be unitary and let $U^*BU = \Lambda + R$ where Λ is diagonal and R is strictly upper–triangular. One obtains that

$$U^*B^mU = \Lambda^m + \tilde{R}$$

where $\tilde{R}$ is strictly upper–triangular. If

$$\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

then

$$\Lambda^m = diag(\lambda_1^m, \lambda_2^m, \ldots, \lambda_n^m) \; .$$

If λ is an eigenvalue of B then λ^m is an eigenvalue of Λ^m. If λ^m is algebraically simple then λ^m comes up only once in the string $\lambda_1^m, \lambda_2^m, \ldots, \lambda_n^m$ and, therefore, λ comes up only once in the string $\lambda_1, \lambda_2, \ldots, \lambda_n$. $\diamond$

The following theorem gives a condition which ensures that the spectral radius $\rho(A)$ of the non–negative matrix A is strictly positive and algebraically simple. Furthermore, the corresponding eigenvector ξ is strictly positive. For positive matrices, these are the properties 1), 2), and 3) of Perron's Theorem.

Theorem 21.3 *Let $A \in \mathbb{R}^{n\times n}, A \geq 0$, and assume that*

$$(I + A)^m > 0$$

for some $m \in \mathbb{N}$. Then $r = \rho(A)$ is positive, and r is an algebraically simple eigenvalue of A. There exists $\xi \in \mathbb{R}^n$ with

$$A\xi = r\xi, \quad \xi > 0 \ .$$

Proof: Set $r = \rho(A)$. The matrix $I + A$ has the spectral radius $1 + r$ and, by the previous theorem, there exists $\xi \in \mathbb{R}^n$ with

$$(I + A)\xi = (1 + r)\xi, \quad \xi \geq 0, \quad \xi \neq 0 \ .$$

It follows that

$$(I + A)^m\xi = (1 + r)^m\xi, \quad \xi > 0 \ .$$

By Perron's Theorem, the eigenvalue $(1+r)^m$ of the positive matrix $(I+A)^m$ is algebraically simple. Using the previous lemma we obtain that the eigenvalue r of A is algebraically simple.

The assumption $(I + A)^m > 0$ implies that A is not zero. We have obtained that $A\xi = r\xi$ and $\xi > 0$. Therefore, $A\xi \neq 0$ and $r > 0$ follows. $\diamond$

Let $A \in \mathbb{R}^{n\times n}$ denote a non–negative matrix. How can we check if the previous theorem applies, i.e., if there exists a strictly positive power, $(I+A)^m > 0$? The directed graph of A is a useful tool.

Directed Graph of a Matrix. Let $A \in \mathbb{C}^{n\times n}$. The directed graph $\mathcal{G} = \mathcal{G}(A)$ of A consists of n nodes $N_1, \ldots, N_n$ with a directed edge from N_i to N_j if and only if $a_{ij} \neq 0$.

The graph $\mathcal{G}$ is called strongly connected if for all nodes N_i, N_j there is a sequence of directed edges from N_i to N_j.

The matrix A is called irreducible if its directed graph is strongly connected. Otherwise, A is called reducible. One can show that A is reducible if and only if there exists a permutation matrix P so that

$$P^T AP = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix} \tag{21.6}$$

where X and Z are square matrices.

Theorem 21.4 *Let $A \in \mathbb{R}^{n\times n}, A \geq 0$.*

a) If A is irreducible then

$$(I + A)^{n-1} > 0 \ .$$

b) If A is reducible then a strictly positive power $(I+A)^m, m \in \mathbb{N}$, does not exist.

Proof: a) We have

$$\Big((I+A)^{n-1}\Big)_{ij} = \sum_{k=0}^{n-1} \begin{pmatrix} n-1 \\ k \end{pmatrix} (A^k)_{ij} \; . \tag{21.7}$$

Clearly, the diagonal of $(I+A)^{n-1}$ is strictly positive.

Let $i \neq j$. By assumption, there exist q distinct indices

$$i_1, i_2, \ldots, i_q \in \{1, 2, \ldots, n\} \setminus \{i, j\}$$

so that

$$a_{ii_1} > 0, \quad a_{i_1 i_2} > 0, \ldots, a_{i_q j} > 0 \; .$$

Here $q \leq n-2$.

We have

$$\begin{aligned} (A^2)_{ii_2} &= \sum_{\alpha=1}^{n} a_{i\alpha} a_{\alpha i_2} > 0 \\ (A^3)_{ii_3} &= \sum_{\alpha=1}^{n} (A^2)_{i\alpha} a_{\alpha i_3} > 0 \end{aligned}$$

etc.

One obtains that

$$(A^{q+1})_{ij} > 0 \quad \text{and} \quad q+1 \leq n-1 \; .$$

The formula (21.7) yields that $(I+A)^{n-1} > 0$.

b) If (21.6) holds then a positive power of $I+A$ does not exist. $\diamond$

Question: Let $A \in \mathbb{R}^{n \times n}$ be irreducible and $A \geq 0$, thus

$$(I+A)^{n-1} > 0 \; .$$

The previous two theorems apply. Is it possible that A has an eigenvalue $\lambda \in \mathbb{C}$ with

$$|\lambda| = r = \rho(A), \quad \lambda \neq r?$$

The answer is yes. The matrix

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

gives a simple example. This shows that property 4) of Perron's Theorem does not hold, in general, for irreducible non–negative matrices.

However, the following holds:

Theorem 21.5 *Let $A \in \mathbb{R}^{n\times n}, A \geq 0$, and assume that $A^m > 0$ for some positive integer m. Then the spectral radius $r = \rho(A)$ is positive and is an algebraically simple eigenvalue of A. Furthermore, if $\lambda \in \mathbb{C}$ is any eigenvalue of A and $\lambda \neq r$, then $|\lambda| < r$.*

Proof: Set $r = \rho(A)$. By Perron's Theorem we have $r^m = \rho(A^m) > 0$, thus $r > 0$. By Schur's Theorem we can write

$$U^*AU = \Lambda + R, \quad U^*U = I \ ,$$

where Λ is diagonal and R is strictly upper triangular. It follows that

$$U^*A^mU = \Lambda^m + \tilde{R}$$

where $\tilde{R}$ is strictly upper triangular. Since r^m is an algebraically simple eigenvalue for A^m the number r^m occurs only one time in Λ^m. Thus r occurs only once in Λ and r is algebraically simple for A.

Let $\lambda \in \sigma(A)$, thus $\lambda^m \in \sigma(A^m)$. Assume that $|\lambda| = r$, thus $|\lambda^m| = r^m$. By Perron's Theorem, we have

$$\lambda^m = r^m \quad \text{or} \quad |\lambda^m| < r^m \ .$$

Clearly, if $|\lambda^m| < r^m$ then $|\lambda| < r$. Thus it remains to discuss the case $\lambda^m = r^m$. If $\lambda \neq r$ but $\lambda^m = r^m$ then the eigenvalue r^m of A^m has geometric multiplicity at least equal to 2, a contradiction. It follows that $\lambda = r$ if $\lambda^m = r^m$. $\diamond$

21.3 Discrete–Time Markov Processes

We let the time variable t evolve in $\{0, 1, 2, \ldots\}$, i.e.,

$$t \in \{0, 1, 2, \ldots\} \ .$$

Let X_t denote a random variable evolving in the finite state space

$$S = \{S_1, S_2, \ldots, S_n\} \ .$$

A Markov process is determined by the probabilities

$$p_{ij} = prob\Big(X_{t+1} = S_i \Big| X_t = S_j\Big) \ .$$

With probability p_{ij} the random variable X_{t+1} is in state S_i under the assumption that X_t is in state S_j.

Clearly, $0 \leq p_{ij} \leq 1$. The $n \times n$ probability matrix $P = (p_{ij})$ satisfies

$$\sum_{i=1}^n p_{ij} = 1 \quad \text{for all} \quad j = 1, 2, \ldots, n \ .$$

If

$$e^T = (1, 1, \ldots, 1)$$

then

$$e^T P = e^T \ .$$

Therefore, the probability matrix P is called column–stochastic; each column sum of P equals one.

Let $q_t \in \mathbb{R}^n$ denote the probability distribution of the random variable X_t, i.e.,

$$(q_t)_j = prob\Big(X_t = S_j\Big) \quad \text{for} \quad j = 1, 2, \ldots, n \ .$$

We have $X_t = S_j$ with probability $(q_t)_j$.

Assuming that $X_t = S_j$ we have $X_{t+1} = S_i$ with probability p_{ij}. Therefore,

$$(q_{t+1})_i = \sum_{j=1}^n p_{ij}(q_t)_j \quad \text{for} \quad i = 1, 2, \ldots, n \ .$$

One obtains the important relation

$$q_{t+1} = P q_t \quad \text{for} \quad t = 0, 1, 2, \ldots$$

for the evolution of the probability density of the random variable X_t.

Application of Perron's Theorem. Assume $P > 0$. Since

$$P^T e = e \quad \text{for} \quad e = (1, 1, \ldots, 1)^T$$

we obtain that $\rho(P) = \rho(P^T) = 1$.

By Perron's Theorem, there exists a unique vector $\xi \in \mathbb{R}^n$ with

$$P\xi = \xi, \quad \sum_{j=1}^n \xi_j = 1 \ .$$

By property 6) of Perron's Theorem we have

$$q_t = P^t q_0 \to \xi \quad \text{as} \quad t \to \infty \ .$$

The normalized Perron vector ξ of P is the unique stationary probability density of the Markov process. Given any initial probability density q_0, the probability density ξ is approached as $t \to \infty$.