Krešimir Veselić

Damped oscillations of linear systems — a mathematical introduction

Wednesday 30th March, 2011

Springer

To my wife.

Foreword

The theory of linear damped oscillations has been studied for more than hundred years and is still of vital interest to researchers in Control Theory, Optimization, and computational aspects. This theory plays a central role in studying the stability of mechanical structures, but it has applications to other fields such as electrical network systems or quantum mechanical systems. We have purposely restricted ourselves to the basic model leaving aside gyroscopic effects and free rigid body motion. In contrast, the case of a singular mass matrix is analysed in some detail. We spend quite a good deal of time discussing underlying spectral theory, not forgetting to stress its limitations as a tool for our ultimate objective — the time behaviour of a damped system. We have restricted ourselves to finite dimension although we have attempted, whenever possible, to use methods which allow immediate generalisation to the infinite-dimensional case.

Our text is intended to be an introduction to this topic and so we have tried to make the exposition as self-contained as possible. This is also the reason why we have restricted ourselves to finite dimensional models. This lowering of the technical barrier should enable the students to concentrate on central features of the phenomenon of damped oscillation.

The introductory chapter includes some paragraphs on the mechanical model in order to accommodate readers with weak or no background in physics.

The text presents certain aspects of matrix theory which are contained in monographs and advanced textbooks but may not be familiar to the typical graduate student. One of them is spectral theory in indefinite product spaces. This topic receives substantial attention because it is a theoretical fundament for our model. We do not address numerical methods especially designed for this model. Instead we limit ourselves to mention what can be done with most common matrix algorithms and to systematically consider the sensitivity, that is, *condition number estimates and perturbation properties* of our topics. In our opinion numerical methods, in particular invariant-subspace reduction for J-symmetric matrices are still lacking and we have tried to make a case for a more intensive research in this direction.

Our intention is to take readers on a fairly sweeping journey from the basics to several research frontiers. This has dictated a rather limited choice of material, once we 'take off' from the basics. The choice was, of course, closely connected with our own research interests. In some cases we present original research (cf. e.g. the chapter on modal approximations), whereas we sometimes merely describe an open problem worth investigating — and leave it unsolved.

This text contains several contributions of our own which may be new. As a rule, we did not give any particular credit for earlier findings due to other authors or ourselves except for more recent ones or when referring to further reading. Our bibliography is far from exhaustive, but it contains several works offering more detailed bibliographical coverage.

The text we present to the reader stems from a synonymous course I have taught for graduate and post-doc students of the Department of Mathematics, University of Osijek, Croatia in the summer term 2007/2008. It took place at the Department of Mathematics at the University of Osijek and was sponsored by the program "Brain gain visitor" from the National Foundation for Science, Higher Education and Technological development of the Republic of Croatia. Due to its origin, it is primarily designed for students of Mathematics but it will be of use also to engineers with enough mathematical background. Even though the text does not cover all aspects of the linear theory of damped oscillations, I hope that it will also be of some help to the researchers in this field.

In spite of a good deal of editing the text still contains some remnants of its oral source. This pertains to the sometimes casual style more suited to a lecture than to a monograph — as was the original aim of this work.

Bibliographical Notes and Remarks are intended to broaden the scope by mentioning some other important directions of research present and past. According to our bibliographical policy, when presenting a topic we usually cite at most one or two related works and refer to their bibliographies.

It is a pleasure to thank the participants in the course, in particular professor N. Truhar and his collaborators for their discussions and support during the lectures in Osijek. The author was very fortunate to have obtained large lists of comments and corrections from a number of individuals who have read this text. These are (in alphabetic order): K. Burazin (Osijek), L. Grubišić (Zagreb), D. Kressner (Zürich), P. Lancaster, (Calgary), H. Langer (Vienna), I. Matić (Osijek), V. Mehrmann (Berlin), M. Miloloža Pandur (Osijek), B. Parlett (Berkeley), I. Nakić (Zagreb), A. Suhadolc (Ljubljana), I. Veselić (Chemnitz), A. Wiegner (Hagen) and, by no means the least, the anonymous referees. Many of them went into great detail and/or depth. This resulted in substantial revisions of the whole text as well as in significant enlargements. To all of them I am deeply obliged. Without their generous support and encouragement this text would hardly deserve to be presented to the public. The financial help of both National Foundation for Science, Higher Education and Technological development of the Republic of Croatia and the Department of Mathematics, University of Osijek, not forgetting the warm hospitality of the latter, is also gratefully acknowledged.

Osijek/Hagen, July 2008 to December 2010, K. Veselić.

Contents

Introduction

Here is some advice to make this book easier to read.

In order to check/deepen the understanding of the material and to facilitate independent work the reader is supplied with some thoroughly worked-out examples as well as a number of exercises. Not all exercises have the same status. Some of them are 'obligatory' because they are quoted later. Such exercises are either easy and straightforward or accompanied by hints or sketches of solution. Some just continue the line of worked examples. On the other end of the scale there are some which introduce the reader to research along the lines of the development. These are marked by the word 'try' in their statement. It is our firm conviction that a student can fully digest these Notes only if he/she has solved a good quantity of exercises.

Besides examples and exercises there are a few theorems and corollaries the proof of which is left to the reader. The difference between a corollary without proof and an exercise without solution lies mainly in their significance in the later text.

We have tried to minimise the interdependencies of various parts of the text.

What can be skipped on first reading? Most of the exercises. Almost none of the examples. Typically the material towards the end of a chapter. In particular

- Chapter 10: Theorem 10.12 –
- Chapter 12: Theorem 12.17 –
- Any of Chapters 19 – 22

Prerequisites and terminology. We require standard facts of matrix theory over the real or complex field Ξ including the following (cf. [31] or [52]).

- Linear (in)dependence, dimension, orthogonality. Direct and orthogonal sums of subspaces.
- Linear systems, rank, matrices as linear maps. We will use the terms *injective/surjective* for a matrix with linearly independent columns/rows.
- Standard matrix norms, continuity and elementary analysis of matrix-valued functions.
- Standard matrix decompositions such as
  - Gaussian elimination and the LU-decomposition $A = LU$, L lower triangular with unit diagonal and U upper triangular.
  - idem with pivoting $A = LU\Pi$, L, U as above and Π a permutation matrix, corresponding to the standard row pivoting.
  - Cholesky decomposition of a positive definite Hermitian matrix $A = LL^*$, L lower triangular.
  - QR-decomposition of an arbitrary matrix $A = QR$, Q unitary and R upper triangular.
  - Singular value decomposition of an arbitrary matrix $A = U\Sigma V^*$, U, V unitary and $\Sigma \geq 0$ diagonal.
  - Eigenvalue decomposition of a Hermitian matrix $A = U\Lambda U^*$, U unitary, Λ diagonal.
  - Simultaneous diagonalisation of a Hermitian matrix pair A, B (the latter positive definite)

$$\Phi^* A \Phi = \Lambda, \quad \Phi^* B \Phi = I, \quad \Lambda \text{ diagonal.}$$

  - Schur (or triangular) decomposition of an arbitrary square $A = UTU^*$, U unitary, T upper triangular.

  - Polar decomposition of an arbitrary $m \times n$-matrix with $m \geq n$

$$A = U\sqrt{A^*A} \text{ with } U^*U = I_n.$$

- Characteristic polynomial, eigenvalues and eigenvectors of a general matrix; their geometric and algebraic multiplicity.
- (desirable but not necessary) Jordan canonical form of a general matrix.

Further general prerequisites are standard analysis in Ξ^n as well as elements of the theory of analytic functions.

Notations. Some notations were implicitly introduced in the lines above. The following notational rules are not absolute, that is, exceptions will be made, if the context requires them.

- Scalar: lower case Greek $\alpha, \lambda, \ldots$.
- Column vector: lower case Latin $x, y, \ldots$, their components: $x_j, y_j, \ldots$ (or $(x)_j, (y)_j, \ldots$)
- Canonical basis vectors: e_j as $(e_j)_k = \delta_{jk}$.
- Matrix order: $m \times n$, that is, the matrix has m rows and n columns, if a matrix is square then we also say it to be of order n.
- The set of all $m \times n$-matrices over the field Ξ: $\Xi^{m,n}$
- Matrix: capital Latin $A, B, X, \ldots$, sometimes capital Greek $\Phi, \Psi, \ldots$; diagonal matrices: Greek $\boldsymbol{\alpha}, \Lambda, \ldots$ (bold face, if lower case). By default the order of a general square matrix will be n, for phase-space matrices governing damped systems this order will mostly be $2n$.
- Identity matrix, zero matrix: I_n, $0_{m,n}$, 0_n; the subscripts will be omitted whenever clear from context.
- Matrix element: corresponding lower case of the matrix symbol $A = (a_{ij})$.
- Block matrix: $A = (A_{ij})$.
- Matrix element in complicated expressions: $ABC = ((ABC)_{ij})$.
- Diagonal matrix: $\mathrm{diag}(a_1, \ldots, a_n)$; block diagonal matrix: $\mathrm{diag}(A_1, \ldots, A_p)$.
- Matrix transpose: $A^T = (a_{ji})$.
- Matrix adjoint: $A^* = (\overline{a_{ji}})$.
- Matrix inverse and transpose/adjoint: $(A^T)^{-1} = A^{-T}, \quad (A^*)^{-1} = A^{-*}$.
- The null space and the range of a matrix as a linear map: $\mathcal{N}(A)$, $\mathcal{R}(A)$.
- Spectrum: $\sigma(A)$.
- Spectral radius: $\mathrm{spr}(A) = \max |\sigma(A)|$.
- Spectral norm: $\|A\| = \sqrt{\mathrm{spr}(A^*A)}$ and condition number $\kappa(A) = \|A\|\|A^{-1}\|$.
- Euclidian norm: $\|A\|_E = \sqrt{\mathrm{Tr}(A^*A)}$ and condition number $\kappa_E(A) = \|A\|_E\|A^{-1}\|_E$. In the literature this norm is sometimes called the Frobenius norm or the Hilbert-Schmidt norm.
- Although a bit confusing the term *eigenvalues* (in plural) will mean any sequence of the complex zeros of the characteristic polynomial, counted with their multiplicity. Eigenvalues as elements of the spectrum will be called *spectral points* or *distinct eigenvalues*.

Chapter 1
The model

Small damped oscillations in the absence of gyroscopic forces are described by the vector differential equation

$$M\ddot{x} + C\dot{x} + Kx = f(t). \tag{1.1}$$

Here $x = x(t)$ is an $\mathbb{R}^n$-valued function of time $t \in \mathbb{R}$; M, C, K are real symmetric matrices of order n. Typically M, K are positive definite whereas C is positive semidefinite. $f(t)$ is a given vector function. The physical meaning of these objects is

$x_j(t)$	position or displacement
M	mass
C	damping
K	stiffness
$f(t)$	external force

while the dots mean time derivatives. So, any triple M, C, K will be called a *damped system*, whereas the solution $x = x(t)$ is called *the motion* or also *the response* of the linear system to the external force $f(t)$.

In this chapter we will describe common physical processes, governed by these equations and give an outline of basic mechanical principles which lead to them. It is hoped that this introduction is self-contained enough to accomodate readers with no background in Physics.

Example 1.1 As a model example consider the spring-mass system like the one in Fig. 1.1. Here $x = \begin{bmatrix} x_1 \cdots x_n \end{bmatrix}^T$ where $x_i = x_i(t)$ is the horizontal displacement of the i-th mass point from its equilibrium position and

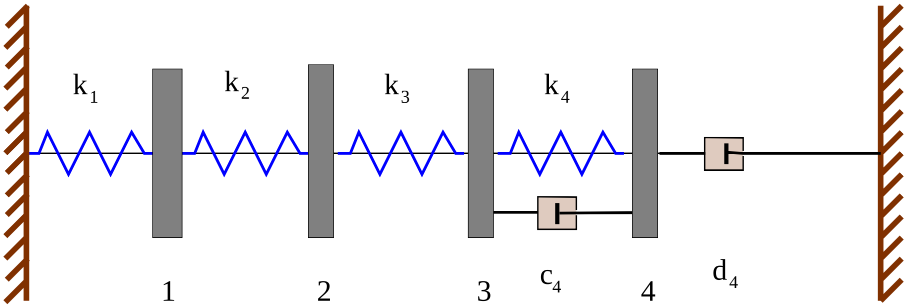

Fig. 1.1 Oscillator ladder

$$M = \operatorname{diag}(m_1, \ldots, m_n), \quad m_1, \ldots, m_n > 0 \tag{1.2}$$

$$K = \begin{bmatrix} k_1 + k_2 & -k_2 & & & \\ -k_2 & k_2 + k_3 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -k_n \\ & & & -k_n & k_n + k_{n+1} \end{bmatrix}, \quad k_1, \ldots k_{n+1} > 0 \tag{1.3}$$

(all void positions are zeros) and

$$C = C_{in} + C_{out}, \tag{1.4}$$

$$C_{in} = \begin{bmatrix} c_1 + c_2 & -c_2 & & & \\ -c_2 & c_2 + c_3 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -c_n \\ & & & -c_n & c_n + c_{n+1} \end{bmatrix}, \quad c_1, \ldots c_{n+1} \geq 0 \tag{1.5}$$

$$C_{out} = \operatorname{diag}(d_1, \ldots, d_n), \quad d_1, \ldots, d_n \geq 0 \tag{1.6}$$

In the case of Fig. 1.1 we have $n = 4$ and

$$k_5 = 0, \quad c_2 = c_3 = 0,$$

$$d_1 = d_2 = d_3 = 0, \quad d_4 > 0.$$

Various dynamical quantities such as 'force', 'work', 'energy' will play an important role in the course of these Notes. We shall next give a short compact overview of the physical background of the equation (1.1).

1.1 Newton's law

We will derive the equations (1.1) from Newton's law for our model example. The coordinate x_i of the i-th mass point is measured from its static equilibrium position, that is, from the point where this mass is at rest and $f = 0$. Newton's second law of motion for the i-th particle reads

$$m_i \ddot{x}_i = f_i^{tot}$$

where f_i^{tot} is the sum of all forces acting upon that mass point. These forces are

- The external force $f_i(t)$.
- The elastic force from the neighbouring springs, negative proportional to the relative displacement.

  $$k_i(x_{i-1} - x_i) + k_{i+1}(x_{i+1} - x_i), \quad i = 1, \ldots, n \tag{1.7}$$

  where k_j is the j-th *spring stiffness.*
- The inner damping force from the neighbouring dampers, negative proportional to the relative velocity

  $$c_i(\dot{x}_{i-1} - \dot{x}_i) + c_{i+1}(\dot{x}_{i+1} - \dot{x}_i), \quad i = 1, \ldots, n. \tag{1.8}$$

- The external damping force, negative proportional to the velocity $-d_i \dot{x}_i$, where d_j, c_j are the respective *inner and external damper viscosities.*

Here to simplify the notation we have set $x_0 = x_{n+1} = 0, \dot{x}_0 = \dot{x}_{n+1} = 0$, these are the fixed end points. Altogether we obtain (1.1) with M, C, K, from (1.2), (1.3) and (1.4). All these matrices are obviously real and symmetric. By

$$x^T M x = \sum_{j=1}^{n} m_j x_j^2 \tag{1.9}$$

and

$$x^T K x = k_1 x_1^2 + \sum_{j=2}^{n} k_j (x_j - x_{j-1})^2 + k_{n+1} x_n^2 \tag{1.10}$$

both M and K are positive definite. By the same argument C_{in} and C_{out} are positive semidefinite.

Exercise 1.2 *How many coefficients k_i in (1.3) may vanish while keeping K positive definite? Interpret the response physically!*

Example 1.3 A typical external force stems from a given movement of the frame in which the vibrating structure is anchored (this is the so-called *inertial force*). On the model example in Fig. 1.1 the system is anchored on two points. The force caused by the horizontal movement of these two points is taken into account by replacing the zero values of x_0, x_{n+1} in (1.7) and (1.8) by given functions $x_0(t), x_{n+1}(t)$, respectively. This does not change the matrices M, C, K whereas $f(t)$ reads

$$f(t) = \begin{bmatrix} k_1 x_0(t) + c_1 \dot{x}_0(t) \\ 0 \\ \vdots \\ 0 \\ k_{n+1} x_{n+1}(t) + c_{n+1} \dot{x}_{n+1}(t) \end{bmatrix}.$$

Exercise 1.4 *Find the equilibrium configuration (i.e. the displacement x at rest) of our model example if the external force f (i) vanishes or (ii) is a constant vector. Hint: the matrix K can be explicitly inverted.*

The model in Example 1.1 is important for other reasons too. It describes a possible discretisation of a continuous damped system (vibrating string). There the parameters c_i come from the inner friction within the material whereas d_i describe the external damping caused by the medium in which the system is moving (air, water) or just artificial devices (dashpots) purposely built in to calm down dangerous vibrations. In the latter case there usually will be few such dashpots resulting in the low rank matrix C_{out}.

It should be mentioned that determining the inner damping matrix for complex vibrating systems in real life may be quite a difficult task, it involves special mathematical methods as well as experimental work.

From the derivations above we see that determining the equilibrium *precedes* any work on oscillations. The equilibrium is the first approximation to the true behaviour of the system, it is found by solving a linear or nonlinear system of equations. The next approximation, giving more detailed information are the linear oscillations around the equilibrium, that is, their movement is governed by a system of linear differential equations. Their linearity will be seen to be due to the assumption that the oscillations have small amplitudes.

1.2 Work and energy

We will now introduce some further relevant physical notions based on Example 1.1. *The work* performed by any force $\phi_j(t)$ on the j-th point mass in the time interval $t_1 \le t \le t_2$ is given by

$$\int_{t_1}^{t_2} \phi_j(t) \dot{x}_j dt,$$

this corresponds to the rule ‘work equals force times distance’. So is the work of the external force:

$$\int_{t_1}^{t_2} f_j(t)\dot{x}_j dt, \tag{1.11}$$

of the damping force:

$$\int_{t_1}^{t_2} [c_j(\dot{x}_{j-1} - \dot{x}_j) + c_{j+1}(\dot{x}_{j+1} - \dot{x}_j)]\dot{x}_j dt - \int_{t_1}^{t_2} d_j \dot{x}_j^2 dt,$$

of the elastic force:

$$\int_{t_1}^{t_2} [k_j(x_{j-1} - x_j) + k_{j+1}(x_{j+1} - x_j)]\dot{x}_j dt.$$

To this we add the work of the so-called 'inertial force' which is given by

$$\int_{t_1}^{t_2} m_j \ddot{x}_j \dot{x}_j dt \tag{1.12}$$

The total work is the sum over all mass points, so from (1.11) – (1.12) we obtain

$$\int_{t_1}^{t_2} \dot{x}^T f(t) dt, \tag{1.13}$$

$$-\int_{t_1}^{t_2} \dot{x}^T C \dot{x} dt,$$

$$-\int_{t_1}^{t_2} \dot{x}^T K x dt = 2(E_p(x(t_1)) - E_p(x(t_2))),$$

$$\int_{t_1}^{t_2} \dot{x}^T M \ddot{x} dt = 2(E_k(x(t_2)) - E_k(x(t_1))), \tag{1.14}$$

as the total work of the external forces, damping forces, elastic forces and inertial forces, respectively; here

$$E_p(x) = \frac{1}{2} x^T K x, \tag{1.15}$$

$$E_k(\dot{x}) = \frac{1}{2} \dot{x}^T M \dot{x} \tag{1.16}$$

are *the potential* and *the kinetic energy* and

$$E(x, \dot{x}) = E_p(x) + E_k(\dot{x}) \tag{1.17}$$

the total energy. Note the difference: in the first two cases the work depends on the whole motion $x(t)$, $t_1 \leq t \leq t_2$ whereas in the second two cases it depends just on the values of E_p, E_k, respectively, taken at the end points of the motion. In the formulae (1.9), (1.10) and (1.13) – (1.14) we observe a property of both potential and kinetic energy: they are 'additive' magnitudes, that is, the kinetic energy is a sum of the kinetic energies of single point masses whereas the potential energy is a sum of the potential energies of single springs.

The relations (1.13) – (1.17) make sense and will be used for general damped systems. By premultiplying (1.1) by $\dot{x}^T$ we obtain *the differential energy balance*

$$\frac{d}{dt}E(x,\dot{x}) + \dot{x}^T C\dot{x} = \dot{x}^T f(t). \tag{1.18}$$

This is obviously equivalent to *the integral energy balance*

$$E(x,\dot{x})\Big|_{t_1}^{t_2} = -\int_{t_1}^{t_2} \dot{x}^T C\dot{x}dt + \int_{t_1}^{t_2} \dot{x}^T f(t)dt. \tag{1.19}$$

for any $t_1 \le t \le t_2$. We see that the work of the damping forces is always non-positive. This effect is called *the energy dissipation.* Thus, (1.19) displays the amount of energy which is transformed into thermal energy. If $f \equiv 0$ then this energy loss is measured by the decrease of the total energy.

If both $f \equiv 0$ and $C = 0$ then the total energy is preserved in time; such systems are called *conservative.*

1.3 The formalism of Lagrange

The next (and last) step in our short presentation of the dynamics principles is to derive the Lagrangian formalism which is a powerful and simple tool in modelling mechanical systems. The position (also called *configuration*) of a mechanical system is described by the generalised coordinates $q_1, \dots, q_s$ as components of the vector q from some region of $\mathbb{R}^s$ which is called *the configuration space.* The term 'generalised' just means that q_i need not be a Euclidian coordinate as in (1.1) but possibly some other measure of position (angle and the like). To this we add one more copy of $\mathbb{R}^s$ whose elements are the *generalised velocities* $\dot{q}$. The dynamical properties are described by the following four functions, defined on $\mathbb{R}^{2s}$,

- the kinetic energy $T = T(q,\dot{q})$, having a minimum equal to zero at $(q,0)$ for any q and with $T''_{\dot{q}}(q,\dot{q})$ positive semidefinite for any q,
- the potential energy $V = V(q)$ with $V'(q_0) = 0$ and $V''(q)$ everywhere positive definite (that is, V is assumed to be uniformly convex),
- the dissipation function $Q = Q(q,\dot{q})$, having a minimum equal to zero at $(q_0,0)$, with $Q''_{\dot{q}}(q,\dot{q})$ positive semidefinite for any q as well as
- the time dependent *generalised external force* $f = f(t) \in \mathbb{R}^s$

Here the first three functions are assumed to be smooth enough (at least twice continuously differentiable) and the symbols $'$ and $''$ denote the vector of first derivatives (gradient) and the matrix of second derivatives (Hessian), respectively. The subscript (if needed) denotes the set of variables with respect to which the derivative is taken. From these functions we construct *the Lagrangian*

$$L = T - V + f^T q.$$

The time development of this system is described by the *Lagrange equations*

$$\frac{d}{dt} L'_{\dot{q}} - L'_q + Q'_{\dot{q}} = 0. \tag{1.20}$$

This is a system of differential equations of at most second order in time. Its solution, under our, rather general, conditions is not easily described. If we suppose that the solution $q = q(t)$ does not depend on time, i.e. the system is at rest then $\dot{q} \equiv 0$ in (1.20) yields

$$-V'(q) = f$$

with the unique solution $q = \hat{q}_0$ (note that here f must also be constant in t and the uniqueness is a consequence of the uniform convexity of V). The point $\hat{q}_0$ is called *the point of equilibrium* of the system. Typical such situations are those in which f vanishes; then $\hat{q}_0 = q_0$ which is the minimum point of the potential energy. The energy balance in this general case is completely analogous to the one in (1.18) and (1.19) and its derivation is left to the reader.

Remark 1.5 If f is constant in time then we can modify the potential energy function into

$$\hat{V}(q) = V(q) - q^T f$$

thus obtaining a system with the vanishing generalised external force. If, in addition, $Q = 0$ then the system is conservative. This shows that there is some arbitrariness in the choice of the potential energy function.

Exercise 1.6 *Using the expressions (1.16), (1.15) for the kinetic and the potential energy, respectively, choose the dissipation function and the generalised external force such that, starting from (1.20) the system (1.1) is obtained.*

Solution. As the dissipation function take

$$Q = \frac{1}{2} \dot{x}^T C \dot{x};$$

then

$$L = \frac{1}{2} \dot{x}^T M \dot{x} - \frac{1}{2} x^T K x + f^T x,$$

$$L'_{\dot{x}} = M, \quad L'_x = -Kx + f, \quad Q'_{\dot{x}} = C$$

and (1.20) immediately yields (1.1).

Now consider small oscillations around q_0, that is, the vectors $q - q_0$ and $\dot{q}$ will be considered small for all times. This will be expected, if the distance from the equilibrium together with the velocity, taken at some initial time as well as the external force $f(t)$ at all times are small enough. This assumed, we will approximate the functions T, V, Q by their Taylor polynomials of second order around the point $(q_0, 0)$:

$$V(q) \approx V(q_0) + (q - q_0)^T V'(q_0) + \frac{1}{2}(q - q_0)^T V''(q_0)(q - q_0),$$

$$T(q, \dot{q}) \approx T(q_0, 0) + \dot{q}^T T'_{\dot{q}}(q_0, 0) + \frac{1}{2}\dot{q}^T T''_{\dot{q}}(q_0, 0)\dot{q}$$

and similarly for Q. Note that, due to the conditions on T the matrix $T''(q_0, 0)$ looks like

$$\begin{bmatrix} 0 & 0 \\ 0 & T''_{\dot{q}}(q_0, 0) \end{bmatrix}$$

(the same for $Q''(q_0, 0)$). Using the properties listed above we have

$$T(q_0, 0) = 0, \quad T'_{\dot{q}}(q_0, 0) = 0,$$

$$Q(q_0, 0) = 0, \quad Q'_{\dot{q}}(q_0, 0) = 0,$$

$$V'(q_0) = 0.$$

With this approximation the Lagrange equations (1.20) yield (1.1) with $x = q$ and

$$M = T''_{\dot{q}}(q_0, 0), \quad C = Q''_{\dot{q}}(q_0, 0), \quad K = V''(q_0). \tag{1.21}$$

This approximation is called *linearisation*[1], because non-linear equations are approximated by linear ones via Taylor expansion.

Under our conditions all three matrices in (1.21) are real and symmetric. In addition, K is positive definite while the other two matrices are positive semidefinite. If we additionally assume that T is also uniformly convex then M will be positive definite as well. We have purposely allowed M to be only positive semidefinite as we will analyse some cases of this kind later.

Example 1.7 Consider a hydraulic model which describes the oscillation of an incompressible homogeneous heavy fluid (e.g. water) in an open vessel. The vessel consists of coupled thin tubes which, for simplicity, are assumed as cylindrical (see Figure 1.7).

The parameters describing the vessel and the fluid are

- h_1, h, h_2: the fluid heights in the vertical tubes,
- s_1, s, s_2: the cross sections of the vertical tubes,
  - D_1, D_2: the lengths,
  - S_1, S_2: the cross sections,

[1] The term 'linearisation' will be used later in a very different sense.

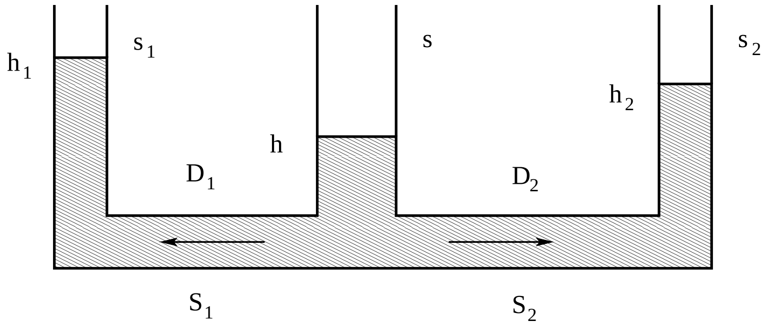

Fig. 1.2 Open vessel

- v_1, v_2: the velocities (in the sense of the arrows),

in the horizontal tubes.

- ρ: the mass density of the fluid

While the values h_1, h, h_2, v_1, v_2 are functions of time t all others are constant. Since the fluid is incompressible the volume conservation laws give

$$s_1\dot{h}_1 = S_1v_1 \tag{1.22}$$

$$S_1v_1 + s\dot{h} + S_2v_2 = 0 \tag{1.23}$$

$$s_2\dot{h}_2 = S_2v_2 \tag{1.24}$$

These equalities imply

$$s_1\dot{h}_1 + s\dot{h} + s_2\dot{h}_2 = 0 \tag{1.25}$$

which, integrated, gives

$$s_1h_1 + sh + s_2h_2 = \mathit{\Gamma} \tag{1.26}$$

where $\mathit{\Gamma}$ is a constant (in fact, $\mathit{\Gamma} + D_1S_1 + D_2S_2$ is the fixed total volume of the fluid). Thus, the movement of the system is described by $h_1 = h_1(t), h_2 = h_2(t)$.

To obtain the Lagrange function we must find the expressions for the kinetic and the potential energy as well as for a dissipation function — as functions of $h_1, h_2, \dot{h}_1, \dot{h}_2$.

In each vertical tube the potential energy equals the mass times half the height (this is the center of gravity) times g, the gravity acceleration. The total potential energy V is given by

$$\frac{V}{\rho g} = \frac{V(h_1,h_2)}{\rho g} = s_1\frac{h_1^2}{2} + s\frac{h^2}{2} + s_2\frac{h_2^2}{2} \tag{1.27}$$

$$= s_1\frac{h_1^2}{2} + \frac{(\mathit{\Gamma} - s_1h_1 - s_2h_2)^2}{2s} + s_2\frac{h_2^2}{2} \tag{1.28}$$

(the contributions from the horizontal tubes are ignored since they do not depend on $h_1, h_2, \dot{h}_1, \dot{h}_2$ and hence do not enter the Lagrange equations). It is readily seen that V takes its minimum at

$$h_1 = h = h_2 = \hat{h} := \frac{\mathit{\Gamma}}{s_1 + s + s_2}. \tag{1.29}$$

The kinetic energy in each tube equals half the mass times the square of the velocity. The total kinetic energy T is given by

$$T/\rho = T(h_1,h_2,\dot{h}_1,\dot{h}_2)/\rho$$

$$= h_1s_1\frac{\dot{h}_1^2}{2} + D_1S_1\frac{v_1^2}{2} + hs\frac{\dot{h}^2}{2} + D_2S_2\frac{v_2^2}{2} + h_2s_2\frac{\dot{h}_2^2}{2} = \tag{1.30}$$

$$\left(h_1s_1 + \frac{D_1s_1^2}{S_1}\right)\frac{\dot{h}_1^2}{2} + (\mathit{\Gamma} - s_1h_1 - s_2h_2)\frac{(s_1\dot{h}_1 + s_2\dot{h}_2)^2}{2s^2} + \left(h_2s_2 + \frac{D_2s_2^2}{S_2}\right)\frac{\dot{h}_2^2}{2} \tag{1.31}$$

where $\dot{h}$ was eliminated by means of (1.25). While T and V result almost canonically from first principles and the geometry of the system the dissipation function allows more freedom. We will assume that the frictional force in each tube is proportional to the velocity of the fluid and to the fluid-filled length of the tube. This leads to the dissipation function

$$Q = Q(h_1,h_2,\dot{h}_1,\dot{h}_2)$$

$$= h_1\theta_1\frac{\dot{h}_1^2}{2} + D_1\mathit{\Theta}_1\frac{v_1^2}{2} + h\theta\frac{\dot{h}^2}{2} + D_2\mathit{\Theta}_2\frac{v_2^2}{2} + h_2\theta_2\frac{\dot{h}_2^2}{2} \tag{1.32}$$

$$= \left(h_1\theta_1 + \frac{D_1\mathit{\Theta}_1s_1^2}{S_1^2}\right)\frac{\dot{h}_1^2}{2} + (\mathit{\Gamma} - s_1h_1 - s_2h_2)\frac{(s_1\dot{h}_1 + s_2\dot{h}_2)^2}{2s^3}\mathit{\Theta} + \tag{1.33}$$

$$\left(h_2\theta_2 + \frac{D_2\mathit{\Theta}_2s_2^2}{S_2^2}\right)\frac{\dot{h}_2^2}{2} \tag{1.34}$$

where $\theta_1, \mathit{\Theta}_1, \theta, \mathit{\Theta}_2, \theta_2$ are positive constants characterising the friction per unit length in the respective tubes. By inserting the obtained T, V, Q into

(1.20) and by taking q as $\begin{bmatrix} h_1 \; h_2 \end{bmatrix}^T$ we obtain two non-linear differential equations of second order for the unknown functions h_1, h_2.

Now comes the linearisation at the equilibrium position which is given by (1.29). The Taylor expansion of second order for T, V, Q around the point $h_1 = h_2 = \hat{h},\ \dot{h}_1 = \dot{h}_2 = 0$ takes into account that all three gradients at that point vanish thus giving

$$V \approx V(\hat{h}, \hat{h}) + \rho g \left(s_1 \frac{\chi_1^2}{2} + \frac{(s_1\chi_1 + s_2\chi_2)^2}{2s} + s_2 \frac{\chi_2^2}{2} \right)$$

$$= V(\hat{h}, \hat{h}) + \frac{1}{2}\chi^T K \chi \tag{1.35}$$

with $\chi = \begin{bmatrix} h_1 - \hat{h} \; h_2 - \hat{h} \end{bmatrix}^T$ and

$$T \approx \rho \left(\left(\hat{h} s_1 + \frac{D_1 s_1^2}{S_1} \right) \frac{\dot{\chi}_1^2}{2} + \hat{h} \frac{(s_1\dot{\chi}_1 + s_2\dot{\chi}_2)^2}{2s} + \left(\hat{h} s_2 + \frac{D_2 s_2^2}{S_2} \right) \frac{\dot{\chi}_2^2}{2} \right) \tag{1.36}$$

$$= \frac{1}{2}\dot{\chi}^T M \dot{\chi} \tag{1.37}$$

$$Q \approx \left(\hat{h}\theta_1 + \frac{D_1 \varTheta_1 s_1^2}{S_1^2} \right) \frac{\dot{\chi}_1^2}{2} + \hat{h} \frac{(s_1\dot{\chi}_1 + sa_2\dot{\chi}_2)^2}{2s^2} \varTheta + \left(\hat{h}\theta_2 + \frac{D_2 \varTheta_2 s_2^2}{S_2^2} \right) \frac{\dot{\chi}_2^2}{2} \tag{1.38}$$

$$= \frac{1}{2}\dot{\chi}^T C \dot{\chi}. \tag{1.39}$$

The symmetric matrices M, C, K are immediately recoverable from (1.35) – (1.38) and are positive definite by their very definition. By inserting these T, V, Q into (1.20) the equations (1.1) result with $f = 0$.

Any linearised model can be assessed in two ways:

- by estimating the error between the true model and the linearised one and
- by investigating the *well-posedness* of the linear model itself: do small forces $f(t)$ and small initial data produce small solutions?

While the first task is completely out of the scope of the present text, the second one will be given our attention in the course of these Notes.

1.4 Oscillating electrical circuits

The equation (1.1) governs some other physical systems, most relevant among them is the electrical circuit system which we will present in the following example.

Example 1.8 n simple wire loops (circuits) are placed in a chain. The j-th circuit is characterised by the current I_j, the impressed electromagnetic force $E_j(t)$, the capacity $C_j > 0$, the resistance $R_j \geq 0$, and the left and the right inductance $L_j > 0$, $M_j > 0$, respectively. In addition, the neighbouring j-th and $j+1$-th inductances are accompanied by the mutual inductance $N_j \geq 0$. The circuits are shown in Fig. 1.3 with $n = 3$. The inductances satisfy the inequality

$$N_j < \sqrt{M_j L_{j+1}}.$$

The equation governing the time behaviour of this system is derived from the fundamental electromagnetic laws and it reads in vector form

$$\mathcal{L}\ddot{I} + \mathcal{R}\dot{I} + \mathcal{K}I = \dot{E}(t).$$

Here

$$\mathcal{L} = \begin{bmatrix} L_1 + M_1 & N_1 & & & \\ N_1 & L_2 + M_2 & N_2 & \ddots & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & N_{n-1} \\ & & & N_{n-1} & L_n + M_n \end{bmatrix}$$

is *the inductance matrix*,

$$\mathcal{R} = \operatorname{diag}(R_1, \ldots, R_n)$$

is *the resistance matrix* and

$$\mathcal{K} = \operatorname{diag}(1/C_1, \ldots, 1/C_n)$$

is *the inverse capacitance matrix.* The latter two matrices are obviously positive definite whereas the positive definiteness of the first one is readily deduced from its structure and the inequality (1.40).

Exercise 1.9 *Show the positive definiteness of the inductance matrix $\mathcal{L}$. Hint: use the fact that each of the matrices*

$$\begin{bmatrix} M_j & N_j \\ N_j & L_{j+1} \end{bmatrix}$$

is positive definite.

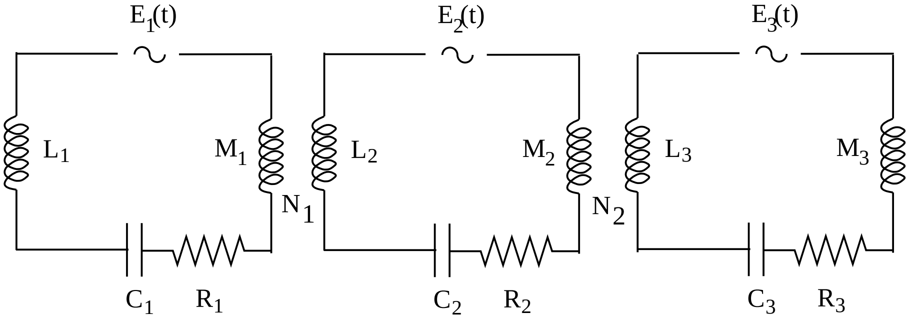

Fig. 1.3 Circuits

Chapter 2

Simultaneous diagonalisation (Modal damping)

In this chapter we describe undamped and modally damped systems. They are wholly explained by the knowledge of the mass and the stiffness matrix. This is the broadly known case and we shall outline it here, not only because it is an important special case, but because it is often used as a starting position in the analysis of general damped systems.

2.1 Undamped systems

The system (1.1) is called *undamped,*, if the damping vanishes: $C = 0$.

The solution of an undamped system is best described by *the generalised eigenvalue decomposition* of the matrix pair K, M:

$$\Phi^T K \Phi = \operatorname{diag}(\mu_1, \dots, \mu_n), \quad \Phi^T M \Phi = I. \tag{2.1}$$

We say that the matrix Φ *reduces the pair* K, M *of symmetric matrices to diagonal form by congruence.* This reduction is always possible, if the matrix M is positive definite. Instead of speaking of the matrix pair one often speaks of *the matrix pencil,* (that is, matrix function) $K - \lambda M$. If $M = I$ then (2.1) reduces to *the (standard) eigenvalue decomposition* valid for any symmetric matrix K, in this case the matrix Φ is orthogonal.

An equivalent way of writing (2.1) is

$$K\Phi = M\Phi \operatorname{diag}(\mu_1, \dots, \mu_n), \quad \Phi^T M \Phi = I. \tag{2.2}$$

or also

$$K\phi_j = \mu_j M \phi_j, \quad \phi_j^T M \phi_k = \delta_{kj}.$$

Thus, the columns ϕ_j of Φ form an M-orthonormal basis of eigenvectors of the generalised eigenvalue problem

$$K\phi = \mu M\phi \tag{2.3}$$

whereas μ_k are the zeros of the characteristic polynomial

$$\det(K - \mu M)$$

of the pair K, M. Hence

$$\mu = \frac{\phi^T K\phi}{\phi^T M\phi}, \text{ in particular, } \mu_k = \frac{\phi_k^T K\phi_k}{\phi_k^T M\phi_k}$$

shows that all μ_k are positive, if *both* K and M are positive definite as in our case. So we may rewrite (2.1) as

$$\Phi^T K\Phi = \Omega^2, \quad \Phi^T M\Phi = I \tag{2.4}$$

with

$$\Omega = \operatorname{diag}(\omega_1, \dots, \omega_n), \quad \omega_k = \sqrt{\mu_k} \tag{2.5}$$

The quantities ω_k will be called *the eigenfrequencies* of the system (1.1) with $C = 0$. The generalised eigenvalue decomposition can be obtained by any common matrix computation package (e.g. by calling eig(K,M) in MATLAB).

The solution of the homogeneous equation

$$M\ddot{x} + Kx = 0 \tag{2.6}$$

is given by the formula

$$x(t) = \Phi \begin{bmatrix} a_1 \cos\omega_1 t + b_1 \sin\omega_1 t \\ \vdots \\ a_n \cos\omega_n t + b_n \sin\omega_n t \end{bmatrix}, \quad a = \Phi^{-1}x_0, \quad \omega_k b_k = (\Phi^{-1}\dot{x}_0)_k \tag{2.7}$$

which is readily verified. The values ω_k are of interest even if the damping C does not vanish and in this context they are called *the undamped frequencies* of the system (1.1).

In physical language the formula (2.7) is oft described by the phrase 'any oscillation is a superposition of harmonic oscillations or eigenmodes' which are

$$\phi_k(a_k \cos\omega_k t + b_k \sin\omega_k t), \quad k = 1, \dots, n.$$

Exercise 2.1 *Show that the eigenmodes are those solutions $x(t)$ of the equation (2.6) in which 'all particles oscillate in the same phase' that is,*

$$x(t) = x_0 T(t),$$

where x_0 is a fixed non-zero vector and $T(t)$ is a scalar-valued function of t (the above formula is also well known under the name 'Fourier ansatz').

The eigenvalues μ_k, taken in the non-decreasing ordering, are given by the known *minimax formula*

$$\mu_k = \max_{S_{n-k+1}} \min_{\substack{x\in S_{n-k+1}\\ x\neq 0}} \frac{x^T K x}{x^T M x} = \min_{S_k} \max_{\substack{x\in S_k\\ x\neq 0}} \frac{x^T K x}{x^T M x} \tag{2.8}$$

where S_j denotes any subspace of dimension j. We will here skip proving these — fairly known — formulae, valid for any pair K, M of symmetric matrices with M positive definite. We will, however, provide a proof later within a more general situation (see Chapter 10 below).

The eigenfrequencies have an important monotonicity property. We introduce the relation called *relative stiffness* in the set of all pairs of positive definite symmetric matrices K, M as follows. We say that the pair $\hat{K}, \hat{M}$ is *relatively stiffer* than K, M, if the matrices $\hat{K} - K$ and $M - \hat{M}$ are positive semidefinite (that is, if stiffness is growing and the mass is falling).

Theorem 2.2 *Increasing relative stiffness increases the eigenfrequencies. More precisely, if $\hat{K} - K$ and $M - \hat{M}$ are positive semidefinite then the corresponding non-decreasingly ordered eigenfrequencies satisfy*

$$\omega_k \leq \hat{\omega}_k.$$

Proof. Just note that

$$\frac{x^T K x}{x^T M x} \leq \frac{x^T \hat{K} x}{x^T \hat{M} x}$$

for all non-vanishing x. Then take first minimum and then maximum and the statement follows from (2.8). Q.E.D.

If in Example 1.1 the matrix $\hat{K}$ is generated by the spring stiffnesses $\hat{k}_j$ then by (1.10) for $\delta K = \hat{K} - K$ we have

$$x^T \delta K x = \delta k_1 x_1^2 + \sum_{j=2}^{n} \delta k_j (x_j - x_{j-1})^2 + \delta k_{n+1} x_n^2 \tag{2.9}$$

where

$$\delta k_j = \hat{k}_j - k_j.$$

So, $\hat{k}_j \geq k_j$ implies the positive semidefiniteness of δK, that is the relative stiffness is growing. The same happens with the masses: take $\delta M = \hat{M} - M$, then

$$x^T \delta M x = \sum_{j=1}^{n} \delta m_j x_j^2, \quad \delta m_j = \hat{m}_j - m_j$$

and $\hat{m}_j \leq m_j$ implies the negative semidefiniteness of δM — the relative stiffness is again growing. Thus, our definition of the relative stiffness has deep physical roots.

The next question is: how do small changes in the system parameters k_j, m_j affect the eigenvalues? We make the term 'small changes' precise as follows

$$|\delta k_j| \le \epsilon k_j, \quad |\delta m_j| \le \eta m_j \tag{2.10}$$

with $0 \le \epsilon, \eta < 1$. This kind of relative error is typical both in physical measurements and in numerical computations, in fact, in floating point arithmetic $\epsilon, \eta \approx 10^{-d}$ where d is the number of significant digits in a decimal number.

The corresponding errors in the eigenvalues will be an immediate consequence of (2.10) and Theorem 2.2. Indeed, from (2.9), (2.10) it follows

$$|x^T \delta K x| \le \epsilon x^T K x, \quad |x^T \delta M x| \le \eta x^T M x. \tag{2.11}$$

Then

$$(1-\epsilon) x^T K x \le x^T \hat{K} x \le (1+\epsilon) x^T K x$$

and

$$(1-\eta) x^T M x \le x^T \hat{M} x \le (1+\eta) x^T M x$$

such that the pairs

$$(1-\epsilon)K,\ (1+\eta)M; \quad \hat{K},\ \hat{M}; \quad (1+\epsilon)K,\ (1-\eta)M$$

are ordered in growing relative stiffness. Therefore by Theorem 2.2 the corresponding eigenvalues

$$\frac{1-\epsilon}{1+\eta}\mu_k, \quad \hat{\mu}_k, \quad \frac{1+\epsilon}{1-\eta}\mu_k$$

satisfy

$$\frac{1-\epsilon}{1+\eta}\mu_k \le \hat{\mu}_k \le \frac{1+\epsilon}{1-\eta}\mu_k \tag{2.12}$$

(and similarly for the respective $\omega_k, \hat{\omega}_k$). In particular, for $\delta\mu_k = \hat{\mu}_k - \mu_k$ the relative error estimates

$$|\delta\mu_k| \le \frac{\epsilon+\eta}{1-\eta}\mu_k \tag{2.13}$$

are valid. Note that both (2.12) and (2.13) are quite general. They depend only on the bounds (2.11), the only requirement is that both matrices K, M be symmetric and positive definite.

In the case $\hat{M} = M = I$ the more commonly known error estimate holds

$$\mu_k + \min \sigma(\hat{K} - K) \le \hat{\mu}_k \le \mu_k + \max \sigma(\hat{K} - K) \tag{2.14}$$

and in particular

$$|\delta\mu_k| \le \|\hat{K} - K\|. \tag{2.15}$$

The proof again goes by immediate application of Theorem 2.2 and is left to the reader.

2.2 Frequencies as singular values

There is another way to compute the eigenfrequencies ω_j. We first make the decomposition

$$K = L_1 L_1^T, \quad M = L_2 L_2^T, \tag{2.16}$$

$$y_1 = L_1^T x, \quad y_2 = L_2^T \dot{x},$$

(here L_1, L_2 may, but need not be Cholesky factors). Then we make the singular value decomposition

$$L_2^{-1} L_1 = U \boldsymbol{\Sigma} V^T \tag{2.17}$$

where U, V are real orthogonal matrices and $\boldsymbol{\Sigma}$ is diagonal with positive diagonal elements. Hence

$$L_2^{-1} L_1 L_1^T L_2^{-T} = U \boldsymbol{\Sigma}^2 U^T$$

or

$$K\Phi = M\Phi\boldsymbol{\Sigma}^2, \quad \Phi = L_2^{-T} U$$

Now we can identify this Φ with the one from (2.4) and $\boldsymbol{\Sigma}$ with Ω. Thus *the eigenfrequencies of the undamped system are the singular values of the matrix* $L_2^{-1} L_1$.[1] The computation of Ω by (2.17) may have advantages over the one by (2.2), in particular, if ω_j greatly differ from each other. Indeed, by setting in Example 1.1 $n = 3$, $k_4 = 0$, $m_i = 1$ the matrix L_1 is directly obtained as

$$L_1 = \begin{bmatrix} \kappa_1 & -\kappa_2 & 0 \\ 0 & \kappa_2 & -\kappa_3 \\ 0 & 0 & \kappa_3 \end{bmatrix}, \quad \kappa_i = \sqrt{k_i}. \tag{2.18}$$

If we take $k_1 = k_2 = 1$, $k_3 \gg 1$ (that is, the third spring is almost rigid) then the way through (2.2) may spoil the lower frequency. For instance, with the value $k_3 = 9.999999 \cdot 10^{15}$ the double-precision computation with Matlab gives the frequencies

```
    sqrt(eig(K,M))              svd(L_2\L_1)

7.962252170181258e-01      4.682131924621356e-01
1.538189001320851e+00      1.510223959022110e+00
1.414213491662415e+08      1.414213491662415e+08
```

[1] Equivalently we may speak of ω_j as *the generalised singular values* of the pair L_1, L_2.

The singular value decomposition gives largely correct low eigenfrequencies. This phenomenon is independent of the eigenvalue or singular value algorithm used and it has to do with the fact that standard eigensolution algorithms compute the lowest eigenvalue of (2.2) with the relative error $\approx \epsilon\kappa(KM^{-1})$, that is, the machine precision ϵ is amplified by the condition number $\kappa(KM^{-1}) \approx 10^{16}$ whereas the same error with (2.17) is $\approx \epsilon\kappa(L_2^{-1}L_1) = \epsilon\sqrt{\kappa(KM^{-1})}$ (cf. e.g [19]). In the second case the amplification is the square root of the first one!

2.3 Modally damped systems

Here we study those damped systems which can be completely explained by their undamped part. In order to do this it is convenient to make a *coordinate transformation*; we set

$$x = \Phi x', \tag{2.19}$$

where Φ is any real non-singular matrix. Thus (1.1) goes over into

$$M'\ddot{x}' + C'\dot{x}' + K'x' = g(t), \tag{2.20}$$

with

$$M' = \Phi^T M\Phi, \quad C' = \Phi^T C\Phi, \quad K' = \Phi^T K\Phi, \quad g = \Phi^T f. \tag{2.21}$$

Choose now the matrix Φ as in the previous section, that is,

$$\Phi^T M\Phi = I, \quad \Phi^T K\Phi = \Omega = \operatorname{diag}(\omega_1^2, \dots, \omega_n^2).$$

(The right hand side $f(t)$ in (2.20) can always be taken into account by the Duhamel's term as in (3.1) so we will mostly restrict ourselves to consider $f = 0$ which corresponds to a 'freely oscillating' system.)

Now, if

$$D = (d_{jk}) = \Phi^T C\Phi \tag{2.22}$$

is diagonal as well then (1.1) is equivalent to

$$\ddot{\xi}_k + d_{kk}\dot{\xi}_k + \omega_k^2\xi_k = 0, \quad x = \Phi\,\xi$$

with the known solution

$$\xi_k = a_k u^+(t, \omega_k, d_{kk}) + b_k u^-(t, \omega_k, d_{kk}), \tag{2.23}$$

$$u^+(t, \omega, d) = e^{\lambda^+(\omega,d)\,t},$$

$$u^-(t, \omega, d) = \begin{cases} e^{\lambda^-(\omega,d)\,t}, & \delta(\omega, d) \neq 0 \\ t\,e^{\lambda^+(\omega,d)\,t}, & \delta(\omega, d) = 0 \end{cases}$$

where

$$\delta(\omega, d) = d^2 - 4\omega^2,$$
$$\lambda^{\pm}(\omega, d) = \frac{-d \pm \sqrt{\delta(\omega, d)}}{2}.$$

The constants a_k, b_k in (2.23) are obtained from the initial data $x_0, \dot{x}_0$ similarly as in (2.7).

However, the simultaneous diagonalisability of the three matrices M, C, K is rather an exception, as shown by the following theorem.

Theorem 2.3 *Let M, C, K be as in (1.1). A non-singular Φ such that the matrices $\Phi^T M\Phi$, $\Phi^T C\Phi$, $\Phi^T K\Phi$ are diagonal exists, if and only if*

$$CK^{-1}M = MK^{-1}C. \tag{2.24}$$

Proof. The "only if part" is trivial. Conversely, using (2.4) the identity (2.24) yields

$$C\Phi\Omega^{-2}\Phi^{-1} = \Phi^{-T}\Omega^{-2}\Phi^T C,$$

hence

$$\Phi^T C\Phi\Omega^{-2} = \Omega^{-2}\Phi^T C\Phi$$

and then

$$\Omega^2\Phi^T C\Phi = \Phi^T C\Phi\Omega^2$$

i.e. the two real symmetric matrices Ω^2 and $D = \Phi^T C\Phi$ commute and Ω^2 is diagonal, so there exists a real orthogonal matrix U such that $U^T\Omega^2 U = \Omega^2$, and $U^T DU = \mathrm{diag}(d_{11}, \ldots, d_{nn})$. Indeed, since the diagonal elements of Ω are non-decreasingly ordered we may write

$$\Omega = \mathrm{diag}(\Omega_1, \ldots, \Omega_p),$$

where $\Omega_1, \ldots, \Omega_p$ are scalar matrices corresponding to distinct spectral points of Ω. Now, $D\Omega^2 = \Omega^2 D$ implies $D\Omega = \Omega D$ and therefore

$$D = \mathrm{diag}(D_1, \ldots, D_p),$$

with the same block partition. The matrices $D_1, \ldots, D_p$ are real symmetric, so there are orthogonal matrices $U_1, \ldots, U_p$ such that all $U_j^T D_j U_j$ are diagonal. By setting $\Phi_1 = \Phi\,\mathrm{diag}(D_1, \ldots, D_p)$ all three matrices $\Phi_1^T M\Phi_1$, $\Phi_1^T C\Phi_1$, $\Phi_1^T K\Phi_1$ are diagonal. Q.E.D.

Exercise 2.4 *Show that Theorem 2.3 remains valid, if M is allowed to be only positive semidefinite.*

Exercise 2.5 *Prove that (2.24) holds, if*

$$\alpha M + \beta C + \gamma K = 0$$

where not all of α, β, γ *vanish (proportional damping). When is this the case with* C *from (1.4), (1.5), (1.6)?*

Exercise 2.6 *Prove that (2.24) is equivalent to* $CM^{-1}K = KM^{-1}C$ *and also to* $KC^{-1}M = MC^{-1}K$*, provided that these inverses exist.*

Exercise 2.7 *Try to find a necessary and sufficient condition that* C *from (1.4) – (1.6) satisfies (2.24).*

If C satisfies (2.24) then we say that the system (1.1) is *modally damped*.

Chapter 3
Phase space

In general, simultaneous diagonalisation of all three matrices M, C, K is not possible and the usual transformation to a system of first order is performed. This transformation opens the possibility of using powerful means of the spectral theory in order to understand the time behaviour of damped systems.

3.1 General solution

If the matrix M is non-singular the existence and uniqueness of the solution of the initial value problem $M\ddot{x} + C\dot{x} + Kx = f(t)$ with the standard initial conditions

$$x(0) = x_0, \quad \dot{x}(0) = \dot{x}_0$$

is insured by the standard theory for linear systems of differential equations which we now review.

We rewrite (1.1) as

$$\ddot{x} = -M^{-1}Kx - M^{-1}C\dot{x} + M^{-1}f(t)$$

and by setting $x_1 = x$, $x_2 = \dot{x}$ as

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = B\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + g(t),$$

$$B = \begin{bmatrix} 0 & 1 \\ -M^{-1}K & -M^{-1}C \end{bmatrix}, \quad g(t) = \begin{bmatrix} 0 \\ M^{-1}f(t) \end{bmatrix}.$$

This equation is solved by the so-called *Duhamel formula*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = e^{Bt}\begin{bmatrix} x_0 \\ \dot{x}_0 \end{bmatrix} + \int_0^t e^{B(t-\tau)}g(\tau)\,d\tau \tag{3.1}$$

with

$$e^{Bt} = \sum_{j=0}^{\infty} \frac{B^n t^n}{n!}, \quad \frac{d}{dt}\, e^{Bt} = \sum_{j=1}^{\infty} \frac{B^n t^{n-1}}{(n-1)!} = Be^{Bt}.$$

For the uniqueness: if x, y solve (1.1) then $u = y - x$, $u_1 = u$, $u_2 = \dot{u}$ satisfy

$$\frac{d}{dt} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = B \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad u_1(0) = u_2(0) = 0.$$

By setting

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = e^{-Bt} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

we obtain

$$\frac{d}{dt} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0, \quad v_1(0) = v_2(0) = 0$$

so, $v_1(t) \equiv v_2(t) \equiv 0$ and by $\left(e^{-Bt}\right)^{-1} = e^{Bt}$ also $u_1(t) \equiv u_2(t) \equiv 0$.

3.2 Energy phase space

The shortcoming of the transformation in Chapter 3.1 is that in the system matrix the symmetry properties of the matrices M, C, K are lost. A more careful approach will try to keep as much structure as possible and this is what we will do in this chapter. An important feature of this approach is that the Euclidian norm on the underlying space (called phase space) is closely related to the total energy of the system.

We start with any decomposition (2.16). By substituting $y_1 = L_1^T y,\ y_2 = L_2^T \dot{y}$ (1.1) becomes

$$\dot{y} = Ay + g(t), \quad g(t) = \begin{bmatrix} 0 \\ L_2^{-1} f(t) \end{bmatrix}, \tag{3.2}$$

$$A = \begin{bmatrix} 0 & L_1^T L_2^{-T} \\ -L_2^{-1} L_1 & -L_2^{-1} C L_2^{-T} \end{bmatrix} \tag{3.3}$$

which is solved by

$$y = e^{At} \begin{bmatrix} y_{10} \\ y_{20} \end{bmatrix} + \int_0^t e^{A(t-\tau)} g(\tau)\, d\tau, \tag{3.4}$$

$$y_{10} = L_1^T x_0, \quad y_{20} = L_2^T \dot{x}_0.$$

The differential equation (3.2) will be referred to as the *evolution equation*. The formula (3.4) is, of course, equivalent to (3.1). The advantage of (3.4) is due to the richer structure of *the phase-space matrix* A in (3.3). It has two important properties:

$$y^T Ay = -(L_2^{-1}y_2)^T CL_2^{-1}y_2 \le 0 \quad \text{for all} \quad y, \tag{3.5}$$

$$A^T = JAJ \tag{3.6}$$

where

$$J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}. \tag{3.7}$$

The property (3.6) is called *J-symmetry*, a general real J-symmetric matrix A has the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ -A_{12}^T & A_{22} \end{bmatrix}, \quad A_{11}^T = A_{11}, \quad A_{22}^T = A_{22},$$

so, A is "symmetric up to signs".

Moreover, for $y = y(t)$ satisfying (1.1)

$$\|y\|^2 = \|y_1\|^2 + \|y_2\|^2 = x^T Kx + \dot{x}^T M\dot{x}. \tag{3.8}$$

Since $x^T Kx/2$ is the potential energy of the system and $\dot{x}^T M\dot{x}/2$ its kinetic energy we see that $\|y\|^2$ is twice the total energy of the system (1.1). That is why we call the representation (3.2), (3.3) of the system (1.1) *the energy phase space representation.* For $g(t) \equiv 0$ (freely oscillating damped system) we have by (3.5)

$$\frac{d}{dt}\|y\|^2 = \frac{d}{dt}y^T y = \dot{y}^T y + y^T \dot{y} = y^T(A + A^T)y = 2y^T Ay \le 0 \tag{3.9}$$

that is, the energy of such a system is a non-increasing function of time. If $C = 0$ then the energy is constant in time.

The property (3.5) of the matrix A is called *dissipativity.*

Exercise 3.1 *Show that the dissipativity of A is equivalent to the* contractivity *of the exponential:*

$$\|e^{At}\| \le 1, \quad t \ge 0.$$

Remark 3.2 There are infinitely many decompositions (2.16). Most common are

- The Cholesky decomposition. This is numerically the least expensive choice.
- $L_1 = \Phi^{-T}\Omega$, $L_2 = \Phi^{-T}$, where Φ, Ω are given by (2.4). Then $L_2^{-1}L_1 = \Omega$ and
$$A = \begin{bmatrix} 0 & \Omega \\ -\Omega & -D \end{bmatrix}, \quad D = \Phi^T C\Phi. \tag{3.10}$$

This is the *modal representation* of the phase-space matrix.

- The positive definite square roots: $L_1 = \sqrt{K}$, $L_2 = \sqrt{M}$.

but some others will also prove useful. Whenever no confusion is expected the matrix D in (3.10) will also be called the damping matrix (in modal coordinates).

In practice one is often given a damped system just in terms of the matrices D and Ω in the modal representation. A common damping matrix is of the form

$$D = \rho\Omega + WW^T$$

where ρ is a small parameter (a few percent) describing the inner damping of the material whereas the matrix W is injective of order $n \times m$ with $m \ll n$. The term WW^T describes a few damping devices (dashpots) built in order to prevent large displacements (see [17]).

Any transition from (1.1) to a first order system with the substitution $y_1 = W_1 x$, $y_2 = W_2 \dot{x}$ where W_1, W_2 are non-singular matrices is called a *linearisation* of the system (1.1). As we have already said we use the term 'linearisation' in two quite different senses, but since both of them are already traditional and very different in their meaning we hope not to cause confusion by using them thus.

Exercise 3.3 *Prove that any two linearisations lead to phase-space matrices which are similar to each other. Investigate under what linearisations the corresponding phase-space matrix will be (i) dissipative, (ii) J-symmetric.*

The transformation of coordinates (2.19) – (2.21) can be made on any damped system (1.1) and with any non-singular transformation matrix Φ thus leading to (2.20). The energy expressions stay invariant:

$$\dot{x}^T M \dot{x} = \dot{x}'^T M' \dot{x}', \quad x^T K x = x'^T K' x'.$$

We now derive the relation between the two phase-space matrices A and A', built by (3.3) from M, C, K and M', C', K', respectively. So we take L_1, L_2 from (2.16) and L_1', L_2' from

$$K' = \Phi^T L_1 L_1^T \Phi = L_1' L_1'^T, \quad M' = \Phi^T L_2 L_2^T \Phi = L_2' L_2'^T.$$

Hence the matrices $U_1 = L_1^T \Phi L_1'^{-T}$ and $U_2 = L_2^T \Phi L_2'^{-T}$ are unitary. Thus,

$$L_2'^{-1} L_1' = U_2^{-1} L_2^{-1} L_1 U_1$$

and

$$L_2'^{-1} C' L_2'^{-T} = U_2^{-1} L_2^{-1} C L_2^{-T} U_2.$$

Altogether

$$A' = \begin{bmatrix} 0 & U_1^{-1} L_1^T L_2^{-T} U_2 \\ -U_2^{-1} L_2^{-1} L_1 U_1 & -U_2^{-1} L_2^{-1} C L_2^{-T} U_2 \end{bmatrix} =$$

$$\begin{bmatrix} U_1^{-1} & 0 \\ 0 & U_2^{-1} \end{bmatrix} \begin{bmatrix} 0 & L_1^T L_2^{-T} \\ -L_2^{-1} L_1 & -L_2^{-1} C L_2^{-T} \end{bmatrix} \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix} =$$

$$U^{-1} A U$$

with

$$U = \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix} \tag{3.11}$$

which is unitary.

Although no truly damped system has a normal phase-space matrix something else is true: among a large variety of possible linearisation matrices the one in (3.3) is *closest to normality.* We introduce *the departure from normality* of a general matrix A of order n as

$$\eta(A) = \|A\|_E^2 - \sum_{i=1}^{n} |\lambda_i|^2$$

where

$$\|A\|_E = \sqrt{\sum_{ij} |a_{ij}|^2}$$

is the Euclidean norm of A and λ_i its eigenvalues. By a unitary transformation of A to triangular form it is seen that $\eta(A)$ is always non-negative. Moreover, $\eta(A)$ vanishes if and only if A is normal. The departure from normality is a measure of sensitivity in computing the eigenvalues and the eigenvectors of a matrix.

Theorem 3.4 *Among all linearisations*

$$y_1 = W_1 x, \quad y_2 = W_2 \dot{x}, \quad W_1, W_2 \quad \textit{real, non-singular} \tag{3.12}$$

of the system (1.1) the one from (3.3) has minimal departure from normality and there are no other linearisations (3.12) sharing this property.

Proof. The linearisation (3.12) leads to the first order system

$$\dot{y} = F y,$$

$$F = \begin{bmatrix} 0 & W_1 W_2^{-1} \\ -W_2 M^{-1} K W_1^{-1} & -W_2 M^{-1} C W_2^{-1} \end{bmatrix} = \mathcal{L} A \mathcal{L}^{-1}$$

with

$$\mathcal{L} = \begin{bmatrix} W_1 L_1^{-T} & 0 \\ 0 & W_2 L_2^T \end{bmatrix}$$

and A from (3.3). The matrices A and F have the same eigenvalues, hence

$$\eta(F) - \eta(A) = \|F\|_E^2 - \|A\|_E^2 = \eta(F_0) + \eta(W_2 L_2^{-T} L_2^{-1} C L_2^{-T} L_2^T W_2^{-1}) \tag{3.13}$$

where A_0, F_0 is the off-block diagonal part of A, F, respectively. The validity of (3.13) follows from the fact that A_0, F_0 are similar and A_0 is normal; the same holds for $L_2^{-1}CL_2^{-T}$ and $W_2L_2^{-T}L_2^{-1}CL_2^{-T}L_2^TW_2^{-1}$. Thus, $\eta(F) \geq \eta(A)$. If the two are equal then we must have both

$$\eta(F_0) = 0 \quad \text{and} \quad \eta(W_2L_2^{-T}L_2^{-1}CL_2^{-T}L_2^TW_2^{-1}) = 0.$$

Thus, both F_0 and $W_2L_2^{-T}L_2^{-1}CL_2^{-T}L_2^{-1}W_2^{-1}$ are normal. Since F_0 has purely imaginary spectrum, it must be skew-symmetric, whereas $W_2L_2^{-T}L_2^{-1}CL_2^{-T}L_2^{-1}W_2^{-1}$ (being similar to $L_2^{-1}CL_2^{-T}$) has real non-negative spectrum and must therefore be symmetric positive semidefinite. This is just the type of matrix A in (3.3). Q.E.D.

Taking $C = 0$ in (3.3) the matrix A becomes skew-symmetric and its eigenvalues are purely imaginary. They are best computed by taking the singular value decomposition (2.17). From the unitary similarity

$$\begin{bmatrix} V^T & 0 \\ 0 & U^T \end{bmatrix} \begin{bmatrix} 0 & L_1^T L_2^{-T} \\ -L_2^{-1}L_1 & 0 \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & U \end{bmatrix} = \begin{bmatrix} 0 & \Omega \\ -\Omega & 0 \end{bmatrix}$$

the eigenvalues of A are seen to be $\pm i\omega_1, \ldots, \pm i\omega_n$.

Exercise 3.5 *Compute the matrix A in the damping-free case and show that e^{At} is unitary.*

Exercise 3.6 *Find the phase-space matrix A for a modally damped system.*

Exercise 3.7 *Show that different choices for $L_{1,2}$ given in Remark 3.2 lead to unitarily equivalent A's.*

Exercise 3.8 *Show that the phase-space matrix A from (3.3) is normal, if and only if $C = 0$.*

Chapter 4
The singular mass case

If the mass matrix M is singular no standard transformation to a first order system is possible. In fact, in this case we cannot prescribe some initial velocities and the phase-space will have dimension less than $2n$. This will be even more so, if the damping matrix, too, is singular; then we could not prescribe even some initial positions. In order to treat such systems we must first separate away these 'inactive' degrees of freedom and then arrive at phase-space matrices which have smaller dimension but their structure will be essentially the same as in the regular case studied before. Now, out of M, C, K only K is supposed to be positive definite while M, C are positive semidefinite.

To perform the mentioned separation it is convenient to simultaneously diagonalise the matrices M and C which now are allowed to be only positive semidefinite.

Lemma 4.1 *If M, C are any real, symmetric, positive semidefinite matrices then there exists a real non-singular matrix Φ such that*

$$\Phi^T M \Phi \quad \textit{and} \quad \Phi^T C \Phi$$

are diagonal.

Proof. Suppose first that $\mathcal{N}(M) \cap \mathcal{N}(C) = \{0\}$. Then $M + C$ is positive definite and there is a Φ such that

$$\Phi^T (M + C) \Phi = I, \quad \Phi^T M \Phi = \boldsymbol{\mu},$$

$\boldsymbol{\mu}$ diagonal. Then

$$\Phi^T C \Phi = I - \boldsymbol{\mu}$$

is diagonal as well. Otherwise, let $u_1, \ldots, u_k$ be an orthonormal basis of $\mathcal{N}(M) \cap \mathcal{N}(C)$. Then there is an orthogonal matrix

$$U = \begin{bmatrix} \hat{U} \; u_1 \; \cdots \; u_k \end{bmatrix}$$

such that

$$U^T M U = \begin{bmatrix} \hat{M} & 0 \\ 0 & 0 \end{bmatrix}, \quad U^T C U = \begin{bmatrix} \hat{C} & 0 \\ 0 & 0 \end{bmatrix},$$

where $\mathcal{N}(\hat{M}) \cap \mathcal{N}(\hat{C}) = \{0\}$ and there is a non-singular $\hat{\Phi}$ such that $\hat{\Phi}^T \hat{M} \hat{\Phi}$ and $\hat{\Phi}^T \hat{C} \hat{\Phi}$ diagonal. Now set

$$\Phi = U \begin{bmatrix} \hat{\Phi} & 0 \\ 0 & 0 \end{bmatrix}.$$

Q.E.D.

We now start separating the 'inactive' variables. Using the previous lemma there is a non-singular Φ such that

$$M' = \Phi^T M \Phi = \begin{bmatrix} M_1' & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$C' = \Phi^T C \Phi = \begin{bmatrix} C_1' & 0 & 0 \\ 0 & C_2' & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

where M_1', C_2' are positive semidefinite (C_2' may be lacking). Then set

$$K' = \Phi^T K \Phi = \begin{bmatrix} K_{11}' & K_{12}' & K_{13}' \\ K_{12}'^T & K_{22}' & K_{23}' \\ K_{13}'^T & K_{23}'^T & K_{33}' \end{bmatrix}$$

where K_{33}' is positive definite (as a principal submatrix of the positive definite K). By setting

$$x = \Phi x', \quad x' = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix}, \quad K' = \Phi^T K \Phi$$

we obtain an equivalent system

$$\begin{aligned} M_1'\ddot{x}_1' + C_1'\dot{x}_1 \qquad + K_{11}'x_1' + K_{12}'x_2' + K_{13}'x_3' &= \varphi_1 \\ C_2'\dot{x}_2' + K_{12}'^T x_1' + K_{22}'x_2' + K_{23}'x_3' &= \varphi_2 \\ K_{13}'^T x_1' + K_{23}'^T x_2' + K_{33}'x_3' &= \varphi_3 \end{aligned} \tag{4.1}$$

with $\varphi = \Phi^T f$. Eliminating x_3' from the third line of (4.1) gives

$$\tilde{M}\ddot{\tilde{x}} + \tilde{C}\dot{\tilde{x}} + \tilde{K}\tilde{x} = \tilde{\varphi}(t) \tag{4.2}$$

with

$$\tilde{x} = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}, \quad \tilde{M} = \begin{bmatrix} M_1' & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.3}$$

$$\tilde{\varphi} = \begin{bmatrix} \varphi_1 - K_{13}' K_{33}^{'-1} \varphi_3 \\ \varphi_2 - K_{23}' K_{33}^{-1} \varphi_3 \end{bmatrix}, \tag{4.4}$$

$$\tilde{C} = \begin{bmatrix} C_1' & 0 \\ 0 & C_2' \end{bmatrix},$$

$$\tilde{K} = \begin{bmatrix} \tilde{K}_{11} & \tilde{K}_{12} \\ \tilde{K}_{12}^T & \tilde{K}_{22} \end{bmatrix} = \begin{bmatrix} K_{11}' - K_{13}' {K'}_{33}^{-1} K_{13}^{'T} & K_{12}' - K_{13}' K_{33}^{'-1} K_{23}^{'T} \\ K_{12}^{'T} - K_{23}' K_{33}^{'-1} K_{13}^{'T} & K_{22}' - K_{23}' K_{33}^{'-1} K_{23}^{'T} \end{bmatrix}.$$

The relation

$$\frac{1}{2}\dot{x}^T M \dot{x} = \frac{1}{2}\dot{x}^{'T} M' \dot{x}' = \frac{1}{2}\dot{\tilde{x}}^T \tilde{M} \dot{\tilde{x}}, \tag{4.5}$$

is immediately verified. Further,

$$\begin{aligned}
\frac{1}{2}x^T K x = \frac{1}{2}x'^T K' x' &= \frac{1}{2}\sum_{j,k=1}^{2} x_j'^T K_{jk}' x_k' + x_1'^T K_{13}' x_3' + x_2'^T K_{23}' x_3' + \frac{1}{2} x_3'^T K_{33}' x_3' \\
&= \frac{1}{2}\sum_{j,k=1}^{2} x_j'^T K_{jk}' x_k' + x_1'^T K_{13}' {K'}_{33}^{-1}(\varphi_3 - K_{13}'^T x_1' - K_{23}'^T x_2') + \\
&+ x_2'^T K_{23}' {K'}_{33}^{-1}(\varphi_3 - K_{13}'^T x_1' - K_{23}'^T x_2') + \\
&+ \frac{1}{2}(\varphi_3^T - x_1'^T K_{13}' - x_2'^T K_{23}') {K'}_{33}^{-1}(\varphi_3 - K_{13}'^T x_1' - K_{23}'^T x_3') \\
&= \frac{1}{2}\tilde{x}^T \tilde{K} \tilde{x} + x_1'^T K_{13}' {K'}_{33}^{-1}\varphi_3 + x_2'^T K_{23}' {K'}_{33}^{-1}\varphi_3 \\
&+ (-x_1'^T K_{13} - x_2'^T K_{23}') {K'}_{33}^{-1}\varphi_3 + \frac{1}{2}\varphi_3^T {K'}_{33}^{-1}\varphi_3 \\
&= \frac{1}{2}\tilde{x}^T \tilde{K} \tilde{x} + \frac{1}{2}\varphi_3^T {K'}_{33}^{-1}\varphi_3
\end{aligned} \tag{4.6}$$

Exercise 4.2 *Show that $\tilde{K}$ is positive definite.*

In particular, if $f = 0$ (free oscillations) then

$$\frac{1}{2}x^T K x = \frac{1}{2}\tilde{x}^T \tilde{K} \tilde{x}.$$

Note that quite often in applications the matrices M and C are already diagonal and the common null space can be immediately separated off. Anyhow, we will now suppose that the system (1.1) has already the property $\mathcal{N}(M) \cap \mathcal{N}(C) = \{0\}$.

For future considerations it will be convenient to simultaneously diagonalise not M and C but M and K. To this end we will use a coordinate transformation Φ as close as possible to the form (2.4).

Proposition 4.3 *In (1.1) let the matrix M have rank $m < n$ and let $\mathcal{N}(M) \cap \mathcal{N}(C) = \{0\}$. Then for any positive definite matrix Ω_2 of order $n - m$ there is a real non-singular Φ such that*

$$\Phi^T M \Phi = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}, \quad \Phi^T K \Phi = \Omega^2 = \begin{bmatrix} \Omega_1^2 & 0 \\ 0 & \Omega_2^2 \end{bmatrix}, \tag{4.7}$$

Ω_1 positive definite. Furthermore, in the matrix

$$D = \Phi^T C \Phi = \begin{bmatrix} D_{11} & D_{12} \\ D_{12}^T & D_{22} \end{bmatrix}. \tag{4.8}$$

the block D_{22} is positive definite. Moreover, if any other Φ' performs (4.7), possibly with different Ω_1', Ω_2' then

$$\Phi' = \Phi \begin{bmatrix} U_{11} & 0 \\ 0 & \Omega_2^{-1} U_{22} \Omega_2' \end{bmatrix} \tag{4.9}$$

where

$$U = \begin{bmatrix} U_{11} & 0 \\ 0 & U_{22} \end{bmatrix} \tag{4.10}$$

is unitary and

$$\Omega_1' = U_{11}^T \Omega_1 U_{11}. \tag{4.11}$$

Proof. There is certainly a $\mathbf{\Phi}$ with

$$\mathbf{\Phi}^T M \mathbf{\Phi} = \begin{bmatrix} M_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{\Phi}^T K \mathbf{\Phi} = I \tag{4.12}$$

where M_1 is positive definite diagonal matrix of rank m; this can be done by the simultaneous diagonalisation (2.1) of the pair M, K of symmetric matrices the second of which is positive definite and $\mathbf{\Phi}$ is said to be K-normalised (the M-normalisation as in (2.1) is not possible M being singular). For given Ω_2 and $\Omega_1 = M_1^{-1/2}$ the matrix

$$\Phi = \mathbf{\Phi} \begin{bmatrix} M_1^{-1/2} & 0 \\ 0 & \Omega_2 \end{bmatrix}$$

is obviously non-singular and satisfies (4.7). Moreover, the matrix D from (4.8) satisfies

$$\mathcal{N}\left(\begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}\right) \cap \mathcal{N}(D) = \{0\},$$

hence D_{22} is positive definite.

Now we prove the relations (4.9) – (4.11). Suppose that we have two transformations Φ and Φ' both performing (4.7) with the corresponding matrices Ω_1, Ω_2, Ω_1', Ω_2', respectively. Then the transformations

$$\mathbf{\Phi} = \Phi \begin{bmatrix} \Omega_1^{-1} & 0 \\ 0 & \Omega_2^{-1} \end{bmatrix}, \quad \mathbf{\Phi}' = \Phi' \begin{bmatrix} \Omega_1'^{-1} & 0 \\ 0 & \Omega_2'^{-1} \end{bmatrix}$$

are K-normalised fulfilling (4.12) with M_1, M_1', respectively. Hence, as one immediately sees,

$$\mathbf{\Phi}' = \mathbf{\Phi} U$$

where

$$U = \begin{bmatrix} U_{11} & 0 \\ 0 & U_{22} \end{bmatrix}$$

is unitary and

$$M_1' = U_{11}^T M_1 U_{11} \text{ or, equivalently } \Omega_1'^{-1} = U_{11}^T \Omega_1^{-1} U_{11}. \tag{4.13}$$

Hence

$$\Phi' \begin{bmatrix} \Omega_1'^{-1} & 0 \\ 0 & \Omega_2'^{-1} \end{bmatrix} = \Phi \begin{bmatrix} \Omega_1^{-1} U_{11} & 0 \\ 0 & \Omega_2^{-1} U_{22} \end{bmatrix}.$$

Since the matrix U_{11} is orthogonal (4.13) is rewritten as $\Omega_1^{-1} U_{11} \Omega_1' = U_{11}$ and we have

$$\begin{aligned} \Phi' &= \Phi \begin{bmatrix} \Omega_1^{-1} U_{11} \Omega_1' & 0 \\ 0 & \Omega_2^{-1} U_{22} \Omega_2' \end{bmatrix}, \\ &= \Phi \begin{bmatrix} U_{11} & 0 \\ 0 & \Omega_2^{-1} U_{22} \Omega_2' \end{bmatrix}. \end{aligned}$$

Q.E.D.

We now proceed to construct the phase-space formulation of (1.1) which, after the substitution $x = \Phi y$, Φ from (4.7), reads

$$\ddot{y}_1 + D_{11}\dot{y}_1 + D_{12}\dot{y}_2 + \Omega_1^2 y_1 = \phi_1, \tag{4.14}$$

$$D_{12}^T \dot{y}_1 + D_{22}\dot{y}_2 + \Omega_2^2 y_2 = \phi_2, \tag{4.15}$$

with

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \Phi^T f.$$

By introducing the new variables

$$z_1 = \Omega_1 y_1, \quad z_2 = \Omega_2 y_2, \quad z_3 = \dot{y}_1$$

the system (4.14), (4.15) becomes

$$\dot{z}_1 = \Omega_1 z_3 \tag{4.16}$$

$$\dot{z}_2 = \Omega_2 D_{22}^{-1} \left(\phi_2 - D_{12}^T z_3 - \Omega_2 z_2 \right) \tag{4.17}$$

$$\dot{z}_3 = \phi_1 - D_{11} z_3 - D_{12} D_{22}^{-1} \left(\phi_2 - D_{12}^T z_3 - \Omega_2 z_2 \right) - \Omega_1 z_1 \tag{4.18}$$

or, equivalently

$$\dot{z} = Az + G, \quad G = \begin{bmatrix} 0 \\ \Omega_2 D_{22}^{-1} \phi_2 \\ \phi_1 - D_{12} D_{22}^{-1} \phi_2 \end{bmatrix}$$

where

$$A = \begin{bmatrix} 0 & 0 & \Omega_1 \\ 0 & -\Omega_2 D_{22}^{-1} \Omega_2 & -\Omega_2 D_{22}^{-1} D_{12}^T \\ -\Omega_1 & D_{12} D_{22}^{-1} \Omega_2 & -\hat{D} \end{bmatrix} \tag{4.19}$$

is of order $n + m$ and

$$\hat{D} = D_{11} - D_{12} D_{22}^{-1} D_{12}^T$$

is positive semidefinite because D itself is such. Conversely, let (4.16) – (4.18) hold and set

$$z_1 = \Omega_1 y_1, \quad z_2 = \Omega_2 y_2.$$

Then by (4.16) we have $\dot{y}_1 = z_3$ and (4.17) implies (4.15) whereas (4.18) and (4.15) imply (4.14).

The total energy identity

$$\|z\|^2 = \|z_1\|^2 + \|z_2\|^2 + \|z_3\|^2 = x^T K x + \dot{x}^T M \dot{x}$$

is analogous to (3.8). As in (3.5), (3.6) we have

$$z^T A z \leq 0 \text{ for all } z, \tag{4.20}$$

and

$$A^T = JAJ \tag{4.21}$$

with

$$J = \begin{bmatrix} I_n & 0 \\ 0 & -I_m \end{bmatrix}. \tag{4.22}$$

There is a large variety of Φ's performing (4.7). However, *the resulting A's in (4.19) are mutually unitarily equivalent.* To prove this it is convenient to go to the inverses. For A from (4.19) and $\Omega = \mathrm{diag}(\Omega_1, \Omega_2)$ the identity

$$A^{-1} = \begin{bmatrix} -\Omega^{-1} D \Omega^{-1} & -F \\ F^T & 0_m \end{bmatrix} \quad \text{with} \quad F = \begin{bmatrix} \Omega_1^{-1} \\ 0 \end{bmatrix} \tag{4.23}$$

is immediately verified. Suppose now that Φ' also performs (4.7) with the corresponding matrices Ω_1', Ω_2', D' A'^{-1}, respectively. Then, according to (4.9), (4.10) in Proposition 4.3,

$$D' = \Phi'^T C \Phi' = \begin{bmatrix} U_{11}^T & 0 \\ 0 & \Omega_2' U_{22}^T \Omega_2^{-1} \end{bmatrix} D \begin{bmatrix} U_{11} & 0 \\ 0 & \Omega_2^{-1} U_{22} \Omega_2' \end{bmatrix}$$

and, using (4.13),

$$\Omega'^{-1}D'\Omega'^{-1} = \begin{bmatrix} \Omega_1'^{-1}U_{11}^T & 0 \\ 0 & U_{22}^T\Omega_2^{-1} \end{bmatrix} D \begin{bmatrix} U_{11}\Omega_1'^{-1} & 0 \\ 0 & \Omega_2^{-1}U_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \Omega_1'^{-1}U_{11}^T & 0 \\ 0 & U_{22}^T\Omega_2^{-1} \end{bmatrix} D \begin{bmatrix} \Omega_1^{-1}U_{11} & 0 \\ 0 & \Omega_2^{-1}U_{22} \end{bmatrix} = U^T\Omega^{-1}D\Omega^{-1}U.$$

Similarly,

$$F' = \begin{bmatrix} \Omega_1'^{-1} \\ 0 \end{bmatrix} = \begin{bmatrix} U_{11}^T\Omega_1^{-1}U_{11} \\ 0 \end{bmatrix} = U^{-1}FU_{11}.$$

Altogether,

$$A'^{-1} = \hat{U}^{-1}A^{-1}\hat{U}$$

and then

$$A' = \hat{U}^{-1}A\hat{U}$$

where the matrix

$$\hat{U} = \begin{bmatrix} U & 0 \\ 0 & U_{11} \end{bmatrix} \tag{4.24}$$

is unitary.

In numerical calculations special choices of Ω_2 might be made, for instance, the ones which possibly decrease the condition number of the transformation Φ. This issue could require further consideration.

Exercise 4.4 *Prove the property (4.20).*

Example 4.5 Our model example from Fig. 1.1 allows for singular masses. We will consider the case with $n = 2$ and

$$m_1 = m > 0, \quad m_2 = 0, \quad k_1, k_2 > 0, k_3 = 0,$$

$$c_i = 0, \quad d_1 = 0, \quad d_2 = d > 0.$$

This gives

$$M = \begin{bmatrix} m & 0 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 0 & d \end{bmatrix}, \quad K = \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix}.$$

Obviously here $\mathcal{N}(M) \cap \mathcal{N}(C) = \{0\}$, so the construction (4.19) is possible. The columns of Φ are the eigenvectors of the generalised eigenvalue problem

$$M\phi = \nu K\phi$$

whose eigenvalues are the zeros of the characteristic polynomial

$$\det(M - \nu K) = k_1k_2\nu^2 - mk_2\nu$$

and they are

$$\nu_1 = \frac{m}{k_1}, \quad \nu_2 = 0.$$

The corresponding eigenvectors are

$$\phi_1 = \alpha_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \phi_2 = \alpha_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

We choose $\alpha_1 = \alpha_2 = 1/\sqrt{m}$. This gives

$$\omega_1 = \sqrt{k_1/m}, \quad \omega_2 = \sqrt{k_2/m}$$

(we replace the symbols $\Omega_1, \Omega_2, D_{ij}$ etc. from (4.19) by their lower case relateds since they are now scalars).

Then

$$\Phi = \begin{bmatrix} \phi_1 \; \phi_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} / \sqrt{m},$$

$$\Phi^T M \Phi = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Phi^T K \Phi = \begin{bmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{bmatrix}.$$

Hence

$$\Phi^T C \Phi = D = \frac{d}{m} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Finally, by $\hat{d} = d_{11} - d_{12}^2/d_{22} = 0$ the phase-space matrix (4.19) reads

$$A = \begin{bmatrix} 0 & 0 & \sqrt{k_1/m} \\ 0 & -k_2/d & -\sqrt{k_2/m} \\ -\sqrt{k_1/m} & \sqrt{k_2/m} & 0 \end{bmatrix}.$$

Exercise 4.6 *If the null space of M is known as*

$$M = \begin{bmatrix} M_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad M_1 \textit{ positive definite}$$

try to form a phase space matrix A in (4.19) without using the spectral decomposition (2.1) and using Cholesky decompositions instead.

Exercise 4.7 *Find out special properties of the matrix A from (4.19) in the modal damping case.*

Exercise 4.8 *Establish the relation between the matrices A from (3.3) and from (4.19). Hint: Replace in (3.3) the matrix M by $M_\epsilon = M + \epsilon I$, $\epsilon > 0$. build the corresponding matrix A_ε and find*

$$\lim_{\varepsilon \to 0} A_\varepsilon^{-1}.$$

Exercise 4.9 *Consider the extreme case of $M = 0$ and solve the differential equation (1.1) by simultaneously diagonalising the matrices C, K.*

If not specified otherwise we shall by default understand the mass matrix M as non-singular.

Chapter 5
”Indefinite metric”

The symmetry property of our phase-space matrices gives rise to a geometrical structure popularly called ’an indefinite metric’ which we now will study in some detail.

In doing so it will be convenient to include complex matrices as well. More precisely, from now on we will consider matrices over the field Ξ of real or complex numbers. Note that all considerations and formulae in Chapters 2 - 4 are valid in the complex case as well. Instead of being real symmetric the matrices M, C, K can be allowed to be Hermitian. Similarly, the vectors $x, y, ...$ may be from Ξ^n and real orthogonal matrices appearing there become unitary. The only change is to replace the transpose ^{T} in $A^T, x^T, y^T, L_1^T, L_2^T, ...$ by the adjoint *. For a general $A \in \Xi^{n,n}$ the dissipativity means

$$\operatorname{Re} y^* A y \leq 0, \quad y \in \mathbb{C}^n. \tag{5.1}$$

While taking complex Hermitian M, C, K does not have direct physical meaning, the phase-space matrices are best studied as complex. Special cases in which only complex or only real matrices are meant will be given explicit mention. (The latter might be the case where in numerical applications one would care to keep real arithmetic for the sake of the computational efficiency.)

We start from the property (3.6) or (4.21). It can be written as

$$[Ax, y] = [x, Ay] \tag{5.2}$$

where

$$[x, y] = y^* J x = (Jx, y) \tag{5.3}$$

and

$$J = \operatorname{diag}(\pm 1). \tag{5.4}$$

More generally we will allow J to be any matrix J with the property

$$J = J^{-1} = J^*. \tag{5.5}$$

Such matrices are usually called *symmetries* and we will keep that terminology in our text hoping to cause not too much confusion with the term symmetry as a matrix property. In fact, without essentially altering the theory we could allow J to be just Hermitian and non-singular. This has both advantages and shortcomings, see Chapter 11 below.

The function $[x, y]$ has the properties

- $[\alpha x + \beta y, z] = \alpha[x, z] + \beta[y, z]$
- $[x, z] = \overline{[z, x]}$
- $[x, y] = 0 \quad \text{for all} \quad y \in \Xi^n, \quad \text{if and only if} \quad x = 0.$

The last property — we say $[\,\cdot\,,\,\cdot\,]$ is *non-degenerate* — is a weakening of the common property $[x, x] > 0$, whenever $x \neq 0$ satisfied by scalar products. This is why we call it an ’indefinite scalar product’ and the geometric environment, created by it an ’indefinite metric’.

The most common form of a symmetry of order n is

$$J = \begin{bmatrix} I_m & 0 \\ 0 & -I_{n-m} \end{bmatrix} \tag{5.6}$$

but variously permuted diagonal matrices are also usual. Common non-diagonal forms are

$$J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & iI \\ -iI & 0 \end{bmatrix}$$

or

$$J = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}.$$

Remark 5.1 Different forms of the symmetry J are a matter of convenience, dictated by computational or theoretical requirements. Anyhow, one may always assume J as diagonal with the values ±1 on its diagonal in any desired order. Indeed, there is a unitary matrix U such that

$$\tilde{J} = U^* J U = \operatorname{diag}(\pm 1)$$

with any prescribed order of signs. If a matrix A was, say, J-Hermitian then its ’unitary map’ $\tilde{A} = U^* A U$ will be $\tilde{J}$-Hermitian. The spectral properties of A can be read-off from those for $\tilde{A}$.

We call vectors x and y *J-orthogonal* and write $x[\perp]y$, if

$$[x, y] = y^* J x = 0. \tag{5.7}$$

Similarly two sets of vectors S_1, S_2 are called J-orthogonal — we then write $S_1[\perp]S_2$ — if (5.7) holds for any $x \in S_1$ and $y \in S_2$. A vector is called

- *J-normalised*, if $|[x,x]| = 1$,
- *J-positive*, if $[x,x] > 0$,
- *J-non-negative*, if $[x,x] \geq 0$ (and similarly for J-negative and J-non-positive vectors),
- *J-definite*, if it is either J-positive or J-negative,
- *J-neutral*, if $[x,x] = 0$.

Analogous names are given to a subspace, if all of its non-zero vectors satisfy one of the conditions above. In addition, a subspace $\mathcal{X}$ is called *J-non-degenerate*, if the only vector from $\mathcal{X}$, J-orthogonal to $\mathcal{X}$ is zero. The synonyms *of positive/negative/neutral type* for J-positive/negative/neutral will also be used.

Proposition 5.2 *A subspace $\mathcal{X}$ is J-definite, if and only if it is either J-positive or J-negative.*

Proof. Let $x_\pm \in \mathcal{X}$, $[x_+, x_+] > 0$, $[x_-, x_-] < 0$. Then for any real t the equation

$$0 = [x_+ + tx_-, x_+ + tx_-] = [x_+, x_+] + 2t\,\mathrm{Re}[x_+, x_-] + t^2[x_-, x_-]$$

has always a solution t producing a J-neutral $x_+ + tx_- \neq 0$. Q.E.D.

A set of vectors $u_1, \ldots, u_p$ is called *J-orthogonal*, if none of them is J-neutral and

$$[u_j, u_k] = 0 \quad \text{for} \quad j \neq k$$

and J-orthonormal, if

$$|[u_j, u_k]| = \delta_{jk}.$$

Exercise 5.3 *J-orthogonal vectors are linearly independent.*

Exercise 5.4 *For any real matrix (3.5) implies (5.1).*

5.1 Sylvester inertia theorem

Non-degenerate subspaces play an important role in the spectral theory of J-Hermitian matrices. To prepare ourselves for their study we will present some facts on Hermitian matrices which arise in the context of the so-called Sylvester inertia theorem. The simplest variant of this theorem says: If A is Hermitian and T non-singular then T^*AT and A have the same numbers of positive, negative and zero eigenvalues. We will here drop the non-singularity of T, it will even not have to be square. At the same time, in tending to make the proof as constructive as possible, we will drop from it the notion of eigenvalue. Inertia is a notion more elementary than eigenvalues and it can be handled by rational operations only.

First we prove the *indefinite decomposition formula* valid for any Hermitian matrix $A \in \Xi^{n,n}$. It reads

$$A = G\boldsymbol{\alpha}G^* \tag{5.8}$$

where $G \in \Xi^{n,n}$ is non-singular and $\boldsymbol{\alpha}$ is diagonal.

A possible construction of $G, \boldsymbol{\alpha}$ is obtained via the eigenvalue decomposition $A = U\Lambda U^*$. Eigenvalue-free construction is given by Gaussian elimination without pivoting which yields

$$A = LDL^*$$

where L is lower triangular with the unit diagonal. This elimination breaks down, if in its course zero elements are encountered on the diagonal. Common row pivoting is forbidden here, because it destroys the Hermitian property. Only simultaneous permutations of both rows and columns are allowed. The process is modified to include block elimination steps. We shall describe one elimination step. The matrix A is given as

$$A = \begin{bmatrix} A^{(k-1)} & 0 \\ 0 & A^{(n-k+1)} \end{bmatrix}, \quad k = 1, \dots, n-1$$

where $A^{(k-1)}$ is of order $k-1$ and is void for $k = 1$. We further partition

$$A^{(n-k+1)} = \begin{bmatrix} E & C^* \\ C & B \end{bmatrix}$$

where E is square of order $s \in \{1, 2\}$ and is supposed to be non-singular. For $s = 1$ the step is *single* and for $s = 2$ *double*. Set

$$X = \begin{bmatrix} I_{k-1} & & 0 \\ 0 & I_s & 0 \\ 0 & CE^{-1} & I_{n-k+1-s} \end{bmatrix}.$$

Then

$$XAX^* = \begin{bmatrix} A^{(k-1)} & 0 & 0 \\ 0 & E & 0 \\ 0 & 0 & A^{(n-k+1-s)} \end{bmatrix}.$$

In order to avoid clumsy indices we will describe the construction by the following algorithm (the symbol := denotes the common assigning operation).

Algorithm 5.5

```
$\Psi := I_n$; $D_0 := I_n$; $k := 1$;
while $k \le n-1$
    Find $j$ such that $|a_{kj}| = \max_{i \ge k} |a_{ki}|$;
    If $a_{kj} = 0$
```

```
        $k := k + 1$;
    End if
    If $|a_{kk}| \geq |a_{kj}|/2 > 0$
        Perform the single elimination step;
        $A := XAX^*$; $\Psi := X^*\Psi$;
        $k := k + 1$;
    Else
        If $|a_{jj}| \geq |a_{kj}|/2 > 0$
            Swap the $k$-th and the $j$-th columns and rows in $A$;
            Swap the $k$-th and the $j$-th columns in $\Psi$;
            Perform the single elimination step;
            $A := XAX^*$; $\Psi := X^*\Psi$;
            $k := k + 1$;
        Else
            Swap the $k+1$-th and the $j$-th columns and rows in $A$;
            Swap the $k+1$-th and the $j$-th columns in $\Psi$;
            Perform the double elimination step;
            $A := XAX^*$; $\Psi := X^*\Psi$;
            $k := k + 2$;
        End if
    End if
End while
```

The choices of steps and swappings in this algorithm secure that the necessary inversions are always possible. On exit a non-singular matrix Ψ is obtained with the property

$$\Psi A \Psi^* = \operatorname{diag}(A_1, \ldots, A_p)$$

where A_s is of order $n_s \in \{1, 2\}$. In the latter case we have

$$A_s = \begin{bmatrix} a & b \\ \bar{b} & c \end{bmatrix}, \quad \text{with } |b| \geq 2\max\{|a|, |c|\}. \tag{5.9}$$

This 2 × 2-matrix is further decomposed as follows:

$$Y_s = \begin{bmatrix} 1 & 0 \\ -\frac{a-c}{2|b|+a-c} & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{b}{|b|} \\ 1 & -\frac{b}{|b|} \end{bmatrix}, \tag{5.10}$$

then

$$Y_s A_s Y_s^* = \begin{bmatrix} 2|b| + a + c & 0 \\ 0 & -\frac{4|b|^2}{2|b|+a-c} \end{bmatrix}.$$

Here again all divisions are possible by virtue of the condition in (5.9).[1] Finally replace Ψ by $\mathrm{diag}(Y_1^*, \ldots, Y_p^*)\Psi$ where any order-one Y_s equals to one. Then

$$\Psi A \Psi^* = \boldsymbol{\alpha}, \tag{5.11}$$

$\boldsymbol{\alpha}$ as in (5.8). Now (5.8) follows with $G = \Psi^{-1}$.

Remark 5.6 The indefinite decomposition as we have presented it is, in fact, close to the common numerical algorithm described in [4]. The factor $1/2$ in appearing in line 7 of Algorithm 5.5 and further on is chosen for simplicity; in practice it is replaced by other values which increase the numerical stability and may depend on the sparsity of the matrix A.

If A is non-singular then all α_i are different from zero and by replacing Ψ by $\mathrm{diag}(|\alpha_1|^{-1/2}, \ldots, |\alpha_n|^{-1/2})\Psi$ in (5.11) we obtain

$$\Psi A \Psi^* = \mathrm{diag}(\pm 1). \tag{5.12}$$

Theorem 5.7 *(Sylvester theorem of inertia) If A is Hermitian then the numbers of the positive, negative and zero diagonal elements of the matrix $\boldsymbol{\alpha}$ in (5.8) depends only on A and not on G. Denoting these numbers by $\iota_+(A), \iota_-(A), \iota_0(A)$, respectively, we have*

$$\iota_\pm(T^*AT) \le \iota_\pm(A)$$

for any T for which the above matrix product is defined. Both inequalities become equalities, if T^* is injective. If T is square and non-singular then also $\iota_0(T^*AT) = \iota_0(A)$.

Proof. Let

$$B = T^*AT \tag{5.13}$$

be of order m. By (5.8) we have $A = G\boldsymbol{\alpha}G^*,\ B = F\boldsymbol{\beta}F^*$ with G, F non-singular and

$$\boldsymbol{\alpha} = \mathrm{diag}(\boldsymbol{\alpha}_+, -\boldsymbol{\alpha}_-, 0_{n_0}), \quad \boldsymbol{\beta} = \mathrm{diag}(\boldsymbol{\beta}_+, -\boldsymbol{\beta}_-, 0_{m_0})$$

where $\boldsymbol{\alpha}_\pm, \boldsymbol{\beta}_\pm$ are positive definite diagonal matrices of order $n_\pm, m_\pm$, respectively.[2] The equality (5.13) can be written as

$$Z^*\boldsymbol{\alpha}Z = \boldsymbol{\beta} \tag{5.14}$$

with $Z = G^*TF^{-*}$. We partition Z as

[1] One may object that the matrix Y_s needs irrational operation in computing $|b|$ for a complex b, this may be avoided by replacing $|b|$ in (5.10) by, say, $|\operatorname{Re} b| + |\operatorname{Im} b|$.

[2] Note that in (5.8) any desired ordering of the diagonals of $\boldsymbol{\alpha}$ may be obtained, if the columns of G are accordingly permuted.

$$Z = \begin{bmatrix} Z_{++} & Z_{+-} & Z_{+0} \\ Z_{-+} & Z_{--} & Z_{-0} \\ Z_{0+} & Z_{0-} & Z_{00} \end{bmatrix}.$$

Here Z_{++} is of order $n_+ \times m_+$ etc. according to the respective partitions in $\boldsymbol{\alpha}, \boldsymbol{\beta}$. By writing (5.14) blockwise and by equating the $1,1$- and the $2,2$-blocks, respectively, we obtain

$$Z_{++}^* \boldsymbol{\alpha}_+ Z_{++} - Z_{-+}^* \boldsymbol{\alpha}_- Z_{-+} = \boldsymbol{\beta}_+, \quad Z_{+-}^* \boldsymbol{\alpha}_+ Z_{+-} - Z_{--}^* \boldsymbol{\alpha}_- Z_{--} = -\boldsymbol{\beta}_-.$$

Thus, the matrices $Z_{++}^* \boldsymbol{\alpha}_+ Z_{++} = \boldsymbol{\beta}_+ + Z_{-+}^* \boldsymbol{\alpha}_- Z_{-+}$ and $Z_{--}^* \boldsymbol{\alpha}_- Z_{--} = \boldsymbol{\beta}_- + Z_{+-}^* \boldsymbol{\alpha}_+ Z_{+-}$ are positive definite. This is possible only if

$$m_+ \le n_+, \quad m_- \le n_-.$$

If T is square and non-singular then Z is also square and non-singular. Hence applying the same reasoning to $Z^{-*} \boldsymbol{\beta} Z^{-1} = \boldsymbol{\alpha}$ yields

$$m_+ = n_+, \quad m_- = n_-$$

(and then, of necessity, $m_0 = n_0$). The proof that the numbers $\iota_+, \iota_-, \iota_0$ do not depend on G in (5.8) is straightforward: in (5.13) we set $T = I$ thus obtaining $\iota_\pm(A) = n_\pm$. The only remaining case is the one with an injective T^* in (5.13). Then TT^* is Hermitian and positive definite and (5.13) implies

$$TBT^* = TT^*ATT^*,$$

hence

$$\iota_\pm(B) \le \iota_\pm(A) = \iota_\pm(TT^*ATT^*) = \iota_\pm(B).$$

The last assertion is obvious. Q.E.D.

The triple

$$\iota(A) = (\iota_+(A), \iota_-(A), \iota_0(A))$$

is called *the inertia* of A. Obviously, $\iota_+(A) + \iota_-(A) = \mathrm{rank}(A)$.

Corollary 5.8 *Let $\hat{A}$ be any principal submatrix of a Hermitian matrix A then*

$$\iota_\pm(\hat{A}) \le \iota_\pm(A).$$

Corollary 5.9 *If A is block diagonal then its inertia is the sum of the inertiae of its diagonal blocks.*

Corollary 5.10 *The inertia equals the number of the positive, negative and zero eigenvalues, respectively.*

Proof. Use the eigenvalue decomposition to obtain (5.8). Q.E.D.

We turn back to the study of non-degenerate subspaces.

Theorem 5.11 *Let $\mathcal{X}$ be a subspace of Ξ^n and $x_1, \ldots, x_p$ its basis and set*

$$X = \begin{bmatrix} x_1 \cdots x_p \end{bmatrix}.$$

Then the following are equivalent.

(i) $\mathcal{X} = \mathcal{R}(X)$ *possesses a J-orthonormal basis.*
(ii) $\mathcal{X}$ *is J-non-degenerate.*
(iii) *The J-Gram matrix of the vectors $x_1, \ldots, x_p$*

$$H = (x_i^* J x_j) = X^* J X$$

is non-singular.
(iv) *The* J-orthogonal companion *of $\mathcal{X}$ defined as*

$$\mathcal{X}_{[\perp]} = \{y \in \Xi^n : x^* J y = 0 \quad \textit{for all} \quad x \in \mathcal{X}\}$$

is a direct complement of $\mathcal{X}$ i.e. we have

$$\mathcal{X} \dot{+} \mathcal{X}_{[\perp]} = \Xi^n$$

and we write

$$\mathcal{X} [+] \mathcal{X}_{[\perp]} = \Xi^n.$$

In this case $\mathcal{X}_{[\perp]}$ is called the J-orthogonal complement *of $\mathcal{X}$.*

Proof. Let $u_1, \ldots, u_p$ be a J-orthonormal basis in $\mathcal{X}$ then $x = \alpha_1 u_1 + \cdots + \alpha_p u_p$ is J-orthogonal to $\mathcal{X}$, if and only if $\alpha_1 = \cdots = \alpha_p = 0$ i.e. $x = 0$. Thus, (i) ⇒ (ii). To prove (ii) ⇔ (iii) note that

$$H\alpha = X^* J X \alpha = 0 \quad \text{for some} \quad \alpha \in \Xi^p, \alpha \neq 0$$

if and only if

$$X^* J x = 0 \quad \text{for some} \quad x \in \mathcal{X}, \quad x \neq 0$$

and this means that $\mathcal{X}$ is J-degenerate.

Now for (ii) ⇔ (iv). The J-orthogonal companion of $\mathcal{X}$ is the (standard) orthogonal complement of $J\mathcal{X}$; since J is non-singular we have $\dim J\mathcal{X} = \dim \mathcal{X} = p$ and so, $\dim \mathcal{X}_{[\perp]} = n - p$ and we have to prove

$$\mathcal{X} \cap \mathcal{X}_{[\perp]} = \{0\}$$

but this is just the non-degeneracy of $\mathcal{X}$. Since the latter is obviously equivalent to the non-degeneracy of $\mathcal{X}_{[\perp]}$ this proves (ii) ⇒ (iv). To prove (iv) ⇔ (iii) complete the vectors $x_1, \ldots, x_p$ to a basis $x_1, \ldots, x_n$ of Ξ^n such that $x_{p+1}, \ldots, x_n$ form a basis in $\mathcal{X}_{[\perp]}$ and set

$$\tilde{X} = \begin{bmatrix} x_1 \cdots x_n \end{bmatrix}.$$

Then the matrix

$$\tilde{H} = \tilde{X}^* J \tilde{X} = \begin{bmatrix} H & 0 \\ 0 & \tilde{H}_{22} \end{bmatrix}$$

is non-singular. The same is then true of H and this is (iii).

It remains to prove e.g. (iii) $\Rightarrow$ (i). By (5.12)

$$H = X^* J X = F_1 J_1 F_1^*$$

because H is non-singular. Then the vectors $u_j = X F_1^{-*} e_j$ from $\mathcal{X}$ are J orthonormal:

$$|u_j^* J u_k| = |e_j^* J_1 e_k| = \delta_{jk}.$$

Q.E.D.

Corollary 5.12 *Any J-orthonormal set can be completed to a J-orthonormal basis in Ξ^n.*

Proof. The desired basis consists of the columns of the matrix U in the proof of Theorem 5.11. Q.E.D.

Theorem 5.13 *Any two J-orthonormal bases in a J-non-degenerate space have the same number of J-positive (and J-negative) vectors.*

Proof. Let two J-orthonormal bases be given as the columns of

$$U = [u_1, \ldots, u_p] \quad \text{and} \quad V = [v_1, \ldots, v_p],$$

respectively. Then

$$U^* J U = J_1, \quad V^* J V = J_2$$

$$J_1 = \mathrm{diag}(\pm 1), \quad J_2 = \mathrm{diag}(\pm 1).$$

On the other hand, by $\mathcal{R}(U) = \mathcal{R}(V)$ we have

$$U = VM, \quad M \quad \text{non-singular},$$

so

$$J_1 = M^* V^* J V M = M^* J_2 M.$$

By the theorem of Sylvester the non-singularity of M implies that the number of positive and negative eigenvalues of J_1 and J_2 — and this is just the number of plus and minus signs on their diagonal — coincide. Q.E.D.

Corollary 5.14 *For any subspace $\mathcal{X} \subseteq \Xi^n$ we define*

$$\iota(\mathcal{X}) = (\iota_+(\mathcal{X}), \iota_0(\mathcal{X}), \iota_-(\mathcal{X})) = \iota(H)$$

where $H = X^* J X$ and $X = [\, x_1, \ldots, x_p \,]$ is any basis of $\mathcal{X}$. Then $\iota(\mathcal{X})$ does not depend on the choice of the basis $x_1, \ldots, x_p \in \mathcal{X}$. If $\mathcal{X}$ is non-degenerate *then obviously* $\iota_0(\mathcal{X}) = \iota_0(\mathcal{X}_{[\perp]}) = 0$ *and*

$$\iota_\pm(\mathcal{X}) + \iota_\pm(\mathcal{X}_{[\perp]}) = \iota_\pm(J). \tag{5.15}$$

Exercise 5.15 *If* $\mathcal{X}_1$ *and* $\mathcal{X}_2$ *are any two non-degenerate, mutually* J*-orthogonal subspaces then their direct sum* $\mathcal{X}$ *is also non-degenerate and*

$$\iota(\mathcal{X}) = \iota(\mathcal{X}_1) + \iota(\mathcal{X}_2)$$

(the addition is understood elementwise).

Chapter 6
Matrices and indefinite scalar products

Having introduced main notions of the geometry based on an indefinite scalar product we will now study special classes of matrices intimately connected with this scalar product. These will be the analogs of the usual Hermitian and unitary matrices. At first sight the formal analogy seems complete but the indefiniteness of the underlying scalar product often leads to big, sometimes surprising, differences.

A matrix $H \in \Xi^{n,n}$ is called *J-Hermitian* (also *J-symmetric*, if real), if

$$H^* = JHJ \quad \Leftrightarrow \quad (JH)^* = JH$$

or, equivalently,

$$[Hx, y] = y^* JHx = y^* H^* Jx = [x, Hy].$$

It is convenient to introduce the J-adjoint $A^{[*]}$ or the J-transpose $A^{[T]}$ of a general matrix A, defined as

$$A^{[*]} = JA^* J, \quad A^{[T]} = JA^T J,$$

respectively. In the latter case the symmetry J is supposed to be real. Now the J-Hermitian property is characterised by

$$A^{[*]} = A$$

or, equivalently by

$$[Ax, y] = [x, Ay] \text{ for all } x, y \in \Xi^n.$$

A matrix $U \in \Xi^{n,n}$ is *J-unitary* (also *J-orthogonal*, if real), if

$$U^{-1} = U^{[*]} = JU^* J \quad \Leftrightarrow \quad U^* JU = J.$$

Obviously all J-unitaries form a multiplicative group and satisfy

$$|\det U| = 1.$$

The J-unitarity can be expressed by the identity

$$[Ux, Uy] = y^* U^* J U x = [x, y].$$

Exercise 6.1 *Prove*

$$\begin{aligned} I^{[*]} &= I \\ (\alpha A + \beta B)^{[*]} &= \overline{\alpha} A^{[*]} + \overline{\beta} B^{[*]} \\ (AB)^{[*]} &= B^{[*]} A^{[*]} \\ (A^{[*]})^{-1} &= (A^{-1})^{[*]} \\ A^{[*]} = A^{-1} &\iff A \textit{ is } J\textit{-unitary} \end{aligned}$$

In the particular case $J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$ a J-Hermitian A looks like

$$A = \begin{bmatrix} A_{11} & A_{12} \\ -A_{12}^* & A_{22} \end{bmatrix}, \quad A_{11}^* = A_{11}, \quad A_{22}^* = A_{22}$$

whereas for $J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$ the J-Hermitian is

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{11}^* \end{bmatrix}, \quad A_{12}^* = A_{12}, \quad A_{21}^* = A_{21}. \tag{6.1}$$

By the unitary invariance of the spectral norm the condition number of a J-unitary matrix U is

$$\kappa(U) = \|U\| \|U^{-1}\| = \|U\| \|JU^*J\| = \|U\| \|U^*\| = \|U\|^2 \geq 1$$

We call a matrix *jointly unitary*, if it is simultaneously J-unitary and unitary. Examples of jointly unitary matrices U are given in (3.11) and (4.24).

Exercise 6.2 *Prove that the following are equivalent*

(i) U *is jointly unitary of order* n.
(ii) U *is* J*-unitary and* $\|U\| = 1$.
(iii) U *is* J*-unitary and* U *commutes with* J.
(iv) U *is unitary and* U *commutes with* J.
(v) U *is* J*-unitary and* $\|U\|_E^2 = \operatorname{Tr} U^* U = n$.

Example 6.3 Any matrix of the form

$$Y = H(W) \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix}, \quad H(W) = \begin{pmatrix} \sqrt{I + WW^*} & W \\ W^* & \sqrt{I + W^*W} \end{pmatrix} \tag{6.2}$$

is obviously J-unitary with J from (5.6); here W is an $m \times (n-m)$-matrix and V_1, V_2 are unitary. As a matter of fact, *any* J-unitary U is of this form. This we will now show. Any J-unitary U, partitioned according to (5.6) is written as

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}.$$

By taking the polar decompositions $U_{11} = W_{11}V_1$, $U_{22} = W_{22}V_2$ with $W_{11} = \sqrt{U_{11}U_{11}^*}$, $W_{22} = \sqrt{U_{22}U_{22}^*}$ we have

$$U = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}.$$

Now in the product above the second factor is J-unitary ($V_{1,2}$ being unitary). Thus, the first factor — we call it H — is also J-unitary, that is, $H^*JH = J$ or, equivalently, $HJH^* = J$. This is expressed as

$$\begin{array}{ll} W_{11}^2 - W_{21}^* W_{21} = I_m & W_{11}^2 - W_{12}W_{12}^* = I_m \\ W_{11}W_{12} - W_{21}^*W_{22} = 0 & W_{11}W_{21}^* - W_{12}W_{22} = 0 \\ W_{12}^*W_{12} - W_{22}^2 = -I_{n-m} & W_{21}W_{21}^* - W_{22}^2 = -I_{n-m} \end{array}$$

This gives (note that W_{11}, W_{22} are Hermitian positive semidefinite)

$$\sqrt{I_m + W_{21}^*W_{21}}W_{12} = W_{21}^*\sqrt{I_{n-m} + W_{21}W_{21}^*}$$

or, equivalently,

$$W_{12}(I_{n-m} + W_{12}^*W_{12})^{-1/2} = (I_m + W_{21}^*W_{21})^{-1/2}W_{21}^*$$

$$= W_{21}^*(I_{n-m} + W_{21}W_{21}^*)^{-1/2} = W_{21}^*(I_{n-m} + W_{12}^*W_{12})^{-1/2}$$

hence $W_{21}^* = W_{12}$. Here the second equality follows from the general identity $Af(BA) = f(AB)A$ which we now assume as known and will address in discussing analytic matrix functions later. Now set $W = W_{12}$ and obtain (6.2).

Jointly unitary matrices are very precious whenever they can be used in computations, because their condition is equal to one.

If A is J-Hermitian and U is J-unitary then one immediately verifies that

$$A' = U^{-1}AU = U^{[*]}AU$$

is again J-Hermitian.

If J is diagonal then the columns (and also the rows) of any J-unitary matrix form a J-orthonormal basis. It might seem odd that the converse is not true. This is so because in our definition of J-orthonormality the order of the vectors plays no role. For instance the vectors

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

are J-orthonormal with respect to

$$J = \begin{bmatrix} 1 & & \\ & -1 & \\ & & 1 \end{bmatrix}$$

but the matrix U built from these three vectors in that order is not J-orthogonal. We rather have

$$U^* J U = J' = \begin{bmatrix} 1 & & \\ & 1 & \\ & & -1 \end{bmatrix}.$$

To overcome such difficulties we call a square matrix J, J'*-unitary* (J, J'*-orthogonal*, if real), if

$$U^* J U = J' \tag{6.3}$$

where J and J' are symmetries. If in (6.3) the orders of J' and J do not coincide we call U *a* J, J'*-isometry.* If both J and J' are unit matrices this is just a standard isometry.

Exercise 6.4 *Any* J, J'*- isometry is injective.*

Proposition 6.5 *If* J, J' *are symmetries and* U *is* J, J'*-unitary then* J *and* J' *are unitarily similar:*

$$J' = V^{-1} J V, \quad V \quad \textit{unitary} \tag{6.4}$$

and

$$U = WV \tag{6.5}$$

where W *is* J*-unitary.*

Proof. The eigenvalues of both J and J' consists of ± ones. By (6.3) U must be non-singular but then (6.3) and the theorem of Sylvester imply that J and J' have the same eigenvalues including multiplicities, so they are unitarily similar. Now (6.5) follows from (6.4). Q.E.D.

If A is J-Hermitian and U is J, J'-unitary then

$$A' = U^{-1} A U$$

is J'-Hermitian. Indeed,

$$A'^* = U^* A^* U^{-*} = U^* A J U J'.$$

Here AJ, and therefore U^*AJU is Hermitian so A'^* is J'-Hermitian.

Using only J-unitary similarities the J-Hermitian matrix

$$A = \begin{bmatrix} 1 & 0 & -5 \\ 0 & 2 & 0 \\ -5 & 0 & 1 \end{bmatrix}, \quad J = \begin{bmatrix} 1 & & \\ & 1 & \\ & & -1 \end{bmatrix} \tag{6.6}$$

cannot be further simplified, but using the J, J'-unitary matrix

$$\Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad J' = \begin{bmatrix} 1 & & \\ & -1 & \\ & & 1 \end{bmatrix}$$

we obtain the more convenient block-diagonal form

$$\Pi^{-1} A \Pi = \begin{bmatrix} 1 & -5 & 0 \\ -5 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

which may have computational advantages. Since any J, J'-unitary matrix U is a product of a unitary matrix and a J-unitary matrix we have

$$\kappa(U) = \|U\|^2 \geq 1$$

where the equality is attained, if and only if U is unitary.

Mapping by a J, J'-unitary matrix U preserves the corresponding "indefinite geometries" e.g.

- if $x' = Ux, \quad y' = Uy$ then $x^*Jy = x'^*J'y', \quad x^*Jx = x'^*J'x'$
- if $\mathcal{X}$ is a subspace and $\mathcal{X}' = U\mathcal{X}$ then $\mathcal{X}'$ is J'-non-degenerate, if and only if $\mathcal{X}$ is J-non-degenerate (the same with 'positive', 'non-negative', 'neutral' etc.)
- The inertia does not change:

$$\iota_\pm(\mathcal{X}) = \iota'_\pm(\mathcal{X}')$$

where $\iota'_\pm$ is related to J'.

The set of J, J'-unitary matrices is not essentially larger than the one of standard J-unitaries but it is often more convenient in numerical computations.

Exercise 6.6 *Find all real J-orthogonals and all complex J-unitaries of order 2 with*

$$J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{or} \quad J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Exercise 6.7 *A matrix is called* jointly Hermitian*, if it is both Hermitian and J-Hermitian. Prove that the following are equivalent*

(i) *H is jointly Hermitian.*
(ii) *H is J-Hermitian and it commutes with J.*
(iii) *H is Hermitian and it commutes with J.*

Chapter 7
Oblique projections

In the course of these lectures we will often have to deal with non-orthogonal (oblique) projections, so we will collect here some useful facts about them.

Any square matrix of order n is called a *projection* (or *projector*), if

$$P^2 = P.$$

We list some obvious properties of projections.

Proposition 7.1 *If P is a projection then the following hold.*

(i) $\mathcal{R}(P) = \{x \in \Xi^n : Px = x\}$.
(ii) $Q = I - P$ *is also a projection and*

$$Q + P = I, \quad PQ = QP = 0,$$

(iii) $\mathcal{R}(P) \dot{+} \mathcal{R}(Q) = \Xi^n, \quad \mathcal{R}(P) = \mathcal{N}(Q), \quad \mathcal{R}(Q) = \mathcal{N}(P)$.
(iv) If $x_1, \dots, x_p$ is a basis in $\mathcal{R}(P)$ and $x_{p+1}, \dots, x_n$ a basis in $\mathcal{R}(Q)$ then $X = \begin{bmatrix} x_1 \cdots x_n \end{bmatrix}$ is non-singular and

$$X^{-1}PX = \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix}.$$

(v) $\operatorname{rank}(P) = \operatorname{Tr}(P)$.
(vi) A set of projections $P_1, \dots, P_q$ is called a decomposition of the identity if

$$P_1 + \cdots + P_q = I, \quad P_i P_j = P_j P_i = P_j \delta_{ij}.$$

To any such decomposition there corresponds the direct sum

$$\mathcal{X}_1 \dot{+} \cdots \dot{+} \mathcal{X}_q = \Xi^n$$

with $\mathcal{X}_i = \mathcal{R}(P_i)$ and vice versa.
(vii) P^* is also a projection and $\mathcal{R}(P^*) = \mathcal{N}(Q)_\perp$.

The proofs are straightforward and are omitted.

Exercise 7.2 *If P_1, P_2 are projections then*

1. *$\mathcal{R}(P_1) \subseteq \mathcal{R}(P_2)$ is equivalent to*
$$P_1 = P_2 P_1; \tag{7.1}$$
2. *$\mathcal{R}(P_1) = \mathcal{R}(P_2)$ is equivalent to*
$$P_1 = P_2 P_1 \quad \& \quad P_2 = P_1 P_2; \tag{7.2}$$
3. *if, in addition, both P_1, P_2 are Hermitian or J-Hermitian then (7.2) implies $P_2 = P_1$.*

Exercise 7.3 *Any two decompositions of the identity $P_1, \ldots, P_q, P_1', \ldots, P_q'$ for which $\operatorname{Tr} P_j = \operatorname{Tr} P_j'$, $j = 1, \ldots, q$, are similar, that is, there exists a non-singular S with*
$$S^{-1} P_i S = P_i', \quad i = 1, \ldots, q$$
if all P_i, P_i' are orthogonal projections then S can be chosen as unitary.

Define $\theta \in (0, \pi/2]$ as the minimal angle between the subspaces $\mathcal{X}$ and $\mathcal{X}'$ as
$$\begin{aligned} \theta &= \min\{\arccos |x^* y|, x \in \mathcal{X}, y \in \mathcal{X}', \|x\| = \|y\| = 1\} && (7.3) \\ &= \arccos(\max\{|x^* y|, x \in \mathcal{X}, y \in \mathcal{X}', \|x\| = \|y\| = 1\}). && (7.4) \end{aligned}$$

A simpler formula for the angle θ is
$$\cos\theta = \|\mathbb{P}\mathbb{P}'\| \tag{7.5}$$
where $\mathbb{P}, \mathbb{P}'$ are the orthogonal projections onto $\mathcal{X}, \mathcal{X}'$, respectively. Indeed,
$$\|\mathbb{P}\mathbb{P}'\| = \max_{x,y \neq 0} \frac{|(\mathbb{P}x)^* \mathbb{P}' y|}{\|x\| \|y\|}$$
and this maximum is taken on a pair $x \in \mathcal{X}, y \in \mathcal{X}'$. In fact, for a general x we have $\|Px\| \leq \|x\|$ and
$$\frac{|(\mathbb{P}x)^* \mathbb{P}' y|}{\|x\| \|y\|} \leq \frac{|(\mathbb{P}x)^* \mathbb{P}' y|}{\|\mathbb{P}x\| \|y\|} = \frac{|(\mathbb{P}(\mathbb{P}x))^* \mathbb{P}' y|}{\|\mathbb{P}x\| \|y\|}$$
and similarly with y. Thus,

$$\|\mathbb{P}\mathbb{P}'\| = \max\left\{\frac{|(\mathbb{P}x)^*\mathbb{P}'y|}{\|\mathbb{P}x\|\|\mathbb{P}'y\|}, \quad \mathbb{P}x, \mathbb{P}'y \neq 0\right\} \tag{7.6}$$

$$= \max\left\{\frac{|u^*v|}{\|u\|\|v\|}, \quad u \in \mathcal{X}, \quad v \in \mathcal{X}', \quad u, v \neq 0\right\} \tag{7.7}$$

$$= \max\{|u^*v|, \quad u \in \mathcal{X}, \quad v \in \mathcal{X}', \quad \|u\| = \|v\| = 1\}. \tag{7.8}$$

which proves (7.5).

Theorem 7.4 *Let P be a projection and let $\mathbb{P}$ be the orthogonal projection onto $\mathcal{R}(P)$. Then for $Q = I - P$ and $\mathbb{Q} = I - \mathbb{P}$ we have*

$$\|P\| = \|Q\| = \frac{1}{\sin\theta}, \tag{7.9}$$

$$\|P^* - P\| = \|Q^* - Q\| = \|\mathbb{P} - P\| = \|\mathbb{Q} - Q\| = \cot\theta, \tag{7.10}$$

where θ is the minimal angle between $\mathcal{R}(P)$ and $\mathcal{R}(Q)$.

Proof. Take any orthonormal basis of $\mathcal{R}(P)$ and complete it to an orthonormal basis of Ξ^n. These vectors are the columns of a unitary matrix U for which

$$\mathbb{P}' = U^*\mathbb{P}U = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

$$P' = U^*PU = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix},$$

where P' is again a projection with $\mathcal{R}(P') = \mathcal{R}(\mathbb{P}')$.
Now apply (7.2). $\mathbb{P}'P' = P'$ gives

$$\begin{bmatrix} P_{11} & P_{12} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \quad \Rightarrow \quad P_{21} = P_{22} = 0,$$

whereas $P'\mathbb{P}' = \mathbb{P}'$ gives

$$\begin{bmatrix} P_{11} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \quad \Rightarrow \quad P_{11} = I.$$

Thus, we can assume that P and $\mathbb{P}$ already are given as

$$P = \begin{bmatrix} I & X \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I \\ 0 \end{bmatrix} \begin{bmatrix} I & X \end{bmatrix} \tag{7.11}$$

$$\mathbb{P} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \tag{7.12}$$

Furthermore,

$$Q = I - P = \begin{bmatrix} 0 & -X \\ 0 & I \end{bmatrix} = \begin{bmatrix} -X \\ I \end{bmatrix} \begin{bmatrix} 0 & I \end{bmatrix}$$

whereas the orthogonal projection onto $\mathcal{N}(P) = \mathcal{R}(Q)$ is

$$\hat{\mathbb{P}} = \begin{bmatrix} -X \\ I \end{bmatrix} (I + X^*X)^{-1} \begin{bmatrix} -X^* & I \end{bmatrix}. \tag{7.13}$$

According to (7.5) we have

$$\begin{aligned} \cos^2\theta = \|\mathbb{P}\hat{\mathbb{P}}\mathbb{P}\| &= \operatorname{spr}\left(\begin{bmatrix} -X \\ 0 \end{bmatrix} (I + X^*X)^{-1} \begin{bmatrix} -X^* & 0 \end{bmatrix} \right) \\ &= \operatorname{spr}((I + X^*X)^{-1} X^*X). \end{aligned}$$

To evaluate the last expression we use the singular value decomposition

$$X = U\xi V^* \tag{7.14}$$

with U, V unitary and ξ diagonal (but not necessarily square). Then $X^*X = V\xi^*\xi V^*$, $\xi^*\xi = \operatorname{diag}(|\xi_1|^2, |\xi_2|^2, \ldots)$ and

$$\begin{aligned} \cos^2\theta = \max_i \frac{|\xi_i|^2}{1 + |\xi_i|^2} &= \frac{\max_i |\xi_i|^2}{1 + \max_i |\xi_i|^2} \\ = \frac{\|X\|^2}{1 + \|X\|^2} &= 1 - \frac{1}{1 + \|X\|^2} \end{aligned}$$

Hence

$$\|X\| = \cot\theta. \tag{7.15}$$

Now

$$\begin{aligned} \|P\|^2 &= \operatorname{spr}\left(\begin{bmatrix} I \\ X^* \end{bmatrix} \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} \begin{bmatrix} I & X \end{bmatrix} \right) \\ &= \operatorname{spr}\left(\begin{bmatrix} I + XX^* & 0 \\ 0 & 0 \end{bmatrix} \right) = 1 + \|X\|^2 \\ &= \frac{1}{\sin^2\theta} \end{aligned}$$

Then obviously $\|P\| = \|Q\|$ and (7.9) holds.
Also

$$\begin{aligned} \|P^* - P\| &= \left\| \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix} \right\| = \|X\| = \cot\theta, \\ \|\mathbb{P} - P\| &= \left\| \begin{bmatrix} 0 & X \\ 0 & 0 \end{bmatrix} \right\| = \cot\theta \end{aligned}$$

and similarly for $\|Q^* - Q\|$ and $\|\mathbb{Q} - Q\|$. Q.E.D.

Corollary 7.5

$$\|P\| = 1 \quad \Leftrightarrow \quad P^* = P$$

i.e. a projection is orthogonal, if and only if its norm equals one.

Theorem 7.6 *If P is a projection and $\mathbb{P}$ the orthogonal projection onto $\mathcal{R}(P)$ then there exists a non-singular S such that*

$$S^{-1}PS = \mathbb{P}$$

and

$$\begin{aligned}\|S\|\|S^{-1}\| &= \frac{1}{2}(2 + \cot^2\theta + \sqrt{4\cot^2\theta + \cot^4\theta}) \\ &= \frac{1}{2}(2 + \|P\|^2 - 1 + \sqrt{\|P\|^2 - 1}\sqrt{3 + \|P\|^2}). \qquad (7.16)\end{aligned}$$

Proof. As in the proof of Theorem 7.4 we may assume that $P, \mathbb{P}$ already have the form (7.11), (7.12), respectively. Set

$$S = \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix}.$$

Then

$$\begin{aligned}S^{-1}PS &= \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}\begin{bmatrix} I & X \\ 0 & 0 \end{bmatrix}\begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = \mathbb{P}.\end{aligned}$$

We compute the condition number of S:

$$S^*S = \begin{bmatrix} I & -X \\ -X^* & I + X^*X \end{bmatrix}.$$

Then using (7.14)

$$S^*S = \begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}\begin{bmatrix} I & -\xi \\ -\xi & I + \xi^*\xi \end{bmatrix}\begin{bmatrix} U & 0 \\ 0 & V^* \end{bmatrix},$$

so those eigenvalues of S^*S which are different from 1 are given by

$$\begin{aligned}\lambda_i^+ &= \frac{2 + \xi_i^2 + \sqrt{4\xi_i^2 + \xi_i^4}}{2} > 1, \\ \lambda_i^- &= \frac{1}{\lambda_i^+} < 1,\end{aligned}$$

where ξ_i are the diagonal elements of ξ. Hence

$$\lambda_{max}(S^*S) = \frac{2 + \|X\|^2 + \sqrt{4\|X\|^2 + \|X\|^4}}{2},$$
$$\lambda_{min}(S^*S) = \frac{1}{\lambda_{max}(S^*S)}$$

and

$$\kappa(S) = \sqrt{\frac{\lambda_{max}(S^*S)}{\lambda_{min}(S^*S)}} =$$
$$= \frac{2 + \cot^2\theta + \sqrt{4\cot^2\theta + \cot^4\theta}}{2}$$

and this proves (ii). Q.E.D.

This theorem automatically applies to any decomposition of the identity consisting of two projections P and $I - P$. It can be recursively applied to any decomposition of the identity $P_1, \dots, P_q$ but the obtained estimates get more and more clumsy with the growing number of projections, and the condition number of the transforming matrix S will also depend on the space dimension n.

Exercise 7.7 *Let P be a projection and $H = I - 2P$. Prove*

(i) $H^2 = I$,
(ii) $\kappa(H) = 1 + 2(\|P\|^2 - 1) + 2\|P\|\sqrt{\|P\|^2 - 1}$.

Hint: use the representation for P from Theorem 7.4.

Chapter 8
J-orthogonal projections

J-orthogonal projections are intimately related to the J-symmetry of the phase space matrices which govern damped systems. In this chapter we study these projections in some detail.

A projection P is called *J-orthogonal* if the matrix P is J-Hermitian i.e. if

$$P^{[*]} = JP^*J = P.$$

A subspace $\mathcal{X}$ is said to possess a J-orthogonal projection P, if $\mathcal{X} = \mathcal{R}(P)$.

Theorem 8.1 *(i) A subspace $\mathcal{X} \subseteq \Xi^n$ is J-non-degenerate, if and only if it possesses a J-orthogonal projection P. (ii) Any J-orthogonal projection P can be represented as*

$$P = UJ'U^*J \tag{8.1}$$

where U is a J, J'-isometry of type $n \times m$ and $\mathcal{R}(P) = \mathcal{R}(U)$. (iii) The J-orthogonal projection P is uniquely determined by $\mathcal{X}$. (iv) We have

$$\iota_\pm(JP) = \iota_\pm(J') = \iota_\pm(\mathcal{X}). \tag{8.2}$$

(v) $Q = I - P$ is the J-orthogonal projection onto $\mathcal{X}_{[\perp]}$.

Proof. (i) Let P be a J-orthogonal projection and $\mathcal{X} = \mathcal{R}(P)$. Then for $x \in \mathcal{X}$ and $z \in \Xi^n$

$$[x, Pz] = [Px, z] = [x, z]$$

holds. Thus, $x[\perp]\mathcal{X}$ is equivalent to $x[\perp]\Xi^n$. Since Ξ^n is non-degenerate the same is true of $\mathcal{X}$. Conversely, let $\mathcal{X}$ be non-degenerate of dimension $p < n$. Then by Theorem 5.11 and Corollary 5.12 there is a J-orthonormal basis $u_1, \ldots, u_n$ of Ξ^n such that $u_1, \ldots, u_p$ spans $\mathcal{X}$. Then P, defined as

$$Pu_j = \begin{cases} u_j, \; j \le p \\ 0, \; j > p \end{cases}$$

is the wanted projection. This proves (i).

(ii) If P is given by (8.1), (6.3) then

$$P^2 = UJ'U^*JUJ'U^*J = UJ'J'J'U^*J = P,$$

$$P^{[*]} = JP^*J = J^2UJ'U^*J = UJ'U^*J = P$$

hence P is a J-orthogonal projection. The injectivity of U follows from (6.3), hence also $\mathrm{rank}(U) = p := \mathrm{rank}(P)$. Conversely, let P be a J-orthogonal projection. Then, according to Theorem 5.11 $\mathcal{X} = \mathcal{R}(P)$ is J-non-degenerate and there is a J-orthonormal basis $u_1, \ldots, u_n$ of Ξ^n such that the first p vectors span $\mathcal{X}$. With $U = [\, u_1, \ldots, u_p \,]$ we have

$$U^*JU = J' = \mathrm{diag}(\pm 1). \tag{8.3}$$

The wanted projection is

$$P = UJ'U^*J \tag{8.4}$$

Obviously $\mathcal{X} \subseteq \mathcal{R}(U)$. This inclusion is, in fact equality because $\mathrm{rank}(U) = p$. This proves (ii).

(iii) If there are two J-orthogonal projections with the same range

$$UJ'U^*J, \quad VJ_2V^*J, \quad \mathcal{R}(V) = \mathcal{R}(U)$$

then $V = UM$ for some non-singular M and from

$$J_2 = V^*JV = M^*U^*JUM = M^*J'M$$

it follows

$$VJ_2V^*J = UMJ_2M^*U^*J = UJ'U^*J.$$

This proves (iii).

(iv) Let $P = UJ'U^*J$. The second equality in (8.2) follows, if in Corollary 5.14 we take X as U. Further, since J is a symmetry,

$$\iota_\pm(JP) = \iota_\pm(UJ'U^*)$$

and since U is injective, by the Sylvester inertia theorem,

$$\iota_\pm(UJ'U^*) = \iota_\pm(J').$$

This proves (iv). The assertion (v) is obvious. Q.E.D.

Let $P_1, \ldots, P_r$ be J-orthogonal projections with the property

$$P_iP_j = \delta_{ij}I$$

then their sum P is again a J-orthogonal projection and we say that the system $P_1, \ldots, P_r$ is a *J-orthogonal decomposition* of P. If $P = I$ then we speak of a *J-orthogonal decomposition of the identity.* To any *J-orthogonal decomposition* there corresponds a J-orthogonal sum of their ranges

$$\mathcal{R}(P) = \mathcal{R}(P_1)[+]\cdots[+]\mathcal{R}(P_r)$$

and vice versa. The proof is straightforward and is left to the reader.

Theorem 8.2 *Let $P_1, \ldots, P_q$ be a J-orthogonal decomposition of the identity and $n_j = \mathrm{Tr}(P_j)$ the dimensions of the respective subspaces. Then there exists J, J^0-unitary U such that*

$$J^0 = \begin{bmatrix} J_1^0 & & \\ & \ddots & \\ & & J_p^0 \end{bmatrix}, \quad \iota_\pm(J_j^0) = \iota_\pm(JP_j) \tag{8.5}$$

and

$$U^{-1}P_jU = P_j^0 = \begin{bmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & I_{n_j} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}. \tag{8.6}$$

Proof. By Theorem 8.1 we may write

$$P_j = U_jJ_j^0U_j^*J, \quad U_j^*JU_j = J_j^0.$$

Then $U = \begin{bmatrix} U_1 & \cdots & U_p \end{bmatrix}$ obviously satisfies

$$U^*JU = \begin{bmatrix} U_1^* \\ \vdots \\ U_p^* \end{bmatrix} J \begin{bmatrix} U_1 & \cdots & U_p \end{bmatrix} = \begin{bmatrix} J_1^0 & & \\ & \ddots & \\ & & J_p^0 \end{bmatrix} =: J^0$$

and

$$U^{-1}P_jU = J'U^*JU_jJ_j^0U_j^*JU =$$

$$= \begin{bmatrix} J_1^0U_1^* \\ \vdots \\ J_p^0U_p^* \end{bmatrix} JU_jJ_j^0U_j^* \begin{bmatrix} JU_1 \cdots JU_p \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ J_j^0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \cdots 0 \; J_j^0 \; 0 \cdots 0 \end{bmatrix} = P_j^0.$$

Q.E.D.

Corollary 8.3 *Any two J-orthogonal decompositions of the identity*

$$P_1, \ldots, P_q \quad \textit{and} \quad P_1', \ldots, P_q'$$

satisfying $\iota_\pm(JP_j) = \iota_\pm(JP_j')$ are J-unitarily similar:

$$P_j' = U^{-1}P_jU, \quad U \ J\textit{-unitary.}$$

Theorem 8.4 *Let P, P' be J-orthogonal projections and $\|P' - P\| < 1$. Then there is a J-unitary U such that*

$$P' = U^{-1}PU. \tag{8.7}$$

Proof. The matrix square root

$$Z = \left[I - (P' - P)^2\right]^{-1/2}$$

is defined by the known binomial series

$$Z = \sum_{k=0}^{\infty} (-1)^k \binom{\alpha}{k} (P' - P)^{2k}, \quad \binom{\alpha}{k} = \frac{\alpha(\alpha - 1) \cdots (\alpha - k + 1)}{k!},$$

which converges because of $\|P' - P\| < 1$. Moreover, Z is J-Hermitian and $(P' - P)^2$ (and therefore Z) commutes with both P and P'. Set

$$U = Z\left[PP' + (I - P)(I - P')\right],$$

then

$$PU = ZPP' = Z\left[PP' + (I - P)(I - P')\right]P' = UP',$$

so (8.7) holds. To prove J-unitarity we compute

$$U^* J U =$$

$$JZ\left[P'P + (I - P')(I - P)\right]\left[PP' + (I - P)(I - P')\right] JZ$$
$$= JZ\left[I - (P' - P)^2\right] Z = J.$$

Q.E.D.

Corollary 8.5 *Under the conditions of the theorem above we have*

$$\iota_\pm(JP') = \iota_\pm(JP).$$

In particular, if JP is positive or negative semidefinite then JP' will be the same. Also, a continuous J-orthogonal-projection valued function $P(t)$ for t from a closed real interval, cannot change the inertia of $JP(t)$.

Proof. Apply the Sylvester inertia theorem to

$$JP' = JU^{-1}PU = U^* JPU.$$

Then cover the interval by a finite number of open intervals such that within each of them $\|P(t) - P(t')\| < 1$ holds and apply Theorem 8.4. Q.E.D.

Exercise 8.6 *Let P_1, P_2 be projections such that P_1 is J-orthogonal and P_2 $J' - orthogonal$. Then they are J, J'-unitarily similar, if and only if*

$$\iota_\pm(J'P_2) = \iota_\pm(JP_1).$$

Exercise 8.7 *A non-degenerate subspace is definite, if and only if it possesses a J-orthogonal projection P with one of the properties*

$$x^* JPx \geq 0 \text{ or } x^* JPx \leq 0 \text{ for all } x \in \Xi^n$$

that is, the matrix JP or $-JP$ is positive semidefinite.

A J-orthogonal projection with one of the properties in Exercise 8.7 will be called *J-positive and J-negative*, respectively.

Exercise 8.8 *Try to prove the following. If P, Q are J-orthogonal projections such that P is J-positive (J-negative) and $PQ = Q$, then*

1. *$QP = Q$,*
2. *Q is J-positive (J-negative),*
3. *$\|Q\| \leq \|P\|$.*

Hint: use Exercise 7.2, the representation (8.1) and Theorem 8.1.

Chapter 9
Spectral properties and reduction of J-Hermitian matrices

Here we start to study the properties of the eigenvalues and eigenvectors of J-Hermitian matrices. Particular attention will be given to the similarities and differences from (standard) Hermitian matrices.

We begin with a list of properties which are more or less analogous to the ones of common Hermitian matrices.

Theorem 9.1 *Let A be J-Hermitian. Then*

1. *If both A and J are diagonal, then A is real.*
2. *The spectrum of A is symmetric with respect to the real axis.*
3. *Any eigenvalue whose eigenvector is not J-neutral, is real.*
4. *If λ and μ are eigenvalues and $\overline{\lambda} \neq \mu$, then the corresponding eigenvectors are J-orthogonal.*
5. *If a subspace $\mathcal{X}$ is invariant under A, then so is its J-orthogonal companion.*
6. *The following are equivalent*

   (i) There is a J-non-degenerate subspace, invariant under A.
   (ii) A commutes with a non-trivial J-orthogonal projection.
   (iii) There is a J, J'- unitary U with

$$U^{-1}AU = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad J' = \mathrm{diag}(\pm 1). \tag{9.1}$$

If the subspace from (i) is J-definite then U can be chosen such that the matrix A_1 in (9.1) is real diagonal and

$$J' = \begin{bmatrix} \pm I & 0 \\ 0 & J_2' \end{bmatrix}, \quad J_2' = \mathrm{diag}(\pm 1). \tag{9.2}$$

Proof. The proofs of 1,2,3,4,5 are immediate and are omitted. To prove 6 let $\mathcal{X}$ be J-non-degenerate. Then there is a J-orthogonal basis in $\mathcal{X}$, represented as the columns of U_1, so

$$U_1^* J U_1 = J_1 = \mathrm{diag}(\pm 1)$$

and

$$AU_1 = U_1 M$$

with

$$M = J_1 U_1^* J A U_1$$

and with $P = U_1 J_1 U_1^* J$ we have $P^2 = P$, $P^{[*]} = P$ and

$$AP = AU_1 J_1 U_1^* J = U_1 J_1 U_1^* J A U_1 J_1 U^* J = PAP.$$

By taking $[*]$-adjoints, we obtain $AP = PA$.

Conversely, if

$$AU_1 J_1 U_1^* J = U_1 J_1 U_1^* J A,$$

then by postmultiplying by U_1,

$$AU_1 = U_1 J_1 U_1^* J A U_1,$$

i.e. $\mathcal{X} = \mathcal{R}(U_1)$ is invariant under A. Thus, (i) and (ii) are equivalent. Now, U from (iii) is obtained by completing the columns of U_1 to a J-orthonormal basis, see Corollary 5.12. Finally, if the subspace is J-definite then by construction (9.2) will hold. Since $U^{-1}AU$ is J'-Hermitian the block A_1 will be Hermitian. Hence there is a unitary V such that $V^{-1}A_1 V$ is real diagonal. Now replace U by

$$U = U \begin{bmatrix} V & 0 \\ 0 & I \end{bmatrix}.$$

Q.E.D.

Example 9.2 Let A be J-Hermitian and $Au = \lambda u$ with $[u, u] = u^* J u \neq 0$. Then the J- orthogonal projection along u, that is, onto the subspace spanned by u is

$$P = \frac{uu^* J}{u^* J u}$$

and it commutes with A. Indeed, since λ is real we have

$$AP = \frac{Auu^* J}{u^* J u} = \lambda \frac{uu^* J}{u^* J u} = \frac{u(Au)^* J}{u^* J u} =$$

$$\frac{uu^* A^* J}{u^* J u} = \frac{uu^* J A}{u^* J u} = PA.$$

Also

$$\|P\|^2 = \|P^* P\| = \frac{\|J u u^* u u^* J\|}{(u^* J u)^2} = \frac{\|u^* u\|^2}{(u^* J u)^2},$$

hence

$$\|P\| = \frac{u^* u}{|u^* J u|}. \tag{9.3}$$

We call a J-Hermitian matrix A *J, J'-unitarily diagonalisable*, if there is a J, J'-unitary matrix U such that the J'-Hermitian matrix

$$A' = U^{-1} A U$$

is diagonal. Note that J' itself need not be diagonal but, if it is then the diagonal elements of A', i.e. the eigenvalues, must be real. The J, J'-unitary block-diagonalisability is defined analogously.

Example 9.3 We consider the one dimensional damped system

$$m\ddot{x} + c\dot{x} + kx = 0, \quad m, c, k > 0. \tag{9.4}$$

The phase-space matrix

$$A = \begin{bmatrix} 0 & \omega \\ -\omega & -d \end{bmatrix}, \quad \omega = \sqrt{k/m},\ d = c/m \tag{9.5}$$

is J-symmetric with

$$J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Any 2×2 real J-orthogonal matrix (up to some signs) is of the form

$$U = U(x) = \begin{bmatrix} \cosh x & \sinh x \\ \sinh x & \cosh x \end{bmatrix}, \tag{9.6}$$

with $U(x)^{-1} = U(-x)$. We have

$$U^{-1} A U = \begin{bmatrix} \omega \sinh 2x + d \sinh^2 x & \omega \cosh 2x + \frac{d}{2} \sinh 2x \\ -\omega \cosh 2x - \frac{d}{2} \sinh 2x & -\omega \sinh 2x - d \cosh^2 x \end{bmatrix}.$$

Requiring the transformed matrix to be diagonal gives

$$\tanh 2x = -\frac{2\omega}{d} = -\sqrt{\frac{4km}{c^2}} \tag{9.7}$$

which is solvable if and only if $c^2 > 4km$. To better understand this, we will find the eigenvalues of A directly:

$$A \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

leads to

$$\begin{aligned} \omega y &= \lambda x \\ -\omega x - dy &= \lambda y \end{aligned}$$

and finally to

$$m\lambda^2 + c\lambda + k = 0$$

which is the characteristic equation of the differential equation (9.4). The roots are

$$\lambda_\pm = \frac{-c \pm \sqrt{c^2 - 4km}}{2m}. \tag{9.8}$$

According to the sign of the discriminant we distinguish three cases.

1. $c^2 > 4km$, the system is 'overdamped'.
   There are two distinct real eigenvalues and the matrix U in (9.6) exists with
   $$U^{-1}AU = \begin{bmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{bmatrix} \tag{9.9}$$
2. $c^2 < 4km$, the system is 'weakly damped'.
   Here diagonalisation is impossible, but we may look for U such that in $U^{-1}AU$ the diagonal elements are equal, which leads to
   $$\tanh 2x = -\frac{d}{2\omega} = -\sqrt{\frac{c^2}{4km}} < 1$$
   and
   $$U^{-1}AU = \begin{bmatrix} \operatorname{Re}\lambda_+ & \operatorname{Im}\lambda_+ \\ -\operatorname{Im}\lambda_+ & \operatorname{Re}\lambda_+ \end{bmatrix} \tag{9.10}$$
   (here, of course, $\lambda_- = \overline{\lambda}_+$). The transformed matrix is not diagonal but its eigenvalues are immediately read-off. Note that in this case we have
   $$|\lambda_\pm|^2 = \frac{k}{m},$$
   that is, as long as the eigenvalues are non-real, they stay on the circle with radius $\omega = \sqrt{k/m}$ around the origin.
3. $c^2 = 4km$, the system is 'critically damped'.
   Here we have only one real eigenvalue $\lambda = -c/(2m) = -\sqrt{k/m}$ and
   $$A = \sqrt{\frac{k}{m}}(-I + N), \quad N = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}.$$
   This matrix is not diagonalisable.

We come back to the first case. To the eigenvalues λ_+, λ_- there corresponds the J-orthogonal decomposition of the identity

$$P_+ = \begin{bmatrix} \cosh x \\ \sinh x \end{bmatrix} \begin{bmatrix} \cosh x & -\sinh x \end{bmatrix}$$

$$P_- = \begin{bmatrix} \sinh x \\ \cosh x \end{bmatrix} \begin{bmatrix} -\sinh x & \cosh x \end{bmatrix}$$

with

$$\|P_\pm\| = \cosh^2 x + \sinh^2 x = \cosh 2x = \frac{1}{\sqrt{1 - \frac{4km}{c^2}}}$$

and also

$$\|U\|^2 = \kappa(U) = e^{2|x|} = \|P_\pm\| + \sqrt{\|P_\pm\|^2 - 1}$$
$$= \sqrt{\frac{1+\theta^2}{\theta^2 - 1}}, \quad \theta = \frac{c}{2\sqrt{km}} = \frac{d}{2\omega}.$$

The matrix U from (9.10) is computed analogously with

$$\kappa(U) = \sqrt{\frac{1+\theta^2}{1-\theta^2}}.$$

Chapter 10
Definite spectra

In this chapter we begin to study J-Hermitian matrices which most resemble the standard Hermitian ones, and hence have some or all real eigenvalues. The most important property of these eigenvalues is that they remain real under small J-Hermitian perturbations. Sometimes, as in the standard Hermitian case these eigenvalues are expressed by minimax formulae.

A real eigenvalue λ of a J-Hermitian matrix is called *J-definite*, if each corresponding non-vanishing eigenvector is J-definite. Since the set of all eigenvectors belonging to λ is a subspace this immediately implies that the whole subspace is either J-positive or J-negative (Proposition 5.2). We then say that λ is *J-positive* or *J-negative*, respectively. The synonyms *of positive/negative/definite type* will correspondingly be used here as well. Otherwise we call λ to be *of mixed type.*

Proposition 10.1 *Let λ be a J-definite eigenvalue of a J-Hermitian matrix A. Then there is a J, J'-unitary U such that*

$$U^{-1}AU = \begin{bmatrix} \lambda I & 0 \\ 0 & A_2' \end{bmatrix}, \quad U^*JU = J' = \begin{bmatrix} \pm I_p & 0 \\ 0 & J_2' \end{bmatrix} \tag{10.1}$$

where A_2' is J_2'-Hermitian, p is the multiplicity of λ and $\lambda \notin \sigma(A_2')$.

Proof. Let λ be, say, J-positive. Then the corresponding eigenspace $\mathcal{X}_\lambda$ is J-positive and therefore J-non-degenerate. Thus, there is a J-orthonormal basis

$$u_1, \dots, u_p, u_{p+1}, \dots, u_n$$

of Ξ^n such that $u_1, \dots, u_p$ span $\mathcal{X}_\lambda$ and $u_1^*Ju_1 = \cdots = u_p^*Ju_p = 1$. Set $J' = \mathrm{diag}(u_1^*Ju_1, \dots, u_n^*Ju_n)$. Then the matrix $U = [u_1 \dots u_n]$ is J, J'-unitary as in (10.1) and

$$AU = [\lambda u_1 \ \dots \ \lambda u_p \ Au_{p+1} \ \dots \ Au_n]$$

and

$$U^{-1}AU = [\lambda e_1 \ \dots \ \lambda e_p \ * \ \dots \ *]$$

where e_i are the canonical basis vectors in Ξ^n. Since $U^{-1}AU$ is J'-Hermitian, it takes the form (10.1). It is clear from the construction that λ cannot be among the eigenvalues of A'_2. Q.E.D.

A J-Hermitian matrix is said to have *definite spectrum*, if each of its eigenvalues is J-definite. Then the reduction in Proposition 10.1 can be continued and we obtain the following corollary.

Corollary 10.2 *Any J-Hermitian matrix with a definite spectrum is J, J'-unitarily diagonalisable with a diagonal J'. The number of the J-positive and J-negative eigenvalues (counting multiplicities) is given by $\iota_\pm(J)$, respectively.*

If A is J-Hermitian and p a real polynomial then obviously $p(A)$ is again J-Hermitian. We call p normalised if its highest-power coefficient is ±1.

Theorem 10.3 *Let A be J-Hermitian. Then it has definite spectrum, if and only if there is a real normalised polynomial p such that $Jp(A)$ is positive definite.*

Proof. If $Jp(A)$ is positive definite then

$$Ax = \lambda x, \quad x \neq 0$$

implies $p(A)x = p(\lambda)x$ and

$$0 < x^* Jp(A)x = p(\lambda)x^* Jx.$$

Hence $x^* Jx \neq 0$ i.e. any eigenvalue is definite. To prove the converse we first assume that both A and J are diagonal:

$$A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \quad J = \mathrm{diag}(j_1, \ldots, j_n)$$

with

$$\lambda_1 \leq \ldots \leq \lambda_n \quad |j_i| = 1.$$

Now, the definiteness of every spectral point implies

$$\lambda_i = \lambda_k \Rightarrow j_i = j_k.$$

Thus, we can partition the sequence $\lambda_1, \ldots, \lambda_n$ into *sign groups* that is, maximal contiguous subsequences having the same sign in J. The sign groups are increasingly ordered. E.g. for

$$A = \mathrm{diag}(-2, 0, 1, 3, 5) \quad J = \mathrm{diag}(1, -1, -1, 1, 1)$$

the sign groups are (-2), $(0, 1)$, $(3, 5)$ and their signs are those on the diagonal of J. We take the polynomial p such that it has a simple zero between each two neighbouring sign groups. When properly signed this polynomial will give

$$j_i p(\lambda_i) > 0, \quad i = 1, \ldots, n$$

i.e. $Jp(A)$ is positive definite. In our example above we could choose $p(\lambda) = (\lambda - 2)(\lambda + 1)$.

For a general A take a J, J'-unitary matrix U such that both $J' = U^* JU$ and $A' = U^{-1}AU$ are diagonal (Corollary 10.2). Now,

$$J'p(A') = U^* JUp(U^{-1}AU) = U^* Jp(A)U$$

is again positive definite by virtue of the Sylvester theorem. Q.E.D.

To any J-Hermitian A with definite spectrum we associate the sequence

$$s(A) = (n_1, \ldots, n_r, \pm)$$

characterising the r sign groups (in increasing order); the first r numbers carry their cardinalities whereas $\pm \in \{-1, 1\}$ indicates the first of the alternating signs. We shall call $s(A)$ the *sign partition* of A or, equivalently, of the pair JA, J. The sign partition is obviously completely determined by any such $p(\lambda)$ which we call *the definitising polynomial.*

If a matrix with definite spectrum varies in a continuous way then the spectrum stays definite and the sign partition remains constant as long as the different sign groups do not collide. To prove this we need the following lemma.

Lemma 10.4 *Let $\mathcal{I} \ni t \mapsto H(t)$ be a Hermitian valued continuous function on a closed interval $\mathcal{I}$. Suppose that all $H(t)$ are non-singular and that $H(t_0)$ is positive definite for some $t_0 \in \mathcal{I}$. Then all $H(t)$ are positive definite.*

Proof. For any t_1 there is an ϵ-neighbourhood in which the inertia of $H(t)$ is constant. This is due to the continuity property of the eigenvalues. By the compactness the whole of $\mathcal{I}$ can be covered by a finite number of such neighbourhoods. This, together with the positive definiteness of $H(t_0)$ implies positive definiteness for all $H(t)$. Q.E.D.

Theorem 10.5 *Let $\mathcal{I} \ni t \mapsto A(t)$ be J-Hermitian valued continuous function on a closed real interval $\mathcal{I}$ such that*

1. *$A(t_0)$ has definite spectrum for some $t_0 \in \mathcal{I}$ and*
2. *there are continuous real valued functions $\mathcal{I} \ni t \mapsto f_k(t)$, $k = 1, \ldots, p-1$ such that*

   a.
   $$\sigma_1 < f_1(t_0) < \sigma_2 < \cdots < f_{p-1}(t_0) < \sigma_p$$
   where $\sigma_1 < \sigma_2 < \cdots < \sigma_p$ are the sign groups of $A(t_0)$ and

   b. *$f_k(t) \cap \sigma(A(t)) = \emptyset$ for all k and t.*

Then $A(t)$ has definite spectrum for all t and $s(A(t))$ is constant in t.

Proof. Let σ_1 be, say, J-negative. Set

$$p_t(A(t)) = (A(t) - f_1(t)I)(A(t) - f_2(t)I)(A(t) - f_3(t)I) \cdots$$

then $H(t) = Jp_t(A(t))$ satisfies the conditions of Lemma 10.4 and the statement follows. Q.E.D.

From the proof of Theorem 10.3 it is seen that the definitising polynomial p can be chosen with the degree one less than the number of sign groups of A. In the case of just two such groups p will be linear and we can without loss of generality assume $p(\lambda)$ as $\lambda - \mu$. Then $Jp(A) = JA - \mu J$ and the matrix A (and also the matrix pair JA, J) is called *J-definitisable* or just *definitisable.* Any μ for which $JA - \mu J$ is positive definite is called a *definitising shift.* Of course, completely analogous properties are enjoyed by matrices for which $JA - \mu J$ is negative definite for some μ (just replace A by $-A$).

Theorem 10.6 *Let A be J-Hermitian. The following are equivalent:*

(i) A is definitisable.
(ii) The spectrum of A is definite and the J-positive eigenvalues are larger then the J-negative ones.

In this case the set of all definitising shifts form an open interval whose ends are eigenvalues; this is called the definiteness interval *of A (or, equivalently, of the* definitisable *Hermitian pair JA, J.*

Proof. If $JA - \mu J$ is positive definite, then

$$Ax = \lambda x$$

or, equivalently, $(JA - \mu J)x = (\lambda - \mu)Jx$ implies

$$x^*(JA - \mu J)x = (\lambda - \mu)x^*Jx.$$

Thus, x is J-positive or J-negative according to whether $\lambda > \mu$ or $\lambda < \mu$ (μ itself is not an eigenvalue). Thus $\sigma_-(A) < \mu < \sigma_+(A)$ where $\sigma_\pm(A)$ denote the set of J-positive/J-negative eigenvalues. Conversely, suppose $\sigma_-(A) < \sigma_+(A)$. The matrix A is J, J'-unitarily diagonalisable:

$$A' = U^{-1}AU = \begin{bmatrix} \lambda_1 I_1 & & \\ & \ddots & \\ & & \lambda_p I \end{bmatrix}, \quad J' = U^*JU = \begin{bmatrix} \epsilon_1 I_1 & & \\ & \ddots & \\ & & \epsilon_p I_p \end{bmatrix}$$

where $\lambda_1, \dots, \lambda_p$ are distinct eigenvalues of A and $\epsilon_i \in \{-1, 1\}$. Take any μ between $\sigma_-(A)$ and $\sigma_+(A)$. Then

$$J'A' - \mu J' = \begin{bmatrix} \epsilon_1(\lambda_1 - \mu)I_1 & & \\ & \ddots & \\ & & \epsilon_p(\lambda_p - \mu)I_p \end{bmatrix} \tag{10.2}$$

and this matrix is positive definite because the product $\epsilon_i(\lambda_i - \mu)$ is always positive. Now,

$$J'A' - \mu J' = U^*JUU^{-1}AU - \mu U^*JU = U^*(JA - \mu J)U$$

and $JA - \mu J$ is positive definite as well. The last assertion follows from (10.2). Q.E.D.
The following theorem is a generalisation of the known fact for standard Hermitian matrices: the eigenvalues are 'counted' by means of the inertia. From now on we will denote the eigenvalues of a definitisable A as

$$\lambda^-_{n_-} \le \cdots \le \lambda^-_1 < \lambda^+_1 \le \cdots \le \lambda^+_{n_+}. \tag{10.3}$$

Theorem 10.7 *Let A be definitisable with the eigenvalues as in (10.3) and the definiteness interval $(\lambda^-_1, \lambda^+_1)$. For any $\lambda > \lambda^+_1$, $\lambda \not\in \sigma(A)$ the quantity $\iota_-(JA - \lambda J)$ equals the number of the J-positive eigenvalues less than λ (and similarly for J-negative eigenvalues).*

Proof. Since the definitisability is invariant under any J, J'-unitary similarity and A is J, J'-unitarily diagonalisable, we may assume A and J to be diagonal, i.e.

$$A = \operatorname{diag}(\lambda^+_{n_+}, \dots, \lambda^-_{n_-}), \quad J = \operatorname{diag}(I_{n_+}, I_{n_-}).$$

Then

$$\iota_-(JA - \lambda J) = \iota_-(\operatorname{diag}(\lambda^+_{n_+} - \lambda, \dots, \lambda^+_1 - \lambda, -(\lambda^-_1 - \lambda), \dots, -(\lambda^-_{n_-} - \lambda)))$$

and the assertion follows. Q.E.D.

The following 'local counting property' may be useful in studying general J-Hermitian matrices.

Theorem 10.8 *Let A be J-Hermitian and let λ_0 be a J-positive eigenvalue of multiplicity p. Take $\epsilon > 0$ such that the open interval $\mathcal{I} = (\lambda_0 - \epsilon, \lambda_0 + \epsilon)$ contains no other eigenvalues of A. Then for*

$$\lambda_0 - \epsilon < \lambda_- < \lambda_0 < \lambda_+ < \lambda_0 + \epsilon$$

we have

$$\iota_+(JA - \lambda_- J) = \iota_+(JA - \lambda_+ J) + p.$$

Proof. By virtue of Proposition 10.1 and (10.1) and the fact that

$$\iota(JA' - \lambda J') = \iota(U^*(JA - \lambda J)U) = \iota(JA - \lambda J)$$

(Sylvester!) we may suppose that A, J have already the form

$$A = \begin{bmatrix} \lambda_0 I & 0 \\ 0 & A_2' \end{bmatrix}, \quad J = \begin{bmatrix} I_p & 0 \\ 0 & J_2' \end{bmatrix}, \quad \mathcal{I} \cap \sigma(A_2') = \emptyset.$$

Now

$$JA - \lambda_\pm J = \begin{bmatrix} (\lambda_0 - \lambda_\pm) I_p & 0 \\ 0 & J_2' A_2' - \lambda_\pm J_2' \end{bmatrix}$$

and

$$\iota_+(JA - \lambda_- J) = p + \iota_+(J_2' A_2' - \lambda_- J_2') =$$

$$p + \iota_+(J_2' A_2' - \lambda_+ J_2') = \iota_+(JA - \lambda_+ J).$$

Here we have used two trivial equalities

$$\iota_+(\lambda_0 - \lambda_-) I_p = p, \quad \iota_+(\lambda_0 - \lambda_+) I_p = 0$$

and the less trivial one

$$\iota_+(J_2' A_2' - \lambda_+ J_2') = \iota_+(JA - \lambda_+ J);$$

the latter is due to the fact that for $\lambda \in \mathcal{I}$ the inertia of $J_2' A_2' - \lambda J_2'$ cannot change since this matrix is non-singular for all these λ. Indeed, by the known continuity of the eigenvalues of $J_2' A_2' - \lambda J_2'$ as functions of λ the matrix $J_2' A_2' - \lambda J_2'$ must become singular on the place where it would change its inertia, this is precluded by the assumption $\mathcal{I} \cap \sigma(A_2') = \emptyset$. Q.E.D.

Theorem 10.9 *Let A be J-Hermitian and $\sigma(A)$ negative. Then $JA - \mu J$ is positive definite, if and only if $-JA^{-1} + J/\mu$ is such.*

Proof. Let $JA - \mu J$ be positive definite. Then $\lambda_1^- < \mu < \lambda_1^+$. Obviously, the matrix $-A^{-1}$ is J-Hermitian as well and it has the eigenvalues

$$0 < -1/\lambda_{n_-}^- \le \cdots \le -1/\lambda_1^- < -1/\mu < -1/\lambda_1^+ \le \cdots \le -1/\lambda_{n_+}^+,$$

where $-1/\lambda_1^+ \le \cdots \le -1/\lambda_{n_+}^+$ are J-positive and $-1/\lambda_{n_-}^- \le \cdots \le -1/\lambda_1^-$ J-negative. The converse is proved the same way. Q.E.D.

Exercise 10.10 *Try to weaken the condition $\sigma(A) < 0$ in the preceding theorem. Produce counterexamples.*

We now consider the boundary of the set of definitisable matrices.

Theorem 10.11 *If $JA - \lambda J$ is positive semidefinite and singular then we have the following alternative.*

1. *The matrix A is definitisable and λ lies on the boundary of the definiteness interval; in this case λ is a J-definite eigenvalue.*
2. *The matrix A is not definitisable or, equivalently, there is no $\lambda' \ne \lambda$ for which $JA - \lambda' J$ would be positive semidefinite. In this case all eigenvalues*

greater than λ are J-positive and those smaller than λ are J-negative whereas the eigenvalue λ itself is not J-definite.

Proof. As in the proof of Theorem 10.6 we write $Ax = \lambda' x$, $\lambda' > \lambda$ as

$$(A - \lambda I)x = (\lambda' - \lambda)x,$$

hence

$$x^*(JA - \lambda J)x = (\lambda' - \lambda)x^* Jx,$$

which implies $x^*Jx \geq 0$. Now, $x^*Jx = 0$ is impossible because $x^*(JA - \lambda J)x = 0$ and the assumed semidefiniteness of $JA - \lambda J$ would imply $(JA - \lambda J)x = 0$ i.e. $Ax = \lambda x$. Thus, $x^*Jx > 0$ and similarly for $\lambda' < \lambda$. So, if λ itself is, say, J-positive, then the condition (ii) of Theorem 10.6 holds and the non-void definiteness interval lies left from λ. Finally, suppose that λ is not definite and that there is, say. $\lambda' < \lambda$ for which $JA - \lambda' J$ would be positive semidefinite. Then, as was shown above, λ would be J-positive which is impossible. Q.E.D.

The definitisability of the pair JA, J can be expressed as

$$[(A - \lambda I)x, x] > 0$$

for some λ and all non-vanishing x.

The eigenvalues of definitisable matrices enjoy extremal properties similar to those for standard Hermitian matrices. Consider the functional

$$x \mapsto r(x) = \frac{x^* JAx}{x^* Jx} = \frac{[Ax, x]}{[x, x]},$$

which is obviously real valued and defined on any non-neutral vector x. It will be called *the Rayleigh quotient* of A (or, which is the same, of the pair JA, J).

Theorem 10.12 *Let $A \in \Xi^{n,n}$ be J-Hermitian. Then A is definitisable, if and only if the values*

$$r_+ = \min_{\substack{x \in \Xi^n \\ x^* Jx > 0}} r(x), \quad r_- = \max_{\substack{x \in \Xi^n \\ x^* Jx < 0}} r(x) \tag{10.4}$$

exist and $r_- < r_+$ holds. In this case (r_-, r_+) is the definiteness interval of the pair JA, J. If $J = I$ then, by convention, $r_- = -\infty$.

Proof. If A is definitisable then by Theorem 10.6 and Corollary 10.2 there is a J, J' unitary such that

$$U^{-1}AU = \mathrm{diag}(\lambda^+_{n_+}, \ldots, \lambda^-_{n_-}),$$

$$J' = \begin{bmatrix} I_{n_+} & 0 \\ 0 & -I_{n_-} \end{bmatrix},$$

and the eigenvalues are given by (10.3). Under the substitution $x = Uy$ we obtain

$$r(x) = \frac{\lambda_{n_+}^+ |y_{n_+}^+|^2 + \cdots + \lambda_1^+ |y_1^+|^2 - \lambda_1^- |y_1^-|^2 - \cdots - \lambda_{n_-}^- |y_{n_-}^-|^2}{|y_{n_+}^+|^2 + \cdots + |y_1^+|^2 - |y_1^-|^2 - \cdots - |y_{n_-}^-|^2}.$$

(the components of y are denoted accordingly). By

$$\lambda_{n_+}^+, \ldots, \lambda_2^+ \geq \lambda_1^+, \quad -\lambda_2^-, \ldots, -\lambda_{n_-}^- \geq -\lambda_1^-$$

and by taking, say, $y^* J y > 0$ we have

$$r(x) \geq \lambda_1^+$$

while the equality is obtained on any eigenvector y for the eigenvalue λ_1^+ and nowhere else. Thus, the left equality in (10.4) is proved (and similarly for the other one).

Conversely let $r(x_0) = r_+$. For any vector h and any real ε we have

$$r(x_0 + \varepsilon h) =$$

$$\frac{(x_0 + \varepsilon h)^* JA (x_0 + \varepsilon h)}{(x_0 + \varepsilon h)^* J (x_0 + \varepsilon h)} = \frac{x_0^* JA x_0 + 2\varepsilon \operatorname{Re} x_0^* JAh + \varepsilon^2 h^* JAh}{x_0^* J x_0 (1 + 2\varepsilon \operatorname{Re} \frac{x_0^* Jh}{x_0^* J x_0} + \varepsilon^2 \frac{h^* Jh}{x_0^* J x_0})} =$$

$$r_+ + \frac{2\varepsilon}{x_0^* J x_0} \operatorname{Re} \left(x_0^* JAh - \frac{x_0^* JA x_0 x_0^* Jh}{x_0^* J x_0} \right) +$$

$$\frac{\varepsilon^2}{x_0^* J x_0} \left[h^* (JA - r_+ J) h - 4 \frac{\operatorname{Re} x_0^* JAh \operatorname{Re} x_0^* Jh}{x_0^* J x_0} + 4 (\operatorname{Re} \frac{x_0^* Jh}{x_0^* J x_0})^2 x_0^* JA x_0 \right] +$$

$$\mathcal{O}(\varepsilon^3).$$

This is a geometric series in ε which has a finite convergence radius (depending on h). Since this has a minimum at $\varepsilon = 0$, the coefficient at ε vanishes i.e.

$$\operatorname{Re}(x_0^* (JA - r_+ J) h) = 0$$

and since h is arbitrary we have $JA x_0 = r_+ J x_0$. Using this we obtain

$$r(x_0 + \varepsilon h) = \frac{\varepsilon^2}{x_0^* J x_0} h^* (JA - r_+ J) h + \mathcal{O}(\varepsilon^3).$$

From this and the fact that x_0 is the minimum point it follows $h^* (JA - r_+ J) h \geq 0$ and this is the positive semidefiniteness of $JA - r_+ J$. The same for $JA - r_- J$ is obtained analogously. By Theorem 10.11 this implies that A

is definitisable and that (r_-, r_+) is its definiteness interval. Q.E.D.

In contrast to the standard Hermitian case the 'outer' boundary of the spectrum of a definitisable A is no extremum for the Rayleigh quotient. As an example take

$$A = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

and

$$x = \begin{bmatrix} \cosh\phi \\ \sinh\phi \end{bmatrix},$$

then $x^*Jx = 1$ and

$$r(x) = 1 + \cosh^2\phi$$

which is not bounded from above. However, all eigenvalues can be obtained by minimax formulae, similar to those for the standard Hermitian case. We will now derive these formulae. We begin with a technical result, which is nonetheless of independent interest.

Lemma 10.13 *(The interlacing property) Let the matrix A be J-Hermitian and definitisable and partitioned as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \text{ and, accordingly, } J = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}. \tag{10.5}$$

Here A_{11} is square of order m. Then A_{11} is J_1-Hermitian and definitisable and its definiteness interval contains that of A. Denote by (10.3) the eigenvalues of A and by

$$\mu^-_{m_-} \le \cdots \le \mu^-_1 < \mu^+_1 \le \cdots \le \mu^+_{m_+}, \quad m_+ + m_- = m,$$

those of A_{11}. Then

$$\lambda^+_k \le \mu^+_k \le \lambda^+_{k+n_+-m_+}$$

$$\lambda^-_k \ge \mu^-_k \ge \lambda^-_{k+n_--m_-}$$

(here an inequality is understood as void whenever an index exceeds its range).

Proof. If $JA - \mu J$ is Hermitian positive definite then so is the submatrix $J_1A_{11} - \mu J_1$. By the J-Hermitian property of A we have $A_{21} = J_2A_{12}^*J_1$ hence

$$JA - \mu J = \begin{bmatrix} J_1A_{11} - \mu J_1 & J_1A_{12} \\ A_{12}^*J_1 & J_2A_{22} - \mu J_2 \end{bmatrix}$$

$$= Z(\mu) \begin{bmatrix} J_1A_{11} - \mu J_1 & 0 \\ 0 & W(\mu) \end{bmatrix} Z(\mu)^*$$

with

$$Z(\mu) = \begin{bmatrix} I_m & 0 \\ A_{12}^*J_1(J_1A_{11} - \mu J_1)^{-1} & I_{n-m} \end{bmatrix},$$

and

$$W(\mu) = J_2 A_{22} - \mu J_2 - A_{12}^* J_1 (J_1 A_{11} - \mu J_1)^{-1} J_1 A_{12}.$$

By the Sylvester inertia theorem,

$$\iota_\pm(JA - \mu J) = \iota_\pm(J_1 A_{11} - \mu J_1) + \iota_\pm(W(\mu)),$$

hence

$$\iota_-(J_1 A_{11} - \mu J_1) \le \iota_-(JA - \mu J) \le \iota_-(J_1 A_{11} - \mu J_1) - n - m. \tag{10.6}$$

Assume now $\mu_k^+ < \lambda_k^+$ for some k. Then there is a μ such that $J_1 A_{11} - \mu J_1$ is non-singular and $\mu_k^+ < \mu < \lambda_k^+$. By Theorem 10.7 we would have

$$\iota_-(J_1 A_{11} - \mu J_1) \ge k, \quad \iota_-(JA - \mu J) < k$$

which contradicts the first inequality in (10.6). Similarly, $\mu_k^+ > \mu > \lambda_{k+n_+-m_+}^+$ would imply

$$\iota_-(J_1 A_{11} - \mu J_1) \ge k, \quad \iota_-(JA - \mu J) < k$$

which contradicts the second inequality in (10.6). Q.E.D.

Theorem 10.14 *For a definitisable J-Hermitian A the following minimax formulae hold:*

$$\lambda_k^\pm = \min_{S_k^\pm} \max_{\substack{x \in S_k^\pm \\ x^* J x \ne 0}} \frac{x^* J A x}{x^* J x} = \min_{S_k^\pm} \max_{\substack{x \in S_k^\pm \\ [x,x] \ne 0}} \frac{[Ax, x]}{[x, x]} \tag{10.7}$$

where $S_k^\pm \subseteq \Xi^n$ is any k-dimensional J-positive/J-negative subspace.

Proof. Let S_k^+ be given and let $u_1, \dots, u_n$ be a J-orthonormal basis of Ξ^n such that $u_1, \dots, u_k$ span S_k^+. Set

$$U^+ = \begin{bmatrix} u_1 \cdots u_k \end{bmatrix}, \quad U = \begin{bmatrix} u_1 \cdots u_n \end{bmatrix} = \begin{bmatrix} U^+ \; U' \end{bmatrix}.$$

Then U is J,J'-unitary with

$$J' = \begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & J_2 \end{bmatrix}, \quad J_2 = \operatorname{diag}(\pm 1)$$

and

$$A' = U^{-1} A U = \begin{bmatrix} A'_{11} & A'_{12} \\ A'_{21} & A'_{22} \end{bmatrix}$$

is J'-Hermitian:

$$A'_{11} = A'^*_{11}, \quad A'_{21} = J_2 A'^*_{12}, \quad A'_{22} = J_2 A'^*_{22} J_2.$$

Moreover, A and A' have the same eigenvalues and the same definiteness interval. By Lemma 10.13 we have the interlacing of the corresponding eigenvalues

$$\lambda_k^+ \le \mu_k^+ \le \lambda_{k+n_+-m_+}^+, \quad k = 1, \dots m_+$$

(the eigenvalues μ_k^- are lacking). The subspace S_k^+ consists of the vectors

$$U \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad y \in \Xi^k$$

thus for arbitrary $x \in S_k^+$ we have

$$\frac{x^* J A x}{x^* J x} = \frac{y^* A'_{11} y}{y^* y}.$$

Since A'_{11} is (standard) Hermitian,

$$\mu_k^+ = \max_{y \ne 0} \frac{y^* A'_{11} y}{y^* y} = \max_{\substack{x \in S_k^+ \\ x \ne 0}} \frac{x^* J A x}{x^* J x} \ge \lambda_k^+$$

and since S_k^+ is arbitrary it follows

$$\inf_{S_k^+} \max_{\substack{x \in S_k^+ \\ x \ne 0}} \frac{x^* J A x}{x^* J x} \ge \lambda_k^+.$$

Now choose S_k^+ as the linear span of J-orthonormal eigenvectors $v_1, \dots, v_k$, belonging to the eigenvalues $\lambda_1^+, \dots, \lambda_k^+$ of A. Then any $x \in S_k^+$ is given as

$$x = \alpha_1 v_1 + \dots + \alpha_k v_k$$

and, with $\alpha = \begin{bmatrix} \alpha_1 & \cdots & \alpha_k \end{bmatrix}^T$,

$$\max_{\substack{x \in S_k^+ \\ x \ne 0}} \frac{x^* J A x}{x^* J x} = \max_{\alpha^* \alpha = 1} \sum_{j=1}^{k} |\alpha_j|^2 \lambda_j^+ = \lambda_k^+.$$

This proves the assertion for J-positive eigenvalues (the other case is analogous). Q.E.D.

Remark 10.15 In the special case $J = I$ Theorem 10.14 gives the known minimax formulae for a Hermitian matrix A of order n:

$$\lambda_k = \min_{S_k} \max_{\substack{x \in S_k \\ x \ne 0}} \frac{x^* A x}{x^* x},$$

where S_k is any k-dimensional subspace and λ_k are non-decreasingly ordered eigenvalues of A. Now we can prove the general formula (2.8): by setting $M = L_2 L_2^*$ and $A = L_2^{-1} K L_2^{-*}$ we obtain

$$\max_{\substack{x\in S_k\\ x\neq 0}} \frac{x^*Ax}{x^*x} = \max_{\substack{y\in L_2^{-*}S_k\\ y\neq 0}} \frac{y^*Ky}{y^*My}$$

and $L_2^{-*}S_k$ again varies over the set of all k-dimensional subspaces. This proves the 'min max' part of (2.8), the other part is obtained by considering the pair $-K, M$. Note, however, that for a general J-Hermitian matrix no 'max min' variant in the formulae (10.7) exists.

The following extremal property could be useful for computational purposes.

Theorem 10.16 *Let A be J-Hermitian. Consider the function*

$$X \mapsto \mathrm{Tr}(X^*JAX) \tag{10.8}$$

defined on the set of all J, J_1-isometries $X \in \Xi^{n,m}$ for a fixed symmetry

$$J_1 = \mathrm{diag}(I_{m_+}, -I_{m_-}), \quad m_\pm \le n_\pm.$$

If A is definitisable with the eigenvalues (10.3) then the function (10.8) takes its minimum on any X of the form

$$X = \begin{bmatrix} X^+ & X^- \end{bmatrix}$$

where $\mathcal{R}(X^\pm)$ is the subspace spanned by $m_\pm$ J-orthonormal eigenvectors for the eigenvalues $\lambda_1^\pm, \ldots, \lambda_{m_\pm}^\pm$ of A, respectively — and nowhere else.

Proof. Without loss of generality we may suppose J to be of the form

$$J = \begin{bmatrix} I_{n_+} & 0 \\ 0 & -I_{n_-} \end{bmatrix}.$$

Indeed, the transition from a general J to the one above is made by unitary similarity as described in Remark 5.1. The trace function (10.8) is invariant under these transformations. Also, by virtue of the Sylvester inertia theorem, applied to $X^*JX = J_1$, we have

$$m_+ \le n_+, \quad m_- \le n_-.$$

First we prove the theorem for the case $m_\pm = n_\pm$ i.e. the matrices X are square and therefore J-unitary.

Since A is definitisable we may perform a J-unitary diagonalisation

$$U^{-1}AU = \Lambda, \quad \Lambda = \mathrm{diag}(\Lambda_+, \Lambda_-), \quad \Lambda_\pm = \mathrm{diag}(\lambda_1^\pm, \ldots \lambda_{n_\pm}^\pm), \quad U^*JU = J.$$

Now

$$\mathrm{Tr}(X^*JAX) = \mathrm{Tr}(X^*JU\Lambda U^{-1}X) = \mathrm{Tr}(X^*U^{-*}J\Lambda U^{-1}X) = \mathrm{Tr}(Y^*J\Lambda Y)$$

where $Y = U^{-1}X$ again varies over the set of all J-unitaries. By using the representation (6.2) we obtain

$$\begin{aligned}&\mathrm{Tr}X^*JAX = \mathrm{Tr}Y^*J\Lambda Y = \mathrm{Tr}H\,(W)\,J\Lambda H\,(W) =\\ &\mathrm{Tr}(\sqrt{I+WW^*}\Lambda_+\sqrt{I+WW^*} - W\Lambda_- W^*)\\ &+\mathrm{Tr}(W^*\Lambda_+ W - \sqrt{I+W^*W}\Lambda_-\sqrt{I+W^*W})\\ &= t_0 + 2(\mathrm{Tr}WW^*\Lambda_+ - \mathrm{Tr}W^*W\Lambda_-)\\ &= t_0 + 2[\mathrm{Tr}WW^*(\Lambda_+ - \mu I) + \mathrm{Tr}W^*W(\mu I - \Lambda_-)],\end{aligned}$$

where μ is a definitising shift and

$$t_0 = \mathrm{Tr}\Lambda_+ - \mathrm{Tr}\Lambda_-.$$

Since $\Lambda_+ - \mu I$ and $\mu I - \Lambda_-$ are positive definite it follows

$$\mathrm{Tr}(X^*JAX) = \mathrm{Tr}Y^*J\Lambda Y$$

$$= t_0 + 2\mathrm{Tr}\left(W^*\left(\Lambda_+ - \mu I\right)W\right) + 2\mathrm{Tr}\left(W\left(\mu I - \Lambda_-\right)W^*\right) \geq t_0, \qquad (10.9)$$

moreover, if this inequality turns to equality then $W = 0$, hence $H(W) = I$ and the corresponding X reads

$$X = U\,\mathrm{diag}(V_1, V_2),$$

V_1, V_2 from (6.2). This is exactly the form of the minimising X in the statement of the theorem.

We now turn to $\mathrm{Tr}X^*JAX$ with a non-square X. Take $C = (X\ \bar{X})$ as a completion of any J, J_1-isometry X such that the columns of C form a J-orthonormal basis, that is,

$$C^*JC = \mathrm{diag}(J_1, J_2) = J'$$

is diagonal (note that the columns of X are J-orthonormal). Then the matrix $A_1 = C^{-1}AC$ is J'-Hermitian and we have

$$A_1 = C^*JAC = \begin{bmatrix} J_1X^*JAX & J_1X^*JA\bar{X} \\ J_2\bar{X}^*JAX & J_2\bar{X}^*JA\bar{X} \end{bmatrix},$$

$$J'A_1 = C^*JAC = \begin{bmatrix} X^*JAX & X^*JA\bar{X} \\ \bar{X}^*JAX & \bar{X}^*JA\bar{X} \end{bmatrix}$$

and the eigenvalues of A_1 are given by (10.3). Then by Lemma 10.13 the eigenvalues $\mu_k^{\pm}$ of the J_1-Hermitian matrix J_1X^*JAX satisfy

$$\lambda_k^+ \leq \mu_k^+, \quad \lambda_k^- \geq \mu_k^-.$$

Using this and applying the first part of the proof for the matrix J_1X^*JAX we obtain

$$\mathrm{Tr} X^* J A X \geq \sum_{i=1}^{m_+} \mu_i^+ - \sum_{j=1}^{m_-} \mu_j^- \geq \sum_{i=1}^{m_+} \lambda_i^+ - \sum_{j=1}^{m_-} \lambda_j^- .$$

This lower bound is attained, if X is chosen as in the statement of the theorem, that is, if

$$A X^{\pm} = X^{\pm} \Lambda^{\pm}, \quad \Lambda^{\pm} = \mathrm{diag}(\lambda_1^{\pm}, \dots, \lambda_{m_{\pm}}^{\pm}).$$

It remains to determine the set of all minimisers. If X is any minimiser then we may apply the formalism of Lagrange with the Lagrange function

$$\mathcal{L} = \mathrm{Tr}(X^* J A X) - \mathrm{Tr}(\Gamma(X^* J X - J_1))$$

where the Lagrange multipliers are contained in the Hermitian matrix Γ of order m. Their number equals the number of independent equations in the isometry constraint $X^* J X = J_1$. By setting the differential of $\mathcal{L}$ to zero we obtain

$$A X = X \Gamma$$

and by premultiplying by $X^* J$,

$$X^* J A X = J_1 \Gamma.$$

Hence Γ commutes with J_1:

$$\Gamma = \mathrm{diag}(\Gamma^+, \Gamma^-).$$

This is the form of the minimiser, stated in the theorem. Q.E.D.

When performing diagonalisation the question of the condition number of the similarity matrix naturally arises. We have

Proposition 10.17 *Let A be J-Hermitian and have a definite spectrum. Then all J, J'- unitaries that diagonalise it with any diagonal J' have the same condition number.*

Proof. Suppose

$$\begin{aligned} U_1^{-1} A U_1 = \Lambda_1, \quad U_1^* J U_1 = J_1 \\ U_2^{-1} A U_2 = \Lambda_2, \quad U_2^* J U_2 = J_2 \end{aligned}$$

where Λ_1, Λ_2 are diagonal and J_1, J_2 are diagonal matrices of signs. The J-definiteness means that there exist permutation matrices Π_1, Π_2 such that

$$\Pi_1^T J_1 \Pi_1 = \Pi_2^T J_2 \Pi_2 = \begin{bmatrix} \epsilon_1 I_{n_1} & & & \\ & \epsilon_2 I_{n_2} & & \\ & & \ddots & \\ & & & \epsilon_p I_{n_p} \end{bmatrix} =: J^0,$$

$$\Pi_1^T \Lambda_1 \Pi_1 = \Pi_2^T \Lambda_2 \Pi_2 = \begin{bmatrix} \lambda_1 I_{n_1} & & & \\ & \lambda_2 I_{n_2} & & \\ & & \ddots & \\ & & & \lambda_p I_{n_p} \end{bmatrix} =: \Lambda,$$

here $\epsilon_i \in \{-1, 1\}$ are the signs of the corresponding subspaces and $\lambda_1, \ldots, \lambda_p$ are the distinct eigenvalues of A with the multiplicities $n_1, \ldots, n_p$. Then both $\tilde{U}_1 = U_1 \Pi_1$ and $\tilde{U}_2 = U_2 \Pi_2$ are J, J^0-unitary and

$$\tilde{U}_1^{-1} A \tilde{U}_1 = \Lambda = \tilde{U}_2^{-1} A \tilde{U}_2 \tag{10.10}$$

or, equivalently,

$$\Lambda V = V \Lambda$$

where the matrix $V = \tilde{U}_1^{-1} \tilde{U}_2$ is J^0-unitary. Since the eigenvalues $\lambda_1, \ldots, \lambda_p$ are distinct, (10.10) implies

$$J^0 V = V J^0$$

that is, V is unitary and

$$U_2 = \tilde{U}_2 \Pi_2^T = \tilde{U}_1 V \Pi_2^T = U_1 \Pi_1 V \Pi_2^T$$

where the matrix $\Pi_1 V \Pi_2^T$ is unitary, so U_2 and U_1 have the same condition numbers. Q.E.D.

Note that the previous theorem applies not only to the standard, spectral norm but to any unitarily invariant matrix norm like e.g. the Euclidian norm.

Exercise 10.18 *Show that the 'if' part of the Theorem 10.12 remains valid, if the symbols min/max in (10.4) are substituted by inf/sup.*

Exercise 10.19 *Try to estimate the condition of the matrix X in (10.9), if the difference* $\mathrm{Tr}(JA) - t_0$ *is known.*

Chapter 11
General Hermitian matrix pairs

Here we briefly overview the general eigenvalue problem $Sx = \lambda Tx$ with two Hermitian matrices S, T and show how to reduce it to the case of a single J-Hermitian matrix A.

The important property, characterised by Theorem 10.14, was expressed more naturally in terms of the Hermitian matrix pair JA, J, than in terms of the single J-Hermitian matrix A. In fact, the eigenvalue problem $Ax = \lambda x$ can be equivalently written as $JAx = \lambda Jx$. More generally, we can consider the eigenvalue problem

$$Sx = \lambda Tx \tag{11.1}$$

where S and T are Hermitian matrices and the determinant of $S - \lambda T$ does not identically vanish. Such pairs are, in fact, essentially covered by our theory. Here we outline the main ideas.

If T is non-singular, then we may apply the decomposition (5.8) to the matrix T, replace there G by $G|\boldsymbol{\alpha}|^{1/2}$ and set $J = \mathrm{sign}(\boldsymbol{\alpha})$, then

$$T = GJG^* \tag{11.2}$$

where G is non-singular and $J = \mathrm{diag}(\pm 1)$. By setting $y = G^* x$, (11.1) is equivalent to

$$Ay = \lambda y, \quad A = JG^{-1}SG^{-*}, \tag{11.3}$$

where A is J-Hermitian. Also,

$$\det(A - \lambda I) = \frac{\det(S - \lambda T)}{\det T}.$$

To any invariant subspace relation

$$AY = Y\Lambda$$

with Y injective, there corresponds

$$SX = TX\Lambda, \quad Y = G^*X$$

and vice versa. The accompanying indefinite scalar product is expressed as

$$y^*Jy' = x^*Tx'$$

with $y = G^*x$, $y' = G^*x'$. Thus, all properties of J-Hermitian matrices studied in the two previous chapters can be appropriately translated into the language of matrix pairs S, T with non-singular T and vice versa. The non-singularity of T will be tacitly assumed in the following. (If T is singular but there is a real μ such that $T - \mu S$ is non-singular then everything said above can be done for the pair S, $T - \mu S$. Now, the equations

$$Sx = \lambda(T - \mu S)x, \quad Sx = \frac{\lambda}{1 - \lambda\mu}Tx$$

are equivalent. Thus the pair S, $T - \mu S$ has the same eigenvectors as S, T while the eigenvalues λ of the former pair are transformed into $\lambda/(1 - \lambda\mu)$ for the latter one.)

The terms 'definite eigenvalues' and 'definitisability' carry over in a natural way to the general pair S, T by substituting the matrix JA for S and the symmetry J for T. Theorems 10.6 - 10.16 can be readily formulated and proved in terms of the matrix pair S, T. This we leave to the reader and, as an example, prove the following

Theorem 11.1 *Let S, T be definitisable and $\iota(T) = (n_+, 0, n_-)$. Then there is a Ψ such that*

$$\Psi^*T\Psi = J = \mathrm{diag}(I_{n_+}, -I_{n_-}),$$
$$\Psi^*S\Psi = \mathrm{diag}(\lambda^+_{n_+}, \ldots, \lambda^+_1, -\lambda^-_1, \ldots, -\lambda^-_{n_-})$$

with $\lambda^+_{n_+} \geq \cdots \geq \lambda^+_1 \geq \lambda^-_1 \geq \cdots \geq \lambda^-_{n_-}$. Moreover, for $m_\pm \leq n_\pm$ and $J_1 = \mathrm{diag}(I_{m_+}, -I_{m_-})$ the function

$$\Phi \mapsto \mathrm{Tr}(\Phi^*S\Phi),$$

defined on the set of all Φ with

$$\Phi^*T\Phi = J_1 \tag{11.4}$$

takes its minimum $\sum_{k=1}^{n_+} \lambda^+_k - \sum_{k=1}^{n_-} \lambda^-_k$ on any

$$\Phi = \begin{bmatrix} \Phi_+ \Phi_- \end{bmatrix}$$

where $\mathcal{R}(\Phi_\pm)$ is spanned by J-orthonormal eigenvectors belonging to the eigenvalues $\lambda^\pm_1, \ldots, \lambda^\pm_{n_\pm}$ — and nowhere else.

Proof. In (11.2) we may take J with diagonals ordered as above. Then with $A = JG^{-1}SG^{-*}$ for any μ we have

$$JA - \mu J = \Psi^*(S - \mu T)\Psi$$

so by the non-singularity of Ψ the definitisability of the pair S, T is equivalent to that of JA, J that is, of the J-Hermitian matrix A. By Corollary 10.2 there is a J-unitary U such that

$$U^{-1}AU = \mathrm{diag}(\lambda^+_{n_+}, \dots, \lambda^+_1, \lambda^-_1, \dots, \lambda^-_{n_-}).$$

By setting $\Psi = G^{-*}U$ we have the diagonalisation

$$\Psi^*S\Psi = U^*JAU = JU^{-1}AU = \mathrm{diag}(\lambda^+_{n_+}, \dots, \lambda^+_1, -\lambda^-_1, \dots, -\lambda_{n_-}).$$

Now set $X = G^*\Phi$, then the 'T, J-isometry condition' (11.4) is equivalent to $X^*JX = J_1$ and we are in the conditions of Theorem 10.16. Thus, all assertions of our theorem follow immediately from those of Theorem 10.16. Q.E.D.

Remark 11.2 A definitisable pair S, T with T non-singular, can be diagonalised by applying the decomposition (2.1) to $K = T$ and $M = S - \mu T$, μ a definitising shift:

$$\Phi^*(S - \mu T)\Phi = I, \quad \Phi^*T\Phi = \mathrm{diag}(\alpha_1, \dots, \alpha_n).$$

Then $\Psi = \Phi\,\mathrm{diag}(|\alpha_1|^{-1/2}, \dots, |\alpha_n|^{-1/2})$ satisfies $\Psi^*T\Psi = J' = \mathrm{diag}(\mathrm{sign}(\alpha_j))$ and

$$\Psi^*S\Psi = \Psi^*(S - \mu T + \mu T)\Psi = J'\,\mathrm{diag}(\frac{1}{\alpha_1} + \mu, \dots, \frac{1}{\alpha_n} + \mu).$$

The desired order of the signs in J' can be obtained by appropriately permuting the columns of Ψ.

The system (1.1) can also be represented as follows. We set $x_1 = x$, $x_2 = \dot{x}$, then (1.1) goes over into

$$\frac{d}{dt}T\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = S\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -f(t) \end{bmatrix} \tag{11.5}$$

with

$$T = \begin{bmatrix} K & 0 \\ 0 & -M \end{bmatrix}, \quad S = \begin{bmatrix} 0 & K \\ K & C \end{bmatrix} \tag{11.6}$$

where the matrix T is non-singular, if both K and M are such.

The representations (3.2), (3.3) and (11.5), (11.6) are connected by the formulae (11.2), (11.3), where

$$G = \begin{bmatrix} L_1 & 0 \\ 0 & L_2 \end{bmatrix}.$$

The choice of the representation is a matter of convenience. A reason for choosing the 'normalised representation' (3.2), (3.3) is the physically natural phase space with its energy norm. Another reason is that here we have to do with a single matrix A while J is usually stored as a sequence of signs which may have computational advantages when dealing with dense matrices. Also the facts about condition numbers are most comfortably expressed in the normalised representation.

On the other hand, if we have to deal with large sparse matrices S, T, then the sparsity may be lost after the transition to JA, J and then the 'non-normalised' representation (11.5), (11.6) is more natural in numerical computations.

Chapter 12
Spectral decomposition of a general J-Hermitian matrix

A general J-Hermitian matrix A may not have a basis of eigenvectors. In this chapter we describe the reduction to a block-diagonal form by a similarity transformation $S^{-1}AS$. We pay particular attention to the problem of the condition number of the transformation matrix S which is a key quantity in any numerical manipulation.

The formal way to solve the initial value problem

$$\dot{y} = Ay, \quad y(0) = y_0$$

is to diagonalise A:[1]

$$S^{-1}AS = \operatorname{diag}(\lambda_1, \dots, \lambda_n). \tag{12.1}$$

Then

$$S^{-1}e^{At}S = \operatorname{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t}).$$

Set

$$S = \begin{bmatrix} s_1 \cdots s_n \end{bmatrix}.$$

Then the solution is obtained in two steps:

- compute y_0' from the linear system

$$Sy_0' = y_0 \tag{12.2}$$

- set

$$y = e^{\lambda_1 t}y_{0,1}'s_1 + \cdots + e^{\lambda_n t}y_{0,n}'s_n. \tag{12.3}$$

This method is not viable, if A cannot be diagonalised or if the condition number of the matrix S is too high. The latter case will typically occur when the matrix A is close to a non-diagonalisable one. High condition of S will

[1] In this chapter we will consider all matrices as complex.

spoil the accuracy of the solution of the linear system (12.2) which is a vital step in computing the solution (12.3).

Things do not look better, if we assume A to be J-Hermitian — with the exception of definite spectra as shown in Chapter 10. An example of a non-diagonalisable J-Hermitian A was produced in (9.5) with $d = 2\omega$ (critical damping).

Instead of the 'full diagonalisation' (12.1) we may seek any reduction to a block-diagonal form

$$S^{-1}AS = \mathrm{diag}(A'_1, \ldots, A'_p).$$

so the exponential solution is split into

$$S^{-1}e^{At}S = \mathrm{diag}(e^{A'_1 t}, \ldots, e^{A'_p t}).$$

A practical value of block-diagonalisation lies in the mere fact that reducing the dimension makes computation easier.

So the ideal would be to obtain as small sized diagonal blocks as possible while keeping the condition number of S reasonably low.

What we will do here is to present the spectral reduction, that is, the reduction in which the diagonal blocks have disjoint spectra. We will pay particular attention to matrices A that are J-Hermitian.

The approach we will take will be the one of complex contour of analytic functions of the complex variable λ. We will freely use the general properties of matrix-valued analytic functions which are completely analogous to those in the scalar case. One of several equivalent definitions of analyticity is the analyticity of the matrix elements.[2] Our present considerations will be based on one such function, *the resolvent* :

$$R(\lambda) = (\lambda I - A)^{-1},$$

which is an analytic, more precisely, rational function in λ as revealed by the Cramer-rule formula

$$R(\lambda) = \frac{A_{\mathrm{adj}}(\lambda)}{\det(\lambda I - A)}$$

where the elements of $A_{\mathrm{adj}}(\lambda)$ are some polynomials in λ.

The fundamental formula of the "analytic functional calculus" is

$$f(A) = \frac{1}{2\pi i}\int_\Gamma f(\lambda)(\lambda I - A)^{-1} d\lambda \tag{12.4}$$

[2] The non-commutativity of matrix multiplication carries only minor, mostly obvious, changes in the standard formulae of the calculus: $(A(\lambda)B(\lambda))' = A'(\lambda)B(\lambda) + A(\lambda)B'(\lambda)$, $(A(\lambda)^{-1})' = -A(\lambda)^{-1}A'(\lambda)A(\lambda)^{-1}$ and the like.

where $f(\lambda)$ is any analytic function in a neighbourhood $\mathcal{O}$ of $\sigma(A)$ ($\mathcal{O}$ need not be connected) and $\varGamma \subseteq \mathcal{O}$ is any contour surrounding $\sigma(A)$. The map $f \mapsto f(A)$ is continuous and has the properties

$$(\alpha f_1 + \beta f_2)(A) = \alpha f_1(A) + \beta f_2(A) \tag{12.5}$$

$$(f_1 \cdot f_2)(A) = f_1(A) f_2(A) \quad (\cdot \text{ is the pointwise multiplication of functions}) \tag{12.6}$$
$$(f_1 \circ f_2)(A) = f_1(f_2(A)) \quad (\circ \text{ is the composition of functions})$$

$$\mathbf{1}(A) = I, \text{ where } \mathbf{1}(\lambda) = 1 \tag{12.7}$$

$$\boldsymbol{\lambda}(A) = A, \text{ where } \boldsymbol{\lambda}(\lambda) = \lambda \tag{12.8}$$
$$\sigma(f(A)) = f(\sigma(A)).$$

All these properties are readily derived by using the resolvent equation

$$R(\lambda) - R(\mu) = (\lambda - \mu) R(\lambda) R(\mu) \tag{12.9}$$

as well as the resulting power series

$$R(\lambda) = \sum_{k=0}^{\infty} \frac{A^k}{\lambda^{k+1}}, \tag{12.10}$$

valid for $|\lambda| > \|A\|$.
We just sketch some of the proofs: (12.5) follows immediately from (12.4), for (12.6) use (12.9) where in computing

$$f_1(A) f_2(A) = \frac{1}{2\pi i} \int_{\varGamma_1} f_1(\lambda)(\lambda I - A)^{-1} d\lambda \int_{\varGamma_2} f_1(\mu)(\mu I - A)^{-1} d\mu.$$

the contour $\varGamma_2$ is chosen so that it is contained in the $\varGamma_1$-bounded neighbourhood of $\sigma(A)$.

(12.7) follows from (12.10) whereas (12.8) follows from (12.7) and the identity
$$\lambda(\lambda I - A)^{-1} = (\lambda I - A)^{-1} - I.$$

The properties (12.5), (12.6), (12.7), (12.8) imply that the definition (12.4) coincides with other common definitions of matrix functions like matrix polynomials, rational functions and convergent power series, provided that the convergence disk contains the whole spectrum. For example,

$$\alpha_0 I + \cdots + \alpha_p A^p = \frac{1}{2\pi i}\int_\Gamma (\alpha_0 + \cdots + \alpha_p \lambda^p)(\lambda I - A)^{-1} d\lambda, \tag{12.11}$$

$$(\mu I - A)^{-1} = \frac{1}{2\pi i}\int_\Gamma \frac{1}{\mu - \lambda}(\lambda I - A)^{-1} d\lambda, \tag{12.12}$$

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!} = \frac{1}{2\pi i}\int_\Gamma e^\lambda (\lambda I - A)^{-1} d\lambda, \tag{12.13}$$

$$A^\alpha = \frac{1}{2\pi i}\int_\Gamma \lambda^\alpha (\lambda I - A)^{-1} d\lambda = \sum_k \binom{\alpha}{k}(A - I)^k. \tag{12.14}$$

Here A^α depends on the choice of the contour Γ and the last equality holds for $\|I - A\| < 1$ (it describes the branch obtained as the analytic continuation starting from $A = I, A^\alpha = I$). We will always have to work with fractional powers of matrices without non-positive real eigenvalues. So, by default the contour Γ is chosen so that it does not intersect the non-positive real axis.

Exercise 12.1 *Find conditions for the validity of the formula*

$$A^\alpha A^\beta = A^{\alpha+\beta}.$$

Another common expression for $f(A)$ is obtained by starting from the obvious property

$$f(S^{-1}AS) = \frac{1}{2\pi i}\int_\Gamma f(\lambda) S^{-1}(\lambda I - A)^{-1} S d\lambda = S^{-1} f(A) S. \tag{12.15}$$

Now, if S diagonalises A i.e. $S^{-1}AS = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ then (12.15) yields

$$f(A) = S \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n)) S^{-1}.$$

Similarly, if S block-diagonalises A i.e. $S^{-1}AS = \operatorname{diag}(A_1', \ldots, A_p')$ then (12.15) yields

$$f(A) = S \operatorname{diag}(f(A_1'), \ldots, f(A_p')) S^{-1}.$$

Exercise 12.2 *Show that the map $A \mapsto f(A)$ is continuous. Hint: for $\hat{A} = A + \delta A$ show that the norm of the series*

$$f(\hat{A}) - f(A) = \sum_{k=1}^{\infty} \frac{1}{2\pi i}\int_\Gamma f(\lambda)(\lambda I - A)^{-1}\left(\delta A(\lambda I - A)^{-1}\right)^k d\lambda \tag{12.16}$$

can be made arbitrarily small, if $\|\delta A\|$ is small enough.

Other fundamental functions f are those which have their values in $\{0, 1\}$, the corresponding $P = f(A)$ are obviously projections. Obviously, any such projection (called *spectral projection*) is given by

$$P = P_\sigma = \frac{1}{2\pi i}\int_\Gamma (\lambda I - A)^{-1} d\lambda \tag{12.17}$$

where Γ separates the subset $\sigma \subseteq \sigma(A)$ from the rest of $\sigma(A)$. Also obvious are the relations

$$P_\sigma P_{\sigma'} = P_{\sigma'} P_\sigma = 0, \quad P_\sigma + P_{\sigma'} = P_{\sigma' \cup \sigma}$$

whenever $\sigma \cap \sigma' = \emptyset$.

Thus, any partition $\sigma_1, \dots, \sigma_p$ of $\sigma(A)$ gives rise to a decomposition of the identity

$$P_{\sigma_1}, \dots, P_{\sigma_p},$$

commuting with A. Taking a matrix S whose columns contain the bases of $\mathcal{R}(P_{\sigma_1}), \dots, \mathcal{R}(P_{\sigma_p})$ we obtain block-diagonal matrices

$$A' = S^{-1}AS = \begin{bmatrix} A'_1 & & \\ & \ddots & \\ & & A'_p \end{bmatrix}, \quad \sigma(A'_j) = \sigma_j, \tag{12.18}$$

$$S^{-1}P_{\sigma_j}S = \begin{bmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & I_{n_j} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix} = P_j^0. \tag{12.19}$$

Here n_j is the dimension of the space $\mathcal{R}(P_{\sigma_j})$. If σ_k is a single point: $\sigma_k = \{\lambda_k\}$ then the space $\mathcal{R}(P_{\sigma_k})$ is called *the root space* belonging to $\lambda \in \sigma(A)$ and is denoted by $\mathcal{E}_\lambda$. If all $\mathcal{R}(P_{\sigma_j})$ are root spaces then we call (12.18) the *spectral decomposition* of A.

Here we see the advantage of the contour integrals in decomposing an arbitrary matrix. The decomposition (12.18) and in particular the spectral decomposition *are stable under small perturbations of the matrix A*. Indeed, the formula (12.16) can be applied to the projections in (12.17): they change continuously, if A changes continuously. The same is then the case with the subspaces onto which they project.

Exercise 12.3 *Show that*

$$\mathrm{Tr}P = \mathrm{Tr}\frac{1}{2\pi i}\int_\Gamma (\lambda I - A)^{-1} d\lambda \tag{12.20}$$

equals the number of the eigenvalues of A within Γ together with their multiplicities whereas

$$\hat{\lambda} = \mathrm{Tr}\frac{1}{2\pi i}\int_\Gamma \lambda(\lambda I - A)^{-1} d\lambda$$

equals their sum. Hint: reduce A to the triangular form.

Exercise 12.4 *Show that $\mathcal{R}(P_\sigma)$ and $\mathcal{R}(P_{\overline{\sigma}})$ have the same dimensions. Hint: use (12.20).*

Exercise 12.5 *Show that for a real analytic function f and a real matrix A the matrix $f(A)$ is also real.*

Exercise 12.6 *Let A, B be any matrices such that the products AB and BA exist. Then*

$$\sigma(AB) \setminus \{0\} = \sigma(BA) \setminus \{0\} \tag{12.21}$$

and

$$Af(BA) = f(AB)A.$$

Hint: Use (12.21) and the identity

$$A(\lambda I - BA)^{-1} = (\lambda I - AB)^{-1}A$$

(here the two identity matrices need not be of the same order).

Exercise 12.7 *Show that any simple real eigenvalue of a J-Hermitian A is J-definite.*

Proposition 12.8 *Let $\mathcal{E}_{\lambda_1}$ be the root space belonging to any eigenvalue λ_1 of A. Then*

$$\mathcal{E}_{\lambda_1} = \{x \in \mathbb{C}^n \,:\, (A - \lambda_1 I)^k x = 0 \textit{ for some } k\,\}. \tag{12.22}$$

Proof. Let λ_1 correspond to the matrix, say, A'_1 in (12.18) and suppose $(A - \lambda_1 I)^k x = 0$ for some k. By setting $x = Sx'$ and noting that the matrices $A'_j - \lambda_1 I_{n_j}$ from (12.18) are non-singular for $j \neq 1$ it follows

$$x' = \begin{bmatrix} x'_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{12.23}$$

that is, $x' \in \mathcal{R}(P_1^{(0)})$ or equivalently $x \in \mathcal{R}(P_1)$.

Conversely, if $x \in \mathcal{R}(P_1)$ that is, (12.23) holds then

$$(A - \lambda_1 I)^k x = S(A' - \lambda_1 I_{n_1})^k x' = S \begin{bmatrix} (A'_1 - \lambda_1 I_{n_1})^k x'_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Now the matrix A'_1 has a single eigenvalue λ_1 of multiplicity n_1. Since we know that λ_1 is a pole of the resolvent $(\lambda I_{n_1} - A'_1)^{-1}$ the function

$$(\lambda - \lambda_1)^k (\lambda I_{n_1} - A'_1)^{-1}$$

will be regular at $\lambda = \lambda_1$ for some k. Thus,

$$(A'_1 - \lambda_1 I_{n_1})^k = \frac{1}{2\pi i} \int_{\Gamma_1} (\lambda - \lambda_1)^k (\lambda I_{n_1} - A'_1)^{-1} d\lambda = 0.$$

This implies $(A - \lambda_1 I)^k x = 0$. Q.E.D.

The matrix

$$N_1 = A'_1 - \lambda_1 I_{n_1},$$

has the property that some power vanishes. Such matrices are called *nilpotent*. We now can make the spectral decomposition (12.18) more precise

$$A' = S^{-1}AS = \begin{bmatrix} \lambda_1 I_{n_1} + N_1 & & \\ & \ddots & \\ & & \lambda_p I_{n_p} + N_p \end{bmatrix} \tag{12.24}$$

with

$$\sigma(A) = \{\lambda_1, \dots, \lambda_p\}, \quad N_1, \dots, N_p \text{ nilpotent.}$$

Obviously, the root space $\mathcal{E}_{\lambda_k}$ contains the eigenspace; if the two are equal the eigenvalue is called *semisimple* or *non-defective*, otherwise it is called *defective*, the latter case is characterised by a non-vanishing nilpotent N_k in (12.24). Similarly a matrix is called defective, if it has at least one defective eigenvalue. A defective matrix is characterised by the fact that its eigenvectors cannot form a basis of the whole space. The Jordan form of a nilpotent matrix will be of little interest in our considerations.

If the matrix A is real then the spectral sets σ_k can be chosen as symmetric with respect to the real axis: $\overline{\sigma_k} = \sigma_k$. Then the curve Γ, surrounding σ_k can also be chosen as $\overline{\Gamma} = \Gamma$ and we have

$$\overline{P}_{\sigma_k} = -\frac{1}{2\pi i} \int_{\Gamma} (\overline{\lambda} I - A)^{-1} d\overline{\lambda} = \frac{1}{2\pi i} \int_{\overline{\Gamma}} (\lambda I - A)^{-1} d\lambda = P_{\sigma_k}.$$

So, all subspaces $\mathcal{R}(P_{\sigma_k})$ are real and the matrix S appearing in (12.18) can be chosen as real.

Another case where this partition is convenient is if A is J-Hermitian (even if complex) since $\sigma(A)$ comes in complex conjugate pairs as well. Here we have

$$(\lambda I - A)^{-*} = J(\overline{\lambda} I - A)^{-1} J$$

and

$$P^*_{\sigma_k} = \frac{1}{-2\pi i} \int_{\Gamma} (\overline{\lambda} I - A)^{-1} d\overline{\lambda} = \frac{1}{2\pi i} \int_{\overline{\Gamma}} J(\lambda I - A)^{-1} J d\lambda = J P_{\sigma_k} J,$$

i.e. the projections P_{σ_k} are J-orthogonal, hence by Theorem 8.2 the matrix S in (12.18) can be taken as J, J'-unitary with

$$J' = S^* J S = \begin{bmatrix} J'_1 & & \\ & \ddots & \\ & & J'_p \end{bmatrix}. \tag{12.25}$$

In this case we call (12.18) a J, J' *unitary decomposition* of a J-Hermitian matrix A. Achieving (12.24) with a J, J'-unitary S is possible if and only if all eigenvalues are real. Indeed, in this case A' from (12.24) is J'-Hermitian, hence each $\lambda_k I_{n_k} + N_k$ is J'_k-Hermitian. Since its spectrum consists of a single point it must be real.

Proposition 12.9 *If the spectral part σ_k consists of J-definite eigenvalues (not necessarily of the same sign) then all these eigenvalues are non-defective and $\iota_\pm(JP_{\sigma_k})$ yields the number of the J-positive and J-negative eigenvalues in σ_k, respectively (counting multiplicities) and each root space is equal to the eigenspace. If all eigenvalues in σ_k are of the same type then in (12.25) we have $J'_k = \pm I_{n_k}$.*

Proof. The non-defectivity follows from Proposition 10.1. All other statements follow immediately from the identity (8.2) and Corollary 10.2 applied to the corresponding A'_k from (12.18). Q.E.D.

Exercise 12.10 *Let A be J-Hermitian and λ_j a J-definite eigenvalue. Then (i) λ_j is non-defective i.e. $N_j = 0$ in (12.26) and (ii) λ_j is a simple pole of the resolvent $(\lambda I - A)^{-1}$.*

The J, J'-unitary decomposition (12.18) can be further refined until each A'_j has either a single real eigenvalue λ_j in which case it reads

$$A'_j = \lambda_j I_{n_j} + N'_j, \quad N'_j \text{ nilpotent} \tag{12.26}$$

(see Proposition 12.8) or its spectrum is $\{\lambda, \overline{\lambda}\}$, $\lambda \neq \overline{\lambda}$. This is the 'most refined' decomposition which can be achieved with a diagonal matrix J', here also the real arithmetic is kept, if A was real. If we allow for complex J, J'-unitary transformations and non-diagonal J' then a further block-diagonalisation is possible.

Proposition 12.11 *Let A be J-Hermitian of order n and $\sigma(A) = \{\lambda, \overline{\lambda}\}$, $\lambda \neq \overline{\lambda}$. Then n is even and there exists a complex J, J'-unitary matrix U such that*

$$U^{-1} A U = \begin{bmatrix} \boldsymbol{\alpha} & 0 \\ 0 & \boldsymbol{\alpha}^* \end{bmatrix}, \quad J' = U^* J U = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix},$$

where $\boldsymbol{\alpha} = \lambda I + N$ and N is a nilpotent matrix of order $n/2$.

Proof. We know that in this case $\mathbb{C}^n = \mathcal{E}_\lambda \dot{+} \mathcal{E}_{\overline{\lambda}}$ and $\dim(\mathcal{E}_\lambda) = \dim(\mathcal{E}_{\overline{\lambda}})$ (Exercise 12.4). Now we need the following fact: *Any root space corresponding to a non-real eigenvalue is J-neutral.* Indeed, let

$$(A - \lambda I)^k x = 0, \quad (A - \mu I)^l y = 0$$

for some $k, l \geq 1$. For $k = l = 1$ we have

$$0 = y^* JAx - \lambda y^* Jx = y^* A^* Jx - \lambda x^* Jx = (2 \operatorname{Im} \lambda) y^* Jx \tag{12.27}$$

hence, $y^* Jx = 0$, which is already known from Theorem 9.1. For $k = 2, l = 1$, we set $\hat{x} = (A - \lambda I)x$, then $\hat{x} \in \mathcal{E}_\lambda$ and $(A - \lambda I)\hat{x} = 0$ and by applying (12.27) to $\hat{x}$ we have

$$0 = y^* J\hat{x} = y^*(A - \lambda I)x = (2 \operatorname{Im} \lambda) y^* Jx.$$

The rest is induction. Q.E.D.

Now let the matrix X_+, X_- carry the basis of $\mathcal{E}_\lambda, \mathcal{E}_{\overline{\lambda}}$, respectively, in its columns. Then

$$X = [X_+ \; X_-]$$

is non-singular and

$$X^* JX = \begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$$

hence B is square and non-singular. By using the polar decomposition

$$B = U(B^* B)^{1/2}$$

and

$$Y_+ = X_+ U (B^* B)^{-1/2} \quad Y_- = X_- (B^* B)^{-1/2}$$

we see that $Y = [\, Y_+ \; Y_- \,]$ satisfies

$$Y^* JY = J' = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$$

and also

$$AY_+ = Y_+ \boldsymbol{\alpha}, \quad AY_- = Y_- \boldsymbol{\alpha}_-$$

i.e.

$$Y^{-1} AY = \begin{bmatrix} \boldsymbol{\alpha} & 0 \\ 0 & \boldsymbol{\alpha}_- \end{bmatrix}.$$

Since the last matrix is J'-Hermitian by (6.1) we have $\boldsymbol{\alpha}_- = \boldsymbol{\alpha}^*$. Q.E.D.

Note that the block-diagonalisation in which each diagonal block corresponds to a *single eigenvalue* cannot be carried out in real arithmetic, if there are non-real eigenvalues hence the matrix $\boldsymbol{\alpha}$ cannot be real.

Thus refined, we call (12.18) a *complex spectral decomposition* of a J-Hermitian matrix A.

If a non-real eigenvalue λ_k is semisimple then we can always choose

$$A'_k = |\lambda_k| \begin{bmatrix} 0 & I_{n_k/2} \\ -I_{n_k/2} & 0 \end{bmatrix}, \quad J'_k = \begin{bmatrix} I_{n_k/2} & 0 \\ 0 & -I_{n_k/2} \end{bmatrix}. \tag{12.28}$$

Exercise 12.12 *Derive the formula (12.28).*

In view of 'the arithmetic of pluses and minuses' in Theorems 8.1, 8.2 and Proposition 12.11 the presence of $2s$ non-real spectral points (including multiplicities) 'consumes' s pluses and s minuses from the inertia of J. Consequently, A has at least

$$|\iota_+(J) - \iota_-(J)| = n - 2\iota_+(A)$$

real eigenvalues.

Example 12.13 Consider the matrix

$$A = \begin{bmatrix} 6 & -3 & -8 \\ 3 & -2 & -4 \\ 8 & -4 & -9 \end{bmatrix}$$

which is J-Hermitian with

$$J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

The matrix

$$U = \begin{bmatrix} 3 & -2 & -2 \\ 2 & -1 & -2 \\ 2 & -2 & -1 \end{bmatrix}$$

is J-unitary and induces the block-diagonalisation

$$U^{-1}AU = \begin{bmatrix} -2 & 1 & 0 \\ -1 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

whereas

$$S = \begin{bmatrix} 3\sqrt{2}/2 - \sqrt{2}i & 3\sqrt{2}/2 + \sqrt{2}i & -2 \\ \sqrt{2} - \sqrt{2}i/2 & \sqrt{2} + \sqrt{2}i/2 & -2 \\ \sqrt{2} - \sqrt{2}i & \sqrt{2} + \sqrt{2}i & -1 \end{bmatrix}$$

is J, J'-unitary with

$$J' = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

and induces the block-diagonalisation (in fact, the full diagonalisation)

$$S^{-1}AS = \begin{bmatrix} -2+i & 0 & 0 \\ 0 & -2-i & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

There is no general information on the non-vanishing powers of the nilpotents appearing in the J, J'-unitary decomposition of a J-Hermitian matrix. More can be said, if the matrix pair JA, J is semidefinite.

Theorem 12.14 *Let A be J-Hermitian, $JA - \lambda J$ positive semidefinite and $JA - \lambda' J$ not positive or negative definite for any other real λ'. Then λ is an eigenvalue and in the decomposition (12.18) for $\lambda_p = \lambda$ we have*

$$A'_p = \lambda_p I_{n_p} + N_p, \quad N_p^2 = 0$$

and $N_j = 0$ for $\lambda_j \neq \lambda_p$.

Proof. According to Theorem 10.6 λ must be an eigenvalue of A — otherwise $JA - \lambda J$ would be positive definite and we would have a non void definiteness interval (points with $JA - \lambda J$ *negative* definite are obviously precluded). Also, with J' from (12.25) the matrix $J'A' - \lambda J'$ and therefore each block $J'_j A'_j - \lambda J'_j$ is positive semidefinite and λ is the only spectral point of A'_j. From the general theory $N_j = A'_j - \lambda I_{n_j}$ must be nilpotent i.e. $N_j^r = 0$ for some $r \leq n$. At the same time $J'_j N_j$ is positive semidefinite. Assuming without loss of generality that r is odd and writing $r = 1 + 2q$ the equality $N_j^r = 0$ implies

$$0 = x^* J N_j^r x = (N_j^q x)^* J N_j N_j^q x$$

which by the positive semidefiniteness of JN_j implies $N_j^q = 0$. By induction this can be pushed down to $N_j^2 = 0$. Q.E.D.

In the theorem above the eigenvalue λ need not be defective, an example is $A = I$.

The developed spectral theory can be used to extend Theorem 10.5 to cases in which only some spectral parts of A are J-definite. They, too, will remain such until they cross eigenvalues of different type. As in Theorem 10.5 we will have to 'erect barriers' in order to prevent the crossing, but now we have to keep out the complex eigenvalues as well.

Theorem 12.15 *Let A be J-Hermitian and $\sigma_0 \subseteq \sigma(A)$ consist of J-positive eigenvalues and let Γ be any contour separating σ_0 from the rest of $\sigma(A)$. Let $\mathcal{I} \ni t \mapsto A(t)$ be J-Hermitian-valued continuous function on a closed interval $\mathcal{I}$ such that $A = A(t_0)$ for some $t_0 \in \mathcal{I}$ and such that*

$$\Gamma \cap \sigma(A(t_0)) = \emptyset.$$

Then the part of $\sigma(A(t))$ within Γ consists of J-positive eigenvalues whose number (with multiplicities) is constant over $\mathcal{I}$ (and analogously in the J-negative case).

Proof. The key fact is the continuity of the total projection

$$P(t) = \frac{1}{2\pi i}\int_\Gamma (\lambda I - A(t))^{-1} d\lambda$$

as a function of t, see Exercise 12.2. We can cover $\mathcal{I}$ by a finite family of open intervals such that within each of them the quantity $\|P(t') - P(t)\|$ is less than $1/2$. Now use Theorem 8.4; it follows that all $P(t)$ are J-unitarily similar. Then the number of the eigenvalues within Γ, which equals $\mathrm{Tr}P(t)$, is constant over $\mathcal{I}$ and all $P(t)$ are J-positive (Corollary 8.5). Q.E.D.

Note that as in Theorem 10.5 the validity of the preceding theorem will persist, if we allow Γ to move continuously in t.

Example 12.16 We shall derive the spectral decomposition of the phase-space matrix corresponding to a modally damped system which is characterised by the property (2.24). Then according to Theorem 2.3 there is a (real) non-singular Φ such that $\Phi^T M \Phi = I$, $\Phi^T K \Phi = \Omega^2 = \mathrm{diag}(\omega_1^2, \dots, \omega_n^2)$, $\omega_i > 0$, $\Phi^T C \Phi = D = \mathrm{diag}(d_{11}, \dots, d_{nn})$, $d_{ii} \ge 0$. The matrices

$$U_1 = L_1^T \Phi \Omega^{-1}, \quad U_2 = L_2^T \Phi \Omega^{-1}$$

are unitary and

$$A = \begin{bmatrix} 0 & U_1 \Omega U_2^{-1} \\ U_2 \Omega U_1^{-1} & U_2 D U_2^{-1} \end{bmatrix} = U A' U^{-1}$$

where

$$U = \begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix}, \quad A' = \begin{bmatrix} 0 & \Omega \\ -\Omega & -D \end{bmatrix}$$

and U is jointly unitary. This A' is essentially block-diagonal — up to a permutation. Indeed, define the permutation matrix V_0, given by $V_0 e_{2j-1} = e_j, V_0 e_{2j} = e_{j+n}$ (this is the so called *perfect shuffling*). For $n = 3$ we have

$$V_0 = \begin{bmatrix} 1&0&0&0&0&0 \\ 0&0&1&0&0&0 \\ 0&0&0&0&1&0 \\ 0&1&0&0&0&0 \\ 0&0&0&1&0&0 \\ 0&0&0&0&0&1 \end{bmatrix}.$$

The matrix V_0 is unitary and J, J'-unitary with

$$J' = \operatorname{diag}(j, \dots, j), \quad j = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

and we have

$$A'' = V_0^{-1} A' V_0 = \operatorname{diag}(A''_1, \dots, A''_n), \tag{12.29}$$

$$A''_i = \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -d_{ii} \end{bmatrix}.$$

Altogether, $V = UV_0$ is J, J'-unitary (and unitary) and we have the real spectral decomposition

$$V^{-1} A V = A''.$$

Exercise 12.17 *Use perfect shuffling and a convenient choice of L_1, L_2 in (2.16) so as to make the phase-space matrix for our model example in Chapter 1 as narrowly banded as possible.*

Exercise 12.18 *Derive the analog of (12.29) for the phase-space matrix (4.19).*

12.1 Condition numbers

We now turn back to the question of the condition number in block-diagonalising a J-Hermitian matrix. More precisely, we ask: are the J, J'-unitaries better than others, if the block-diagonalised matrix was J-Hermitian? The answer is positive. This makes J, J' unitaries even more attractive in numerical applications. The rest of this chapter will be devoted to this question.

Most interesting is the case in which in (12.18), (12.19) the spectra of A'_i are disjoint and symmetric with respect to the real axis, i.e. $\sigma(A)$ is partitioned as

$$\sigma_k = \sigma(A'_k), \quad \sigma(A) = \sigma_1 \cup \dots \cup \sigma_p, \quad \overline{\sigma}_j = \sigma_j, \ j = 1, \dots, p,$$

$$\sigma_i \cap \sigma_j = \emptyset \text{ if } \ i \neq j.$$

Then the relation (8.6) is equivalent to the block-diagonalisation (12.18) and it is sufficient to study the transformation (8.6).

Theorem 12.19 *Let $P_1, \dots P_p$ be any J-orthogonal decomposition of the identity and S non-singular with*

$$S^{-1} P_j S = P_j^{(0)}, \ j = 1, \dots, p \tag{12.30}$$

with $P_j^{(0)}$ from (12.19). Then there is a J, J'-unitary U such that

$$U^{-1}P_jU = P_j^{(0)},\ J' = \begin{bmatrix} J_1' & & \\ & \ddots & \\ & & J_p' \end{bmatrix},\ J_k' = \mathrm{diag}(I_{n_k^+}, -I_{n_k^-})$$

and

$$\kappa(U) \le \kappa(S).$$

The same inequality holds for the condition number measured in the Euclidian norm.

Proof. Any S satisfying (12.30) will be called *block-diagonaliser.* By representing S as

$$S = \left[\, S_1 \cdots\ S_p \,\right]$$

(the partition as in (12.18)) we can write (12.30) as

$$S_kP_j = \delta_{kj}P_k.$$

Then S^*JS is block-diagonal:

$$S_j^*JS_k = S_j^*JP_kS_k = S_j^*P_k^*JS_k = (P_kS_j)^*JS_k = \delta_{kj}S_j^*JS_k.$$

Hence each matrix $S_k^*JS_k$ is Hermitian and non-singular; its eigenvalue decomposition can be written as

$$S_k^*JS_k = W_k\Lambda_kW_k^* = W_k|\Lambda_k|^{1/2}J_k|\Lambda_k|^{1/2}W_k^*, \quad k = 1, \dots, p$$

where Λ_k is diagonal and non-singular, W_k is unitary and $J_k = \mathrm{sign}(\Lambda_k)$. By setting

$$\Lambda = \mathrm{diag}(\Lambda_1, \dots, \Lambda_p), \quad J' = \mathrm{diag}(J_1, \dots, J_p), \quad W = \mathrm{diag}(W_1, \dots, W_p)$$

the matrix

$$U = SW|\Lambda|^{-1/2}$$

is immediately seen to be J, J'-unitary. Now, since Λ and J' commute,

$$\kappa(S) = \|U|\Lambda|^{1/2}W^*\|\|W|\Lambda|^{-1/2}U^{-1}\| =$$

$$\|U|\Lambda|^{1/2}\|\|J'|\Lambda|^{-1/2}U^*J\| = \|U|\Lambda|^{1/2}\|\||\Lambda|^{-1/2}U^*\|$$

$$\ge \|UU^*\| = \kappa(U).$$

The proof for the Euclidian norm also uses the property

$$\kappa_E(F) = \|F\|_E\|F^{-1}\|_E = \|F\|_E\|J_1F^*J\|_E = \|F\|_E\|F^*\|_E = \|F\|_E^2, \tag{12.31}$$

valid for any J, J'unitary F. Thus,

$$\kappa_E(S)^2 = \mathrm{Tr}(S^*S)\mathrm{Tr}(S^{-1}S^{-*}) = \mathrm{Tr}(F^*F|\Lambda|)\mathrm{Tr}(|\Lambda|^{-1}F^*F)$$

$$= \left(\sum_{k}^{n} |\lambda_k| \mathrm{Tr}(F^* F)_{kk} \right) \left(\sum_{k}^{n} \mathrm{Tr}(F^* F)_{kk} / |\lambda_k| \right) \geq \mathrm{Tr}(F^* F)^2$$

due to the inequality

$$\sum_k p_k \phi_k \sum_k \phi_k / p_k \geq \sum_k \phi_k^2$$

valid for any positive p_k and non-negative ϕ_k. Q.E.D.

According to Theorem 12.19 and Proposition 10.17 if the spectrum is definite then the best condition number of a block-diagonalising matrix is achieved, by any J, J'-unitary among them. If the spectrum is not definite or if the spectral partition contains spectral subsets consisting of eigenvalues of various types (be these eigenvalues definite or not[3]) then the mere fact that a block-diagonalising matrix is J, J'-unitary does not guarantee that the condition number is the best possible. It can, indeed, be arbitrarily large. But we know that a best-conditioned block-diagonaliser should be sought among the J, J'-unitaries. The following theorem contains a construction of an optimal block-diagonaliser.

Theorem 12.20 *Let $P_1, \ldots, P_p \in \Xi^{n,n}, J'$ be as in Theorem 12.19. Then there is a J, J'-unitary block-diagonaliser $F \in \Xi^{n,n}$ with*

$$\kappa_E(F) \leq \kappa_E(\hat{S})$$

for any block-diagonaliser $\hat{S}$.

Proof. We construct F from an arbitrary given block-diagonaliser S as follows. By setting $S = \begin{bmatrix} S_1 \cdots S_p \end{bmatrix}$ and using the properties established in the proof of Theorem 12.19 we may perform the generalised eigenvalue decomposition

$$\Psi_k^* S_k^* S_k \Psi_k = \mathrm{diag}(\alpha_1^k, \ldots, \alpha_{n_k^+}^k, \beta_1^k, \ldots, \beta_{n_k^-}^k), \tag{12.32}$$

$$\Psi_k^* S_k^* J S_k \Psi_k = J_k' = \mathrm{diag}(I_{n_k^+}, -I_{n_k^-}), \tag{12.33}$$

$k = 1, \ldots, p$. An optimal block-diagonaliser is given by

$$F = \begin{bmatrix} F_1 \cdots F_p \end{bmatrix}, \quad F_k = S_k \Psi_k. \tag{12.34}$$

It is clearly J, J'-unitary by construction. To prove the optimality first note that the numbers $n_\pm^k$ and α_j^k, β_j^k do not depend on the block-diagonaliser S. Indeed, any block-diagonaliser $\hat{S}$ is given by

$$\hat{S} = \begin{bmatrix} S_1 \Gamma_1 \cdots S_p \Gamma_p \end{bmatrix}, \quad \Gamma_1, \ldots, \Gamma_p \text{ non-singular.}$$

[3] Such coarser partition may appear necessary, if a more refined partition with J-definite spectral sets would yield highly conditioned block-diagonaliser.

By the Sylvester theorem $n_\pm^k$ is given by the inertia of $S_k^* J S_k$ which is the same as the one of $\hat{S}_k^* J \hat{S}_k = \Gamma_k^* S_k^* J S_k \Gamma_k$. Furthermore, the numbers $\alpha_j^k, -\beta_j^k$ are the generalised eigenvalues of the matrix pair $S_k^* S_k$, $S_k^* J S_k$ and they do not change with the transition to the congruent pair $\Gamma_k^* S_k^* S_k \Gamma_k, \Gamma_k^* S_k^* J S_k \Gamma_k$. In particular, the sums

$$t_0^k = \alpha_1^k + \cdots + \alpha_{n_+}^k + \beta_1^k + \cdots + \beta_{n_-}^k, \quad t_0 = \sum_{k=1}^p t_0^k$$

are independent of S. By Theorem 12.19 there is a J, J'-unitary block-diagonaliser U with

$$\kappa_E(U) \le \kappa_E(S). \tag{12.35}$$

If we now do the construction (12.32) – (12.34) with S replaced by U then (12.33) just means that Ψ_k is J_k'-unitary. From Theorem 10.16 and from the above mentioned invariance property of t_0^k it follows

$$t_0^k = \mathrm{Tr}(F_k^* F_k) \le \mathrm{Tr}(U_k^* U_k)$$

and, by summing over k,

$$t_0 = \mathrm{Tr}(F^* F) \le \mathrm{Tr}(U^* U).$$

By (12.31) this is rewritten as

$$t_0 = \kappa_E(F)^2 \le \kappa_E(U)^2.$$

This with (12.35) and the fact that S was arbitrary gives the statement. Q.E.D.

The numbers α_i^k, β_j^k appearing in the proof of the previous theorem have a deep geometrical interpretation which is given in the following theorem.

Theorem 12.21 *Let F be any optimal block-diagonaliser from Theorem 12.20. Then the numbers α_i^k, β_j^k from (12.32) are the non-vanishing singular values of the projection P_k. Moreover,*

$$\kappa_E(F) = \|F\|_E^2 = \sum_{k=1}^p \left(\sum_{i=1}^{n_k^+} \alpha_i^k + \sum_{i=1}^{n_k^-} \beta_i^k \right) = \sum_{k=1}^p \mathrm{Tr}\sqrt{P_k P_k^*}. \tag{12.36}$$

Proof. We revisit the proof of Theorem 12.20. For any fixed k we set $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_1^k, \ldots, \beta_{n_k^-}^k)$. Since the numbers $\alpha_1^k, \ldots, \beta_{n_k^-}^k$ do not depend on the block diagonaliser S we may take the latter as J, J'-unitary:

$$S = U = \begin{bmatrix} U_1 \cdots U_p \end{bmatrix}, \quad U_j^* J U_k = \delta_{kj} J_k' \quad P_k = U_k J U_k^* J_k'$$

Now (12.32), (12.33) read

$$\Psi_k^* U_k^* U_k \Psi_k = \boldsymbol{\alpha}, \quad \Psi_k^* J_k' \Psi_k = J_k'$$

and we have $F_k = U_k \Psi_k$. It is immediately verified that the matrix $V = U_k \Psi_k \boldsymbol{\alpha}^{-1/2}$ is an isometry. Moreover, its columns are the eigenvectors to the non-vanishing eigenvalues of $P_k P_k^*$. To prove this note the identity

$$P_k P_k^* = (U_k J_k' U_k^*)^2.$$

On the other hand, using $\Psi_k J_k' \Psi_k^* = J_k$ gives

$$U_k J_k' U_k^* V = U_k \Psi_k J_k' \Psi_k^* U_k^* U_k \Psi_k \boldsymbol{\alpha}^{-1/2} = U_k \Psi_k J_k' \boldsymbol{\alpha}^{1/2}$$

and, since both $\boldsymbol{\alpha}$ and J_k' are diagonal,

$$U_k J_k' U_k^* V = V J_k' \boldsymbol{\alpha},$$

so the diagonal of $J_k' \boldsymbol{\alpha}$ consists of all non-vanishing eigenvalues of the Hermitian matrix $U_k J_k' U_k^*$. Hence the diagonal of $\boldsymbol{\alpha}$ consists of the non-vanishing singular values of P_k. Then also

$$\|F_k\|_E^2 = \mathrm{Tr}\boldsymbol{\alpha} = \sum_{i=1}^{n_k^+} \alpha_i^k + \sum_{i=1}^{n_k^-} \beta_i^k = \mathrm{Tr}\sqrt{P_k P_k^*},$$

so (12.36) holds as well. Q.E.D.

The value $\mathrm{Tr}\sqrt{P_k P_k^*}$ (this is, in fact, a matrix norm) may be seen as a condition number of the projection P_k. Indeed, we have

$$n_k = \mathrm{Tr} P_k \le \mathrm{Tr}\sqrt{P_k P_k^*} \tag{12.37}$$

where equality is attained, if and only if the projection P_k is orthogonal in the ordinary sense.

To prove (12.37) we write the singular value decomposition of P_k as

$$P_k = U \boldsymbol{\alpha} V^*$$

where U, V are isometries of type $n \times n_k$. The identity $P_k^2 = P_k$ means (note that $\boldsymbol{\alpha}$ is diagonal positive definite)

$$\boldsymbol{\alpha} V^* U = I_{n_k}.$$

Now

$$n_k = \mathrm{Tr} P_k = \mathrm{Tr}(U \boldsymbol{\alpha} V^*) = (V \boldsymbol{\alpha}^{1/2}, U \boldsymbol{\alpha}^{1/2})_E$$

where

$$(T, S)_E = \mathrm{Tr}(S^* T)$$

denotes the Euclidian scalar product on matrices. Using the Cauchy-Schwartz inequality and the fact that U, V are isometries we obtain

$$n_k = \mathrm{Tr} P_k \le \|V\boldsymbol{\alpha}^{1/2}\|_E \|U\boldsymbol{\alpha}^{1/2}\|_E \le \|\boldsymbol{\alpha}^{1/2}\|_E^2 = \mathrm{Tr}\boldsymbol{\alpha} = \mathrm{Tr}\sqrt{P_k P_k^*}.$$

If this inequality turns to an equality then the matrices $V\boldsymbol{\alpha}^{1/2}$, $U\boldsymbol{\alpha}^{1/2}$, and therefore V, U must be proportional, which implies that P_k is Hermitian and therefore orthogonal in the standard sense. If this is true for all k then by (12.36) the block-diagonaliser F is jointly unitary.

There are simple damped systems in which no reduction like (12.18) is possible. One of them was produced as the critical damping case in Example 9.3. Another is given by the following

Example 12.22 Take the system in (1.2) – (1.4) with $n = 2$ and

$$m_1 = 1, \; m_2 = \frac{25}{4},$$

$$k_1 = 0, \; k_2 = 5, \; k_3 = \frac{5}{4},$$

$$C = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}.$$

Tedious, but straightforward calculation (take L_1, L_2 as the Cholesky factors) shows that in this case the matrix (3.3) looks like

$$A = \begin{bmatrix} 0 & 0 & \sqrt{5} & -2/\sqrt{5} \\ 0 & 0 & 0 & 1/\sqrt{5} \\ -\sqrt{5} & 0 & -4 & 0 \\ 2/\sqrt{5} & -1/\sqrt{5} & 0 & 0 \end{bmatrix} = -I + N \qquad (12.38)$$

where $N^3 \neq 0$, $N^4 = 0$. So, in (12.18) we have $p = 1$ and no block-diagonalisation is possible.[4] Analogous damped systems can be constructed in any dimension.

Exercise 12.23 *Prove (12.38).*

Exercise 12.24 *Modify the parameters m_2, k_1, k_2, k_3, c in Example 12.22 in order to obtain an A with a simple real spectrum (you may make numerical experiments).*

Exercise 12.25 *Find a permutation P such that $P^T A P$ (A from (12.38)) is tridiagonal.*

[4] This intuitive fact is proved by constructing the Jordan form of the matrix N which we omit here.

Chapter 13
The matrix exponential

The matrix exponential, its properties and computation play a vital role in a study of damped systems. We here derive some general properties, and then touch a central concern of oscillations: the exponential decay which is crucial in establishing stability of vibrating structures.

The decomposition (12.18) can be used to simplify the computation of the matrix exponential needed in the solution (3.4) of the system (3.2). We have

$$e^{At} = Se^{A't}S^{-1} = S \begin{bmatrix} e^{A'_1 t} & & \\ & \ddots & \\ & & e^{A'_p t} \end{bmatrix} S^{-1}, \tag{13.1}$$

with a J, J'-unitary S and J' from (12.25). As was said a 'most refined' decomposition (13.1) is obtained, if $\sigma(A'_j) = \lambda_j$ is real, with

$$A'_j = \lambda_j I_{n_j} + N_j, \quad N_j \quad \text{nilpotent}$$

or $\sigma(A'_j) = \{\lambda_j, \overline{\lambda_j}\}$ non-real with

$$J'_j = \begin{bmatrix} 0 & I_{n_j/2} \\ I_{n_j/2} & 0 \end{bmatrix}, \quad A'_j = \begin{bmatrix} \boldsymbol{\alpha}_j & 0 \\ 0 & \boldsymbol{\alpha}_j^* \end{bmatrix}$$

and

$$\boldsymbol{\alpha}_j = \lambda_j I_{n_j} + M_j, \quad M_j \quad \text{nilpotent.}$$

Thus, the computation of $e^{A't}$ reduces to exponentiating $e^{(\lambda I + N)t}$ with N nilpotent:

$$e^{(\lambda I + N)t} = e^{\lambda t} e^{Nt} = e^{\lambda t} \sum_{k=0}^{r} \frac{t^k}{k!} N^k \tag{13.2}$$

where N^r is the highest non-vanishing power of N. (In fact, the formulae (13.1), (13.2) hold for any matrix A, where S has no J-unitarity properties, it is simply non-singular.)

Example 13.1 We compute the matrix exponential to the matrix A from (9.5), Example 9.3. First consider $D = d^2 - 4\omega^2 > 0$ and use (9.7) – (9.9) to obtain

$$e^{At} = e^{-dt/2}\left(\cosh(\sqrt{D}t/2)I + \frac{\sinh(\sqrt{D}t/2)}{\sqrt{D}}\begin{bmatrix} d & 2\omega \\ -2\omega & -d\end{bmatrix}\right). \tag{13.3}$$

A similar formula holds for $D < 0$. This latter formula can be obtained from (13.3) by taking into account that e^{At} is an entire function in each of the variables ω, d, t. We use the identities

$$\cosh(\sqrt{D}t/2) = \cos(\sqrt{-D}t/2), \quad \frac{\sinh(\sqrt{D}t/2)}{\sqrt{D}} = \frac{\sin(\sqrt{-D}t/2)}{\sqrt{-D}}$$

thus obtaining

$$e^{At} = e^{-dt/2}\left(\cos(\sqrt{-D}t/2)I + \frac{\sin(\sqrt{-D}t/2)}{\sqrt{-D}}\begin{bmatrix} d & 2\omega \\ -2\omega & -d\end{bmatrix}\right) \tag{13.4}$$

which is more convenient for $D < 0$. For $D = 0$ using the L'Hospital rule in either (13.3) or (13.4) we obtain

$$e^{At} = e^{-dt/2}\left(I + \omega t\begin{bmatrix} 1 & 1 \\ -1 & -1\end{bmatrix}\right).$$

It does not seem to be easy to give an exponential bound for the norm of (13.3) and (13.4) which would be both simple and tight. By introducing the functions

$$\operatorname{sinc}(\tau) = \sin(\tau)/\tau, \quad \operatorname{sinhc}(\tau) = \sinh(\tau)/\tau$$

we can express both exponentials by the 'special matrix function' $e2(\tau, \theta)$ given as

$$e2(\tau,\theta) = \exp\left(\begin{bmatrix} 0 & 1 \\ -1 & -2\theta\end{bmatrix}\tau\right) =$$

$$\begin{cases} e^{-\theta\tau}\left(\cos(\sqrt{1-\theta^2}\tau)I + \tau\operatorname{sinc}(\sqrt{1-\theta^2}\tau)\begin{bmatrix} \theta & 1 \\ -1 & -\theta\end{bmatrix}\right) & \theta < 1 \\ e^{-\theta\tau}\left(\cosh(\sqrt{\theta^2-1}\tau)I + \tau\operatorname{sinhc}(\sqrt{\theta^2-1}\tau)\begin{bmatrix} \theta & 1 \\ -1 & -\theta\end{bmatrix}\right) & \theta > 1 \\ e^{-\tau}\left(I + \tau\begin{bmatrix} 1 & 1 \\ -1 & -1\end{bmatrix}\right) & \theta = 1 \end{cases}$$

defined for any real τ and any non-negative real θ. So, we have

$$e^{At} = e2(\omega t, \frac{d}{2\omega}).$$

On Fig. 13.1 we have plotted the function $\tau \mapsto \|e2(\tau,\theta)\|$ for different values of θ shown on each graph. We see that different viscosities may produce locally erratic behaviour of $\|e2(\tau,\theta)\|$ within the existing bounds for it. An 'overall best behaviour' is apparent at $\theta = 1$ (critical damping, bold line).

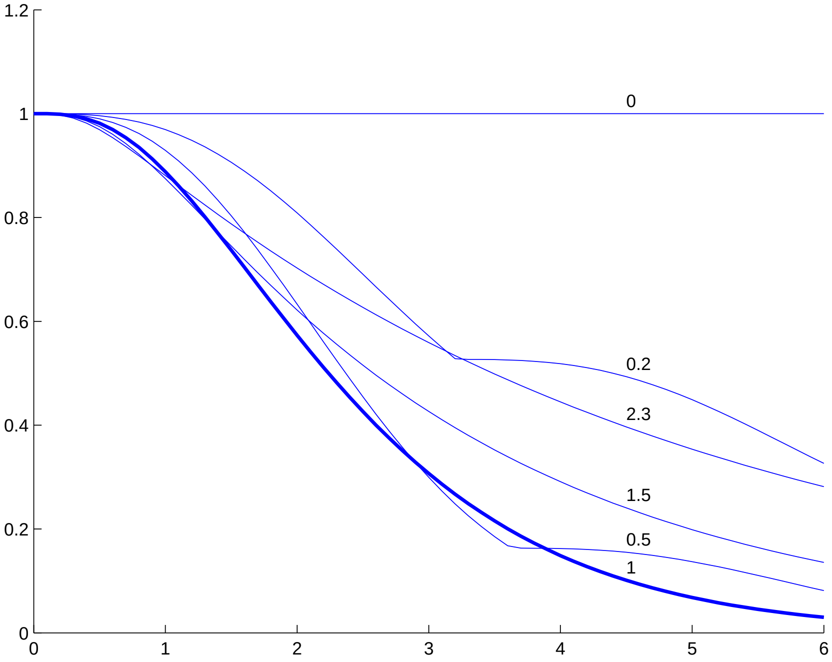

Fig. 13.1 Norm decay of $e2$

The 2×2 case studied in Example 13.1 is unexpectedly annoying: the formulae (13.3) or (13.4) do not allow a straightforward and simple estimate of the exponential decay of the matrix exponential. What can be done is to compute

$$\max_{t\geq 0} e^{\nu t}\|e2(\omega t,\theta)\| = \max_{\tau\geq 0} e^{\nu\tau/\omega}\|e2(\tau,\theta)\| = f2(\kappa,\theta)$$

with

$$f2(\kappa,\theta) = \max_{\tau\geq 0} e^{\kappa\tau}\|e2(\tau,\theta)\|.$$

The function $f2$ can be numerically tabulated and we have

$$\|e2(\omega t)\| \leq e^{-\nu t} f2(\frac{\nu}{\omega},\theta).$$

Remark 13.2 The function $f2$ is defined whenever $\theta > 0$ and $0 < \kappa = \kappa_0$, where

$$\kappa_0 = \begin{cases} \theta - \sqrt{\theta^2 - 1}, & \theta > 1 \\ \theta, & \theta < 1 \end{cases}$$

is the negative spectral abscissa of the 'normalised' matrix

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

— with one exception at $\theta = 1; \kappa = \kappa_0$, where $f2$ is infinite.

The two dimensional matrix exponential studied above shows another annoying phenomenon, characteristic for damped systems: the spectral condition number, that is the condition of the similarity which diagonalises the phase-space matrix A, may become infinite without any pathology at all by the matrix exponential. This is the case when the system is critically damped and the phase-space matrix defective. However, at this point the system is optimal in the sense that its spectral abscissa is the lowest there (see also the fat line in Fig. 13.1). This means that the method of spectral decomposition *may be insufficient to describe the solution of the differential equation (1.1) just on configurations that exhibit an optimal behaviour.*

Exercise 13.3 *Prove the identity*

$$\|e2(\tau, 1)\| = e^{-\tau}(\tau + \sqrt{1 + \tau^2}).$$

Exercise 13.4 *Compute e^{At} for any real J-symmetric matrix*

$$A = \begin{bmatrix} a_{11} & a_{12} \\ -a_{12} & a_{22} \end{bmatrix}.$$

Hint: Consider $A - a_{11}I$ or $A - a_{22}I$ and use Example 13.1.

Exercise 13.5 *Using (13.2) compute e^{At} for A from (12.38).*

We turn back to A from (13.1). We have

$$\|e^{At}\| \le \kappa(S)\|e^{A't}\| \tag{13.5}$$

where

$$\|e^{A't}\| \le e^{\lambda_a t} p(t), \tag{13.6}$$

and

$$\lambda_a = \lambda_a(A) = \max \operatorname{Re}(\sigma(A))$$

is *the spectral abscissa* of A and $p(t)$ is a polynomial of degree $\le n$ whose coefficients are built up from the powers of the nilpotents N_j, M_j. This is not very useful since (i) the formulae are inconvenient both for computing and for estimating,[1] (ii) there is no simple connection to the input data contained in the matrices M, C, K and (iii) there is no *a priori* control on $\kappa(S)$. Anyway, from (13.1) – (13.2) it follows that for any $\epsilon > 0$ there is a $C_\epsilon > 0$ such that

$$\|e^{At}\| \le C_\epsilon e^{(\lambda_a+\epsilon)t} \tag{13.7}$$

where naturally one would be interested to find a best (that is, the smallest) C_ϵ for a given $\epsilon > 0$.

If

$$\|e^{At}\| \le Fe^{-\nu t} \tag{13.8}$$

holds with some $\nu > 0, F \ge 1$ and all $t \ge 0$ then we say that the matrix A is *asymptotically stable* and e^{At} *exponentially decaying*. The following theorem is fundamental.

Theorem 13.6 *For a general square matrix A the following are equivalent:*

(i) *A is asymptotically stable.*
(ii) $\operatorname{Re}\sigma(A) < 0$.
(iii) $e^{At} \to 0, \quad t \to \infty$.
(iv) *For some (and then for each) positive definite B there is a unique positive definite X such that*

$$A^*X + XA = -B. \tag{13.9}$$

(v) $\|e^{At}\| < 1$, *for some* $t = t_1 > 0$.

In this case we have

$$\|e^{At}\| \le \kappa(X)e^{-t/\beta} \tag{13.10}$$

with

$$\beta = \max_y \frac{y^*Xy}{y^*By} = \|B^{-1/2}XB^{-1/2}\|.$$

Proof. The equivalences (i) $\Longleftrightarrow$ (ii) $\Longleftrightarrow$ (iii) obviously follow from (13.1) – (13.2). The equation (13.9) is known as *the Lyapunov equation* and it can be understood as a linear system of n^2 equations with as many unknowns.

Thus, let (i) hold. Then the integral

$$X = \int_0^\infty e^{A^*t}Be^{At}dt \tag{13.11}$$

(which converges by (13.8)) represents a positive definite matrix which solves (13.9). Indeed, X is obviously Hermitian and

[1] This difficulty is somewhat alleviated, if the block-diagonalisation went up to diagonal blocks of dimension one or two only; then we may use the formulae derived in Example 13.1.

$$y^*Xy = \int_0^\infty y^* e^{A^*t} B e^{At} y \; dt \geq 0$$

where the equality sign is attained, if and only if

$$Be^{At}y \equiv 0$$

which implies $y = 0$. The matrix X solves the Lyapunov equation. Indeed, by partial integration we obtain

$$A^*X = \int_0^\infty \frac{d}{dt} e^{A^*t} B e^{At} \; dt = e^{A^*t} B e^{At}\big|_0^\infty -$$

$$-\int_0^\infty e^{A^*t} B \frac{d}{dt} e^{At} \; dt = -B - XA.$$

The unique solvability of the Lyapunov equation under the condition (ii) is well-known; indeed the more general *Sylvester equation*

$$CX + XA = -B$$

is uniquely solvable, if and only if $\sigma(C) \cap \sigma(-A) = \emptyset$ and this is guaranteed, if $C^* = A$ is asymptotically stable. Thus (i) implies (iv).

Now, assume that a positive definite X solves (13.9). Then

$$\frac{d}{dt}\left(y^* e^{A^*t} X e^{At} y\right) = y^* e^{A^*t} (A^*X + XA) e^{At} y =$$

$$-y^* e^{A^*t} B e^{At} y \leq -y^* e^{A^*t} X e^{At} y/\beta$$

or

$$\frac{d}{dt} \ln \left(y^* e^{A^*t} X e^{At} y\right) \leq -\frac{1}{\beta}.$$

We integrate this from zero to s and take into account that $e^{A^*0} = I$:

$$y^* e^{A^*s} X e^{As} y \leq y^* X y e^{-s/\beta}$$

which implies

$$y^* e^{A^*s} e^{As} y \leq y^* X y \|X^{-1}\| e^{-s/\beta}.$$

Hence

$$\|e^{As}\|^2 \leq \|X\| \|X^{-1}\| e^{-s/\beta}$$

and (13.10) follows. This proves (iv) ⇒ (iii). Finally, (iii) ⇒ (v) is obvious; conversely, any t can be written as $t = kt_1 + \tau$ for some $k = 0, 1, \ldots$ and some $0 \leq \tau \leq t_1$. Thus,

$$\|e^{At}\| = \|(e^{At_1})^k e^{A\tau}\| \leq \|e^{At_1}\|^k \max_{0 \leq \tau \leq t_1} \|e^{A\tau}\| \to 0, \quad t \to \infty.$$

Thus, (v) ⇒ (iii). Q.E.D.

Exercise 13.7 *Suppose that A is dissipative and $\alpha = \|e^{At_1}\| < 1$, for some $t_1 > 0$. Show that (13.8) holds with*

$$F = 1/\alpha, \quad \nu = -\frac{\ln \alpha}{t_1}.$$

Corollary 13.8 *If A is asymptotically stable then X from (13.11) solves (13.9) for any matrix B.*

Exercise 13.9 *Show that for an asymptotically stable A the matrix X from (13.11) is the only solution of the Lyapunov equation (13.9) for an arbitrary B. Hint: supposing $A^*Y + YA = 0$ compute the integral*

$$\int_0^\infty e^{A^*t}(A^*Y + YA)e^{At}dt.$$

Exercise 13.10 *Show that already the positive semidefiniteness of B almost always (i.e. up to rare special cases) implies the positive definiteness of X in (13.9). Which are the exceptions?*

The matrix X from (13.9) will be shortly called *the Lyapunov solution.*

Exercise 13.11 *Show that the Lyapunov solution X satisfies*

$$2\lambda_a \le -\frac{1}{\beta}. \tag{13.12}$$

where λ_a is the spectral abscissa. Hint: compute y^*By on any eigenvector y of A.

If A were normal, then (13.8) would simplify to

$$\|e^{At}\| = e^{\lambda_a t}$$

as is immediately seen when A is unitarily diagonalised, i.e. here the *spectrum alone* gives the full information on the decay. In this case for $B = I$ (13.12) goes over into an equality because the Lyapunov solution reads

$$X = -(A + A^*)^{-1}$$

and $2\sigma(X) = -1/\operatorname{Re}\sigma(A)$.

Each side of the estimate (13.12) is of a different nature: the right hand side depends on the scalar product in Ξ^n (because it involves the adjoint of A) whereas the left hand side does not. Taking the matrix $A_1 = H^{1/2}AH^{-1/2}$ with H positive definite (13.9) becomes

$$A_1^*X_1 + X_1A_1 = -B_1$$

with

$$X_1 = H^{-1/2} X H^{-1/2}, \quad B_1 = H^{-1/2} B H^{-1/2}.$$

This gives rise to a β_1 which again satisfies (13.12). It is a non-trivial fact that by taking $B = I$ and varying H arbitrarily *the value* $-1/\beta_1$ *comes arbitrarily close to* $2\lambda_a$. The proof is rather simple, it is similar to the one that $\|H^{1/2} A H^{-1/2}\|$ is arbitrarily close to $\operatorname{spr}(A)$ if H varies over the set of all positive definite matrices. Here it is.

Since we are working with spectra and spectral norms we can obviously assume A to be upper triangular. Set $A' = DAD^{-1}$, $D = \operatorname{diag}(1, M, \ldots, M^{n-1})$, $M > 0$. Then $A' = (a'_{ij})$ is again upper triangular, has the same diagonal as A and

$$a'_{ij} = a_{ij}/M^{j-i}$$

hence $A' \to A_0 = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ as $M \to \infty$. The matrix A_0 is normal and the Lyapunov solution obviously depends continuously on A. This proves the statement.

On the other hand we always have the lower bound

$$\|e^{At}\| \geq e^{\lambda_a t}, \tag{13.13}$$

Indeed, let λ be the eigenvalue of A for which $\operatorname{Re}\lambda = \lambda_a$. Then $Ax = \lambda x$ implies $\|e^{At}x\| = e^{\lambda_a t}\|x\|$ and (13.13) follows.

Exercise 13.12 *Show that whenever the multiplicity p of an eigenvalue is larger than the rank r of the damping matrix C then this eigenvalue must be defective. Express the highest non-vanishing power of the corresponding nilpotent by means of r and p.*

Exercise 13.13 *Replace the matrix C in Example 12.22 by*

$$C = \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix}$$

Compute numerically the spectral abscissa for $2 < c < 6$ and find its minimum.

Chapter 14
The quadratic eigenvalue problem

Thus far we have studied the damped system through its phase space matrix A by taking into account its J-Hermitian structure and the underlying indefinite metric. Now we come back to the fact that this matrix stems from a second order system which carries additional structure commonly called 'the quadratic eigenvalue problem'. We study the spectrum of A and the behaviour of its exponential over time. A special class of so-called overdamped systems will be studied in some detail.

If we set $f = 0$ in (3.2), and make the substitution $y(t) = e^{\lambda t}y$, y a constant vector, we obtain

$$Ay = \lambda y \tag{14.1}$$

and similarly, if in the homogeneous equation (1.1) we insert $x(t) = e^{\lambda t}x$, x constant we obtain

$$(\lambda^2 M + \lambda C + K)x = 0 \tag{14.2}$$

which is called the quadratic eigenvalue problem, attached to (1.1), λ is an eigenvalue and x a corresponding eigenvector. We start here a detailed study of these eigenvalues and eigenvectors constantly keeping the connection with the corresponding phase-space matrix.

Here, too, we can speak of the 'eigenmode' $x(t) = e^{\lambda t}x$ but the physical appeal is by no means as cogent as in the undamped case (Exercise 2.1) because the proportionality is a complex one. Also, the general solution of the homogeneous equation (1.1) will not always be given as a superposition of the eigenmodes.

The equations (14.1) and (14.2) are immediately seen to be equivalent via the substitution (for generality we keep on having complex M, C, K)

$$y = \begin{bmatrix} L_1^* x \\ \lambda L_2^* x \end{bmatrix}, \ K = L_1 L_1^*, \ M = L_2 L_2^*. \tag{14.3}$$

(14.2) may be written as

$$Q(\lambda)x = 0$$

where $Q(\cdot)$, defined as

$$Q(\lambda) = \lambda^2 M + \lambda C + K \tag{14.4}$$

is the *quadratic matrix pencil* associated with (1.1). The solutions of the equation $Q(\lambda)x = 0$ are referred to as the eigenvalues and the eigenvectors of the pencil $Q(\cdot)$. For $C = 0$ the eigenvalue equation reduces to (2.3) with $\lambda^2 = -\mu$. Thus, to two different eigenvalues $\pm i\omega$ of (14.2) there corresponds only one linearly independent eigenvector. In this case the pencil can be considered as linear.

Exercise 14.1 *Show that in each of the cases*

1. $C = \alpha M$
2. $C = \beta K$

the eigenvalues lie on a simple curve in the complex plane. Which are the curves?

The set $\mathcal{X}_\lambda = \{x : Q(\lambda)x = 0\}$ is *the eigenspace* for the eigenvalue λ of the pencil $Q(\cdot)$, attached to (1.1). In fact, (14.3) establishes a one-to-one linear map from $\mathcal{X}_\lambda$ onto the eigenspace

$$\mathcal{Y}_\lambda = \{y \in \mathbb{C}^{2n} : Ay = \lambda y\},$$

in particular, $dim(\mathcal{Y}_\lambda) = dim(\mathcal{X}_\lambda)$. In other words, "the eigenproblems (14.1) and (14.2) have the same geometric multiplicity."
As we shall see soon, the situation is the same with the algebraic multiplicities. As is well known, the algebraic multiplicity of an eigenvalue of A is equal to its multiplicity as the root of the polynomial $\det(\lambda I - A)$ or, equivalently, the dimension of the root space $\mathcal{E}_\lambda$. With J from (3.7) we compute

$$JA - \lambda J = \begin{bmatrix} I & 0 \\ -L_2^{-1}L_1/\lambda & L_2^{-1} \end{bmatrix} \begin{bmatrix} -\lambda & 0 \\ 0 & \frac{\lambda^2 M + \lambda C + K}{\lambda} \end{bmatrix} \begin{bmatrix} I & -L_1^* L_2^{-*}/\lambda \\ 0 & L_2^{-*} \end{bmatrix} \tag{14.5}$$

where L_1, L_2 are from (14.3). Thus, after some sign manipulations,

$$\det(\lambda I - A) = \det(L_2^{-1}(\lambda^2 M + \lambda C + K)L_2^{-*}) = \det Q(\lambda)/\det M.$$

Hence the roots of the equation $\det(\lambda I - A) = 0$ coincide with those of

$$\det(\lambda^2 M + \lambda C + K) = 0$$

including their multiplicities. This is what is meant by saying that the algebraic multiplicity of an eigenvalue of A is equal to its algebraic multiplicity as an eigenvalue of $Q(\cdot)$.

The situation is similar with the phase-space matrix A from (4.19). The linear relation

$$x = \Phi y, \quad z_1 = \Omega_1 y_1, \quad z_2 = \Omega_2 y_2, \quad z_3 = \lambda y_1 \tag{14.6}$$

establishes a one-to-one correspondence between the eigenspace Z_λ for A and the eigenspace $\mathcal{X}_\lambda$ for (14.2) (here, too, we allow complex Hermitian M, C, K). To compute a decomposition analogous to (14.5), we consider the inverse in (4.23):

$$\begin{aligned} -JA^{-1} + \frac{1}{\lambda}J &= \begin{bmatrix} \Omega^{-1}D\Omega^{-1} + \frac{1}{\lambda}I_n & F \\ F^* & -\frac{1}{\lambda}I_m \end{bmatrix} \\ &= W \begin{bmatrix} Z & 0 \\ 0 & -\frac{1}{\lambda}I_m \end{bmatrix} W^* \end{aligned} \tag{14.7}$$

where λ is real and

$$W = \begin{bmatrix} I_n & -\lambda F \\ 0 & I_m \end{bmatrix}$$

and

$$\begin{aligned} Z &= \frac{\lambda\Omega^{-1}D\Omega^{-1} + I_n + \lambda^2 FF^*}{\lambda} \\ &= \frac{\Omega^{-1}}{\lambda}\left(\lambda^2 \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} + \lambda D + \Omega^2\right)\Omega^{-1} \\ &= \frac{\Omega^{-1}}{\lambda}\Phi^* Q(\lambda)\Phi\Omega^{-1}. \end{aligned}$$

Now, using this, (14.7) as well as the equality

$$\det A = \det(\Omega)^2 (\det D_{22})^{-1} (-1)^{n+m} \tag{14.8}$$

we obtain

$$\det(\lambda I - A) = \det\left(-J\lambda\left(\frac{J}{\lambda} - JA^{-1}\right)A\right) = (\det D_{22})^{-1}(\det\Phi)^2 \det Q(\lambda)$$

where Φ is from (4.7) and D_{22} from (4.8). From this it again follows that the algebraic multiplicity of an eigenvalue of A equals the algebraic multiplicity of the same eigenvalue of the pencil $Q(\cdot)$.

That is why we have called the phase-space matrix a 'linearisation': the quadratic eigenvalue problem is replaced by one which is linear and equivalent to it.

Let us now concentrate on the eigenvalue equation (14.2). It implies

$$\lambda^2 x^* M x + \lambda x^* C x + x^* K x = 0, \ x \neq 0 \tag{14.9}$$

and hence

$$\lambda \in \{\, p_+(x),\ p_-(x)\,\} \tag{14.10}$$

with

$$p_\pm(x) = \frac{-x^*Cx \pm \sqrt{\Delta(x)}}{2x^*Mx}, \tag{14.11}$$

$$\Delta(x) = (x^*Cx)^2 - 4x^*Kxx^*Mx. \tag{14.12}$$

The functions $p_\pm$ are homogeneous:

$$p_\pm(\alpha x) = p_\pm(x)$$

for any scalar α and any vector x, both different from zero. If the discriminant $\Delta(x)$ is non-negative, then $p_\pm(x)$ is real and negative. If x is an eigenvector then by (14.10) the corresponding eigenvalue λ is also real and negative. Now for the corresponding y from (14.3) we have

$$\begin{aligned} y^*Jy &= x^*Kx - \lambda^2 x^*Mx \\ &= 2x^*Kx + \lambda x^*Cx \\ &= \frac{4x^*Mxx^*Kx - (x^*Cx)^2 \pm x^*Cx\sqrt{\Delta(x)}}{2x^*Mx} \\ &= \frac{\Delta(x) \pm x^*Cx\sqrt{\Delta(x)}}{2x^*Mx} = \mp\lambda\sqrt{\Delta(x)} \end{aligned} \tag{14.13}$$

where we have used the identity (14.9) and the expressions (14.11), (14.12). Since $-\lambda$ is positive, we see that the eigenvector y of A is J-positive or J-negative according to the sign taken in (14.11). The same is true of A from (4.19) and z from (14.6). Indeed, with J from (4.22), we have

$$z^*Jz = y_1^T\Omega_1 y_1 + y_2^T\Omega_2 y_2 - \lambda^2 y_1^T y_1 =$$

$$y^T\Omega y - \lambda^2 y^T \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} y = x^TKx - \lambda^2 x^TMx$$

as in (14.13). We summarise:

Proposition 14.2 *The following statements are equivalent*

1. *λ is an eigenvalue of (14.2) and every corresponding eigenvector satisfies $\Delta(x) > 0$.*
2. *λ is a J-definite eigenvalue of the phase-space matrix A from either (3.3) or (4.19).*

In this case λ is J-positive (J-negative) if and only if $\lambda = p_+(x)$ ($\lambda = p_-(x)$) for any corresponding eigenvector x.

The appellation positive/negative/definite type will be naturally extended to the eigenvalues of $Q(\cdot)$.

The reader will notice that we do not use the term "overdamped" for such eigenvalues (as we did for the one-dimensional oscillator in Example 9.3).

The overdampedness property in several dimensions is stronger than that as will be seen in the sequel.

If all eigenvalues of (14.2) are real and simple then they are of definite type (cf. Exercise 12.7).

We call the family $Q(\cdot)$ (or, equivalently, the system (1.1)) *overdamped*, if there is a μ such that

$$Q(\mu) = \mu^2 M + \mu C + K$$

is negative definite.

Proposition 14.3 *For λ negative we have*

$$\iota_+(JA - \lambda J) = \iota_-(Q(\lambda)) + n$$

and consequently

$$\iota_-(JA - \lambda J) = \iota_+(Q(\lambda)).$$

for A, J from (3.3), (3.7), respectively. In particular, $Q(\lambda)$ is negative definite, if and only if $JA - \lambda J$ is positive definite.

Proof. The assertion immediately follows from (14.5) and the Sylvester inertia theorem (Theorem 5.7). Q.E.D.

Corollary 14.4 *Let $Q(\cdot)$ be overdamped with eigenvalues*

$$\lambda^-_{n^-} \le \cdots \le \lambda^-_1 < \lambda^+_1 \le \cdots \le \lambda^+_{n^+}$$

and the definiteness interval $(\lambda^-_1, \lambda^+_1)$. Then for any $\lambda > \lambda^+_1$, with $Q(\lambda)$ non-singular the quantity $\iota_+(Q(\lambda))$ is equal to the number of the J-positive eigenvalues less than λ (and similarly for the J-negative eigenvalues).

Proof. The statement follows from Proposition 14.3 and Theorem 10.7. Q.E.D.

Thus, the overdampedness of the system M, C, K is equivalent to the definitisability of its phase-space matrix. By Theorem 10.6 the eigenvalues are split into two groups: the J-positive eigenvalues on the right and the J-negative ones on the left. They are divided by the definiteness interval which will go under the same name for the pencil $Q(\cdot)$ as well.

The overdampedness condition means

$$x^*Mx\lambda^2 + x^*Cx\lambda + x^*Kx \equiv x^*Mx(\lambda - p_-(x))(\lambda - p_+(x)) < 0 \qquad (14.14)$$

for any fixed λ from the definiteness interval and for all non-vanishing x, that is

$$p_-(x) < \lambda, \quad p_+(x) > \lambda,$$

in particular,

$$p_-(x) < p_+(x),$$

or equivalently

$$\Delta(x) > 0 \tag{14.15}$$

for every $x \neq 0$. *Prima facie* the condition (14.15) is weaker than (14.14). It is a non-trivial fact that the two are indeed equivalent as will be seen presently.

The overdampedness is not destroyed, if we increase the damping and/or decrease the mass and the stiffness, that is, if we replace M, C, K by M', C', K' with

$$x^*M'x \leq x^*Mx, \quad x^*C'x \geq x^*Cx, \quad x^*K'x \leq x^*Kx \text{ for all } x.$$

In this situation we say that the system M', C', K' is *more viscous* than M, C, K. The viscosity is, in fact, an order relation on the set of all damped systems.

We now formulate a theorem collecting together several equivalent definitions of the overdampedness.

Theorem 14.5 *The following properties are equivalent:*

(i) The system (1.1) is overdamped
(ii) The discriminant $\Delta(x)$ from (14.12) is positive for all non-vanishing $x \in \mathbb{C}^n$.
(iii) The pair JA, J or, equivalently, S, T from (11.6) is definitisable.

Proof. The relations (i) $\iff$ (iii), (i) $\implies$ (ii) are obvious. To prove (ii) $\implies$ (i) we will use the following, also obvious, facts:

- The property (ii) is preserved under increased viscosity.
- The property (i) is preserved under increased viscosity, moreover the new definiteness interval contains the old one.
- The set of all semidefinite matrices is closed.
- The set of all overdamped systems is open, that is, small changes in the matrices M, C, K do not destroy overdampedness.

Suppose, on the contrary, that (ii) $\implies$ (i) does not hold. Then there would exist an $\epsilon_0 \geq 1$ such that the pencils

$$Q(\lambda, \epsilon) = \lambda^2 M + \lambda\epsilon C + K$$

are overdamped for all $\epsilon > \epsilon_0$ but not for $\epsilon = \epsilon_0$ (obviously the overdampedness holds, if ϵ is large enough). Since the corresponding phase-space matrix A_ϵ) depends continuously — and the definiteness intervals monotonically — on $\epsilon > \epsilon_0$ there is a real point $\lambda = \lambda_0$, contained in all these intervals and this is an eigenvalue of A_{ϵ_0}— otherwise $JA_{\epsilon_0} - \lambda_0 J$ would be not only positive semidefinite (which it must be), but positive definite which is precluded. Thus $JA_{\epsilon_0} - \lambda_0 J$ is singular positive semidefinite, but by (ii) the eigenvalue

λ_0 is of definite type. By Theorem 10.11 the pair JA_{ϵ_0}, J is definitisable i.e. $JA_{\epsilon_0} - \lambda J$ is positive definite for some λ in the neighbourhood of λ_0. This proves (ii) $\Longrightarrow$ (i). Q.E.D.

Exercise 14.6 *Let the system M, C, K be overdamped and let C be perturbed into $C' = C + \delta C$ with*

$$|x^T \delta C x| \le \epsilon x^T C x \quad \textit{for all } x,$$

$$\epsilon < 1 - \frac{1}{d}, \quad d = \min_x \frac{x^T C x}{2\sqrt{x^T M x x^T K x}}.$$

Show that the system M, C', K is overdamped.

Exercise 14.7 *Let M, C, K be real symmetric positive definite matrices. Show that $\Delta(x) > 0$ for all* real $x \neq 0$ *is equivalent to $\Delta(x) > 0$ for all* complex $x \neq 0$.

Exercise 14.8 *Let the damping be proportional: $C = \alpha M + \beta K$, $\alpha, \beta \ge 0$. Try to tell for which choices of α, β the system is overdamped. Show that this is certainly the case, if $\alpha\beta > 1$.*

Exercise 14.9 *Prove the following statement. If μ is negative and $Q(\mu)$ nonsingular then A has at least*

$$2(n - \iota_-(Q(\mu)))$$

real eigenvalues (counting multiplicities).

Exercise 14.10 *Prove the following statement. Let $Q(\cdot)$ have p eigenvalues $\lambda_1, \ldots, \lambda_p$ (counting multiplicities). If all these eigenvalues are of positive type then corresponding eigenvectors $x_1, \ldots, x_p$ can be chosen as linearly independent. Hint: Use the fact that the J-Gram matrix $X^* J X$ for the corresponding phase space eigenvectors*

$$y_j = \begin{bmatrix} L_1^* x_j \\ \lambda_j L_2^* x_j \end{bmatrix}$$

is positive definite.

Exercise 14.11 *Prove the identity (14.8). Check this identity on the system from Example 4.5.*

Chapter 15
Simple eigenvalue inclusions

Although, as we have seen, the eigenvalues alone may not give sufficient information on the behaviour of the damped system, knowledge of them may still be useful. In this chapter we will derive general results on the location of the eigenvalues of a damped system.

In fact, the parameters obtained in such 'bounds' or 'inclusions' may give direct estimates for the matrix exponential which are often more useful than (13.5), (13.6). An example is *the numerical range* or *the field of values*

$$r(A) = \{x^* Ax : \; x \in \mathbb{C}^n, \; \|x\| = 1\}$$

which contains $\sigma(A)$. This is seen, if the eigenvalue equation

$$Ax - \lambda x = 0$$

is premultiplied with x^* which gives

$$\lambda = \frac{x^* Ax}{x^* x}.$$

Now (3.9) can be strengthened to

$$\frac{d}{dt}\|y\|^2 = \frac{d}{dt} y^* y = \dot{y}^* y + y^* \dot{y} = y^*(A^* + A)y =$$

$$2 \operatorname{Re} y^* Ay \le 2 \max \operatorname{Re}(r(A)) \|y\|^2 = \lambda_a \|y\|^2.$$

By integrating the inequality

$$\frac{\frac{d}{dt}\|y\|^2}{\|y\|^2} = \frac{d}{dt}\frac{1}{\|y\|^2} \le 2 \max \operatorname{Re}(r(A))$$

we obtain

$$\|y\|^2 \le e^{2 \max \operatorname{Re}(r(A))} \|y(0)\|^2.$$

Similarly, the inequality

$$\frac{d}{dt}\|y\|^2 = 2\,\mathrm{Re}\,y^*Ay \geq 2\min\mathrm{Re}(r(A))\|y\|^2$$

implies

$$\|y\|^2 \geq e^{2\min\mathrm{Re}(r(A))}\|y(0)\|^2.$$

Altogether we obtain

$$e^{\min\mathrm{Re}(r(A))t} \leq \|e^{At}\| \leq e^{\max\mathrm{Re}(r(A))t}. \tag{15.1}$$

This is valid for any matrix A. If A is our phase-space matrix then $\max\mathrm{Re}(r(A)) = 0$ which leads to the known estimate $\|e^{At}\| \leq 1$ whereas the left hand side inequality in (15.1) yields

$$\|e^{At}\| \geq e^{-\|C\|t}.$$

The decay information obtained in this way is rather poor. A deeper insight comes from the quadratic eigenvalue equation (14.2). We thus define *the numerical range of the system M, C, K*, also called *the quadratic numerical range* as

$$w(M,C,K) = \{\lambda \in \mathbb{C} :\ \lambda^2 x^*Mx + \lambda x^*Cx + x^*Kx = 0,\ x \in \mathbb{C}^n \setminus \{0\}\}$$

or, equivalently,

$$w(M,C,K) = \mathcal{R}(p_+) \cup \mathcal{R}(p_-)$$

where the symbol $\mathcal{R}$ denotes the range of a map and $p_\pm$ is given by (14.11); for convenience we set a non-real p_+ to have the positive imaginary part. By construction, the quadratic numerical range contains all eigenvalues.[1] Due to the obvious continuity and homogeneity of the functions $p_\pm$ the set $w(M,C,K)$ is compact. (That is, under our standard assumption that M be positive definite. If M is singular, then $w(M,C,K)$ is unbounded. This has to do with the fact that the 'missing' eigenvalues in this case can be considered as infinite.)

The non-real part of $w(M,C,K)$ is confined to certain regions of the left half plane:

Proposition 15.1 *Let $\lambda \in w(M,C,K)$ be non real. Then*

$$c_- \leq -Re\lambda \leq c_+ \tag{15.2}$$

$$\omega_1 \leq |\lambda| \leq \omega_n \tag{15.3}$$

where

$$c_- = \min\frac{x^TCx}{2x^TMx},\ c_+ = \sup\frac{x^TCx}{2x^TMx}, \tag{15.4}$$

[1] This is a special case of the so called *polynomial numerical range*, analogously defined for general matrix polynomials.

$$\omega_1^2 = \min \frac{x^T K x}{x^T M x},\ \omega_n^2 = \sup \frac{x^T K x}{x^T M x}. \tag{15.5}$$

Here infinity as sup value is allowed, if M has a non-trivial null space and only those x are taken for which $x^T M x > 0$. In particular, any non-real eigenvalue λ satisfies (15.2), (15.3).

Proof. If λ is non-real, then

$$\lambda = \frac{-x^* C x \pm i\sqrt{4x^* K x x^* M x - (x^* C x)^2}}{2x^* M x}$$

and

$$\operatorname{Re} \lambda = -\frac{x^* C x}{2x^* M x}$$

from which (15.2) follows. Similarly,

$$|\lambda|^2 = \frac{x^* K x}{x^* M x}$$

from which (15.3) follows. Q.E.D.

Note that for M non-singular the suprema in (15.4), (15.5) turn to maxima.

The eigenvalue inclusion obtained in Proposition 15.1 can be strengthened as follows.

Theorem 15.2 *Any eigenvalue λ outside of the region (15.2), (15.3) is real and of definite type. More precisely, any $\lambda < \max\{-c_-, -\omega_1\}$ is of negative type whereas any $\lambda > \min\{-c_+, -\omega_n\}$ is of positive type.*

Proof. The complement of the region (15.2), (15.3) contains only real negative eigenvalues λ. Hence for any eigenpair λ, x of (14.2) we have $\Delta(x) \geq 0$ and $\lambda \in \{p_+(x), p_-(x)\}$.

Suppose

$$\lambda = p_+(x) < -c_-$$

that is,

$$-\frac{x^* C x}{2x^* M x} + \sqrt{\left(\frac{x^* C x}{2x^* M x}\right)^2 - \frac{x^* K x}{x^* M x}} < -\max_y \frac{y^* C y}{2y^* M y}.$$

Then

$$\frac{x^* C x}{2x^* M x} > \max_y \frac{y^* C y}{2y^* M y}.$$

— a contradiction. Thus, $\lambda \neq p_+(x)$ and $\lambda = p_-(x)$, hence $\Delta(x) > 0$ and λ is of negative type. The situation with $\lambda = p_-(x) > -c_+$ is analogous.

Suppose now

$$\lambda = p_+(x) < -\omega_n$$

that is

$$\sqrt{\left(\frac{x^*Cx}{2x^*Mx}+\sqrt{\frac{x^*Kx}{x^*Mx}}\right)\left(\frac{x^*Cx}{2x^*Mx}-\sqrt{\frac{x^*Kx}{x^*Mx}}\right)} < \frac{x^*Cx}{2x^*Mx}-\sqrt{\max_y\frac{y^*Ky}{y^*My}}$$

$$\leq \frac{x^*Cx}{2x^*Mx}-\sqrt{\frac{x^*Kx}{x^*Mx}}.$$

So, we would have

$$\frac{x^*Cx}{2x^*Mx}+\sqrt{\frac{x^*Kx}{x^*Mx}} < \frac{x^*Cx}{2x^*Mx}-\sqrt{\frac{x^*Kx}{x^*Mx}}$$

— a contradiction. Hence $\lambda \neq p_+(x)$, $\lambda = p_-(x)$, $\Delta(x) > 0$ and by Proposition 14.2 λ is of negative type. Suppose

$$-\omega_1 < \lambda = p_-(x).$$

Then

$$\frac{1}{p_-(x)} = -\frac{x^*Cx}{2x^*Kx}+\sqrt{\left(\frac{x^*Cx}{2x^*Kx}\right)^2-\frac{x^*Mx}{x^*Kx}} < -\max_y\frac{y^*My}{y^*Ky}.$$

or again

$$\sqrt{\left(\frac{x^*Cx}{2x^*Kx}+\sqrt{\frac{x^*Mx}{x^*Kx}}\right)\left(\frac{x^*Cx}{2x^*Kx}-\sqrt{\frac{x^*Mx}{x^*Kx}}\right)} < \frac{x^*Cx}{2x^*Kx}-\sqrt{\max_y\frac{y^*My}{y^*Ky}}.$$

This leads to a contradiction similarly as before. Q.E.D.

Remark 15.3 Another set of eigenvalue inclusions is obtained if we note that the *adjoint eigenvalue problem*

$$(\mu^2K+\mu C+M)x=0$$

has the same eigenvectors as (14.2) and the inverse eigenvalues $\mu = 1/\lambda$. Analogously to (15.2) non-real eigenvalues λ are contained in the set difference of two circles

$$\left\{\mu_1+i\mu_2:\ \left(\mu_1\ +\frac{1}{2\gamma_-}\right)^2+\mu_2^2\leq\left(\frac{1}{2\gamma_-}^2\right)\right\}\setminus$$

$$\left\{\mu_1+i\mu_2:\ \left(\mu_1\ +\frac{1}{2\gamma_+}\right)^2+\mu_2^2\leq\left(\frac{1}{2\gamma_+}\right)^2\right\}$$

with

$$\gamma_+ = \max_x \frac{x^*Cx}{2x^*Kx}, \quad \gamma_- = \min_x \frac{x^*Cx}{2x^*Kx}$$

(the analog of (15.3) gives nothing new).

Proposition 15.4 *Let M be non-singular and n eigenvalues (counting multiplicities) of A are placed on one side of the interval*

$$[-c_+, -c_-] \cap [-\omega_n, -\omega_1] \tag{15.6}$$

with $c_\pm$ as in Proposition 15.1. Then the system is overdamped.

Proof. Let n eigenvalues $\lambda_1^+, \dots, \lambda_n^+$ be placed, say, in

$$(-\min\{c_-, \omega_1\}, 0).$$

Then their eigenvectors are J-positive and all these eigenvalues are J-positive. Therefore to them there correspond n eigenvectors of A: $y_1^+, \dots, y_n^+$ with $y_j^* J y_k = \delta_{jk}$ and they span a J-positive subspace $\mathcal{X}_+$, invariant for A. Then the J-orthogonal complement $\mathcal{X}_-$ is also invariant for A and is J-negative (note that $n = \iota_+(J) = \iota_-(J)$) hence other eigenvalues are J-negative and they cannot belong to the interval (15.6). Thus we have the situation from Theorem 10.6. So there is a μ such that $JA - \mu J$ is positive definite and the system is overdamped. Q.E.D.

The systems having one or more *purely imaginary* eigenvalues in spite of the damping have special properties:

Proposition 15.5 *An eigenvalue λ in (14.2) is purely imaginary, if and only if for the corresponding eigenvector x we have $Cx = 0$.*

Proof. If $Cx = 0$ then λ is obviously purely imaginary. Conversely, let $\lambda = i\omega$ with ω real i.e.

$$-\omega^2 Mx + i\omega Cx + Kx = 0$$

hence

$$-\omega^2 x^* Mx + i\omega x^* Cx + x^* Kx = 0.$$

By taking the imaginary part we have $x^*Cx = 0$ and, by the positive semidefiniteness of C, $Cx = 0$. Q.E.D.

The presence of purely imaginary eigenvalues allows to split the system into an orthogonal sum of an undamped system and an exponentially decaying one. We have

Proposition 15.6 *Whenever there are purely imaginary eigenvalues there is a non-singular Φ such that*

$$\Phi^* Q(\lambda) \Phi = \begin{bmatrix} K_1(\lambda) & 0 \\ 0 & K_2(\lambda) \end{bmatrix},$$

$$K_1(\lambda) = \lambda^2 M_1 + K_1, \quad K_2(\lambda) = \lambda^2 M_2 + \lambda C_2 + K_2$$

where the system M_2, C_2, K_2 has no purely imaginary eigenvalues.

Proof. Take any eigenvector of $Q(\cdot)$ for an eigenvalue $i\omega$. Then by completing it to an M-orthonormal basis in $\varXi^n$ we obtain a matrix $\varPhi_1$ such that

$$\varPhi_1^* Q(\lambda) \varPhi_1 = \begin{bmatrix} \lambda^2 - \omega^2 & 0 \\ 0 & \hat{K}(\lambda) \end{bmatrix}$$

and, if needed, repeat the procedure on

$$\hat{K}(\lambda) = \lambda^2 \hat{M} + \lambda \hat{C} + \hat{K}.$$

and finally set $\varPhi = \varPhi_1 \varPhi_2 \cdots$. Q.E.D.

Corollary 15.7 *The phase-space matrix A from (3.3) or (4.19) is asymptotically stable, if and only if the form x^*Cx is positive definite on any eigenspace of the undamped system. Any system which is not asymptotically stable is congruent to an orthogonal sum of an undamped system and an asymptotically stable one.*

Proof. According to Proposition 15.5 A is asymptotically stable, if and only if $Cx \neq 0$ for any non-vanishing x from any undamped eigenspace. By the positive semidefiniteness of C the relation $Cx \neq 0$ is equivalent to $x^*Cx > 0$. The last statement follows from Proposition 15.6. Q.E.D.

If $i\omega$ is an undamped eigenvalue of multiplicity m then C must have rank at least m in order to move it completely into the left half-plane. (That eigenvalue may then, of course, split into several complex eigenvalues.) In particular, if C has rank one then the asymptotic stability can be achieved, only if all undamped eigenvalues are simple.

In spite of their elegance and intuitiveness the considerations above give no further information on the exponential decay of e^{At}.

Exercise 15.8 *Show that the system from Example 1.1 is asymptotically stable, if any of c_i or d_i is different from zero.*

Chapter 16
Spectral shift

Spectral shift is a simple and yet powerful tool to obtain informations on the decay in time of a damped system.

If in the differential equation

$$\dot{y} = Ay$$

we substitute

$$y = e^{\mu t}u, \quad \dot{y} = \mu e^{\mu t}u + e^{\mu t}\dot{u}$$

then we obtain the 'spectral-shifted' system

$$\dot{u} = \tilde{A}u, \quad \tilde{A} = A - \mu I.$$

Now, if we know that $\tilde{A}$ has an exponential which is bounded in time and $\mu < 0$ this property would yield an easy decay bound

$$\|e^{At}\| \le e^{\mu t} \sup_{t \ge 0} \|e^{\tilde{A}t}\|.$$

In the case of our phase-space matrix A it does not seem to be easy to find or to estimate the quantity $\sup_{s \ge 0} \|e^{\tilde{A}t}\|$ just from the spectral properties of A. However, if an analogous substitution is made on the second order system (1.1) the result is much more favourable. So, by setting $x = e^{\mu t}z$ and using

$$\dot{x} = \mu e^{\mu t}z + e^{\mu t}\dot{z}$$

$$\ddot{x} = \mu^2 e^{\mu t}z + 2\mu e^{\mu t}\dot{z} + e^{\mu t}\ddot{z}$$

the system (1.1) (with $f \equiv 0$) goes over into

$$M\ddot{z} + C(\mu)\dot{z} + Q(\mu)z = 0$$

with $Q(\mu)$ from (14.4) and

$$C(\mu) = K'(\mu) = 2\mu M + C.$$

As long as $C(\mu), Q(\mu)$ stay positive definite this is equivalent to the phase-space representation

$$\dot{w} = \hat{A}w, \quad \hat{A} = \begin{bmatrix} 0 & Q(\mu)^{1/2}M^{-1/2} \\ -M^{-1/2}Q(\mu)^{1/2} & -M^{-1/2}C(\mu)M^{-1/2} \end{bmatrix}, \tag{16.1}$$

$$w = \begin{bmatrix} Q(\mu)^{1/2}z \\ M^{1/2}\dot{z} \end{bmatrix},$$

whose solution is given by $e^{\hat{A}t}$. We now connect the representations (16.1) and (3.2). We have

$$y_1 = K^{1/2}x = e^{\mu t}K^{1/2}z = \mu e^{\mu t}K^{1/2}Q(\mu)^{-1/2}w_1,$$

$$y_2 = M^{1/2}\dot{x} = \mu e^{\mu t}M^{1/2}Q(\mu)^{-1/2}w_1 + e^{\mu t}w_2.$$

Thus,

$$y = e^{\mu t}\mathcal{L}(\mu)w,$$

$$\mathcal{L}(\mu) = \begin{bmatrix} K^{1/2}Q(\mu)^{-1/2} & 0 \\ \mu M^{1/2}Q(\mu)^{-1/2} & I \end{bmatrix}.$$

This, together with the evolution equations for y, w gives

$$\mathcal{L}(\mu)\hat{A} = (A - \mu I)\,\mathcal{L}(\mu) \tag{16.2}$$

or, equivalently

$$e^{At} = \mathcal{L}^{-1}(\mu)e^{\hat{A}+\mu t I}\mathcal{L}(\mu).$$

Hence an elegant decay estimate

$$\|e^{At}\| \le \|\mathcal{L}(\mu)\|\|\mathcal{L}(\mu)^{-1}\|e^{\mu t}. \tag{16.3}$$

This brings exponential decay only, if there are such $\mu < 0$ for which $C(\mu), Q(\mu)$ stay positive definite, in fact, $C(\mu)$ may be only positive semidefinite. By continuity, such μ will exist, if and only if C is positive definite. That is, *positive definite C insures the exponential decay.*

The next task is to establish the optimal decay factor μ in (16.3). We set

$$\gamma = \max_{x \ne 0} \operatorname{Re} p_+(x) = \max \operatorname{Re} w(M, C, K),$$

this maximum is obtained because $w(M, C, K)$ is compact.

Theorem 16.1 *Let all of M, C, K be all positive definite. Then*

$$\gamma \ge -\inf_x \frac{x^*Cx}{2x^*Mx}, \tag{16.4}$$

moreover, γ is the infimum of all μ for which both $Q(\mu)$ and $C(\mu)$ are positive definite.

Proof. The relation (16.4) is obvious. To prove the second assertion take any $\mu > \gamma$. Then

$$x^*Q(\mu)x = \tag{16.5}$$

$$x^*Mx(\mu - p_-(x))(\mu - p_+(x)) \ge x^*Mx(\mu - \operatorname{Re} p_+(x))^2 \ge x^*Mx(\mu - \gamma)^2$$

(note that $p_-(x) \le p_+(x)$ whenever $\Delta(x) \ge 0$). Thus, $Q(\mu)$ is positive definite. By (16.4) $C(\mu)$ is positive definite also. By continuity, both $Q(\gamma)$ and $C(\gamma)$ are positive semidefinite. Conversely, suppose that, say, $C(\gamma)$ is positive definite; we will show that $Q(\gamma)$ must be singular. Indeed, take a unit vector x_0 such that

$$\operatorname{Re} p_+(x_0) = \gamma.$$

Then by inserting in (16.5) $\mu = \gamma$ we obtain $x^*Q(\gamma)x_0 = 0$, i.e. $Q(\gamma)$ is singular. Q.E.D.

The previous theorem gives a simple possibility to compute γ by iterating the Cholesky decomposition of $Q(\mu)$ on the interval

$$-\inf_x \frac{x^*Cx}{2x^*Mx} \le \mu < 0$$

and using bisection.

Exercise 16.2 *Show that the constants in the bound (16.3) do not depend on the transformation of coordinates of type (2.19).*

Exercise 16.3 *Compute the constants in (16.3) for the matrices A from the Example 9.3. Show that they are poor in the overdamped case and try to improve them. Hint: reduce A to the diagonal form.*

Exercise 16.4 *Compute the constants in (16.3) for the matrices A from the Example 12.22.*

Exercise 16.5 *If $Q(\mu)$ is negative definite for some μ derive a formula analogous to (16.2) in which $\hat{A}$ is symmetric.*

Exercise 16.6 *Investigate the behaviour of the departure from normality of the matrix $\hat{A}$ from (16.1) as a function of μ.*

Solution. The function $\eta(A)$ is shift invariant; this and the fact that the eigenvalues of $A - \mu I$ and $\hat{A}$ coincide imply

$$\eta(\hat{A}) - \eta(A) = \|\hat{A}\|_E^2 - \|A - \mu I\|_E^2 =$$

$$2\operatorname{Tr}\left(M^{-1}Q(\mu)\right) + \operatorname{Tr}\left(M^{-1}C(\mu)\right) -$$

$$
\begin{aligned}
&-\mathrm{Tr}\left(\mu^2 I\right) - 2\mathrm{Tr}\left(M^{-1}K\right) - \mathrm{Tr}\left((\mu I + M^{-1/2}CM^{-1/2})^2\right)\\
&= 2\mathrm{Tr}\left(\mu^2 I + \mu M^{-1}C + M^{-1}K\right) + \mathrm{Tr}\left(4\mu^2 I + 4\mu M^{-1}C + (M^{-1}C)^2\right)\\
&-\mathrm{Tr}\left(4\mu^2 I\right) - 2\mathrm{Tr}\left(M^{-1}K\right) - \mathrm{Tr}\left(\mu^2 I + 2\mu M^{-1}C + (M^{-1}C)^2\right)\\
&= 4n\mu^2 + 4\mu\mathrm{Tr}\left(M^{-1/2}CM^{-1/2}\right)
\end{aligned}
$$

which is minimal at

$$
\mu = \mu_0 = -\frac{\mathrm{Tr}\left(M^{-1/2}CM^{-1/2}\right)}{2n} = -\frac{\gamma_1 + \cdots + \gamma_n}{2n} \le \gamma
$$

where $\gamma_1, \ldots, \gamma_n$ are the eigenvalues of $M^{-1/2}CM^{-1/2}$ and the last inequality is due to (16.4).

Chapter 17
Resonances and resolvents

We have thus far omitted considering the inhomogeneous differential equation (1.1) since its solution can be obtained from the one of the homogeneous equation. There are, however, special types of right hand side f in (1.1) the solution of which is particularly simply obtained and can, in turn, yield valuable information on the damped system itself. So they deserve a closer study.

Suppose that the external force f is *harmonic*, that is,

$$f(t) = f_a \cos \omega t + f_b \sin \omega t \tag{17.1}$$

where ω is a real constant and f_a, f_b are real constant vectors. So called steady-state vibrations in which the system is continuously excited by forces whose amount does not vary much in time are oft approximated by harmonic forces.

With the substitution

$$x(t) = x_a \cos \omega t + x_b \sin \omega t \tag{17.2}$$

(1.1) gives the linear system (here we keep M, C, K real)

$$\begin{bmatrix} -\omega^2 M + K & \omega C \\ -\omega C & -\omega^2 M + K \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix} = \begin{bmatrix} f_a \\ f_b \end{bmatrix}. \tag{17.3}$$

The function (17.2) is called *the harmonic response* to the harmonic force (17.1). By introducing complex quantities

$$x_0 = x_a - i x_b, \quad f_0 = f_a - i f_b$$

the system (17.3) is immediately seen to be equivalent to

$$Q(i\omega) x_0 = f_0. \tag{17.4}$$

By Proposition 15.5 we have the alternative

- The system is asymptotically stable or, equivalently, the system (17.4) is uniquely solvable.
- The system (17.4) becomes singular for some $\omega \in \sigma(\Omega)$, Ω from (2.5).

Exercise 17.1 *Show that $Q(i\omega)$ is non-singular whenever $\omega \notin \sigma(\Omega)$.*

Exercise 17.2 *Show that (17.4) is equivalent to*

$$(i\omega I - A)y_0 = F_0 \tag{17.5}$$

with

$$y_0 = \begin{bmatrix} L_1^T x_0 \\ i\omega L_2^T x_0 \end{bmatrix}, \quad F_0 = \begin{bmatrix} 0 \\ L_2^{-1} f_0 \end{bmatrix}$$

and $y(t) = y_0 e^{i\omega t}$ is a solution of $\dot{y} = Ay + F_0 e^{i\omega t}$.

Exercise 17.3 *Which harmonic response corresponds to a general right hand side vector F_0 in (17.5)?*

From (17.5) we see that the harmonic response in the phase-space is given by the resolvent of A taken on the imaginary axis. In analogy we call the function $Q(\lambda)^{-1}$ *the resolvent* of the quadratic pencil (14.4) or simply *the quadratic resolvent.* In the damping-free case $Q(i\omega)$ will be singular at every undamped frequency $\omega = \omega_j$.

Proposition 17.4 *If the system is not asymptotically stable then there is an $f(t)$ such that (1.1) has an unbounded solution.*

Proof. There is a purely imaginary eigenvalue $i\omega$ and a real vector $x_0 \neq 0$ such that $Kx_0 = \omega^2 M x_0$ and $Cx_0 = 0$. Take $f(t) = \alpha x_0 e^{i\omega t}$ and look for a solution $x(t) = \xi(t)x_0$. Substituted in (1.1) this gives

$$\ddot{\xi} + \omega^2 \xi = \alpha e^{i\omega t}$$

which is known to have unbounded solutions. Q.E.D.

Exercise 17.5 *Prove the identity*

$$(\lambda I - A)^{-1} = \begin{bmatrix} \frac{1}{\lambda} I - \frac{L_1^T Q(\lambda)^{-1} L_1}{\lambda} & L_1^T Q(\lambda)^{-1} L_2 \\ -L_2^T Q(\lambda)^{-1} L_1 & L_2^T Q(\lambda)^{-1} L_2 \end{bmatrix}. \tag{17.6}$$

Exercise 17.6 *Show that the singularity at $\lambda = 0$ in (17.6) is removable by conveniently transforming the* 1, 1- *block.*

Any system in which the frequency of the harmonic force corresponds to a purely imaginary eigenvalue as in Proposition 17.4 is said to be in the state of *resonance.* The same term is used, if an eigenvalue of (14.4) is close to the imaginary axis. In this case the amplitude is expected to have a high peak at the point ω equal to the imaginary part of the close eigenvalue. In practice, high peaks are just as dangerous as true singularities; both result in large displacements which may break the structure or, to alter the elasticity properties of the material with equally disastrous consequences. In any case, large displacements do not warrant the use of a linearised theory anymore.

Given a linear system of equations like that in (17.4) one is interested in its accurate numerical solution which involves the condition number of the matrix $Q(i\omega)$. One would wish to have an efficient algorithm exploiting this structure. Remember that the non-singularity of these matrices, as that of $Q(i\omega)$ is generally a consequence of a rather subtle interplay of M, C, K. Most interesting values of ω lie between ω_1 and ω_n, so $K - \omega^2 M$ will typically be a non-definite symmetric matrix.

Exercise 17.7 *Show that the condition number of the coefficient matrix in (17.3) equals that of $Q(i\omega)$.*

We can measure the size of the harmonic response by the square of its norm, averaged over the period $\tau = 2\pi/\omega$ which equals

$$\frac{1}{\tau}\int_0^\tau \|x(t)\|^2 dt = \frac{\|x_a\|^2 + \|x_b\|^2}{2} = \frac{\|x_0\|^2}{2}$$

$$= \frac{f_0^* Q(-i\omega)^{-1} Q(i\omega)^{-1} f_0}{2}.$$

The square root of this quantity is *the average displacement amplitude.* Another measure is the total energy

$$\frac{1}{\tau}\int_0^\tau \frac{\|y(t)\|^2}{2} dt = \frac{x_0^*(\omega^2 M + K)x_0}{2} =$$

$$= \frac{f_0^* Q(-i\omega)^{-1}(\omega^2 M + K) Q(i\omega)^{-1} f_0}{2}.$$

This is the *average energy amplitude.*

Exercise 17.8 *Compute the harmonic response of a modally damped system and its average energy amplitude. Hint: go over to the transformed system with the transformation matrix Φ from (3.10).*

The free oscillations (those with $f = 0$) are also called *transient* oscillations whereas those with harmonic or similar forces which produce responses not largely varying in time are called *steady state* oscillations.

Chapter 18
Well-posedness

Having studied the asymptotic stability in previous chapters we are now in a position to rigorously state the problem of well-posedness, that is, the dependence of the solution of a damped system on the right hand side as well as on the initial conditions.

We will say that the system (1.1) or, equivalently its phase-space formulation (3.2) is *well posed*, if there is a constant $\mathcal{C}$ such that

$$\sup_{t\geq 0}\|y(t)\| \leq \mathcal{C}(\sup_{t\geq 0}\|g(t)\| + \|y_0\|) \tag{18.1}$$

for all initial data y_0.

Theorem 18.1 *The system (3.2) is well-posed, if and only if it is asymptotically stable.*

Proof. Let the system be asymptotically stable. From (3.1), (13.8) as well as the contractivity of e^{At} we have

$$\|y(t)\| \leq \|y_0\| + Fe^{-\nu t}\sup_{t\geq 0}\|g(t)\|\int_0^t e^{\nu\tau}d\tau =$$

$$\|y_0\| + F\sup_{t\geq 0}\|g(t)\|\frac{1-e^{\nu t}}{\nu},$$

so (18.1) holds with $\mathcal{C} = 1 + F/\nu$. Conversely, if the system is not asymptotically stable then Proposition 17.4 provides an unbounded solution. Q.E.D.

Of course, in applications the mere existence of the constant $\mathcal{C} = 1 + F/\nu$ is of little value, if one does not know its size. To this end the bounds (13.10) and (16.3) may prove useful.

If the system is not asymptotically stable we may modify the definition of well-posedness so that *only the homogeneous system* is considered in which case no $g(t)$ appears in (18.1). Then any damped system is trivially well-posed with $\mathcal{C} = 1$.

Chapter 19
Modal approximation

Some, rather rough, facts on the positioning of the eigenvalues were given in Chapter 15. Further, more detailed, information can be obtained, if the system is close to one with known properties. Such techniques go under the common name of 'perturbation theory'. The simplest 'thoroughly known' system is the undamped one. Next to this lie the modally damped systems which were studied in Chapter 2. Both cases are the subject of the present chapter. In particular, estimates in the sense of Gershgorin will be obtained.

A straightforward eigenvalue estimate for a general matrix A close to a matrix A_0 is based on the invertibility of any matrix $I + Z$ with $\|Z\| < 1$. Thus,

$$A - \lambda I = \left(I + (A - A_0)(A_0 - \lambda I)^{-1}\right)(A_0 - \lambda I)$$

which leads to

$$\sigma(A) \subseteq \mathcal{G}_1 = \{\lambda : \ \|(A - A_0)(A_0 - \lambda I)^{-1}\| < 1\} \tag{19.1}$$

Obviously $\mathcal{G}_1 \subseteq \mathcal{G}_2$ with

$$\mathcal{G}_2 = \{\lambda : \ \|(A_0 - \lambda I)^{-1}\|^{-1} \leq \|A - A_0\|\}.$$

where we have used the obvious estimate

$$\|(A - A_0)(A_0 - \lambda I)^{-1}\| \leq \|A - A_0\| \|(A_0 - \lambda I)^{-1}\|.$$

If A is block-diagonal

$$A = \operatorname{diag}(A_{11}, \ldots, A_{pp})$$

then $\|(A_0 - \lambda I)^{-1}\|$ simplifies to $\max_i \|(A_{ii} - \lambda I_{n_i})^{-1}\|$ and

$$\sigma(A) \subseteq \{\lambda : \ \max_i \|(A_{ii} - \lambda I_i)^{-1}\| \leq \|A - A_0\|\}.$$

This is valid for any matrices A, A_0. We will first consider the phase-space matrix A from (3.10). The matrix A_0 (this is the undamped approximation) is obtained from A by setting $D = 0$. Thus,

$$A - A_0 = \begin{bmatrix} 0 & 0 \\ 0 & -D \end{bmatrix}.$$

The matrix A_0 is skew-symmetric and therefore normal, so $\|(A_0 - \lambda I)^{-1}\|^{-1} = \operatorname{dist}(\lambda, \sigma(A_0))$ hence

$$\mathcal{G}_2 = \{\lambda : \ \operatorname{dist}(\lambda, \sigma(A_0)) \le \|A - A_0\|\} \tag{19.2}$$

where

$$\|A - A_0\| = \|D\| = \|L_2^{-1} C L_2^{-T}\| = \max \frac{x^T C x}{x^T M x} \tag{19.3}$$

is the largest eigenvalue of the matrix pair C, M. We may say that here 'the size of the damping is measured relative to the mass'.

Thus, the perturbed eigenvalues are contained in the union of the disks of radius $\|D\|$ around $\sigma(A_0)$.

Exercise 19.1 *Show that $\sigma(A)$ is also contained in the union of the disks*

$$\{\lambda : \ |\lambda \mp i\omega_j| \le R_j\}$$

where ω_j are the undamped frequencies and

$$R_j = \sum_{k=1}^{n} |d_{kj}|. \tag{19.4}$$

Hint: Replace the spectral norm in (19.1) by the norm $\|\cdot\|_1$.

The bounds obtained above are, in fact, too crude, since we have not taken into account the structure of the perturbation $A - A_0$ which has so many zero elements.

Instead of working in the phase-space we may turn back to the original quadratic eigenvalue problem in the representation in the form (see (2.4) and (3.10))

$$\det(\lambda^2 I + \lambda D + \Omega^2) = 0.$$

The inverse

$$(\lambda^2 I + \lambda D + \Omega^2)^{-1} =$$
$$(\lambda^2 I + \Omega^2)^{-1}(I + \lambda D(\lambda^2 I + \Omega^2)^{-1})^{-1}$$

exists, if

$$\|D(\lambda^2 I + \Omega^2)^{-1}\||\lambda| < 1$$

which is implied by

$$\|(\lambda^2 I + \Omega^2)^{-1}\|\|D\||\lambda| = \frac{\|D\||\lambda|}{\min_j(|\lambda - i\omega_j||\lambda + i\omega_j|)} < 1.$$

Thus,

$$\sigma(A) \subseteq \cup_j \mathcal{C}(i\omega_j, -i\omega_j, \|D\|), \tag{19.5}$$

where the set

$$\mathcal{C}(\lambda_+, \lambda_-, r) = \{\lambda : |\lambda - \lambda_+||\lambda - \lambda_-| \le |\lambda| r\} \tag{19.6}$$

will be called *stretched Cassini ovals* with foci $\lambda_\pm$ and extension r. This is in analogy with the standard Cassini ovals where on the right hand side instead of $|\lambda|r$ one has just r^2. (The latter also appear in eigenvalue bounds in somewhat different context.) We note the obvious relation

$$\mathcal{C}(\lambda_+, \lambda_-, r) \subset \mathcal{C}(\lambda_+, \lambda_-, r'), \quad \text{whenever } r < r'. \tag{19.7}$$

The stretched Cassini ovals are qualitatively similar to the standard ones; they can consist of one or two components; the latter case occurs when r is sufficiently small with respect to $|\lambda_+ - \lambda_-|$. In this case the ovals in (19.5) are approximated by the disks

$$|\lambda \pm i\omega_j| \le \frac{\|D\|}{2}$$

and this is one half of the bound in (19.2), (19.3).

Exercise 19.2 *Show that $\sigma(A)$ is also contained in the union of the ovals*

$$\mathcal{C}(i\omega_j, -i\omega_j, \|\Omega^{-1} D \Omega^{-1}\| \omega_j^2).$$

Hint: Instead of inverting $\lambda^2 I + \lambda D + \Omega^2$ invert $\lambda^2 \Omega^{-2} + \lambda \Omega^{-1} D \Omega^{-1} + I$.

Exercise 19.3 *Show that $\sigma(A)$ is contained in the union of the ovals*

$$\mathcal{C}(i\omega_j, -i\omega_j, R_j)$$

and also

$$\mathcal{C}(i\omega_j, -i\omega_j, \rho_j \omega_j^2)$$

with

$$\rho_j = \sum_{\substack{k=1 \\ k \ne j}}^{n} \frac{|d_{kj}|}{\omega_k \omega_j}.$$

Exercise 19.4 *Show that any of the inequalities*

$$\min_j (2\omega_j - R_j) > 0, \tag{19.8}$$

$$\min_j 2\omega_j - \|D\| > 0,$$

with R_j from (19.4) and

$$\min_j \frac{2}{\omega_j} - \|\Omega^{-1} D \Omega^{-1}\| > 0$$

precludes real eigenvalues of (14.2).

Solution. If an eigenvalue λ is real then according to Exercise 19.3 we must have

$$\lambda^2 + \omega_j^2 \leq |\lambda| R_j$$

for some j and this implies $R_j^2 - 4\omega_j^2 \geq 0$. Thus, under (19.8) no real eigenvalues can occur (and similarly in other two cases).

The just considered undamped approximation was just a prelude to the main topic of this chapter, namely the modal approximation. The modally damped systems are so much simpler than the general ones that practitioners often substitute the true damping matrix by some kind of 'modal approximation'. Most typical such approximations in use are of the form

$$C_{prop} = \alpha M + \beta K \tag{19.9}$$

where α, β are chosen in such a way that C_{prop} be in some sense as close as possible to C, for instance,

$$\mathrm{Tr}\,[(C - \alpha M - \beta K) W (C - \alpha M - \beta K)] = \min,$$

where W is some convenient positive definite weight matrix. This is a *proportional approximation.* In general such approximations may go quite astray and yield thoroughly false predictions. We will now assess them in a more systematic way.

A modal approximation to the system (1.1) is obtained by first representing it in modal coordinates by the matrices D, Ω and then by replacing D by its diagonal part

$$D^0 = \mathrm{diag}(d_{11}, \ldots, d_{nn}). \tag{19.10}$$

The off-diagonal part

$$D' = D - D^0 \tag{19.11}$$

is considered a perturbation. Again we can work in the phase-space or with the original quadratic eigenvalue formulation. In the first case we can make perfect shuffling to obtain

$$A = (A_{ij}), \quad A_{ii} = \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -d_{ii} \end{bmatrix}, \quad A_{ij} = \begin{bmatrix} 0 & 0 \\ 0 & -d_{ij} \end{bmatrix} \tag{19.12}$$

$$A_0 = \mathrm{diag}(A_{11}, \ldots, A_{nn}).$$

So, for $n = 3$

$$A = \left[\begin{array}{cc|cc|cc} 0 & \omega_1 & 0 & 0 & 0 & 0 \\ -\omega_1 & -d_{11} & 0 & -d_{12} & 0 & -d_{13} \\ \hline 0 & 0 & 0 & \omega_2 & 0 & 0 \\ 0 & -d_{12} & -\omega_2 & -d_{22} & 0 & -d_{23} \\ \hline 0 & 0 & 0 & 0 & 0 & \omega_3 \\ 0 & -d_{13} & 0 & -d_{23} & -\omega_3 & -d_{33} \end{array}\right].$$

Then

$$\|(A_0 - \lambda I)^{-1}\|^{-1} = \max_j \|(A_{jj} - \lambda I_j)^{-1}\|^{-1}.$$

Even for 2×2-blocks any common norm of $(A_{jj} - \lambda I_j)^{-1}$ seems complicated to express in terms of disks or other simple regions, unless we diagonalise each A_{jj} as

$$S_j^{-1} A_{jj} S_j = \begin{bmatrix} \lambda_+^j & 0 \\ 0 & \lambda_-^j \end{bmatrix}, \quad \lambda_\pm^j = \frac{-d_{jj} \pm \sqrt{d_{jj}^2 - 4\omega_j^2}}{2}. \tag{19.13}$$

As we know from Example 9.3 we have

$$\kappa(S_j) = \sqrt{\frac{1 + \mu_j^2}{|1 - \mu_j^2|}}, \quad \mu_j = \frac{d_{jj}}{2\omega_j}.$$

Set $S = \operatorname{diag}(S_{11}, \dots, S_{nn})$ and

$$A' = S^{-1} A S = A_0' + A'', \quad A'' = S^{-1}(A - A_0)S$$

then

$$A_0' = \operatorname{diag}(\lambda_\pm^1, \dots, \lambda_\pm^n),$$
$$A_{jk}'' = S_j^{-1} A_{jk} S_k.$$

Now the general perturbation bound (19.2), applied to A_0', A'', gives

$$\sigma(A) \subseteq \cup_{j,\pm} \{\lambda : \ |\lambda - \lambda_\pm^j| \le \kappa(S) \|D'\|\}. \tag{19.14}$$

There is a related 'Gershgorin-type bound'

$$\sigma(A) \subseteq \cup_{j,\pm} \{\lambda : \ |\lambda - \lambda_\pm^j| \le \kappa(S_j) r_j\} \tag{19.15}$$

with

$$r_j = \sum_{\substack{k=1 \\ k \ne j}}^{n} \|d_{jk}\|. \tag{19.16}$$

To show this we replace the spectral norm $\| \cdot \|$ in (19.1) by the norm $||| \cdot |||_1$, defined as

$$|||A|||_1 := \max_j \sum_k \|A_{kj}\|$$

where the norms on the right hand side are spectral. Thus, (19.1) will hold, if

$$\max_j \sum_k \|(A - A_0)_{kj}\| \|(A_{jj} - \lambda I_j)^{-1}\| < 1.$$

Taking into account the equality

$$\|(A - A_0)_{kj}\| = \begin{cases} |d_{kj}|, \; k \neq j \\ \quad 0 \; k = j \end{cases}$$

$\lambda \in \sigma(A)$ implies

$$r_j \geq \|(A_{jj} - \lambda I)^{-1}\| \geq \frac{\min\{|\lambda - \lambda_+^j|, |\lambda - \lambda_-^j|\}}{\kappa(S_j)}$$

and this is (19.15).

The bounds (19.14) and (19.15) are poor whenever the modal eigenvalue approximation is close to a critically damped eigenvalue. Better bounds are expected, if we work directly with the quadratic eigenvalue equation. The inverse

$$(\lambda^2 I + \lambda D + \Omega^2)^{-1} =$$
$$(\lambda^2 I + \lambda D^0 + \Omega^2)^{-1}(I + \lambda D'(\lambda^2 I + \lambda D^0 + \Omega^2)^{-1})^{-1}$$

exists, if

$$\|D'(\lambda^2 I + \lambda D^0 + \Omega^2)^{-1}\| |\lambda| < 1,$$

D' from (19.11), which is insured if

$$\|(\lambda^2 I + \lambda D^0 + \Omega^2)^{-1}\| \|D'\| |\lambda| = \frac{\|D'\| |\lambda|}{\min_j(|\lambda - \lambda_+^j| |\lambda - \lambda_-^j|)} < 1.$$

Thus,

$$\sigma(A) \subseteq \cup_j \mathcal{C}(\lambda_+^j, \lambda_-^j, \|D'\|). \tag{19.17}$$

These ovals will always have both foci either real or complex conjugate. If $r = \|D'\|$ is small with respect to $|\lambda_+^j - \lambda_-^j| = \sqrt{|d_{jj}^2 - 4\omega_j^2|}$ then either $|\lambda - \lambda_+^j|$ or $|\lambda - \lambda_-^j|$ is small. In the first case the inequality $|\lambda - \lambda_+^j||\lambda - \lambda_-^j| \leq |\lambda| r$ is approximated by

$$|\lambda - \lambda_+^j| \leq \frac{|\lambda_+^j| r}{|\lambda_+^j - \lambda_-^j|} = r \begin{cases} \frac{\omega_j}{\sqrt{d_{jj}^2 - 4\omega_j^2}} & d_{jj} < 2\omega_j \\ \\ \frac{d_{jj} - \sqrt{d_{jj}^2 - 4\omega_j^2}}{\sqrt{d_{jj}^2 - 4\omega_j^2}} & d_{jj} > 2\omega_j \end{cases}$$

and in the second

$$|\lambda - \lambda_-^j| \le \frac{|\lambda_-^j| r}{|\lambda_+^j - \lambda_-^j|} = r \begin{cases} \frac{\omega_j}{\sqrt{d_{jj}^2 - 4\omega_j^2}} & d_{jj} < 2\omega_j \\ \\ \frac{d_{jj} + \sqrt{d_{jj}^2 - 4\omega_j^2}}{\sqrt{d_{jj}^2 - 4\omega_j^2}} & d_{jj} > 2\omega_j \end{cases}.$$

This is again a union of disks. If $d_{jj} \approx 0$ then their radius is $\approx r/2$. If $d_{jj} \approx 2\omega_j$ i.e. $\lambda_- = \lambda_+ \approx -d_{jj}/2$ the ovals look like a single circular disk. For large d_{jj} the oval around the absolutely larger eigenvalue is $\approx r$ (the same behaviour as with (19.15)) whereas the smaller eigenvalue has the diameter $\approx 2r\omega_j^2/d_{jj}^2$ which is drastically better than (19.15).

In the same way as before the Gershgorin type estimate is obtained

$$\sigma(A) \subseteq \cup_j \mathcal{C}(\lambda_+^j, \lambda_-^j, r_j). \tag{19.18}$$

We have called D^0 *a* modal approximation to D because the matrix D is not uniquely determined by the input matrices M, C, K. Different choices of the transformation matrix Φ give rise to different modal approximations D^0 but the differences between them are mostly non-essential. To be more precise, let Φ and $\tilde{\Phi}$ both satisfy (2.4). Then

$$M = \Phi^{-T}\Phi^{-1} = \tilde{\Phi}^{-T}\tilde{\Phi}^{-1},$$

$$K = \Phi^{-T}\Omega^2\Phi^{-1} = \tilde{\Phi}^{-T}\Omega^2\tilde{\Phi}^{-1}$$

implies that $U = \Phi^{-1}\tilde{\Phi}$ is an orthogonal matrix which commutes with Ω which we now write in the form (by labelling ω_j differently)

$$\Omega = \mathrm{diag}(\omega_1 I_{n_1}, \dots, \omega_s I_{n_s}), \quad \omega_1 < \cdots < \omega_s. \tag{19.19}$$

Denote by $D = (D_{ij})$ the corresponding partition of the matrix D and set

$$U^0 = \mathrm{diag}(U_{11}, \dots, U_{ss}),$$

where each U_{jj} is an orthogonal matrix of order n_j from (19.10). Now,

$$\tilde{D} := \tilde{\Phi}^T C \tilde{\Phi} = U^T \Phi^* C \Phi U = U^T D U,$$

$$\tilde{D}_{ij} = U_{ii}^T D_{ij} U_{jj}$$

and hence

$$\tilde{D}' = U^T D' U.$$

Now, if the undamped frequencies are all simple, then U is diagonal and the estimates (19.3) or (19.5) – (19.6) remain unaffected by this change of coordinates. Otherwise we replace $\mathrm{diag}(d_{11}, \dots, d_{nn})$ by

$$D^0 = \mathrm{diag}(D_{11}, \ldots, D_{ss}) \tag{19.20}$$

where D^0 commutes with Ω. In fact, a general definition of any *modal approximation* is that it

1. is block-diagonal part of D and
2. commutes with Ω.

The modal approximation with the coarsest possible partition — this is the one whose block dimensions equal the multiplicities in Ω — is called a *maximal modal approximation.* Accordingly, we say that $C^0 = \Phi^{-1} D^0 \Phi^{-T}$ is a modal approximation to C (and also M, C^0, K to M, C, K).

Proposition 19.5 *Each modal approximation to C is of the form*

$$C^0 = \sum_{k=1}^{s} P_k^* C P_k$$

where $P_1, \ldots . P_s$ is an M-orthogonal decomposition of the identity (that is $P_k^* = M P_k M^{-1}$) and P_k commute with the matrix

$$\sqrt{M^{-1}K} = M^{-1/2}\sqrt{M^{-1/2} K M^{-1/2}} M^{1/2}.$$

Proof. Use the formula

$$D^0 = \sum_{k=1}^{s} P_k^0 D P_k^0$$

with

$$P_k^0 = \mathrm{diag}(0, \ldots, I_{n_k}, \ldots, 0), \quad D = \Phi^T C \Phi, \quad D^0 = \Phi^T C^0 \Phi$$

and set $P_k = \Phi P_k^0 \Phi^{-1}$. Q.E.D.

Exercise 19.6 *Characterise the set of all maximal modal approximations, again in terms of the original matrices M, C, K.*

It is obvious that the maximal approximation is the best among all modal approximations in the sense that

$$\|D - D^0\|_E \le \|D - \hat{D}^0\|_E, \tag{19.21}$$

where

$$\hat{D}^0 = \mathrm{diag}(\hat{D}_{11}, \ldots, \hat{D}_{zz}) \tag{19.22}$$

and $D = (\hat{D}_{ij})$ is any block partition of D which is finer than that in (19.20). It is also obvious that D^0 is better than any diagonal damping matrix e.g. the one given by (19.9).

We will now prove that the inequality (19.21) is valid for the spectral norm also. We shall need the following

Proposition 19.7 *Let $H = (H_{ij})$ be any partitioned Hermitian matrix such that the diagonal blocks H_{ii} are square. Set*

$$H^0 = \operatorname{diag}(H_{11}, \dots, H_{ss}), \quad H' = H - H^0.$$

Then

$$\lambda_k(H) - \lambda_n(H) \le \lambda_k(H') \le \lambda_k(H) - \lambda_1(H) \tag{19.23}$$

where $\lambda_k(\cdot)$ denotes the non-decreasing sequence of the eigenvalues of any Hermitian matrix.

Proof. The estimate (2.14) yields

$$\lambda_k(H) - \max_j \max \sigma(H_{jj}) \le \lambda_k(H') \le \lambda_k(H) - \min_j \min \sigma(H_{jj}).$$

By the interlacing property,

$$\lambda_1(H) \le \sigma(H_{jj}) \le \lambda_n(H).$$

Altogether we obtain (19.23). Q.E.D.

From (19.23) some simpler estimates immediately follow:

$$\|H'\| \le \lambda_n(H) - \lambda_1(H) =: \operatorname{spread}(H) \tag{19.24}$$

and, if H is positive (or negative) semidefinite

$$\|H'\| \le \|H\|. \tag{19.25}$$

Hence, in addition to (19.21) we also have

$$\|D - D^0\| \le \|D - \hat{D}^0\|.$$

So, a best bound in (19.17) is obtained, if $D^0 = D - D'$ is a maximal modal approximation.

If D^0 is block-diagonal and the corresponding $D' = D - D^0$ is inserted in (19.17) then the values d_{jj} from (19.13) should be replaced by the corresponding eigenvalues of the diagonal blocks D_{jj}. But in this case we can further transform $\varOmega$ and D by a unitary similarity

$$U = \operatorname{diag}(U_1, \dots, U_s)$$

such that each of the blocks D_{jj} becomes diagonal ($\varOmega$ stays unchanged). With this stipulation we may retain the formula (19.17) unaltered. This shows that taking just the diagonal part D^0 of D covers, in fact, *all possible modal approximations*, when $\varPhi$ varies over all matrices performing (2.4).

Similar extension can be made to the bound (19.18) but then no improvements in general can be guaranteed although they are more likely to occur than not.

By the usual continuity argument it is seen that the number of eigenvalues in each component of $\cup_i \mathcal{C}(\lambda_+^i, \lambda_-^i, r_i)$ is twice the number of involved diagonals. In particular, if we have the maximal number of $2n$ components, then each of them contains exactly one eigenvalue.

A strengthening in the sense of Brauer ([8]) is possible here also. We will show that the *spectrum is contained in the union of double ovals*, defined as

$$\mathcal{D}(\lambda_+^p, \lambda_-^p, \lambda_+^q, \lambda_-^q, r_p r_q) =$$

$$\{\lambda : |\lambda - \lambda_+^p||\lambda - \lambda_-^p||\lambda - \lambda_+^q||\lambda - \lambda_-^q| \le r_p r_q |\lambda|^2\},$$

where the union is taken over all pairs $p \neq q$ and $\lambda_\pm^p$ are the solutions of $\lambda^2 + d_{pp}\lambda + \omega_p^2 = 0$ and similarly for $\lambda_\pm^q$. The proof just mimics the standard Brauer's one. The quadratic eigenvalue problem is written as

$$(\lambda^2 + \lambda d_{pp} + \omega_p^2) x_p = -\lambda \sum_{\substack{j=1 \\ j \neq p}}^n d_{pj} x_j, \tag{19.26}$$

$$(\lambda^2 + \lambda d_{qq} + \omega_q^2) x_q = -\lambda \sum_{\substack{j=1 \\ j \neq q}}^n d_{qj} x_j, \tag{19.27}$$

where $|x_p| \ge |x_q|$ are the two absolutely largest components of x. If $x_q = 0$ then $x_j = 0$ for all $j \neq p$ and trivially $\lambda \in \mathcal{D}(\lambda_+^p, \lambda_-^p, \lambda_+^q, \lambda_-^q, r_p r_q)$. If $x_q \neq 0$ then multiplying the equalities (19.26) and (19.27) yields

$$|\lambda - \lambda_+^p||\lambda - \lambda_-^p||\lambda - \lambda_+^q||\lambda - \lambda_-^q||x_p||x_q| \le$$

$$|\lambda|^2 \sum_{\substack{j=1 \\ j \neq p}}^n \sum_{\substack{k=1 \\ k \neq q}}^n |d_{pj}||d_{qk}||x_j||x_k|.$$

Because in the double sum there is no term with $j = k = p$ we always have $|x_j||x_k| \le |x_p||x_q|$, hence the said sum is bounded by

$$|\lambda|^2 |x_p||x_q| \sum_{\substack{j=1 \\ j \neq p}}^n |d_{pj}| \sum_{\substack{k=1 \\ k \neq q}}^n |d_{qk}|.$$

Thus, our inclusion is proved. As it is immediately seen, *the union of all double ovals is contained in the union of all stretched Cassini ovals.*

The simplicity of the modal approximation suggests to try to extend it to as many systems as possible. A close candidate for such extension is any system with tightly clustered undamped frequencies, that is, Ω is close to an Ω^0 from (19.19). Starting again with

$$(\lambda^2 I + \lambda D + \Omega^2)^{-1} =$$

$$(\lambda^2 I + \lambda D^0 + (\Omega^0)^2)^{-1}(I + (\lambda D' + Z) + (\lambda^2 I + \lambda D^0 + (\Omega^0)^2)^{-1})^{-1}$$

with $Z = \Omega^2 - (\Omega^0)^2$ we immediately obtain

$$\sigma(A) \subseteq \cup_j \hat{\mathcal{C}}(\lambda_+^j, \lambda_-^j, \|D'\|, \|Z\|). \tag{19.28}$$

where the set

$$\hat{\mathcal{C}}(\lambda_+, \lambda_-, r, q) = \{\lambda : |\lambda - \lambda_+||\lambda - \lambda_-| \leq |\lambda| r + q\}$$

will be called *modified Cassini ovals* with foci $\lambda_\pm$ and extensions r, q.

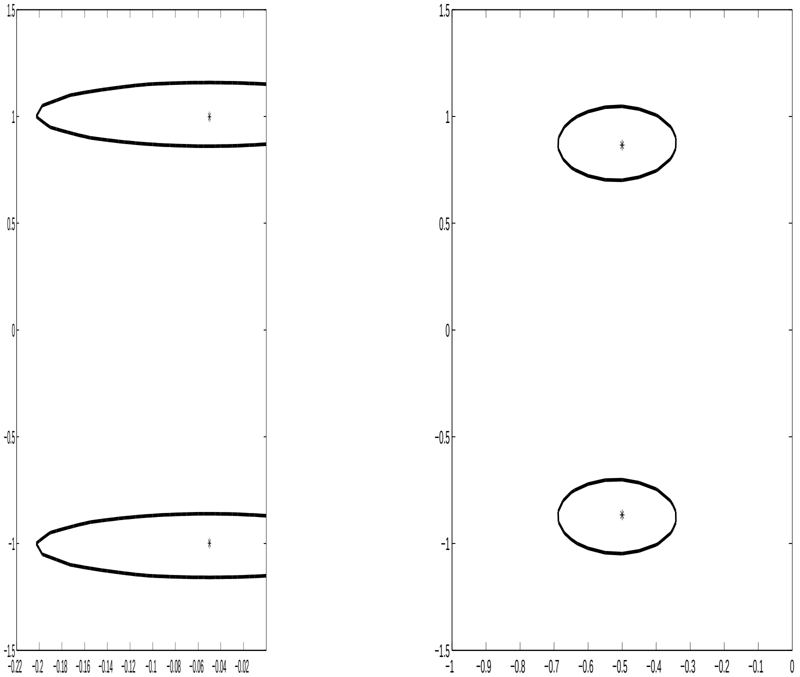

Fig. 19.1 Ovals for $\omega = 1; d = 0.1, 1;\ r = 0.3$

Remark 19.8 The basis of any modal approximation is the diagonalisation of the matrix pair M, K. Now, an analogous procedure with similar results can be performed by diagonalising the pair M, C or C, K. The latter would be recommendable in Example 1.8 because there the corresponding matrices

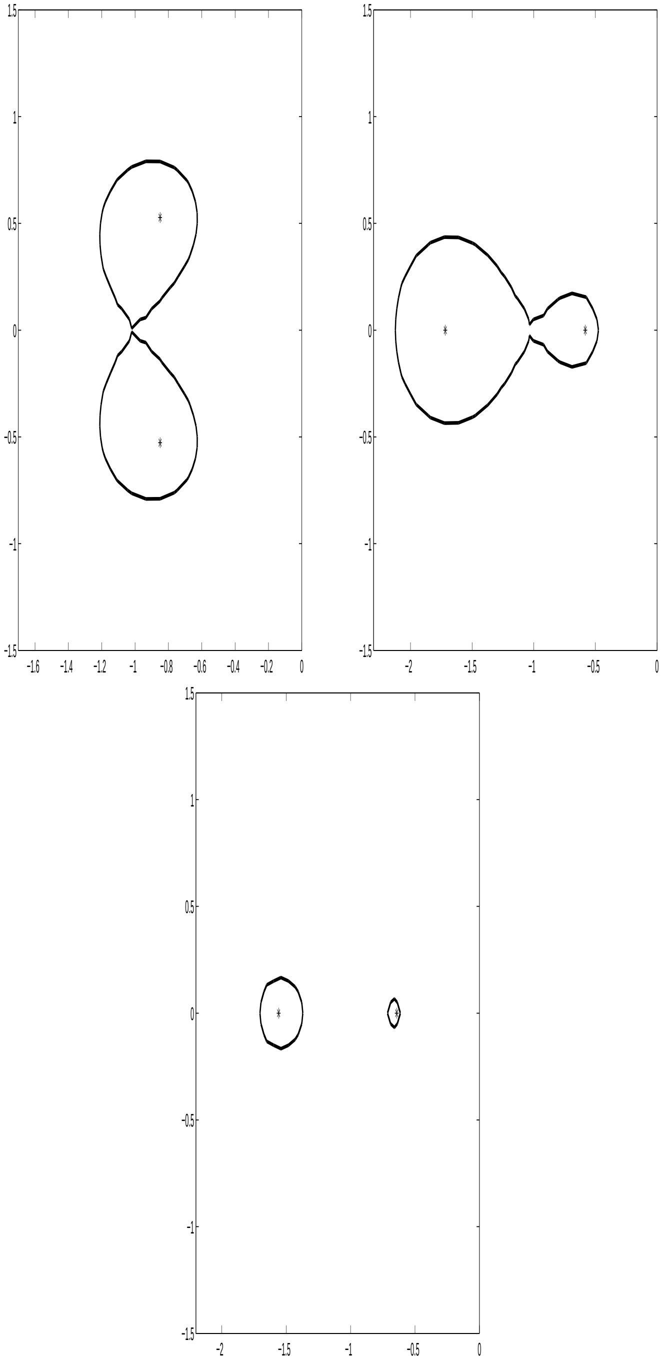

Fig. 19.2 Ovals for $\omega = 1$; $d = 1.7, 2.3, 2.2$; $r = 0.3, 0.3, 0.1$

$\mathcal{R}$ and $\mathcal{K}$ are already diagonal. There is a third possibility: to approximate the system M, C, K by some M_0, C_0, K_0 which is modally damped. Here we change *all three system matrices* and expect to get closer to the original system and therefore to obtain better bounds. For an approach along these lines see [46].

Exercise 19.9 *Try to apply the techniques of this chapter to give bounds to the harmonic response considered in Chapter 18.*

Chapter 20
Modal approximation and overdampedness

If the systems in the previous chapter are all overdamped then estimates are greatly simplified as complex regions become just intervals. But before going into this a more elementary — and more important — question arises: Can the modal approximation help to decide the overdampedness of a given system? After giving an answer to the latter question we will turn to the perturbation of the overdamped eigenvalues themselves.

We begin with some obvious facts the proofs of which are left to the reader.

Proposition 20.1 *If the system M, C, K is overdamped, then the same is true of the* projected system

$$M' = X^* M X, \quad C' = X^* C X, \quad K' = X^* K X$$

where X is any injective matrix. Moreover, the definiteness interval of the former is contained in the one of the latter.

Proposition 20.2 *Let*

$$M = \operatorname{diag}(M_{11}, \ldots, M_{ss})$$

$$C = \operatorname{diag}(C_{11}, \ldots, C_{ss})$$

$$K = \operatorname{diag}(K_{11}, \ldots, K_{ss}).$$

Then the system M, C, K is overdamped, if and only if each of the systems M_{jj}, C_{jj}, K_{jj} is overdamped and their definiteness intervals have a non trivial intersection (which is then the definiteness interval of M, C, K)

Corollary 20.3 *If the system M, C, K is overdamped, then the same is true of any of its modal approximations.*

Exercise 20.4 *If a maximal modal approximation is overdamped, then so are all others.*

In the following we shall need some well known sufficient conditions for negative definiteness of a general Hermitian matrix $A = (a_{ij})$; these are:

$$a_{jj} < 0$$

for all j and either

$$\|A - \mathrm{diag}(a_{11}, \ldots, a_{nn})\| < -\max_j a_{jj}$$

(norm-diagonal dominance) or

$$\sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{kj}| < -a_{jj} \text{ for all } j$$

(Gershgorin-diagonal dominance).

Theorem 20.5 *Let Ω, D, r_j be from (2.5), (2.22), (19.16), respectively and*

$$D^0 = \mathrm{diag}(d_{11}, \ldots, d_{nn}), \quad D' = D - D^0.$$

Let

$$\Delta_j = (d_{jj} - \|D'\|)^2 - 4\omega_j^2 > 0 \textit{ for all } j \tag{20.1}$$

and

$$q_- := \max_j \frac{-d_{jj} + \|D'\| - \sqrt{\Delta_j}}{2} < \min_j \frac{-d_{jj} + \|D'\| + \sqrt{\Delta_j}}{2} =: q_+ \tag{20.2}$$

or

$$\hat{\Delta}_j = (d_{jj} - r_j)^2 - 4\omega_j^2 > 0 \textit{ for all } j \tag{20.3}$$

and

$$\hat{q}_- := \max_j \frac{-d_{jj} + r_j - \sqrt{\hat{\Delta}_j}}{2} < \min_j \frac{-d_{jj} + r_j + \sqrt{\hat{\Delta}_j}}{2} =: \hat{q}_+. \tag{20.4}$$

Then the system M, C, K is overdamped. Moreover, the interval (q_-, q_+), $(\hat{q}_-, \hat{q}_+)$, respectively, is contained in the definiteness interval of M, C, K.

Proof. Let $q_- < \mu < q_+$. The negative definiteness of

$$\mu^2 I + \mu D + \Omega^2 = \mu^2 I + \mu D^0 + \Omega^2 + \mu D'$$

will be insured by norm-diagonal dominance, if

$$-\mu\|D'\| < -\mu^2 - \mu d_{jj} - \omega_j^2 \text{ for all } j,$$

that is, if μ lies between the roots of the quadratic equation

$$\mu^2 + \mu(d_{jj} - \|D'\|) + \omega_j^2 = 0 \text{ for all } j$$

and this is insured by (20.1) and (20.2). The conditions (20.3) and (20.4) are treated analogously. Q.E.D.

We are now prepared to adapt the spectral inclusion bounds from the previous chapter to overdamped systems. Recall that in this case the definiteness interval divides the $2n$ eigenvalues into two groups: J-negative and J-positive.

Theorem 20.6 *If (20.1) and (20.2) hold then the J-negative/J-positive eigenvalues of $Q(\cdot)$ are contained in*

$$\cup_j(\mu^j_{--}, \mu^j_{-+}), \quad \cup_j(\mu^j_{+-}, \mu^j_{++}),$$

respectively, with

$$\mu^j_{\substack{++\\--}} = \frac{-d_{jj} - \|D'\| \pm \sqrt{(d_{jj} + \|D'\|)^2 - 4\omega_j^2}}{2} \tag{20.5}$$

$$\mu^j_{\substack{+-\\-+}} = \frac{-d_{jj} + \|D'\| \pm \sqrt{(d_{jj} - \|D'\|)^2 - 4\omega_j^2}}{2}. \tag{20.6}$$

An analogous statement holds, if (20.3) and (20.4) hold and $\mu^j_{\substack{++\\--}}$, $\mu^j_{\substack{+-\\-+}}$ is replaced by $\hat{\mu}^j_{\substack{++\\--}}$, $\hat{\mu}^j_{\substack{+-\\-+}}$ where in (20.5,20.6) $\|D'\|$ is replaced by r_j.

Proof. First note that the function $0 > z \mapsto z + \sqrt{z^2 - 4\omega_j^2}$ is decreasing, so $\mu^j_{+-} < \mu^j_{++}$ and similarly $\mu^j_{--} < \mu^j_{-+}$.

Take first $r = \|D'\|$. All spectra are real and negative, so we have to find the intersection of $\mathcal{C}(\lambda^j_+, \lambda^j_-, r)$ with the real line the foci λ^j_+, λ^j_- from (19.13) being also real. This intersection will be a union of two intervals. For $\lambda < \lambda^j_-$ and also for $\lambda > \lambda^j_+$ the j-th ovals are given by

$$(\lambda^j_- - \lambda)(\lambda^j_+ - \lambda) \leq -\lambda r$$

i.e.

$$\lambda^2 - (\lambda^j_+ + \lambda^j_- - r)\lambda + \lambda^j_+\lambda^j_- \leq 0$$

where $\lambda^j_+ + \lambda^j_- = -d_{jj}$ and $\lambda^j_+\lambda^j_- = \omega_j^2$. Thus, the left and the right boundary point of the real ovals are $\mu^j_{\substack{++\\--}}$.

For $\lambda^j_- < \lambda < \lambda^j_+$ the ovals will not contain λ, if

$$(\lambda - \lambda^j_-)(\lambda^j_+ - \lambda) \leq -\lambda r$$

i.e.

$$\lambda^2 + (d_{jj} - r)\lambda + \omega_j^2 < 0$$

with the solution

$$\mu^j_{-+} < \lambda < \mu^j_{+-}.$$

The same argument goes with $r = r_j$. Q.E.D.

Note the inequality

$$(\mu^j_{--}, \mu^j_{-+}) < (\mu^k_{+-}, \mu^k_{++})$$

for all j, k.

The results obtained in this chapter can be partially extended to non-overdamped systems which have some J-definite eigenvalues. The idea is to use Theorem 12.15.

Theorem 20.7 *Suppose that there is a connected component $\mathcal{C}_0$ of the ovals (19.17) containing only foci λ_j^+ with $d_{jj}^2 \geq 4\omega_j^2$. Then the eigenvalues of $Q(\cdot)$ contained in $\mathcal{C}_0$ are all J-positive and their number (with multiplicities) equals the number of the foci in $\mathcal{C}_0$. For these eigenvalues the estimates of Theorem 20.6 hold (the same for J-negatives).*

Proof. Let A be from (3.10) and $D = D^0 + D'$ as in (19.10),

$$A = \begin{bmatrix} 0 & \Omega \\ -\Omega & -D \end{bmatrix}, \quad A^0 = \begin{bmatrix} 0 & \Omega \\ -\Omega & -D^0 \end{bmatrix}.$$

$$\hat{A}(t) = \begin{bmatrix} 0 & \Omega \\ -\Omega & -D^0 - tD' \end{bmatrix}, \quad 0 \leq t \leq 1$$

where D (and therefore $D^0 + tD'$) is positive semidefinite. The set of all such $\hat{A}(t)$ satisfies the requirements in Theorem 12.15. In particular, by the property (19.7) the ovals of each $\hat{A}(t)$ are contained in the ones of A and by taking a contour Γ which separates $\mathcal{C}_0$ from the rest of the ovals in (19.17) the relation $\Gamma \cap \sigma(\hat{A}(t))$ will hold and Theorem 12.15 applies. Now the ovals from $\mathcal{C}_0$ become intervals and Theorem 20.6 applies as well. Q.E.D.

Monotonicity-based bounds. As it is known for symmetric matrices monotonicity-based bounds for the eigenvalues ((2.14), (2.15) for a single matrix or (2.12) for a matrix pair) have an important advantage over Gershgorin-based bounds: While the latter are merely inclusions, that is, the eigenvalue is contained in a union of intervals the former tell more: there each interval contains 'its own eigenvalue', even if it intersects other intervals. In this chapter we will derive bounds of this kind for overdamped systems.

A basic fact is the following theorem

Theorem 20.8 *With overdamped systems the eigenvalues move asunder under growing viscosity. More precisely, let*

$$\lambda_n^- \le \cdots \le \lambda_1^- < \lambda_1^+ \le \cdots \le \lambda_n^+ < 0$$

be the eigenvalues of an overdamped system M, C, K. *If* $\hat{M}, \hat{C}, \hat{K}$ *is more viscous than the original one then its corresponding eigenvalues* $\hat{\lambda}_k^\pm$ *satisfy*

$$\hat{\lambda}_k^- \le \lambda_k^-, \quad \lambda_k^+ \le \hat{\lambda}_k^+$$

A possible way to prove this theorem is to use Duffin's minimax formulae [20]. Denoting the eigenvalues as in (10.3) the following formulae hold

$$\lambda_k^+ = \min_{S_k} \max_{x \in S_k} p_+(x), \quad \lambda_k^- = \max_{S_k} \min_{x \in S_k} p_-(x).$$

where S_k is any k-dimensional subspace and $p_\pm$ is defined in (14.11). Now the proof of Theorem 20.8 is immediate, if we observe that

$$\hat{p}_+(x) \ge p_+(x), \quad \hat{p}_-(x) \le p_-(x)$$

for any x ($\hat{p}_\pm$ is the functional (14.11) for the system $\hat{M}, \hat{C}, \hat{K}$). Note that in this case the representation

$$p_+(x) = \frac{-2x^*Kx}{x^*Cx + \sqrt{\Delta(x)}}$$

is more convenient.

As a natural relative bound for the system matrices we assume

$$|x^*\delta Mx| \le \epsilon x^*Mx, \quad |x^*\delta Cx| \le \epsilon x^*Cx, \quad |x^*\delta Kx| \le \epsilon x^*Kx,$$

with

$$\delta M = \hat{M} - M, \quad \delta C = \hat{C} - C, \quad \delta H = \hat{K} - K, \quad \epsilon < 1.$$

We suppose that the system M, C, K is overdamped and modally damped. As in Exercise 14.6 one sees that the overdampedness of the perturbed system $\hat{M}, \hat{C}, \hat{K}$ is insured, if

$$\epsilon < \frac{d-1}{d+1}, \quad d = \min_x \frac{x^*Cx}{2\sqrt{x^*Mxx^*Kx}}.$$

So, the following three overdamped systems

$$(1+\epsilon)M, (1-\epsilon)C, (1+\epsilon)K; \quad \hat{M}, \hat{C}, \hat{K}; \quad (1-\epsilon)M, (1+\epsilon)C, (1-\epsilon)K$$

are ordered in growing viscosity. The first and the last system are overdamped and also modally damped while their eigenvalues are known and given by

$$\lambda_k^\pm\left(\frac{1-\epsilon}{1+\epsilon}\right), \quad \lambda_k^\pm\left(\frac{1+\epsilon}{1-\epsilon}\right),$$

respectively, where

$$\lambda_k^\pm(\eta) = \frac{-d_{jj}\eta \pm \sqrt{d_{jj}^2\eta^2 - 4\omega_j^2}}{2},$$

ω_j, d_{jj} as in (19.12), are the eigenvalues of the system $M, \eta C, K$. We suppose that the unperturbed eigenvalues $\lambda_k^\pm = \lambda_k^\pm(1)$ are ordered as

$$\lambda_n^- \le \cdots \le \lambda_1^- < \lambda_1^+ \le \cdots \le \lambda_n^+.$$

By the monotonicity property the corresponding eigenvalues are bounded as

$$\tilde{\lambda}_k^+\left(\frac{1-\epsilon}{1+\epsilon}\right) \le \hat{\lambda}_k^+ \le \tilde{\lambda}_k^+\left(\frac{1+\epsilon}{1-\epsilon}\right),$$

where $\tilde{\lambda}_k^+(\eta)$ are obtained by permuting $\lambda_k^+(\eta)$ such that

$$\tilde{\lambda}_1^+(\eta) \le \cdots \le \tilde{\lambda}_n^+(\eta)$$

for all $\eta > 0$. It is clear that each $\tilde{\lambda}_k^+(\eta)$ is still non-decreasing in η. An analogous bound holds for $\hat{\lambda}_k^-$ as well.

The proof of the minimax formulae which were basic for the needed monotonicity is rather tedious. There is another proof of Theorem 20.8 which uses the analytic perturbation theory. This approach is of independent interest and we will sketch it here. The basic fact from the analytic perturbation theory is this.

Theorem 20.9 *Let λ_0 be, say, a J-positive eigenvalue of $Q(\cdot)$ and let m be its multiplicity. Set*

$$\lambda^2(M + \epsilon M_1) + \lambda(C + \epsilon C_1), +K + \epsilon K_1 \tag{20.7}$$

M_1, C_1, K_1 arbitrary Hermitian. Choose any

$$\mathcal{U}_r = \{\lambda \in \mathbb{C};\ |\lambda - \lambda_0| < r)\}$$

with

$$(\sigma(Q(\cdot)) \setminus \{\lambda_0\}) \cap \mathcal{U}_r = \emptyset.$$

Then there is a real neighbourhood $\mathcal{O}$ of $\epsilon = 0$ such that for $\epsilon \in \mathcal{O}$ the eigenvalues of $Q(\cdot, \epsilon)$ within $\mathcal{U}_r$ together with the corresponding eigenvectors can be represented as real analytic functions

$$\lambda_1(\epsilon), \ldots, \lambda_m(\epsilon), \tag{20.8}$$

$$x_1(\epsilon), \dots, x_m(\epsilon), \tag{20.9}$$

respectively. All these functions can be analytically continued along the real axis, and their J-positivity is preserved until any of these eigenvalues meets another eigenvalue of different type.

The proof of this theorem is not much simpler than that of the minimax formulae (see [26]). But it is plausible and easy to memorise: the eigenpairs are analytic as long as the eigenvalues do not get mixed. Once one stipulates this the monotonicity is easy to derive, we sketch the key step. Take any pair λ, x from (20.8), (20.9), respectively (for simplicity we suppress the subscript and the variable ϵ). Then differentiate the identity

$$Q(\lambda, \epsilon)x = 0$$

with respect to ϵ and premultiply by x^* thus obtaining

$$\lambda' = -\frac{x^*(\lambda^2 M_1 + \lambda C_1 + K_1)x}{2\lambda x^*(M + \epsilon M_1)x + x^*(C + \epsilon C_1)x}$$

$$= -\frac{x^*(\lambda^2 M_1 + \lambda C_1 + K_1)x}{\sqrt{\Delta}} \tag{20.10}$$

where

$$\Delta = \Delta(\epsilon) = (x^*(C + \epsilon C_1)x)^2 - 4x^*(M + \epsilon M_1)xx^*(K + \epsilon K_1)x$$

is positive by the assumed J-positivity. By choosing $-M_1, C_1, -K_1$ as positive semidefinite, the right hand side of (20.10) is non-negative (note that λ is negative). So, λ grows with ϵ.

If the system is overdamped then the proof of the monotonicity in Theorem 20.8 is pretty straightforward. If the system $\hat{M}, \hat{C}, \hat{K}$ is more viscous than M, C, K then by setting

$$M_1 = \hat{M} - M, \quad C_1 = \hat{C} - C, \quad K_1 = \hat{K} - K,$$

we obtain positive semidefinite matrices $-M_1, C_1, -K_1$. Now form $Q(\lambda, \epsilon)$ as in (20.7), apply Theorem 20.9 and the formula (20.10) and set $\epsilon = 0, 1$.

A further advantage of this approach is its generality: the monotonicity, holds 'locally' just for *any eigenvalue of definite type*, the pencil itself need not be overdamped. That is, with growing viscosity an eigenvalue of positive type moves to the right (and an eigenvalue ov the negative type moves to the left); this state of affairs lasts until an eigenvalue of different type is met.

Exercise 20.10 *Try to extend the criteria in Theorem 20.5 for the overdampedness to the case where some undamped frequencies are tightly clustered — in analogy to the bounds (19.28).*

Exercise 20.11 *Try to figure out what happens, if in (20.10) the matrices* $-M_1, C_1, -K_1$ *are positive semidefinite and the right hand side vanishes, say, for* $\epsilon = 0$.

Exercise 20.12 *Using (20.10) show that for small and growing damping the eigenvalues move into the left plane.*

Chapter 21
Passive control

Large oscillations may be dangerous to the vibrating system and are to be avoided in designing such systems. In many cases the stiffness and the mass matrix are determined by the static requirements so the damping remains as the regulating factor.

How to tell whether a given system is well damped? Or, which quantity has to be optimised to obtain a best damped system within a given set of admissible systems (usually determined by one or several parameters)? These are the questions which we pose and partly solve in this chapter.

A common criterion says: *take the damping which produces the least spectral abscissa* (meaning that its absolute value should be maximal). This criterion is based on the bounds (13.8) and (13.6).

Consider once more the one-dimensional oscillator from Example 9.3. The minimal spectral abscissa is attained at the critical damping $d = 2\omega$ and its optimal value is

$$\lambda_{ao} = -\frac{d}{2}.$$

This value of d gives a best asymptotic decay rate in (13.7) as well as a 'best behaviour' at finite times as far as it can be qualitatively seen on Fig. 13.1 in Chapter 13. However, in general, optimising the spectral abscissa gives no control for $\|e^{At}\|$ for all times.

Another difficulty with the spectral abscissa is that it is not a smooth function of the elements of A and its minimisation is tedious, if using common optimisation methods.

In this chapter we will present a viable alternative curing both shortcomings mentioned above. We say: *an optimal damping extracts as much energy as possible from the system.* A possible rigorous quantification of this requirement is to ask

$$\int_0^\infty y(t)^T B y(t) dt = \int_0^\infty y_0^T e^{A^T t} B e^{At} y_0 dt = y_0^T X y_0 = \min \qquad (21.1)$$

where B is some positive semidefinite matrix serving as a weight and X is the solution of the Lyapunov equation (13.9). This is the total energy, averaged over the whole time history of the free oscillating system.[1] In order to get rid of the dependence on the initial data y_0 we average once more over all y_0 with the unit norm. We thus obtain the penalty function, that is, the function to be minimised

$$\mathcal{E} = \mathcal{E}(A, \rho) = \int_{\|y_0\|=1} y_0^T X y_0 d\rho$$

where $d\rho$ is a given non-negative measure on the unit sphere in $\mathbb{R}^n$.[2] This measure, like the matrix B, has the role of a weight, provided by the user. Such measures can be nicely characterised as the following shows.

Proposition 21.1 *For any non-negative measure $d\rho$ on the unit sphere in $\mathbb{R}^n$ there is a unique symmetric positive semidefinite matrix S such that*

$$\int_{\|y_0\|=1} y_0^T X y_0 d\rho = \mathrm{Tr}(XS) \tag{21.2}$$

for any real symmetric matrix X.

Proof. Consider the set of all symmetric matrices as a vector space with the scalar product

$$\langle X, Y \rangle = \mathrm{Tr}(XY).$$

The map

$$X \mapsto \int_{\|y_0\|=1} y_0^T X y_0 d\rho \tag{21.3}$$

is obviously a linear functional and hence is represented as

$$\langle X, S \rangle = \mathrm{Tr}(XS).$$

where S is a uniquely determined symmetric matrix. To prove its positive semidefiniteness recall that by (21.2) the expression $\mathrm{Tr}(XS)$ is non-negative whenever X is positive semidefinite. Suppose that S has a negative eigenvalue λ and let P_λ be the projection onto the corresponding eigenspace. Now take

$$X = P_\lambda + \epsilon(I - P_\lambda), \quad \epsilon > 0.$$

Then X is positive definite and

$$\mathrm{Tr}(XS) = \lambda P_\lambda + \epsilon(I - P_\lambda)S$$

[1] The integral in (21.1) has the dimension of 'action' (energy × time) and so the principle (21.1) can be called a 'minimum action principle' akin to other similar principles in Mechanics.

[2] In spite of the fact that we will make some complex field manipulations with the matrix A we will, until the rest of this chapter, take A as real.

would be negative for ϵ small enough — a contradiction. Q.E.D.

Thus, we arrive at the following optimisation problem: find the minimum

$$\min_{A\in\mathcal{A}} \left\{ \mathrm{Tr}(XS) : \ A^T X + XA = -B \right\} \tag{21.4}$$

where B and S are given positive semidefinite real symmetric matrices and $\mathcal{A}$ is the set of allowed matrices A. Typically $\mathcal{A}$ will be determined by the set of allowed dampings C in (3.3). A particularly distinguished choice is $S = I/(2n)$, which correponds to the Lebesgue measure (this is the measure which 'treats equally' any vector from the unit sphere).

Remark 21.2 The previous proposition does not quite settle the uniqueness of S as we might want it here. The question is: do all asymptotically stable A (or does some smaller relevant set of A-s over which we will optimise) uniquely define the functional (21.3)? We leave this question open.

Exercise 21.3 *Let a real A be asymptotically stable and let (13.9) hold. Then*

$$\mathrm{Tr}(XS) = \mathrm{Tr}(YB)$$

where Y solves

$$AY + YA^T = -S.$$

Exercise 21.4 *For the system from Example 1.1 determine the matrix B in (21.1) from the requirement*

1. *Minimise the average kinetic energy of the k-th mass point.*
2. *Minimise the average potential energy of the k-th spring.*

Minimising the trace can become quite expensive, if one needs to solve a series of Lyapunov equations in course of a minimisation process even on medium size matrices. In this connection we note an important economy in computing the gradient of the map $X \mapsto \mathrm{Tr}(SX)$. We set $A(\alpha) = A + \alpha A_1$, A_1 arbitrary and differentiate the corresponding Lyapunov equation:

$$A(\alpha)^T X'(\alpha) + X'(\alpha)A(\alpha) = -A'(\alpha)^T X(\alpha) - X(\alpha)A'(\alpha) =$$

$$-A_1^T X(\alpha) - X(\alpha)A_1.$$

Now, this is another Lyapunov equation of type (13.9), with $A = A(\alpha)$ and $B = A_1^T X(\alpha) + X(\alpha)A_1$ hence, using (13.11),

$$\mathrm{Tr}(SX(\alpha))' = \int_0^\infty \mathrm{Tr}(e^{At}Se^{A^T t}A_1X(\alpha))dt + \int_0^\infty \mathrm{Tr}(e^{At}Se^{A^T t}X(\alpha)A_1)dt$$

and the directional derivative along A_1 reads

$$\frac{d\mathrm{Tr}(SX)}{dA_1} = \mathrm{Tr}(SX(\alpha))'_{\alpha=0} = \mathrm{Tr}(YA_1^T X) + \mathrm{Tr}(XA_1Y) = 2\mathrm{Tr}(XA_1Y), \tag{21.5}$$

where Y solves the equation

$$AY + YA^T = -S.$$

Thus, after solving just two Lyapunov equations we obtain all directional derivatives of the trace for free. A similar formula exists for the Hessian matrix as well.

21.1 More on Lyapunov equations

The Lyapunov equation plays a central role in our optimisation approach. We will review here some of its properties and ways to its solution. As we have said the Lyapunov equation can be written as a system of n^2 linear equations with as many unknowns. The corresponding linear operator

$$\mathcal{L} = A^T \cdot \ + \ \cdot A, \tag{21.6}$$

is known under the name of *the Lyapunov operator* and it maps the vector space of matrices of order n into itself.

Thus, Gaussian eliminations would take some $(n^2)^3 = n^6$ operations which is feasible for matrices of very low order — not much more than $n = 50$ on present day personal computers. The situation is different, if the matrix A is safely diagonalised:

$$A = S\Lambda S^{-1}, \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots).$$

Then the Lyapunov equation (13.9) is transformed into

$$\overline{\Lambda} Z + Z\Lambda = -F, \quad Z = S^T X S, \quad F = S^T B S$$

with the immediate solution

$$Z_{ij} = \frac{-F_{ij}}{\overline{\lambda_i} + \lambda_j}.$$

The diagonalisation may be impossible or numerically unsafe if S is ill-conditioned. A much more secure approach goes via the triangular form. If the matrix A is of, say, upper triangular form then the first row of (13.9) reads

$$(\overline{a_{11}} + a_{11})x_{1j} + a_{12}x_{2j} + \cdots + a_{1j}x_{jj} = b_{1j}$$

from which the first row of X is immediately computed; other rows are computed recursively. The solution with a lower triangular T is analogous.

A general matrix is first reduced to the upper triangular form as

$$T = U^* A U, \quad U \text{ unitary, possibly complex;}$$

then (13.9) is transformed into

$$T^* Y + Y A = -V, \quad Y = U^* X U, \quad V = U^* B U.$$

which is then solved as above. (This is also another proof of the fact that the Lyapunov equation with an asymptotically stable A possesses a unique solution.) Both steps to solve a general Lyapunov equation take some $25n^3$ operations, where the most effort is spent to the triangular reduction.

Thus far we have been optimising the transient behaviour only. It is important to know how this optimisation affects the steady state behaviour of the system. A key result to this effect is contained in the following

Theorem 21.5 *Let A be real dissipative; then the integral*

$$Y = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} (\overline{\lambda} I - A^T)^{-1} B (\lambda I - A)^{-1} d\lambda \tag{21.7}$$

exists, if A is asymptotically stable. In this case Y equals X from (13.9).

Proof. For sufficiently large $|\lambda|$ we have

$$\|(\overline{\lambda} I - A^T)^{-1} B (\lambda I - A)^{-1}\| \le \|B\| \left(\sum_{k=0}^{\infty} \frac{\|A\|^k}{|\lambda|^{k+1}} \right)^2$$

which is integrable at infinity. Thus, the integral in (21.7) exists, if there are no singularities on the imaginary axis, that is, if A is asymptotically stable. In this case we have

$$A^T Y = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \left[\overline{\lambda} (\overline{\lambda} I - A^T)^{-1} - I \right] B (\lambda I - A)^{-1} d\lambda,$$

$$Y A = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} (\overline{\lambda} I - A^T)^{-1} B \left[\lambda (\lambda I - A)^{-1} - I \right] d\lambda.$$

By using the identities

$$\lim_{\eta \to \infty} \int_{-i\eta}^{i\eta} (\lambda I - A)^{-1} d\lambda = i\pi I, \quad \lim_{\eta \to \infty} \int_{-i\eta}^{i\eta} (\overline{\lambda} I - A^T)^{-1} d\lambda = i\pi I,$$

we obtain

$$A^T Y + Y A = -B + \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} (\overline{\lambda} + \lambda)(\overline{\lambda} I - A^T)^{-1} B (\lambda I - A)^{-1} d\lambda = -B.$$

Since the Lyapunov equation (13.9) has a unique solution it follows $Y = X$. Q.E.D.

Hence

$$\mathrm{Tr}(XS) = \int_{\|y_0\|=1} y_0^T X y_0 d\rho =$$

$$\int_{\|y_0\|=1} d\rho \int_0^\infty \|B^{1/2} e^{At} y_0\|^2 dt = \frac{1}{2\pi} \int_{\|y_0\|=1} d\rho \int_{-\infty}^\infty \|B^{1/2}(i\omega I - A)^{-1} y_0\|^2 d\omega$$

This is an important connection between the 'time-domain' and the 'frequency-domain' optimisation: the energy average over time and all unit initial data of the free vibration is equal to some energy average of harmonic responses taken over all frequencies and all unit external forces.

Exercise 21.6 *Use the reduction to triangular form to show that a general Sylvester equation $CX + XA = -B$ is uniquely solvable, if and only if $\sigma(C) \cap \sigma(-A) = \emptyset$. Hint: reduce A to the upper triangular form and C to the lower triangular form.*

Exercise 21.7 *Give the solution of the Lyapunov equation if a non-singular S is known such that $S^{-1}AS$ is block-diagonal. Are there any simplifications, if S is known to be J-unitary and A J-Hermitian?*

Exercise 21.8 *The triangular reduction needs complex arithmetic even if the matrix A is real. Modify this reduction, as well as the subsequent recursive solution of the obtained Lyapunov equation with a block-triangular T.*

Exercise 21.9 *Let $A = -I + N$, $N^{p-1} \neq 0$, $N^p = 0$. Then the solution of the Lyapunov equation (13.9) is given by the formula*

$$X = \sum_{n=0}^{2p-1} \frac{1}{2^{k+1}} \sum_{k=0}^{n} \binom{n}{k} (N^*)^k B N^{n-k}.$$

21.2 Global minimum of the trace

The first question in assessing the trace criterion (21.4) is: if we let the damping vary just over all positive semidefinite matrices, where will the minimum be taken? The answer is: the minimum is taken at $D = 2\Omega$ in the representation (3.10) that is, at the 'modal critical damping' — a very appealing result which speaks for the appropriateness of the trace as a measure of stability. Note that without loss of generality we may restrict ourselves to the modal representation (3.10). As we know, the matrix A from (3.10) is unitarily similar to the one from (3.3), the same then holds for the respective solutions X of the Lyapunov equation

$$A^T X + XA = -I, \tag{21.8}$$

so the trace of X is the same in both cases.

First of all, by the continuity property of the matrix eigenvalues it is clear that the set $\mathcal{C}_s$ of all symmetric damping matrices with any fixed M and K and an asymptotically stable A is open. We denote by $\mathcal{D}_s^+$ the connected component of $\mathcal{C}_s$ containing the set of positive semidefinite matrices D. It is natural to seek the minimal trace within the set $\mathcal{D}_s^+$ as damping matrices.

Theorem 21.10 *Let Ω be given and let in the modal representation (3.10) the matrix D vary over the set $\mathcal{D}_s^+$. Then the function $D \mapsto \mathrm{Tr}(X)$ where $X = X(D)$ satisfies (21.8) has a unique minimum point $D = D_0 = 2\Omega$.*

The proof we will offer is not the simplest possible but it sheds some light on the geometry of asymptotically stable damped systems. The difficulty is that the trace is not generally a convex function which precludes the use of the usual uniqueness argument. Our proof will consist of several steps and end up with interpolating the trace function on the set of the stationary points by a function which is strictly convex and takes its minimum on the modal damping matrix, thus capturing the uniqueness.

Lemma 21.11 *If $\|C\|^2 \geq 4\|K\|\|M\|$ then*

$$\max \operatorname{Re} \sigma(A) \geq -\frac{2\|K\|}{\|C\|}. \tag{21.9}$$

Proof. Let x be a unit vector with $Cx = \|C\|x$. Then

$$x^T(\lambda_+^2 M + \lambda_+ C + K)x = 0$$

with

$$\lambda_+ = -\frac{2x^T Kx}{\|C\| + \sqrt{\|C\|^2 - 4x^T Kx x^T Mx}} \geq -\frac{2\|K\|}{\|C\|}.$$

Since $\lambda^2 M + \lambda C + K$ is positive definite for λ negative and close to zero, there must be an eigenvalue of A in the interval $[\lambda_+, 0)$ i.e. $\max \operatorname{Re} \sigma(A) \geq \lambda_+$. Q.E.D.

The Lyapunov equation $A^T X + XA = -I$ for the phase-space matrix A from (3.10) reads

$$(A_0 - BDB^T)^T X + X(A_0 - BDB^T) = -I, \tag{21.10}$$

where

$$A_0 = \begin{bmatrix} 0 & \Omega \\ -\Omega & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

The so–called dual Lyapunov equation is given by

$$(A_0 - BDB^T)Y + Y(A_0 - BDB^T)^T = -I. \tag{21.11}$$

To emphasise the dependence of X and Y on D we write $X(D)$ and $Y(D)$. By the J-symmetry of the matrix A it follows

$$Y(D) = JX(D)J. \tag{21.12}$$

Let us define the function $f : \mathcal{D}_s^+ \to \mathbb{R}$ by

$$f(D) = \mathrm{Tr}(X(D)), \text{ where } X(D) \text{ solves (21.10)}. \tag{21.13}$$

Lemma 21.12 $D_0 \in \mathcal{D}_s^+$ *is a stationary point of* f*, if and only if*

$$B^T X(D_0)Y(D_0)B = 0. \tag{21.14}$$

Proof of the lemma. By setting in (21.5)

$$A_1 = BD_1B^T$$

where D_1 is any real symmetric matrix the point D_0 is stationary, if and only if

$$\mathrm{Tr}(X(D_0)BD_1B^TY(D_0)) = \mathrm{Tr}(D_1B^TY(D_0)X(D_0)B) = 0.$$

Now, by (21.12) the matrix $B^TY(D_0)X(D_0)B$ is symmetric:

$$B^TY(D_0)X(D_0)B = B^TJX(D_0)JX(D_0)B,$$

$$(JX(D_0)JX(D_0))_{22} = -(X(D_0)JX(D_0))_{22}$$

and (in sense of the trace scalar product) orthogonal to all symmetric matrices D_1, so it must vanish. Q.E.D.

Proof of Theorem 21.10. By (13.12), Lemma 21.11 and the continuity of the map $A \mapsto \lambda_a(A)$ the values of $\mathrm{Tr}X$ become infinite on the boundary of $\mathcal{D}_s^+$. Hence there exists a minimiser and it is a stationary point. Let now D be any such stationary point. By $Y = JXJ$ (21.14) is equivalent to

$$X_{22}^2 = X_{12}^TX_{12}, \tag{21.15}$$

where $X = X(D) = \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^T & X_{22} \end{bmatrix}$. Let us decompose (21.10) into components. We get

$$-\Omega X_{12}^T - X_{12}\Omega = -I, \tag{21.16}$$

$$\Omega X_{11} - DX_{12}^T - X_{22}\Omega = 0, \tag{21.17}$$

$$\Omega X_{12} + X_{12}^T\Omega - X_{22}D - DX_{22} = -I. \tag{21.18}$$

From (21.16) it follows

$$X_{12} = \frac{1}{2}(I - S)\Omega^{-1}, \tag{21.19}$$

where S is skew–symmetric. From the equation (21.17) and the previous formula it follows

$$X_{11} = \frac{1}{2}\Omega^{-1}D\Omega^{-1} + \frac{1}{2}\Omega^{-1}D\Omega^{-1}S + \Omega^{-1}X_{22}\Omega.$$

Since $\mathrm{Tr}(\Omega^{-1}D\Omega^{-1}S) = 0$ we have

$$\mathrm{Tr}(X) = \mathrm{Tr}(X_{11}) + \mathrm{Tr}(X_{22}) = \frac{1}{2}\mathrm{Tr}(\Omega^{-1}D\Omega^{-1}) + 2\mathrm{Tr}(X_{22}). \tag{21.20}$$

From (21.15) it follows

$$X_{12} = UX_{22} \tag{21.21}$$

where U is an orthogonal matrix. Note that the positive definiteness of X implies that X_{22} is positive definite. Now (21.18) can be written as

$$-X_{22}D - DX_{22} = -I - \Omega U X_{22} - X_{22}U^T\Omega.$$

Hence D is the solution of a Lyapunov equation, and we have

$$D = \int_0^\infty e^{-X_{22}t}(I + \Omega U X_{22} + X_{22}U^T\Omega)e^{-X_{22}t}\mathrm{d}t.$$

This implies

$$D = \frac{1}{2}X_{22}^{-1} + \int_0^\infty e^{-X_{22}t}\Omega U e^{-X_{22}t}\mathrm{d}t\, X_{22} + X_{22}\int_0^\infty e^{-X_{22}t}U^T\Omega e^{-X_{22}t}\mathrm{d}t.$$

From (21.19) and (21.21) follows

$$\Omega U = \frac{1}{2}\left(X_{22}^{-1} - X_{22}^{-1}U^T S U\right), \tag{21.22}$$

hence

$$D = X_{22}^{-1} - \frac{1}{2}X_{22}^{-1}\int_0^\infty e^{-X_{22}t}U^T S U e^{-X_{22}t}\mathrm{d}t\, X_{22} + \frac{1}{2}X_{22}\int_0^\infty e^{-X_{22}t}U^T S U e^{-X_{22}t}\mathrm{d}t\, X_{22}^{-1}.$$

So, we have obtained

$$D = X_{22}^{-1} - \frac{1}{2}X_{22}^{-1}WX_{22} + \frac{1}{2}X_{22}WX_{22}^{-1}, \tag{21.23}$$

where W is the solution of the Lyapunov equation

$$-X_{22}W - WX_{22} = -U^T SU. \tag{21.24}$$

Since $U^T SU$ is skew–symmetric, so is W.

From (21.24) it follows

$$X_{22}WX_{22}^{-1} = -W + U^T SUX_{22}^{-1},$$

which together with (21.23) implies

$$D = X_{22}^{-1} + \frac{1}{2}U^T SUX_{22}^{-1} - \frac{1}{2}X_{22}^{-1}U^T SU.$$

Now taking into account relation (21.22) we obtain

$$D = \Omega U + U^T \Omega,$$

hence

$$\Omega^{-1}D\Omega^{-1} = U\Omega^{-1} + \Omega^{-1}U^T. \tag{21.25}$$

From (21.19) and (21.21) we obtain

$$S = I - 2UX_{22}\Omega, \tag{21.26}$$

and from this and the fact that S is skew–symmetric it follows that

$$\Omega X_{22}U^T + UX_{22}\Omega = I. \tag{21.27}$$

This implies

$$\Omega^{-1}U^T = \Omega^{-1}X_{22}^{-1}\Omega^{-1} - \Omega^{-1}X_{22}^{-1}\Omega^{-1}UX_{22}\Omega.$$

From the previous relation it follows that

$$\mathrm{Tr}(\Omega^{-1}U^T) = \mathrm{Tr}(\Omega^{-1}X_{22}^{-1}\Omega^{-1}) - \mathrm{Tr}(U\Omega^{-1}).$$

Now (21.20), (21.25) and the previous relation imply

$$\mathrm{Tr}(X(D)) = g(X_{22}(D))$$

with

$$g(Z) = \frac{1}{2}\mathrm{Tr}(\Omega^{-1}Z^{-1}\Omega^{-1}) + 2\mathrm{Tr}(Z).$$

for any D satisfying (21.14). Obviously, g is positive-valued as defined on the set of all symmetric positive definite matrices Z. We compute g', g'':

$$g'(Z)(\hat{Y}) = \mathrm{Tr}(-\frac{1}{2}\Omega^{-1}Z^{-1}\hat{Y}Z^{-1}\Omega^{-1} + 2\hat{Y}) \tag{21.28}$$

$$g''(Z)(\hat{Y}, \hat{Y}) = \mathrm{Tr}(\Omega^{-1}Z^{-1}\hat{Y}Z^{-1}\hat{Y}Z^{-1}\Omega^{-1}) > 0$$

for any symmetric positive definite Z and any symmetric $\hat{Y} \neq 0$. (The above expressions are given as linear/quadratic functionals; an explicit vector/matrix representation of the gradient and the Hessian would be clumsy whereas the functional representations are sufficient for our purposes.)

By (21.28) the equation $g'(Z) = 0$ means

$$\mathrm{Tr}(-\frac{1}{2}\Omega^{-1}Z^{-1}\hat{Y}Z^{-1}\Omega^{-1} + 2\hat{Y}) = 0$$

for each symmetric $\hat{Y}$. This is equivalent to

$$-\frac{1}{2}Z^{-1}\Omega^{-2}Z^{-1} + 2I = 0$$

i.e. $Z = Z_0 = \frac{1}{2}\Omega^{-1}$ and this is the unique minimiser with the minimum value $2\mathrm{Tr}(\Omega^{-1})$ (here, too, one easily sees that $g(Z)$ tends to infinity as Z approaches the boundary of the set of positive definite symmetric matrices). Now, $\frac{1}{2}\Omega^{-1} = X_{22}(2\Omega)$ and $D_0 = 2\Omega$ is readily seen to satisfy (21.14). Hence the value

$$2\mathrm{Tr}(\Omega^{-1}) = g(X_{22}(D_0)) = \mathrm{Tr}(X(D_0))$$

is strictly less than any other stationary value $\mathrm{Tr}(X(D))$. Q.E.D.

As we see the optimal damping is always positive definite; this is in contrast to the spectral abscissa criterion, where examples will be given in which the minimal spectral abscissa is actually taken at an indefinite damping matrix that is, outside of the physically allowed region.

Exercise 21.13 *Prove that to the matrix $D_0 = 2\Omega$ there corresponds the damping matrix*

$$C = C_0 = 2L_2\sqrt{L_2^{-1}KL_2^{-T}}L_2^T. \tag{21.29}$$

Exercise 21.14 *Prove the formula*

$$X(D_0) = \begin{bmatrix} \frac{3}{2}\Omega^{-1} & \frac{1}{2}\Omega^{-1} \\ \\ \frac{1}{2}\Omega^{-1} & \frac{1}{2}\Omega^{-1} \end{bmatrix}.$$

21.3 Lyapunov trace vs. spectral abscissa

We will make some comparisons of the two optimality criteria: the spectral abscissa and the Lyapunov trace. It is expected that the first will give a better behaviour at $t = \infty$ whereas the second will take into account finite times as well.

In comparisons we will deal with the *standard Lyapunov equation*

$$A^T X + XA = -I.$$

With this X (13.10) reads

$$\|e^{As}\| \le \kappa(X) e^{-t/\|X\|}$$

whereas (13.12) reads

$$2\lambda_a \le -\frac{1}{\|X\|}.$$

A simple comparison is made on the matrix

$$A = \begin{bmatrix} -1 & \alpha \\ 0 & -1 \end{bmatrix} \text{ with } e^{At} = e^{-t} \begin{bmatrix} -1 & \alpha t \\ 0 & -1 \end{bmatrix}.$$

Here $\lambda_a(A) = -1$ independently of the parameter α which clearly influences $\|e^{At}\|$. This is seen from the standard Lyapunov solution which reads

$$X = \begin{bmatrix} 1/2 & \alpha/4 \\ \alpha/4 & 1/2 + \alpha^2/4 \end{bmatrix}.$$

If A is a phase-space matrix of a damped system no such drastic differences between $\lambda_a(A)$ and the growth of $\|e^{At}\|$ seem to be known.[3]

Example 21.15 For the matrix A from Example 9.3 and $B = I$ the Lyapunov equation reads

$$\begin{bmatrix} 0 & -\omega \\ \omega & -d \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} + \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} 0 & \omega \\ -\omega & -d \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

By solving four linear equations with four unknowns we obtain

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{d} + \frac{d}{2\omega^2} & \frac{1}{2\omega} \\ \frac{1}{2\omega} & \frac{1}{d} \end{bmatrix}$$

where

$$\mathrm{Tr}(X) = \frac{2}{d} + \frac{d}{2\omega^2}$$

has a unique minimum at the critical damping $d = 2\omega$.

The spectral abscissa and the Lyapunov trace as functions of the damping d are displayed in Fig. 21.1 (an additive constant sees that both curves better fit to one coordinate frame). The minimum position is the same; the trace is smooth whereas the spectral abscissa has a cusp at the minimum.

Example 21.16 Reconsider Example 13.13 by replacing the spectral abscissa by the Lyapunov trace as a function of the parameter c.

[3] This seems to be an interesting open question: how far away can be the quantities $2\lambda_a(A)$ and $-1/\|X\|$ on general damped systems.

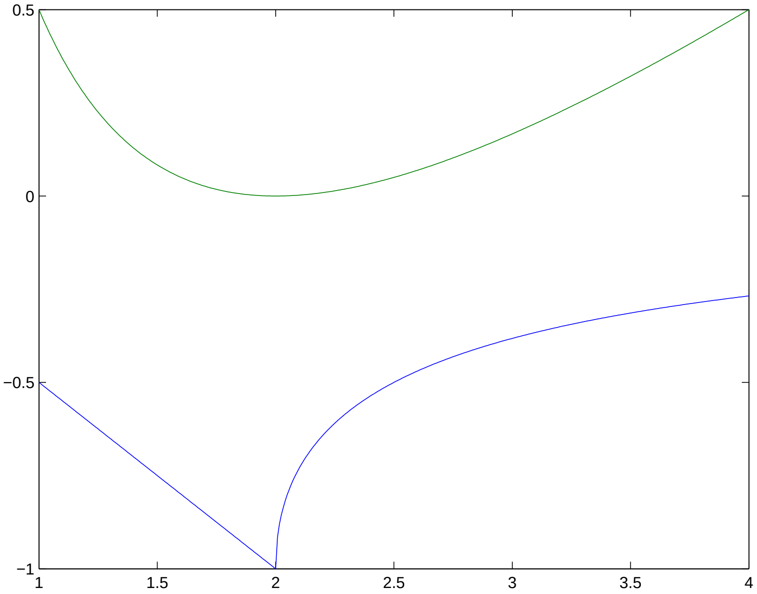

Fig. 21.1 Spectral abscissa and Lyapunov trace (1D)

Concerning optimal spectral abscissae the following theorem holds

Theorem 21.17 *([23]) Set*

$$\tau(C) = \lambda_a(A)$$

where A is from (3.3), M, K are arbitrary but fixed whereas C varies over the set of all real symmetric matrices of order n. Then $\tau(C)$ attains its minimum which is equal to

$$\tau_0 = \left(\frac{\det(K)}{\det(M)}\right)^{\frac{1}{2n}}.$$

The minimiser is unique, if and only if the matrices K and M are proportional.

The proof is based on a strikingly simple fact with an even simpler and very illuminating proof:

Proposition 21.18 *Let $\mathcal{A}$ be any set of matrices of order $2n$ with the following properties*

1. $\det(A)$ *is constant over* $\mathcal{A}$,
2. *the spectrum of each $A \in \mathcal{A}$ is symmetric with respect to the real axis, counting multiplicities,*
3. $\operatorname{Re}\sigma(A) < 0$ *for* $A \in \mathcal{A}$.

Denote by $\mathcal{A}_0$ the subset of $\mathcal{A}$ consisting of all matrices whose spectrum consists of a single point. If $\mathcal{A}_0$ is not empty then

$$\min_{A \in \mathcal{A}} \lambda_a(A)$$

is attained on $\mathcal{A}_0$ and nowhere else.

Proof. Let $\lambda_1, \ldots, \lambda_{2n}$ be the eigenvalues of A. Then

$$0 < \det(A) = \prod_{k=1}^{2n} \lambda_k = \prod_{k=1}^{2n} |\lambda_k| \geq \prod_{k=1}^{2n} |\operatorname{Re}(\lambda_k)| \geq (\lambda_a(A))^{2n} . \qquad (21.30)$$

Thus,

$$\lambda_a(A) \geq -\left(\det(A)\right)^{\frac{1}{2n}} \qquad (21.31)$$

By inspecting the chain of inequalities in (21.30) it is immediately seen that the equality in (21.31) implies the equalities in (21.30) and this is possible, if and only if

$$\lambda_k = \lambda_a(A), \quad k = 1, \ldots, 2n.$$

Q.E.D.

The set $\mathcal{A}$, of all asymptotically stable phase-space matrices A with fixed M, K obviously fulfills the conditions of the preceding proposition. Indeed, $\det(A) = \tau_0^{2n}$, hence the only thing which remains to be proved is: for given M, K find a symmetric damping matrix C such that the spectrum of A consists of a single point. This is a typical *inverse spectral problem* where one is asked to construct matrices from a given class having prescribed spectrum. This part of the proof is less elementary, it borrows from the inverse spectral theory and will be omitted. As an illustration we bring a simple inverse spectral problem:

Example 21.19 Consider the system in Example 4.5. We will determine the constants m, k_1, k_2, c by prescribing the eigenvalues

$$\lambda_1, \lambda_2, \lambda_3$$

with $\operatorname{Re}(\lambda_k) < 0$ and, say, λ_1 real. Since the eigenvalues of (14.2) do not change, if the equation is multiplied by a constant, we may assume that $m = 1$. The remaining constants are determined from the identity

$$\det \begin{bmatrix} \lambda^2 + k_1 + k_2 & -k_2 \\ -k_2 & \lambda c + k_2 \end{bmatrix} = c\lambda^3 + k_2\lambda^2 + c(k_1 + k_2)\lambda + k_1 k_2 =$$

$$c\left(\lambda^3 - (\lambda_1 + \lambda_2 + \lambda_3)\lambda^2 + (\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1)\lambda - \lambda_1\lambda_2\lambda_3\right) .$$

The comparison of the coefficients yields

$$k_2 = -c(\lambda_1 + \lambda_2 + \lambda_3),$$

$$k_1 + k_2 = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1,$$

$$k_1 k_2 = -c\lambda_1\lambda_2\lambda_3.$$

This system is readily solved with the unique solution

$$k_1 = \frac{\lambda_1\lambda_2\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3},$$

$$k_2 = \frac{\lambda_1^2(\lambda_2 + \lambda_3) + \lambda_2^2(\lambda_1 + \lambda_3) + \lambda_3^2(\lambda_1 + \lambda_2) + 2\lambda_1\lambda_2\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3},$$

$$c = -\frac{k_2}{\lambda_1 + \lambda_2 + \lambda_3}.$$

Here we can apply Proposition 21.18 by taking first, say, $\lambda_1 = \lambda_2 = \lambda_3 = -1$ thus obtaining

$$k_1 = 1/3, \quad k_2 = 8/3, \quad c = 8/9.$$

Since λ_2, λ_3 are either real or complex conjugate the values k_1, k_2, c are all positive.

Now with the fixed values k_1, k_2 as above we let $c > 0$ be variable. The so obtained set $\mathcal{A}$ of phase-space matrices from Example 4.5 again satisfies the conditions of Proposition 21.18 and $c = 8/9$ yields the optimal spectral abscissa.

For numerical comparison between the two criteria we will take an example from [23]. Take

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$$

with the optimal-abscissa damping

$$C = \frac{1}{k_1^{1/2} + k_2^{1/2}} \begin{bmatrix} 4k_1^{3/4}k_2^{1/4} & \pm(k_1^{1/2} - k_2^{1/2})^2 \\ \pm(k_1^{1/2} - k_2^{1/2})^2 & 4k_1^{1/4}k_2^{3/4} \end{bmatrix}.$$

Here the different sign choices obviously lead to unitarily equivalent both damping and phase-space matrices, so we will take the first one. According to Theorem 21.10 our best damping in this case is

$$C_d = \begin{bmatrix} 2\sqrt{k_1} & 0 \\ 0 & 2\sqrt{k_2} \end{bmatrix}.$$

For comparing time histories it is convenient to observe the quantity

$$\|e^{At}\|_E^2 = \mathrm{Tr}(e^{A^T t} e^{At})$$

because (i) it dominates the square of the spectral norm of e^{At} and (ii) it equals the average over all unit vectors y_0 of the energy $\|e^{At}y_0\|^2$ at each time t. These histories can be seen in Fig. 21.2. The dashed line corresponds to the optimal trace and the full one to the optimal abscissa. The values we have taken are

$$k_1 = 1, \quad k_2 = 4,\ 16,\ 33,\ 81$$

The optimal abscissa history is better at large times but this difference is almost invisible since both histories are then small. At k_2 about 33 the matrix C ceases to be positive semidefinite; the history for $k_2 = 81$ is even partly increasing in time and this becomes more and more drastic with growing k_2. In such situations the theory developed in [23] does not give information on an optimal-abscissa within positive-semidefinite dampings.

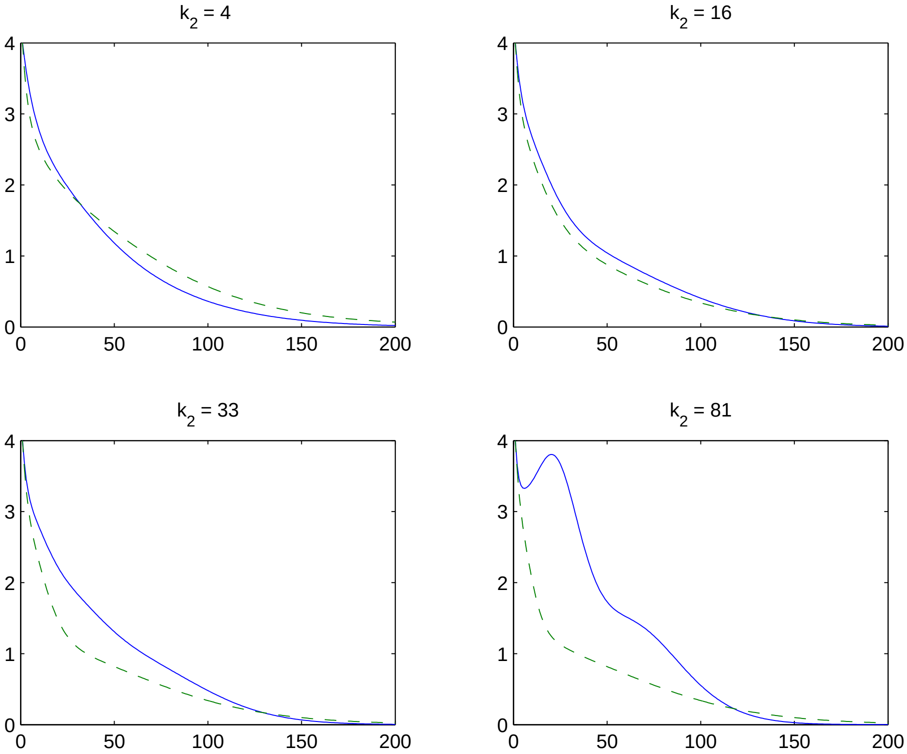

Fig. 21.2 Time histories

Chapter 22
Perturbing matrix exponential

Modal approximation appears to be very appealing to treat matrix exponential by means of perturbation techniques. We will now derive some abstract perturbation bounds and apply them to phase-space matrices of damped systems.

The starting point will be the formula which is derived from (3.4)

$$e^{Bt} - e^{At} = \int_0^t e^{B(t-s)}(B-A)e^{As}ds = \int_0^t e^{Bs}(B-A)e^{A(t-s)}ds. \tag{22.1}$$

A simplest bound is obtained, if the exponential decay of the type (13.8) is known for both matrices. So, let (13.8) hold for A and B with the constants F, ν, F_1, ν_1, respectively. Then by (22.1) we immediately obtain

$$\|e^{Bt} - e^{At}\| \le \|B-A\| \int_0^t e^{-\nu s} e^{-\nu_1(t-s)} F F_1 ds \tag{22.2}$$

$$= \|B-A\| \begin{cases} FF_1 \frac{e^{-\nu t} - e^{-\nu_1 t}}{\nu_1 - \nu}, \ \nu_1 \ne \nu \\ \\ FF_1 t e^{-\nu t}, \, \nu_1 = \nu. \end{cases}$$

If B is only known to be dissipative ($F_1 = 1, \nu_1 = 0$) then

$$\|e^{Bt} - e^{At}\| \le \|B-A\| \frac{F}{\nu}(1 - e^{-\nu t}) \le \|B-A\| \frac{F}{\nu}. \tag{22.3}$$

If A, too, is just dissipative ($F = 1, \nu = 0$) then

$$\|e^{Bt} - e^{At}\| \le \|B-A\| t. \tag{22.4}$$

The perturbation estimates obtained above are given in terms of general asymptotically stable matrices. When applying this to phase-space matrices A, B we are interested to see how small will $B - A$ be, if M, C, K are subject to small changes. This is called 'structured perturbation' because it respects the structure of the set of phase-space matrices within which the perturbation takes place. An example of structured perturbation was given in Chapter 19 on modal approximations. If B is obtained from A by changing C into $\hat{C}$ then as in (19.3) we obtain the following estimate

$$\|(B-A)\| = \max_x |\frac{x^T(\hat{C}-C)x}{x^T Mx}|.$$

Exercise 22.1 *Let C be perturbed as*

$$|x^T(\hat{C}-C)x| \le \epsilon x^T Cx.$$

Prove

$$\|B-A\| \le \epsilon \max_x \frac{x^T Cx}{x^T Mx}.$$

The dependence of A on M and K is more delicate. This is because M and K define the phase-space itself and by changing them we are changing the underlying geometry of the system.

First of all, recall that A is defined by M, C, K only up to (jointly) unitary equivalence. So, it may happen that in order to obtain best estimates we will have to take phase-space realisations not enumerated in (3.10). We consider the case in which K is perturbed into $\hat{K}$ and M, C remain unchanged, that is,

$$\hat{K} = K + \delta K = \hat{L}_1 \hat{L}_1^T$$

with

$$\hat{L}_1 = L_1\sqrt{I+Z}, \quad Z = L_1^{-1}\delta K L_1^{-T}$$

where we have assumed $\|Z\| \le 1$. Now,

$$L_2^{-1}(\hat{L}_1 - L_1) = L_2^{-1}L_1 \sum_{k=1}^{\infty} \binom{1/2}{k} Z^k =$$

$$\sum_{k=1}^{\infty} \binom{1/2}{k} L_2^{-1}\delta K L_1^{-T} Z^{k-1} = L_2^{-1}\delta K L_1^{-T} f(Z)$$

with

$$f(\zeta) = \begin{cases} \frac{(1+\zeta)^{1/2}-1}{\zeta}, \, \zeta \ne 0 \\ 0, \, \zeta = 0 \end{cases}$$

and $|\zeta| \le 1$. By

$$f(-1) = 1, \quad f'(\zeta) = -\frac{((1+\zeta)^{1/2}+1)^2}{2\zeta^2(1+\zeta)^{1/2}} \le 0$$

and by the fact that Z is a real symmetric matrix we have $\|f(Z)\| \le 1$ and

$$\|B - A\| = \|L_2^{-1}(\hat{L}_1 - L_1)\| \leq \|L_2^{-1}\delta K L_1^{-T}\|.$$

Thus (22.3) yields

$$\|e^{Bt} - e^{At}\| \leq \frac{F}{\nu}\|L_2^{-1}\delta K L_1^{-T}\|. \tag{22.5}$$

Note that here $\hat{L}_1$ will not be the Cholesky factor, if L_1 was such. If we had insisted $\hat{L}_1$ to be the Cholesky factor then much less sharp estimate would be obtained.

We now derive a perturbation bound which will assume the dissipativity of both A and B but no further information on their exponential decay — there might be none, in fact. By $\mathcal{T}(A)$ we denote the set of all *trajectories*

$$S = \left\{x = e^{At}x_0, t \geq 0\right\}, \text{ for some vector } x_0.$$

Lemma 22.2 *Let A, B be arbitrary square matrices. Suppose that there exist trajectories $S \in \mathcal{T}(A)$, $T \in \mathcal{T}(B^*)$ and an $\varepsilon > 0$ such that for any $y \in S$, $x \in T$*

$$|x^*(B - A)y|^2 \leq \varepsilon^2 \operatorname{Re}(-x^*Bx)\operatorname{Re}(-y^*Ay). \tag{22.6}$$

Then for all such x, y

$$|x^*(e^{Bt} - e^{At})y| \leq \frac{\varepsilon}{2}\|x\|\|y\|.$$

(Note that in (22.6) it is tacitly assumed that the factors on the right hand side are non-negative. Thus, we assume that A, B are 'dissipative along a trajectory'.)

Proof. Using (22.1), (22.6) and the Cauchy-Schwarz inequality we obtain

$$|x^*(e^{Bt} - e^{At})y|^2 \leq$$

$$\int_0^t |(e^{B^*s}x)^*(B - A)e^{A(t-s)}y|ds\ ^2 \leq$$

$$\varepsilon^2 \int_0^t \sqrt{\operatorname{Re}(-x^*e^{Bs}Be^{B^*s}x)\operatorname{Re}(-y^*e^{A^*(t-s)}Ae^{A(t-s)}y)}ds\ ^2 \leq$$

$$\varepsilon^2 \int_0^t \operatorname{Re}(-x^*e^{Bs}Be^{B^*s}x)ds \int_0^t \operatorname{Re}(-y^*e^{A^*s}Ae^{As}y)ds$$

By partial integration we compute

$$\mathcal{I}(A, y, t) = \int_0^t Re(-y^*e^{A^*s}Ae^{As}y)ds =$$

$$-\|e^{As}y\|^2\Big|_0^t - \mathcal{I}(A, y, t),$$

$$\mathcal{I}(A,y,t) = \frac{1}{2}\left(\|y\|^2 - \|e^{At}y\|^2\right)$$

(note that $\mathcal{I}(A,y,t) = \mathcal{I}(A^*,y,t)$). Obviously

$$0 \le \mathcal{I}(A,y,t) \le \frac{1}{2}\|y\|^2$$

and $\mathcal{I}(A,y,t)$ increases with t. Thus, there exist limits

$$\|e^{At}y\|^2 \searrow P(A,y), \quad t \to \infty$$

$$\mathcal{I}(A,y,t) \nearrow \mathcal{I}(A,y,\infty) = \frac{1}{2}\,\|y\|^2 - P(A,y)\big)\,, \quad t \to \infty$$

with

$$0 \le \mathcal{I}(A,y,\infty) \le \frac{1}{2}\|y\|^2.$$

(and similarly for B). Altogether

$$\begin{aligned} |x^*(e^{Bt} - e^{At})y|^2 &\le \frac{\varepsilon^2}{4}\left(\|x\|^2 - P(B^*,x)\right)\left(\|y\|^2 - P(A,y)\right) \\ &\le \frac{\varepsilon^2}{4}\|x\|^2\|y\|^2. \end{aligned}$$

Q.E.D.

Corollary 22.3 *Suppose that (22.6) holds for all y from some $S \in \mathcal{T}(A)$ and all x. Then*

$$\|\left(e^{Bt}y - e^{At}y\right)\| \le \frac{\varepsilon}{2}\|y\|.$$

Proposition 22.4 *For any dissipative A we have*

$$P(A,y) = y^*P(A)y,$$

where the limit

$$P(A) = \lim_{t\to\infty} e^{A^*t}e^{At} \tag{22.7}$$

exists, is Hermitian and satisfies $0 \le y^*P(A)y \le y^*y$ for all y.

Proof. The existence of the limit follows from the fact that any bounded, non decreasing sequence of real numbers — and also of Hermitian matrices — is convergent. All other statements are then obvious. Q.E.D.

Theorem 22.5 *If both A and B are dissipative and (22.6) holds for all x, y, then*

$$|\left(x,(e^{Bt} - e^{At})y\right)|^2 \le \frac{\varepsilon^2}{4}\left(\|x\|^2 - x^*P(B)x\right)\left(\|y\|^2 - y^*P(A)y\right).$$

In particular,

$$\|e^{Bt} - e^{At}\| \leq \frac{\varepsilon}{2}. \tag{22.8}$$

Exercise 22.6 *With A dissipative prove that $P(A)$ from (22.7) is an orthogonal projection and find its range.*

Corollary 22.7 *If in Theorem 22.5 we have $\varepsilon < 2$ then the asymptotic stability of A implies the same for the B and vice versa.*

Proof. Just recall that the asymptotic stability follows, if $\|e^{At}\| < 1$ for some $t > 0$. Q.E.D.

We now apply the key result of this chapter, contained in Theorem 22.5 to phase-space matrices A, B having common masses and stiffnesses and different damping matrices C, C_1, respectively. As is immediately seen then the bound (22.6) is equivalent to

$$|x^*(C_1 - C)y|^2 \leq \varepsilon^2 x^* C x y^* C_1 y. \tag{22.9}$$

Example 22.8 Consider the damping matrix $C = C_{in}$ from (1.4,1.5) and let C_1 be of the same type with parameters $c_j^{(1)}$ and let the two sets of parameters be close in the sense

$$|c_j^{(1)} - c_j| \leq \epsilon \sqrt{c_j^{(1)} c_j}.$$

Then (everything is real)

$$x^T(C_1 - C)y = (c_1^{(1)} - c_1)x_1 y_1 + \sum_{j=2}^{n}(c_j^{(1)} - c_j)(x_j - x_{j-1})(y_j - y_{j-1})$$

$$+(c_{n+1}^{(1)} - c_{n+1})x_n y_n$$

and, by the Cauchy-Schwartz inequality,

$$|x^*(C_1 - C)y| \leq$$

$$\epsilon \left[\sqrt{c_1^{(1)} c_1}|x_1 y_1| + \sum_{j=2}^{n} \sqrt{c_j^{(1)} c_j}|x_j - x_{j-1}||y_j - y_{j-1}| + \sqrt{c_{n+1}^{(1)} c_{n+1}}|x_n y_n| \right]$$

$$\leq \epsilon \sqrt{x^T C_1 x y^T C y}.$$

This shows how natural the bound (22.6) is.

Exercise 22.9 *Do the previous example for the case of perturbed C_{out} in the analogous way.*

Note the difference between (22.6), (22.8) and (22.2) – (22.4): the latter use the usual norm measure for $B - A$ but need more information on the decay of

e^{At} whereas the former need no decay information at all (except, of course, the dissipativity for both), in fact, neither A nor B need be asymptotically stable. On the other hand, the right hand side of (22.2) goes to zero as $t \to 0$ while the right hand side of (22.4) diverges with $t \to \infty$.

Exercise 22.10 *How does the bound (22.9) affect the phase-space matrices of the type (4.19)?*

We finally apply our general theory to the modal approximation. Here the true damping C and its modal approximation C^0 are not equally well known that is, the difference between them will be measured by the modal damping alone, for instance,

$$|x^*(C - C^0)x| \le \eta x^* C^0 x, \quad \eta < 1. \tag{22.10}$$

As is immediately seen this is equivalent to the same estimate for the damping matrix $D = \Phi^T C \Phi$ and its modal approximation $D^0 = \Phi^T C^0 \Phi$; therefore we have

$$\|D''\| \le \eta$$

with

$$d''_{ij} = \begin{cases} 0, & i = j \text{ or } d_{ii}d_{jj} = 0 \\ \frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}}, & \text{otherwise.} \end{cases}$$

In other words,

$$|x^T D'' y| \le \eta \sqrt{x^T D^0 x y^T D^0 y}$$

for any x, y. By

$$x^T D^0 x = x^T D x + x^T (D^0 - D) x \le x^T D x + \eta x^T D^0 x$$

we obtain

$$x^T D^0 x \le \frac{x^T D x}{1 - \eta}$$

and finally, for $D' = D - D^0$

$$|x^T D' y| \le \frac{\eta}{\sqrt{1 - \eta}} \sqrt{x^T D x y^T D^0 y}.$$

Thus, (22.10) implies (22.9) and then (22.6) with

$$\epsilon = \frac{\eta}{\sqrt{1 - \eta}}$$

and the corresponding matrix exponentials are bounded by

$$\|e^{Bt} - e^{At}\| \le \frac{\eta}{2\sqrt{1 - \eta}}.$$

Chapter 23
Notes and remarks

General. The best known monographs on our topic are Lancaster's book [51] and the work by Müller and Schiehlen [72]; the former is more mathematical than the latter.

Chapter 1

1. Suitable sources on the mechanical background are Goldstein's book [30] as well as Landau-Livshitz treatise, in particular [60] and [61] where some arguments for the symmetry of the damping matrix are given. A more mathematically oriented (and deeper diving) text is Arnold [3].
2. Engineering points of view on vibration theory may be found in the monographs [42], [12], [71]. Rather exhaustive recent works on construction of damping matrices are [76] (with a rich collection of references) and [78].

Chapters 2 – 4.

1. Accurate computation of low (undamped) eigenfrequencies is an important issue, among others because in approximating continuum structures the highest frequencies are present just because one increases the dimension in order to get the lower ones more accurately. This increases the condition number — an old dilemma with any discretisation procedure. We will illustrate this phenomenon on a simple example. Undamped oscillations of a string of length l are described by the eigenvalue equation

$$-(a(x)y')' = \lambda\rho(x)y, \quad y(0) = y(l) = 0 \tag{23.1}$$

where $a(x), \rho(x)$ are given positive functions. By introducing the equidistant mesh

$$x_j = jh, \quad j = 1, \dots, n+1, \quad h = \frac{l}{n+1}$$

and the abbreviations

$$y(x_j) = z_j, \quad a(x_j) = a_j, \quad \rho(x_j) = \rho_j, \quad z_0 = z_{n+1} = 0$$

we approximate the derivatives by differences

$$(a(x)y')'_{x=x_j} \approx \frac{a_{j+1}(z_{j+1} - z_j) - a_j(z_j - z_{j-1})}{h^2}$$

$$= -\frac{(n+1)^2}{l^2}(-a_j z_{j-1} + (a_j + a_{j+1})z_j - a_{j+1}z_{j+1}).$$

Thus, (23.1) is approximated by the matrix eigenvalue problem

$$Kz = \tilde{\lambda} Mz \tag{23.2}$$

where K is from (1.3) with $k_j = a_j$ while M is from (1.2) with $m_j = \frac{\rho_j l^2}{(n+1)^2}$. For a homogeneous string with $a(x) \equiv 1$, $\rho(x) \equiv 1$ both the continuous system (23.1) and its discretisation (23.2) have known eigenvalues

$$\lambda_j = \frac{j^2\pi^2}{l^2}, \quad \tilde{\lambda}_j = 4\frac{(n+1)^2}{l^2}\sin^2\frac{j\pi}{2n+2}.$$

For small j/n we have $\tilde{\lambda}_j \approx \lambda_j$, but the approximation is quite bad for j close to n. So, on one hand, n has to be large in order to insure good approximation of low eigenvalues but then

$$\kappa(K) = \frac{\tilde{\lambda}_n}{\tilde{\lambda}_1} \approx (n+1)^2 \tag{23.3}$$

which may jeopardise the computational accuracy. With strongly inhomogeneous material (great differences among k_j or m_j) the condition number is likely to be much worse than the one in (23.3).

2. High accuracy may be obtained by (i) using an algorithm which computes the eigenfrequencies with about the same relative error as warranted by the given matrix elements and by (ii) choosing the matrix representation on which small relative errors in the matrix elements cause about the same error in the eigenfrequencies. The first issue was treated rather exhaustively in [21] and the literature cited there. The second was addressed in [2], for numerical developments along these lines see, for instance, [77], [1], [7]. To illustrate the idea of [2] we represent the stiffness matrix K of order n from Example 1.1 as $K = L_1 L_1^T$ with

$$L_1 = \begin{bmatrix} \kappa_1 & -\kappa_2 & 0 & 0 & 0 \\ 0 & \kappa_2 & -\kappa_3 & & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \kappa_n & \kappa_{n+1} \end{bmatrix}, \quad \kappa_i = \sqrt{k_i}.$$

The matrix L_1 — called *the natural factor* — is not square but its nonzero elements are close to the physical parameters: square root operation takes place on the stiffnesses (where it is accurate) instead on the stiffness matrix whose condition number is the square of that of L_1. The eigenfrequencies are the singular values of $L_1^T M^{-1/2}$. From what was said in Chapter 2.2 it is plausible that this way will yield more accurate low frequencies. Not much seems to be known on these issues in the presence of damping.

3. The phase-space construction is even more important for vibrating continua, described by partial differential equations of Mathematical Physics of which our matrix models can be regarded as a finite dimensional approximation. The reader will observe that throughout the text we have avoided the use of the norm of the phase-space matrix since in continuum models this norm is infinite that is, the corresponding linear operator is unbounded. In the continuum case the mere existence and uniqueness of the solution (mostly in terms of semigroups) may be less trivial to establish. Some references for continuous damped systems, their semigroup realisations and properties are [47], [48], [49], [39], [40], [73], [13], [90]. The last article contains an infinite dimensional phase space construction which allows for very singular operator coefficients M, C, K. This includes our singular mass matrix case in Sect. 5. In [13] one can find a systematic functional-analytic treatment with an emphasis on the exponential decay and an ample bibliography.
4. The singular matrix case is an example of a *differential algebraic system*. We have resolved this system by a careful elimination of 'non-differentiated' variables. Such elimination is generally not easy and a direct treatment is required, see e.g. [10].
5. Our phase-space matrix A is a special linearisation of the second order equation (1.1). There are linearisations beyond the class considered in our Theorem 3.4. In fact, even the phase-space matrices obtained in Chapter 16 are a sort of linearisations not accounted for in Theorem 3.4. A rather general definition of linearisation is the following: We say that a matrix pair S, T or, equivalently, the pencil $S - \lambda T$ is *a linearisation* of the pencil $Q(\lambda)$, if

$$\begin{bmatrix} Q(\lambda) & 0 \\ 0 & I \end{bmatrix} = E(\lambda)(S - \lambda T)F(\lambda)$$

where $E(\lambda), F(\lambda)$ are some matrix functions of λ whose determinants are constant in λ. A main feature of any linearisation is that it has the same eigenvalues as the quadratic pencil, including the multiplicities (see [80] and the literature cited there).[1] The phase-space matrices, obtained in Sect. 16 do not fall, strictly speaking, under this definition, one has to undo the spectral shift first.

[1] Here, too, special care is required, if the mass is singular, see [57].

The advantage of our particular linearising (1.1) is that it not only enables us to use the structure of dissipativity and J-symmetry but also gives a proper physical framework of the energy phase space.

6. There is a special linearisation which establishes a connection with models in Relativistic Quantum Mechanics. The equation (1.1) can be written in the form

$$\left(\frac{d}{dt}+\frac{C_1}{2}\right)^2 y + \left(K-\frac{C_1^2}{4}\right) y = f_1$$

with

$$C_1 = L_2^{-1}CL_2^{-T}, \quad K_1 = L_2^{-1}KL_2^{-T}, \quad y = L_2^T x, \quad f_1 = L_2^{-1}f.$$

The substitution

$$y_1 = y, \quad y_2 = \left(\frac{d}{dt}+\frac{C_1}{2}\right) y$$

leads to

$$\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} -\frac{C_1}{2} & I \\ -K_1 + \frac{C_1^2}{4} & -\frac{C_1}{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ f_1 \end{bmatrix}$$

Here the phase-space matrix is J-symmetric with

$$J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}.$$

This is formally akin to the Klein-Gordon equation which describes spinless relativistic quantum particles and reads (in matrix environment)

$$\left(\left(i\frac{d}{dt}+V\right)^2 - H\right) y = 0$$

where V, H are symmetric matrices and H is positive definite. The same linearisation leads here to the phase-space equation

$$i\frac{d}{dt}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathcal{H}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \mathcal{H} = \begin{bmatrix} V & I \\ H & V \end{bmatrix}.$$

Note the main difference: here the time is 'imaginary' which, of course, changes the dynamics dramatically. Nevertheless, the spectral properties of phase-space matrices are analogous: the overdampedness condition means in the relativistic case the spectral separation (again to the plus and the minus group) which implies the uniform boundedness of the matrix exponential

$$e^{-i\mathcal{H}t}, \quad -\infty < t < \infty$$

which is here J-unitary — another difference to our damped system.

Chapters 5 – 10.

1. There are authoritative monographs on the geometry and the spectral theory in the indefinite product spaces. These are Mal'cev [67] and Gohberg, Lancaster and Rodman [26] and more recently, [29].
2. Our Theorem 12.15 has an important converse: if all matrices from a J-Hermitian neighbourhood of A produce only real spectrum near an eigenvalue λ_0 of A then this eigenvalue is J-definite. For a thorough discussion of this topic see [26].
3. Related to Exercise 10.10 is the recent article [35] where the spectral groups of the same sign are connected through the point infinity thus enlarging the class of simply tractable cases.
4. Theorem 10.7 may be used to compute the eigenvalues by bisection which is numerically attractive if the decomposition $JA - \mu J = G^* J' G$ can be quickly computed for various values of μ. In fact, bisection finds real eigenvalues on a general, not necessarily definitisable A whenever the decomposition $JA - \mu J = G^* J' G$ shows a change of inertia. In order to compute *all of them* this way one has to know a definitising shift, or more generally, the roots of a definitising polynomial in Theorem 10.3. Indeed, in the latter case the eigenvalues between two adjacent roots are of the same type and by Theorem 10.8 all of them are captured by bisection.
5. Theorem 10.16 is borrowed from the work [45] which contains some further results in this direction.

Chapter 12

1. Jordan canonical forms for J-Hermitian matrices are given in [56] or [79]. Numerical methods for block-diagonalising are described e.g. in [31].
2. J-orthogonal similarities were accepted as natural tool in computing J-symmetric eigenvalues and eigenvectors long ago, see e.g. [11], [9], [38], [85], [86] and in particular [37]. The latter article brings a new short proof of the fact that for any J-orthogonal matrix U with a diagonal symmetry J the inequality
$$\kappa(D_1 U D_2) \leq \kappa(U)$$
holds for any diagonal matrices D_1, D_2 (optimal scaling property). This result is akin to our results in Proposition 10.17 and Theorem 12.19. These are taken from [91] where some further results of this type can be found.
3. The computation of an optimal block-diagonaliser, as given in Theorem 12.20 is far from satisfactory because its 'zeroth' step, the construction of the initial block-diagonaliser S, goes via the Schur decomposition which uses a unitary similarity which then destroys the J-Hermitian property of the initial matrix A; J-unitarity of the block diagonaliser is only recovered in the last step. To do the block-diagonalisation in preserving the structure one would want to begin by reducing A by a *jointly* J, J'-*unitary* transformation to a 'triangular-like' form. Such form does exist (cf. [75]) and it looks like

$$A' = \begin{bmatrix} T & B \\ 0 & T^* \end{bmatrix}, \quad J' = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$$

with T, say, upper triangular and B Hermitian. This form displays all eigenvalues and is attractive to solve the Lyapunov equation, say $A'^* X + XA' = -I_n$ blockwise as

$$\begin{aligned} T^* X_{11} + X_{11} T &= -I_{n/2} \\ T^* X_{12} + X_{12} T^* &= -X_{11} - B \\ T X_{22} + X_{22} T^* &= -I_{n/2} - BX_{12} - X_{12}^* B. \end{aligned}$$

These are triangular Lyapunov/Sylvester equations of half order which are solved in that succession. But there are serious shortcomings: (i) this reduction is possible only in complex arithmetic (except in rare special cases), (ii) no real eigenvalues are allowed, unless they have even multiplicities, (iii) there are no algorithms for this structure which would beat the efficiency of the standard Schur-form reduction (cf. [22] and the literature cited there). To make a breakthrough here is a hard but important line of research.

Chapter 13.

1. Some further informations on the matrix exponential are included in the general monograph [34].
2. The fact that the knowledge of the spectrum alone is not enough for describing the behaviour of the differential equation was known long ago. A possibility to make the spectrum 'more robust' is to introduce so-called 'pseudospectra' of different kinds. Rather than a set in $\mathbb{C}$ a pseudospectrum is a family of sets. A possible pseudospectrum of a matrix A is

   $$\sigma_\epsilon(A) = \{\lambda \in \mathbb{C} : \ \|(\lambda I - A)^{-1}\| > 1/\epsilon\}.$$

   With $\epsilon \to 0$ we have $\sigma_\epsilon(A) \to \sigma(A)$ but the *speed* of this convergence may be different for different matrices and different eigenvalues. So, sometimes for quite small ϵ the component of the set $\sigma_\epsilon(A)$ containing some spectral point may still be rather large. The role of the pseudospectra in estimating decay and perturbation bounds is broadly discussed e.g. in [81] and the literature cited there. See also [24], [25].[2]
3. As it could be expected, our bounds for the exponential decay are mostly much better than those for general matrices (e.g. [44], [64]) because of the special structure of our phase-space matrix. Further improved bounds are given in [74] and [89]. Some comparisons can be found in [89].

Chapter 14

[2] The Russian school uses the term *spectral portrait* instead of pseudospectrum.

1. The literature on the quadratic eigenvalue problem is sheer enormous. Numerous works are due to Peter Lancaster and his school and co-workers. Few of all these works are cited in our bibliography; a good source for both the information and the biography is the monograph by Lancaster, Gohberg and Rodman, [28]. More recent presentations [65] and [80] also include numerical and approximation issues as well as selection of applications which lead to various types of the matrices M, C, K and accordingly to various structural properties of the derived phase-space matrices. Another general source is [68]. For the properties of J-definite eigenvalues see also [59].
2. Numerical methods for our problems have recently been systematically studied under the 'structural aspect' that is, one avoids to form any phase-space matrix and sticks at the original quadratic pencil, or else identifies the types of the symmetry of the phase-space matrix and develops 'structured' numerical methods and corresponding sensitivity theory. Obviously, 'structure' can be understood at various levels, from coarser to finer ones. This research has been done intensively by the school of Volker Mehrmann, substantial information, including bibliography, can be found e.g. in [66], [22] and [41]. Several perturbation results in the present volume are of this type, too.
3. For quadratic/polynomial numerical ranges see e.g. [69], [58]. A related notion is the 'block-numerical range', see e.g. [63], [62], [82]. All these numerical ranges contain the spectrum just like Gershgorin circles or our stretched Cassini ovals in Sect. 19.
4. Concerning overdamped systems see the pioneering work by Duffin [20] as well as [51]. Recent results concerning criteria for overdampedness can be found in [36], [33].
5. (Exercise 14.10) More on linearly independent eigenvectors can be found in [28].

Chapter 20. A rigorous proof of Theorem 20.9 is found in [26].

Chapter 21

1. For results on the sensitivity (condition number) of the Lyapunov operator (21.6) see [6] and references there. A classical method for solving Lyapunov and Sylvester equations with general coefficient matrices is given in [5]. For more recent results see [43], [50].
2. Theorem 21.10 is taken from [16] where also some interesting phenomena about the *positioning* of dampers are observed. The problem of optimal positioning of dampers seems to be quite hard and it deserves more attention. Further results of optimal damping are found in [14], [15]. Related to this is the so called 'eigenvalue assignment' problem, see [18] and the literature cited there.

   The statement of Theorem 21.10 can be generalised to the penalty function

$$X \mapsto \mathrm{Tr}(XS),$$

where S is positive semidefinite and commutes with Ω, see the dissertation [73] which also contains an infinite dimensional version and applications to continuous damped systems.

3. For references on inverse spectral theory, applied to damped systems see e.g. the works of Lancaster and coauthors [23], [55], [54]. A result on an inverse spectral problem for the oscillator ladder with just one damper in Example 1.1 is given in [87], [88]. In this special case it was experimentally observed, (but yet not proved) that the spectral abscissa and the minimal trace criterion (21.4) yield the same minimiser (independently of the number of mass points).
4. An efficient algorithm for solving the Lyapunov equation with rank-one damping is given in [88]. Efficient solving and trace-optimising of the Lyapunov equation is still under development, see e.g. [83], [84] and the literature cited there.

Chapter 22 Our perturbation bounds are especially designed for dissipative or a least asymptotically stable matrices. As could be expected most common bounds are not uniform in t (see e.g. [64]).

References

1. Amestoy, P. R; Duff, I. S.; and Puglisi, C. Multifrontal QR factorization in a multiprocessor environment. Technical Report TR/PN94/09, ENSEEIHT, Toulouse, France 1994.
2. Argyris, J. H., Brønlund, O. E., The natural factor formulation of stiffness for the matrix deplacement method, Comp. Methods Appl. Mech. Engrg. **40** (1975) 97-119.
3. Arnol´d, V. I., Mathematical methods of classical mechanics, Springer-Verlag, New York, 1989.
4. Ashcraft, C., Grimes, R., Lewis, J., Accurate Symmetric Indefinite Linear Equation Solvers, SIMAX **20** (1999), 513–561.
5. Bartels, R. H., Stewart, G. W., A solution of the matrix equation AX + XB = C, Comm. ACM **15** (1972) 820–826.
6. Bhatia, R., A note on the Lyapunov equation, LAA **259** (1997) 71–76.
7. Boman, E. G., Hendrickson, B., Vavasis, S., Solving elliptic finite element systems in near-linear time with support preconditioners, arXiv:cs/0407022v4 [cs.NA] 4 Apr 2008
8. Brauer, A. Limits for the characteristic roots of a matrix, Duke Math. J. **13** (1946) 387–395.
9. Brebner, M. A., Grad, J., Eigenvalues of $Ax = \lambda Bx$ for real symmetric matrices A and B computed by reduction to a pseudosymmetric form on the HR process, LAA **43** (1982) 99–118.
10. Brenan, K. E., Campbell, S. L. and Petzold, L. R., Numerical Solution of Initial-Value Problems in Differential Algebraic Equations. Elsevier, North Holland, New York, 1989.
11. Bunse-Gerstner, A., An analysis of the HR algorithm for computing the eigenvalues of a matrix, LAA **35** (1981) 155–173.
12. Clough, R. W., Penzien, J., Dynamics of structures, McGraw-Hill, NY 1993.
13. Chen, G., Zhou, J., Vibration and Damping in Distributed Systems I, II CRC Press 1993.
14. Cox, S. J., Designing of optimal energy absorption, I: Lumped parameter systems, ASMA J. Vib. and Acoustics, **120** (1998) 339–345.
15. Cox, S. J., Designing for optimal energy absorption, III: numerical minimization of the spectral abscissa, Structural and Multidisciplinary Optimization, **13** [1997) 17–22.
16. Cox, S. J., Nakić, I., Rittmann, A., Veselić, K., Lyapunov optimization of a damped system, Systems & Control Letters, **53** (2004) 187–194.

17. Danisch, R., Delinić, K., Veselić, K., Quadratic eigenvalue solver for modal response analysis of non-proportionally damped elastic structures, Trans. 7-th Int. Conf. on Struct. Mech. in Reactor Technology, Chicago 1983, K 3/1 North-Holland 1983.
18. Datta, B. N., Numerical methods for linear control systems. Design and analysis. Elsevier 2004.
19. Demmel, J., Veselić, K., Jacobi's method is more accurate than QR, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.
20. Duffin, R. J., A minimax theory for overdamped networks, J. Rational Mech. Anal. **4** (1955), 221–233.
21. Drmač, Z., Veselić, K., New fast and accurate Jacobi algorithm I, SIMAX 299 (2008) 1322-1342; II, SIMAX 30 (2008) 1343-1362.
22. Fassbender, H., Kressner, D., Structured eigenvalue problems, GAMM-Mitt. **29** (2006) no. 2, 297–318.
23. Freitas, P., Lancaster, P., Optimal decay of energy for a system of linear oscillators, SIAM J. Matrix Anal. Appl. 21 (1999), no. 1, 195–208.
24. Godunov, S. V., Kiriljuk, O. P., Kostin, I. V., Spectral portraits of matrices (Russian), AN SSSR Siber. Otd. Novosibirsk preprint 1990.
25. Godunov, S. V., Lectures on modern aspects of Linear Algebra (Russian), N. Kniga 2002.
26. Gohberg, I., Lancaster, P., Rodman, L., Matrices and indefinite scalar products, Birkhäuser, Basel 1983.
27. Gohberg, I., Lancaster, P., Rodman, L., Quadratic matrix polynomials with a parameter. Adv. in Appl. Math. **7** (1986), 253–281.
28. Gohberg, I., Lancaster, P., Rodman, L., Matrix polynomials, Academic Press, New York, 1982.
29. Gohberg, I., Lancaster, P., Rodman, L., Indefinite linear algebra and applications, Springer 2005.
30. Goldstein, H., Classical Mechanics, Addison-Wesley 1959.
31. Golub, G. H., van Loan, Ch. F., Matrix computations, J. Hopkins University Press, Baltimore 1989.
32. Guang-Da Hu and Guang-Di Hu, A relation between the weighted logarithmic norm of a matrix and the Lyapunov equation, BIT **40** (2000) 606–610.
33. Guo, Chun-Hua, Lancaster, P., Algorithms for hyperbolic quadratic eigenvalue problems, Math. Comp. **74** (2005) 1777–1791.
34. Higham, N. J., Functions of Matrices: Theory and Computation. SIAM 2008.
35. Higham, N. J., Mackey, S. D., Tisseur, F., Definite matrix polynomials and their linearization by definite pencils, University of Manchester MIMS EPrint 2007.97.
36. Higham, N. J., van Dooren, P., Tisseur, F., Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem and associated nearness problems, Linear Algebra Appl. **351-352** (2002) 455–474.
37. Higham, N. J., J-orthogonal matrices: properties and generation, SIAM R. (2003) 504–519.
38. Hoppe, P., Jacobi-ähnliche Blockverfahren zur numerischen Behandlung des J-symmetrischen Eigenwertproblems, PhD Fernuniversität Hagen 1984. http://www.fernuni-hagen.de/MATHPHYS/veselic/doktorandi.html
39. Hryniv, R., Shkalikov, A., Operator models in elasticity theory and hydrodynamics and associated analytic semigroups, Vestnik Moskov. Univ. Ser. I Mat. Mekh. 1999, No. 5, 5–14.
40. Hryniv, R., Shkalikov, A., Exponential stability of semigroup s associated with some operator models in mechanics. (Russian) Mat. Zametki 73 (2003), No. 5, 657–664.

41. Hwang, T-M., Lin, W-W., Mehrmann, V., Numerical solution of quadratic eigenvalue problems with structure-preserving methods, SIAM J. Sci. Comput. **24** (2003) 1283–1302.
42. Inman, D. J., Vibration with Control Measurement and Stability, Prentice Hall, 1989.
43. Jonsson, I., Kågström, B., Recursive blocked algorithm for solving triangular systems. I. One-sided and coupled Sylvester-type matrix equations, ACM Trans. Math. Software, **28** (2002) 392–415.
44. Kågstrom, B., Bounds and Perturbation Bounds for the Matrix Exponential, BIT **17** 39–57 (1977).
45. Kovač-Striko, M., Veselić, K., Trace minimization and definiteness of symmetric matrix pairs, LAA **216** (1995) 139–158.
46. Knowles, J. K., On the approximation of damped linear systems, Struct. Control Health Monit. **13** (1995) 324–335.
47. Kreĭn, M. G., Langer, H., On some mathematical principles in the linear theory of damped oscillations of continua. I. Translated from the Russian by R. Troelstra, Integral Equations Operator Theory 1 (1978), no. 4, 539–566.
48. Kreĭn, M. G., Langer, H., On some mathematical principles in the linear theory of damped oscillations of continua. II. Translated from the Russian by R. Troelstra, Integral Equations Operator Theory 1 (1978), no. 3,364–399.
49. Kreĭn, M. G., Langer, G. K., Certain mathematical principles of the linear theory of damped vibrations of continua. (Russian) 1965 Appl. Theory of Functions in Continuum Mechanics (Proc. Internat. Sympos., Tbilisi, 1963), Vol. II, Fluid and Gas Mechanics, Math. Methods (Russian) pp. 283–322 Izdat. "Nauka", Moscow.
50. Kressner, D., Block variants of Hammarling's method for solving Lyapunov equations, ACM Trans. Math. Software **34** (2008) 1–15.
51. Lancaster, P., Lambda-matrices and vibrating systems, Pergamon Press Oxford 1966, Dover 2002.
52. Lancaster, P., Tismenetsky, M., The theory of matrices, Academic Press (1985)
53. Lancaster, P., Quadratic eigenvalue problems, LAA **150** (1991) 499–506.
54. Lancaster, P., Prells, U., Inverse problems for damped vibrating systems, J. Sound Vibration **283** (2005) 891-914.
55. Lancaster, P., Ye, Q., Inverse spectral problems for linear and quadratic matrix pencils. Linear Algebra Appl. **107** (1988), 293309.
56. Lancaster, P., Rodman, L., Canonical forms for hermitian matrix pairs under strict equivalence and congruence, SIAM Review, bf 47 (2005) 407–443.
57. Lancaster, P., Linearization of regular matrix polynomials, Electronic J. Lin. Alg. **17** (2008) 21–27.
58. Lancaster, P., Psarrakos, P., The Numerical Range of Self-Adjoint Quadratic Matrix Polynomials, SIMAX bf 23 (2001) 615–631.
59. Lancaster, P., Model updating for self-adjoint quadratic eigenvalue problems, LAA **428** (2008)
60. Landau, L. D., Lifshitz, E. M., Course of theoretical physics. Vol. 1. Mechanics. Pergamon Press, Oxford-New York-Toronto, Ont., 1976.
61. Landau, L. D., Lifshitz, E. M., Course of theoretical physics. Vol. Vol. 5: Statistical physics. Pergamon Press, Oxford-Edinburgh-New York 1968.
62. Langer, H., Matsaev,V., Markus, A., Tretter, C., A new concept for block operator matrices: the quadratic numerical range, Linear Algebra Appl. **330** (2001) 89–112.
63. Langer, H., Tretter, C., Spectral decomposition of some block operator matrices, J. Operator Theory **39** (1998) 1–20.
64. van Loan, Ch., The Sensitivity of the Matrix Exponential, SIAM J. Num. Anal. **13** (1977) 971–981.

65. Mackey, D. S., Mackey, N., Mehl, C., Mehrmann, V., Vector Spaces of Linearizations for Matrix Polynomials, SIMAX **28** (2006) 971–1004.
66. Mackey, D. S., Mackey, N., Mehl, C., Mehrmann, V., Structured Polynomial Eigenvalue Problems: Good Vibrations from Good Linearizations, SIMAX **28** (2006) 1029–1051.
67. Mal'cev, A. I., Foundations of linear algebra. Translated from the Russian by Thomas Craig Brown; edited by J. B. Roberts W. H. Freeman & Co., San Francisco, Calif.-London 1963 xi+304 pp.
68. Markus, A. S., Introduction to the spectral theory of polynomial operator pencils, AMS 1988.
69. Markus, A. S., Rodman, L. Some results on numerical ranges and factorizations of matrix polynomials, Linear and Multilinear Algebra **42** (1997), 169–185.
70. Mehrmann, V., The autonomous linear quadratic control problem : theory and numerical solution, Lecture notes in control and information sciences **163** Springer 1977.
71. Meirovich, L., Elements of vibration analysis, McGraw-Hill, NY 1984.
72. Müller, P. C., Schiehlen, W. O., Lineare Schwingungen, Akademische Verlagsgesellschaft Wiesbaden 1976.
73. Nakić, I., Optimal damping of vibrational systems, Dissertation, Fernuniversität Hagen 2003.
    www.fernuni-hagen.de/MATHPHYS/veselic/doktorandi.html
74. Nechepurenko, Yu. M., On norm bounds for the matrix exponential, Doklady Mathematics, **63** (2001) 233–236.
75. Paige, C., van Loan, C., A Schur decomposition for Hamiltonian matrices, LAA, **41** (1981), 11–32.
76. Pilkey, D., Computation of a damping matrix for finite element model updating, Ph. D. dissertation 1998.
    citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.5658
77. Plemmons. R. J., Minimum norm solutions to linear elastic analysis problems, Int. J. Num. Meth. Engrg. **20** (2005) 983–998.
78. Srikantha Phania, A., Woodhouse, J., Viscous damping identification in linear vibration, Journal of Sound and Vibration **303** (2007), 475–500.
79. Thompson, R. C., Pencils of complex and real symmetric and skew matrices, Linear Algebra Appl. **147** (1991) 323–371.
80. Tisseur, F., Meerbergen, K., The Quadratic Eigenvalue Problem, SIAM R. **43** 235–286 (2001).
81. Trefethen, L. N., Embree, M., Spectra and pseudospectra, Princeton University Press 2005.
82. Tretter, Ch., Spectral Theory of Block Operator Matrices and Applications, Imperial College Pr. 2008,
83. Truhar, N., An efficient algorithm for damper optimization for linear vibrating systems using Lyapunov equation, J. Comp. Appl. Math. **172** (2004) 169–182.
84. Truhar, N., Veselić, K., An efficient method for estimating the optimal dampers' viscosity for linear vibrating systems using Lyapunov equation, 2007, to appear in SIMAX.
85. Veselić, K., Global Jacobi method for a symmetric indefinite problem $Sx = \lambda Tx$, Comp. Meth. Appl. Mech. Engrg. **38** (1983) 273–290.
86. Veselić, K., A Jacobi eigenreduction algorithm for definite matrix pairs, Numer. Math. **64** (1993) 241–269.
87. Veselić, K., On linear vibrational systems with one dimensional damping, Appl. Anal. **29** (1988) 1–18.
88. Veselić, K., On linear vibrational systems with one dimensional damping II, Integral Eq. Operator Th. **13** (1990) 883–897.
89. Veselić, K., Bounds for exponentially stable semigroups, Linear Algebra Appl. **358** (2003) 309–333.

90. Veselić, K., Bounds for contractive semigroups and second order systems, Operator Theory: Advances and Applications, **162** (2005) 293-308.
91. Veselić, K:, Slapničar, I., On spectral condition of J-Hermitian operators, Glasnik Mat. **35(55)** (2000) 3–23.

Index