Chapter 1

Introduction

Before we start with the subject of this notes we want to show how one actually arrives at large eigenvalue problems in practice. In the following, we restrict ourselves to problems from physics [4, 7] and computer science.

1.1 What makes eigenvalues interesting?

In physics, eigenvalues are usually connected to vibrations. Objects like violin strings, drums, bridges, sky scrapers can swing. They do this at certain frequencies. And in some situations they swing so much that they are destroyed. On November 7, 1940, the Tacoma narrows bridge collapsed, less than half a year after its opening. Strong winds excited the bridge so much that the platform in reinforced concrete fell into pieces. A few years ago the London millennium footbridge started wobbling in a way that it had to be closed. The wobbling had been excited by the pedestrians passing the bridge. These are prominent examples of vibrating structures.

But eigenvalues appear in many other places. Electric fields in cyclotrones, a special form of particle accelerators, have to vibrate in a precise manner, in order to accelerate the charged particles that circle around its center. The solutions of the Schrödinger equation from quantum physics and quantum chemistry have solutions that correspond to vibrations of the, say, molecule it models. The eigenvalues correspond to energy levels that molecule can occupy.

Many characteristic quantities in science *are* eigenvalues:

- decay factors,
- frequencies,
- norms of operators (or matrices),
- singular values,
- condition numbers.

In the sequel we give a number of examples that show why computing eigenvalues is important. At the same time we introduce some notation.

1.2 Example 1: The vibrating string

1.2.1 Problem setting

Let us consider a string as displayed in Fig. 1.1. The string is clamped at both ends,

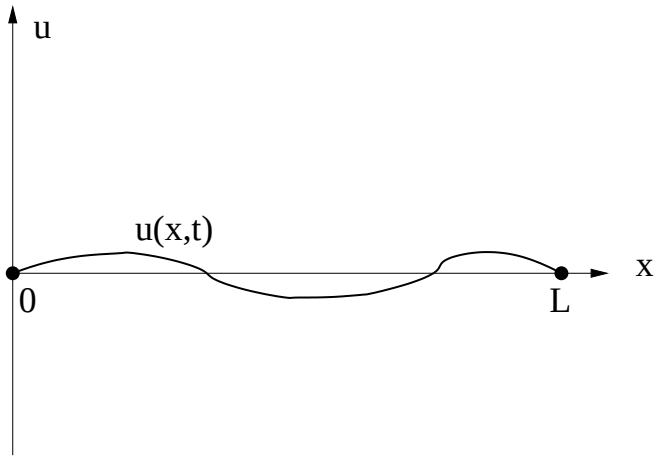


Figure 1.1: A vibrating string clamped at both ends.

at $x = 0$ and $x = L$. The x-axis coincides with the string's equilibrium position. The displacement of the rest position at x, $0 < x < L$, and time t is denoted by $u(x, t)$.

We will assume that the spatial derivatives of u are not very large:

$$\left|\frac{\partial u}{\partial x}\right| \quad \text{is small.}$$

This assumption entails that we may neglect terms of higher order.

Let $v(x, t)$ be the velocity of the string at position x and at time t. Then the kinetic energy of a string section ds of mass $dm = \rho\, ds$ is given by

$$dT = \frac{1}{2} dm\, v^2 = \frac{1}{2}\rho\, ds\, \left(\frac{\partial u}{\partial t}\right)^2. \tag{1.1}$$

From Fig. 1.2 we see that $ds^2 = dx^2 + \left(\frac{\partial u}{\partial x}\right)^2 dx^2$ and thus

$$\frac{ds}{dx} = \sqrt{1 + \left(\frac{\partial u}{\partial x}\right)^2} = 1 + \frac{1}{2}\left(\frac{\partial u}{\partial x}\right)^2 + \text{ higher order terms.}$$

Plugging this into (1.1) and omitting also the second order term (leaving just the number 1) gives

$$dT = \frac{\rho\, dx}{2}\left(\frac{\partial u}{\partial t}\right)^2.$$

The kinetic energy of the whole string is obtained by integrating over its length,

$$T = \int_0^L dT(x) = \frac{1}{2}\int_0^L \rho(x)\left(\frac{\partial u}{\partial t}\right)^2 dx$$

The potential energy of the string has two components

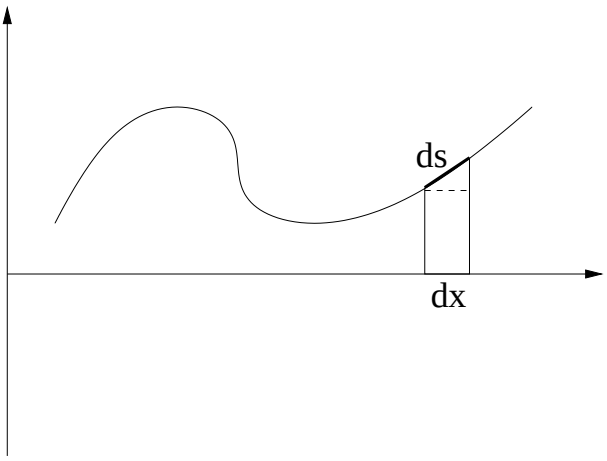


Figure 1.2: A vibrating string, local picture.

1. the stretching times the exerted strain τ.

$$\tau \int_0^L ds - \tau \int_0^L dx = \tau \int_0^L \left(\sqrt{1 + \left(\frac{\partial u}{\partial x} \right)^2} - 1 \right) dx$$
$$= \tau \int_0^L \left(\frac{1}{2} \left(\frac{\partial u}{\partial x} \right)^2 + \text{ higher order terms} \right) dx$$

2. exterior forces of density f

$$- \int_0^L f u dx$$

Summing up, the kinetic energy of the string becomes

$$V = \int_0^L \left(\frac{\tau}{2} \left(\frac{\partial u}{\partial x} \right)^2 - f u \right) dx \tag{1.2}$$

To consider the motion (vibration) of the string in a certain time interval $t_1 \leq t \leq t_2$ we form the integral

$$\begin{aligned} I(u) &= \int_{t_1}^{t_2} (T - V)\, dt \\ &= \frac{1}{2} \int_{t_1}^{t_2} \int_0^L \left[\rho(x) \left(\frac{\partial u}{\partial t} \right)^2 - \tau \left(\frac{\partial u}{\partial x} \right)^2 - f u \right] dx\, dt \end{aligned} \tag{1.3}$$

Here functions $u(x,t)$ are admitted that are differentiable with respect to x and t and satisfy the **boundary conditions (BC)** that correspond to the clamping,

$$u(0,t) = u(L,t) = 0, \qquad t_1 \leq t \leq t_2, \tag{1.4}$$

as well as given **initial conditions** and **end conditions**,

$$\begin{aligned} u(x,t_1) &= u_1(x), \\ u(x,t_2) &= u_2(x), \end{aligned} \qquad 0 < x < L. \tag{1.5}$$

According to the **principle of Hamilton** a mechanical system with kinetic energy T and potential energy V behaves in a time interval $t_1 \le t \le t_2$ for given initial and end positions such that

$$I = \int_{t_1}^{t_2} L\,dt, \qquad L = T - V,$$

is minimized.

Let $u(x,t)$ be such that $I(u) \le I(w)$ *for all* w, that satisfy the initial, end, and boundary conditions. Let $w = u + \varepsilon\, v$ with

$$v(0,t) = v(L,t) = 0, \qquad v(x,t_1) = v(x,t_2) = 0.$$

v is called a *variation*. We now consider $I(u + \varepsilon\, v)$ as a function of ε. Then we have the equivalence

$$I(u) \text{ minimal} \quad \Longleftrightarrow \quad \boxed{\frac{dI}{d\varepsilon}(u) = 0 \text{ for all admitted } v.}$$

Plugging $u + \varepsilon\, v$ into eq. (1.3) we obtain

$$\begin{aligned} I(u+\varepsilon\, v) &= \frac{1}{2}\int_{t_1}^{t_2}\int_{0}^{L}\left[\rho(x)\left(\frac{\partial(u+\varepsilon\, v)}{\partial t}\right)^2 - \tau\left(\frac{\partial(u+\varepsilon\, v)}{\partial x}\right)^2 - 2f(u+\varepsilon\, v)\right] dx\,dt \\ &= I(u) + \varepsilon\int_{t_1}^{t_2}\int_{0}^{L}\left[\rho(x)\frac{\partial u}{\partial t}\frac{\partial v}{\partial t} - \tau\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + 2fv\right] dx\,dt + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{1.6}$$

Thus,

$$\frac{\partial I}{\partial \varepsilon} = \int_{t_1}^{t_2}\int_{0}^{L}\left[-\rho\frac{\partial^2 u}{\partial t^2} + \tau\frac{\partial^2 u}{\partial x^2} + 2\,f\right] v\,dx\,dt = 0$$

for all admissible v. Therefore, the bracketed expression must vanish,

$$-\rho\frac{\partial^2 u}{\partial t^2} + \tau\frac{\partial^2 u}{\partial x^2} + 2\,f = 0. \tag{1.7}$$

This last differential equation is named **Euler-Lagrange equation**.

Next we want to solve a differential equation of the form

$$\boxed{\begin{gathered} -\rho(x)\frac{\partial^2 u}{\partial t^2} + \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right) + q(x)u(x,t) = 0. \\ u(0,t) = u(1,t) = 0 \end{gathered}} \tag{1.8}$$

which is a generalization of the Euler-Lagrange equation (1.7) Here, $\rho(x)$ plays the role of a mass density, $p(x)$ of a locally varying elasticity module. We do not specify initial and end conditions for the moment.

From physics we know that $\rho(x) > 0$ and $p(x) > 0$ for all x. These properties are of importance also from a mathematical view point! For simplicity, we assume that $\rho(x) = 1$.

1.2.2 The method of separation of variables

For the solution u in (1.8) we make the *ansatz*

$$u(x,t) = v(t)w(x). \tag{1.9}$$

Here, v is a function that depends only on the time t, while w depends only on the spacial variable x. With this ansatz (1.8) becomes

$$v''(t)w(x) - v(t)(p(x)w'(x))' + q(x)v(t)w(x) = 0. \tag{1.10}$$

Now we *separate* the variables depending on t from those depending on x,

$$\frac{v''(t)}{v(t)} = \frac{1}{w(x)}(p(x)w'(x))' + q(x).$$

This equation holds for any t and x. We can vary t and x independently of each other without changing the value on each side of the equation. Therefore, each side of the equation must be equal to a constant value. We denote this value by $-\lambda$. Thus, from the left side we obtain the equation

$$-v''(t) = \lambda v(t). \tag{1.11}$$

This equation has the well-known solution $v(t) = a \cdot \cos(\sqrt{\lambda}t) + b \cdot \sin(\sqrt{\lambda}t)$ where $\lambda > 0$ is assumed. The right side of (1.10) gives a so-called **Sturm-Liouville problem**

$$\boxed{-(p(x)w'(x))' + q(x)w(x) = \lambda w(x), \qquad w(0) = w(1) = 0} \tag{1.12}$$

A value λ for which (1.12) has a *non-trivial* solution w is called an **eigenvalue**; w is a corresponding **eigenfunction**. It is known that all eigenvalues of (1.12) are positive. By means of our ansatz (1.9) we get

$$u(x,t) = w(x)\left[a \cdot \cos(\sqrt{\lambda}t) + b \cdot \sin(\sqrt{\lambda}t)\right]$$

as a solution of (1.8). It is known that (1.12) has infinitely many real positive eigenvalues $0 < \lambda_1 \le \lambda_2 \le \cdots$, $(\lambda_k \underset{k\to\infty}{\longrightarrow} \infty)$. (1.12) has a non-zero solution, say $w_k(x)$ *only* for these particular values λ_k. Therefore, the general solution of (1.8) has the form

$$u(x,t) = \sum_{k=0}^{\infty} w_k(x)\left[a_k \cdot \cos(\sqrt{\lambda_k}\,t) + b_k \cdot \sin(\sqrt{\lambda_k}\,t)\right]. \tag{1.13}$$

The coefficients a_k and b_k are determined by initial and end conditions. We could, e.g., require that

$$u(x,0) = \sum_{k=0}^{\infty} a_k w_k(x) = u_0(x),$$
$$\frac{\partial u}{\partial t}(x,0) = \sum_{k=0}^{\infty} \sqrt{\lambda_k}\, b_k w_k(x) = u_1(x),$$

where u_0 and u_1 are given functions. It is known that the w_k form an orthogonal basis in the space of square integrable functions $L_2(0,1)$. Therefore, it is not difficult to compute the coefficients a_k and b_k.

In concluding, we see that the difficult problem to solve is the eigenvalue problem (1.12). Knowing the eigenvalues and eigenfunctions the general solution of the time-dependent problem (1.8) is easy to form.

Eq. (1.12) can be solved analytically only in very special situation, e.g., if all coefficients are constants. In general a *numerical method* is needed to solve the Sturm-Liouville problem (1.12).

1.3 Numerical methods for solving 1-dimensional problems

In this section we consider three methods to solve the Sturm-Liouville problem.

1.3.1 Finite differences

We approximate $w(x)$ by its values at the discrete points $x_i = ih$, $h = 1/(n+1)$, $i = 1, \ldots, n$.

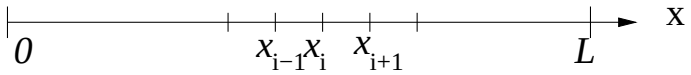


Figure 1.3: Grid points in the interval $(0, L)$.

At point x_i we approximate the derivatives by **finite differences**. We proceed as follows. First we write

$$\frac{d}{dx}g(x_i) \approx \frac{g(x_{i+\frac{1}{2}}) - g(x_{i+\frac{1}{2}})}{h}.$$

For $g = p\frac{dw}{dx}$ we get

$$g(x_{i+\frac{1}{2}}) = p(x_{i+\frac{1}{2}})\frac{w(x_{i+1}) - w(x_i)}{h}$$

and finally, for $i = 1, \ldots, n$,

$$\begin{aligned} -\frac{d}{dx}\left(p\frac{dw}{dx}(x_i)\right) &\approx -\frac{1}{h}\left[p(x_{i+\frac{1}{2}})\frac{w(x_{i+1}) - w(x_i)}{h} - p(x_{i-\frac{1}{2}})\frac{w(x_i) - w(x_{i-1})}{h}\right] \\ &= \frac{1}{h^2}\left[p(x_{i-\frac{1}{2}})w_{i-1} + (p(x_{i-\frac{1}{2}}) + p(x_{i+\frac{1}{2}}))w_i - p(x_{i+\frac{1}{2}})w_{i+1}\right]. \end{aligned}$$

Note that at the interval endpoints $w_0 = w_{n+1} = 0$.

We can collect all equations in a matrix equation,

$$\begin{bmatrix} \frac{p(x_{\frac{1}{2}})+p(x_{\frac{3}{2}})}{h^2} + q(x_1) & -p(x_{\frac{3}{2}}) & & \\ -p(x_{\frac{3}{2}}) & \frac{p(x_{\frac{3}{2}})+p(x_{\frac{5}{2}})}{h^2} + q(x_2) & -p(x_{\frac{5}{2}}) & \\ & -p(x_{\frac{5}{2}}) & \ddots & \ddots \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} = \lambda \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix}$$

or, briefly,

$$A\mathbf{w} = \lambda\mathbf{w}. \tag{1.14}$$

By construction, A is symmetric and tridiagonal. One can show that it is positive definite as well.

1.3.2 The finite element method

We write (1.12) in the form

Find a twice differentiable function w with $w(0) = w(1) = 0$ such that

$$\int_0^1 \left[-(p(x)w'(x))' + q(x)w(x) - \lambda w(x)\right]\phi(x)dx = 0$$

for all smooth functions ϕ that satisfy $\phi(0) = \phi(1) = 0$.

To relax the requirements on w we integrate by parts and get the new so-called weak form of the problem:

Find a differentiable function w with $w(0) = w(1) = 0$ such that

$$\int_0^1 \left[p(x)w(x)'\phi'(x) + q(x)w(x)\phi(x) - \lambda w(x)\phi(x)\right] dx = 0 \tag{1.15}$$

for all differentiable functions ϕ that satisfy $\phi(0) = \phi(1) = 0$.

Remark: Requiring differentiability is too strong and does not lead to a mathematically suitable formulation. In particular, the test functions that will be used below are not differentiable in the classical sense. It is more appropriate to require w and ϕ to be *weakly* differentiable. In terms of Sobolev spaces: $w, \phi \in H_0^1([0,1])$. An introduction to Sobolev spaces is, however, beyond the scope of these notes.

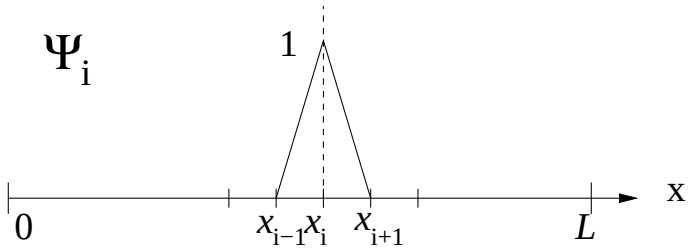


Figure 1.4: A basis function of the finite element space: a hat function.

We now write w as the linear combination

$$w(x) = \sum_{i=1}^{n} \xi_i \, \Psi_i(x), \tag{1.16}$$

where

$$\Psi_i(x) = \left(1 - \frac{|x - x_i|}{h}\right)_+ = \max\{0,\ 1 - \frac{|x - x_i|}{h}\}, \tag{1.17}$$

is the function that is linear in each interval (x_i, x_{i+1}) and satisfies

$$\Psi_i(x_k) = \delta_{ik} := \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases}$$

An example of such a basis function, a so-called *hat function*, is given in Fig. 1.4.

We now replace w in (1.15) by the linear combination (1.16), and replace testing 'against all ϕ' by testing against all Ψ_j. In this way (1.15) becomes

$$\int_0^1 \left(-p(x)(\sum_{i=1}^{n} \xi_i \, \Psi_i'(x))\Psi_j'(x) + (q(x) - \lambda) \sum_{i=1}^{n} \xi_i \, \Psi_i(x)\Psi_j(x) \right) dx, \quad \text{for all } j,$$

or,

$$\boxed{\sum_{i=1}^{n} \xi_i \int_0^1 \left(p(x)\Psi_i'(x)\Psi_j'(x) + (q(x) - \lambda)\Psi_i(x)\Psi_j(x)\right) dx = 0, \quad \text{for all } j.} \tag{1.18}$$

These last equations are called the **Rayleigh–Ritz–Galerkin** equations. Unknown are the n values ξ_i and the eigenvalue λ. In matrix notation (1.18) becomes

$$A\mathbf{x} = \lambda M\mathbf{x} \tag{1.19}$$

with

$$a_{ij} = \int_0^1 \left(p(x)\Psi_i'\Psi_j' + q(x)\Psi_i\Psi_j\right)\,dx \quad \text{and} \quad m_{ij} = \int_0^1 \Psi_i\Psi_j\,dx$$

For the specific case $p(x) = 1 + x$ and $q(x) = 1$ we get

$$a_{kk} = \int_{(k-1)h}^{kh} \left[(1+x)\frac{1}{h^2} + \left(\frac{x-(k-1)h}{h}\right)^2\right] dx + \int_{kh}^{(k+1)h} \left[(1+x)\frac{1}{h^2} + \left(\frac{(k+1)h-x}{h}\right)^2\right] dx = 2(n+1+k) + \frac{2}{3}\frac{1}{n+1}$$

$$a_{k,k+1} = \int_{kh}^{(k+1)h} \left[(1+x)\frac{1}{h^2} + \frac{(k+1)h-x}{h}\cdot\frac{x-kh}{h}\right] dx = -n - \frac{3}{2} - k + \frac{1}{6}\frac{1}{n+1}$$

In the same way we get

$$M = \frac{1}{6(n+1)}\begin{bmatrix} 4 & 1 & & \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{bmatrix}$$

Notice that both matrices A and M are symmetric tridiagonal and positive definite.

1.3.3 Global functions

Formally we proceed as with the finite element method. But now we choose the $\Psi_k(x)$ to be functions with global support. We could, e.g., set

$$\Psi_k(x) = \sin k\pi x,$$

functions that are differentiable and satisfy the homogeneous boundary conditions. The Ψ_k are eigenfunctions of the nearby problem $-u''(x) = \lambda u(x)$, $u(0) = u(1) = 0$ corresponding to the eigenvalue $k^2\pi^2$. The elements of matrix A are given by

$$a_{kk} = \int_0^1 \left[(1+x)k^2\pi^2\cos^2 k\pi x + \sin^2 k\pi x\right]\,dx = \frac{3}{4}k^2\pi^2 + \frac{1}{2},$$

$$a_{kj} = \int_0^1 \left[(1+x)kj\pi^2\cos k\pi x\cos j\pi x + \sin k\pi x\sin j\pi x\right]\,dx = \frac{kj(k^2+j^2)((-1)^{k+j}-1)}{(k^2-j^2)^2}, \quad k \neq j.$$

1.3.4 A numerical comparison

We consider the above 1-dimensional eigenvalue problem

$$-((1+x)w'(x))' + w(x) = \lambda w(x), \qquad w(0) = w(1) = 0, \tag{1.20}$$

and solve it with the finite difference and finite element methods as well as with the global functions method. The results are given in Table 1.1.

Clearly the global function method is the most powerful of them all. With 80 basis functions the eigenvalues all come right. The convergence rate is exponential.

With the finite difference and finite element methods the eigenvalues exhibit quadratic convergence rates. If the mesh width h is reduced by a factor of $q = 2$, the error in the eigenvalues is reduced by the factor $q^2 = 4$.

1.4 Example 2: The heat equation

The instationary temperature distribution $u(\mathbf{x}, t)$ in an insulated container satisfies the equations

$$\begin{aligned} \frac{\partial u(\mathbf{x},t)}{\partial t} - \Delta u(\mathbf{x},t) &= 0, && \mathbf{x} \in \Omega,\ t > 0, \\ \frac{\partial u(\mathbf{x},t)}{\partial n} &= 0, && \mathbf{x} \in \partial\Omega,\ t > 0, \\ u(\mathbf{x},0) &= u_0(\mathbf{x}), && \mathbf{x} \in \Omega. \end{aligned} \tag{1.21}$$

Here Ω is a 3-dimensional domain[1]with boundary $\partial\Omega$. $u_0(\mathbf{x}), \mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$, is a given bounded, sufficiently smooth function. $\Delta u = \sum \frac{\partial^2 u}{\partial {x_i}^2}$ is called the *Laplace operator* and $\frac{\partial u}{\partial n}$ denotes the derivative of u in direction of the outer normal vector $\mathbf{n}$. To solve the heat equation the **method of separation of variables** is employed. We write u in the form

$$u(\mathbf{x}, t) = v(t)w(\mathbf{x}). \tag{1.22}$$

If a constant λ can be found such that

$$\boxed{\begin{aligned} \Delta w(\mathbf{x}) + \lambda w(\mathbf{x}) = 0, \quad w(\mathbf{x}) \neq 0, \quad & \mathbf{x} \text{ in } \Omega, \\ \frac{\partial w(\mathbf{x},t)}{\partial n} = 0, \quad & \mathbf{x} \text{ on } \partial\Omega, \end{aligned}} \tag{1.23}$$

then the product $u = vw$ is a solution of (1.21) if and only if

$$\frac{dv(t)}{dt} + \lambda v(t) = 0, \tag{1.24}$$

the solution of which has the form $a \cdot \exp(-\lambda t)$. By separating variables, the problem (1.21) is divided in two subproblems that are hopefully easier to solve. A value λ, for which (1.23) has a *nontrivial* (i.e. a nonzero) solution is called an *eigenvalue*; w then is called a *corresponding eigenfunction*.

If λ_n is an eigenvalue of problem (1.23) with corresponding eigenfunction w_n, then

$$e^{-\lambda_n t} w_n(\mathbf{x})$$

is a solution of the first two equations in (1.21). It is known that equation (1.23) has *infinitely many* real eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots, (\lambda_n \underset{t\to\infty}{\longrightarrow} \infty)$. Multiple eigenvalues are counted according to their multiplicity. An arbitrary bounded piecewise continuous function can be represented as a linear combination of the eigenfunctions $w_1, w_2, \ldots$. Therefore, the solution of (1.21) can be written in the form

$$u(\mathbf{x}, t) = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t} w_n(\mathbf{x}), \tag{1.25}$$

[1]In the sequel we understand a domain to be bounded and simply connected.

Finite difference method				
k	$\lambda_k(n=10)$	$\lambda_k(n=20)$	$\lambda_k(n=40)$	$\lambda_k(n=80)$
1	15.245	15.312	15.331	15.336
2	56.918	58.048	58.367	58.451
3	122.489	128.181	129.804	130.236
4	206.419	224.091	229.211	230.580
5	301.499	343.555	355.986	359.327
6	399.367	483.791	509.358	516.276
7	492.026	641.501	688.398	701.185
8	578.707	812.933	892.016	913.767
9	672.960	993.925	1118.969	1153.691
10	794.370	1179.947	1367.869	1420.585

Finite element method				
k	$\lambda_k(n=10)$	$\lambda_k(n=20)$	$\lambda_k(n=40)$	$\lambda_k(n=80)$
1	15.447	15.367	15.345	15.340
2	60.140	58.932	58.599	58.511
3	138.788	132.657	130.979	130.537
4	257.814	238.236	232.923	231.531
5	426.223	378.080	365.047	361.648
6	654.377	555.340	528.148	521.091
7	949.544	773.918	723.207	710.105
8	1305.720	1038.433	951.392	928.983
9	1702.024	1354.106	1214.066	1178.064
10	2180.159	1726.473	1512.784	1457.733

Global function method				
k	$\lambda_k(n=10)$	$\lambda_k(n=20)$	$\lambda_k(n=40)$	$\lambda_k(n=80)$
1	15.338	15.338	15.338	15.338
2	58.482	58.480	58.480	58.480
3	130.389	130.386	130.386	130.386
4	231.065	231.054	231.053	231.053
5	360.511	360.484	360.483	360.483
6	518.804	518.676	518.674	518.674
7	706.134	705.631	705.628	705.628
8	924.960	921.351	921.344	921.344
9	1186.674	1165.832	1165.823	1165.822
10	1577.340	1439.083	1439.063	1439.063

Table 1.1: Numerical solutions of problem (1.20)

where the coefficients c_n are determined such that

$$u_0(\mathbf{x}) = \sum_{n=1}^{\infty} c_n w_n(\mathbf{x}). \tag{1.26}$$

The smallest eigenvalue of (1.23) is $\lambda_1 = 0$ with $w_1 = 1$ and $\lambda_2 > 0$. Therefore we see from (1.25) that

$$u(\mathbf{x}, t) \underset{t\to\infty}{\longrightarrow} c_1. \tag{1.27}$$

Thus, in the limit (i.e., as t goes to infinity), the temperature will be constant in the whole container. The convergence rate towards this equilibrium is determined by the smallest *positive* eigenvalue λ_2 of (1.23):

$$\begin{aligned}\|u(\mathbf{x}, t) - c_1\| = \|\sum_{n=2}^{\infty} c_n e^{-\lambda_n t} w_n(\mathbf{x})\| &\le \sum_{n=2}^{\infty} |e^{-\lambda_n t}| \|c_n w_n(\mathbf{x})\| \\ &\le e^{-\lambda_2 t} \sum_{n=2}^{\infty} \|c_n w_n(\mathbf{x})\| \le e^{-\lambda_2 t} \|u_0(\mathbf{x})\|.\end{aligned}$$

Here we have assumed that the value of the constant function $w_1(\mathbf{x})$ is set to unity.

1.5 Example 3: The wave equation

The air pressure $u(\mathbf{x}, t)$ in a volume with acoustically "hard" walls satisfies the equations

$$\frac{\partial^2 u(\mathbf{x}, t)}{\partial t^2} - \Delta u(\mathbf{x}, t) = 0, \qquad \mathbf{x} \in \Omega,\ t > 0, \tag{1.28}$$

$$\frac{\partial u(\mathbf{x}, t)}{\partial n} = 0, \qquad \mathbf{x} \in \partial\Omega,\ t > 0, \tag{1.29}$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega, \tag{1.30}$$

$$\frac{\partial u(\mathbf{x}, 0)}{\partial t} = u_1(\mathbf{x}), \qquad \mathbf{x} \in \Omega. \tag{1.31}$$

Sound propagates with the speed $-\nabla\mathbf{u}$, i.e. along the (negative) gradient from high to low pressure.

To solve the wave equation we proceed as with the heat equation in section 1.4: separation of u according to (1.22) leads again to equation (1.23) but now together with

$$\frac{d^2 v(t)}{dt^2} + \lambda v(t) = 0. \tag{1.32}$$

We know this equation from the analysis of the vibrating sting, see (1.11). From there we know that the general solution of the wave equation has the form

$$u(x, t) = \sum_{k=0}^{\infty} w_k(x) \left[a_k \cdot \cos(\sqrt{\lambda_k}\, t) + b_k \cdot \sin(\sqrt{\lambda_k}\, t) \right]. \tag{1.13}$$

where the w_k, $k = 1, 2, \ldots$ are the eigenfunctions of the eigenvalue problem (1.23). The coefficients a_k and b_k are determined by eqrefeq:wave3 and eqrefeq:wave4.

If a harmonic oscillation is forced on the system, an *inhomogeneous* problem

$$\frac{\partial^2 u(\mathbf{x},t)}{\partial t^2} - \Delta u(\mathbf{x},t) = f(\mathbf{x},t), \tag{1.33}$$

is obtained. The boundary and initial conditions are taken from (1.28)–(1.31). This problem can be solved by setting

$$\begin{aligned} u(\mathbf{x},t) &:= \sum_{n=1}^{\infty} \tilde{v}_n(t) w_n(\mathbf{x}), \\ f(\mathbf{x},t) &:= \sum_{n=1}^{\infty} \phi_n(t) w_n(\mathbf{x}). \end{aligned} \tag{1.34}$$

With this approach, $\tilde{v}_n$ has to satisfy equation

$$\frac{d^2 \tilde{v}_n}{dt^2} + \lambda_n \tilde{v}_n = \phi_n(t). \tag{1.35}$$

If $\phi_n(t) = a \sin \omega t$, then the solution becomes

$$\tilde{v}_n = A_n \cos \sqrt{\lambda_n} t + B_n \sin \sqrt{\lambda_n} t + \frac{1}{\lambda_n - \omega^2} a \sin \omega t. \tag{1.36}$$

A_n and B_n are real constants that are determined by the initial conditions. If ω gets close to $\sqrt{\lambda_1}$, then the last term can be very large. In the limit, if $\omega = \sqrt{\lambda_n}$, $\tilde{v}_n$ gets the form

$$\tilde{v}_n = A_n \cos \sqrt{\lambda_n} t + B_n \sin \sqrt{\lambda_n} t + at \sin \omega t. \tag{1.37}$$

In this case, $\tilde{v}_n$ is not bounded in time anymore. This phenomenon is called *resonance*. Often resonance is not desirable; it may, e.g., mean the blow up of some structure. In order to prevent resonances eigenvalues have to be known. Possible remedies are changing the domain (the structure).

Remark 1.1. Vibrating membranes satisfy the wave equation, too. In general the boundary conditions are different from (1.29). If the membrane (of a drum) is fixed at its boundary, the condition

$$u(\mathbf{x},t) = 0 \tag{1.38}$$

is imposed. This boundary conditions is called *Dirichlet boundary conditions*. The boundary conditions in (1.21) and (1.29) are called *Neumann boundary conditions*. Combinations of these two can occur. □

1.6 Numerical methods for solving the Laplace eigenvalue problem in 2D

In this section we again consider the eigenvalue problem

$$-\Delta u(\mathbf{x}) = \lambda u(\mathbf{x}), \qquad \mathbf{x} \in \Omega, \tag{1.39}$$

with the more general boundary conditions

$$u(\mathbf{x}) = 0, \qquad \mathbf{x} \in C_1 \subset \partial\Omega, \tag{1.40}$$

$$\frac{\partial u}{\partial n}(\mathbf{x}) + \alpha(\mathbf{x})u(\mathbf{x}) = 0, \qquad \mathbf{x} \in C_2 \subset \partial\Omega. \tag{1.41}$$

Here, C_1 and C_2 are *disjoint* subsets of $\partial\Omega$ with $C_1 \cup C_2 = \partial\Omega$. We restrict ourselves in the following on *two-dimensional* domains and write (x, y) instead of (x_1, x_2).

In general it is not possible to solve a problem of the Form (1.39)–(1.41) exactly (analytically). Therefore one has to resort to numerical approximations. Because we cannot compute with infinitely many variables we have to construct a finite-dimensional eigenvalue problem that represents the given problem as well as possible, i.e., that yields good approximations for the desired eigenvalues and eigenvectors. Since finite-dimensional eigenvalue problem only have a finite number of eigenvalues one cannot expect to get good approximations for all eigenvalues of (1.39)–(1.41).

Two methods for the discretization of eigenvalue problems of the form (1.39)–(1.41) are the *Finite Difference Method* [1, 6] and the *Finite Element Method (FEM)* [5, 8]. We deal with these methods in the following subsections.

1.6.1 The finite difference method

In this section we just want to mediate some impression what the finite difference method is about. Therefore we assume for simplicity that the domain Ω is a square with sides of length 1: $\Omega = (0, 1) \times (0, 1)$. We consider the eigenvalue problem

$$\begin{aligned} -\Delta u(x, y) &= \lambda u(x, y), && 0 < x, y < 1 \\ u(0, y) = u(1, y) = u(x, 0) &= 0, && 0 < x, y < 1, \\ \frac{\partial u}{\partial n}(x, 1) &= 0, && 0 < x < 1. \end{aligned} \tag{1.42}$$

This eigenvalue problem occurs in the computation of eigenfrequencies and eigenmodes of a homogeneous quadratic membrane with three fixed and one free side. It can be solved analytically by separation of the two spatial variables x and y. The eigenvalues are

$$\lambda_{k,l} = \left(k^2 + \frac{(2l-1)^2}{4}\right)\pi^2, \quad k, l \in \mathbb{N},$$

and the corresponding eigenfunctions are

$$u_{k,l}(x, y) = \sin k\pi x \sin \frac{2l-1}{2}\pi y.$$

In the finite difference method one proceeds by defining a rectangular grid with grid points $(x_i, y_j), 0 \le i, j \le N$. The coordinates of the grid points are

$$(x_i, y_j) = (ih, jh), \qquad h = 1/N.$$

By a Taylor expansion one can show for sufficiently smooth functions u that

$$\begin{aligned} -\Delta u(x, y) = \frac{1}{h^2}&(4u(x, y) - u(x - h, y) - u(x + h, y) - u(x, y - h) - u(x, y + h)) \\ &+ O(h^2). \end{aligned}$$

It is therefore straightforward to replace the differential equation $\Delta u(x, y) + \lambda u(x, y) = 0$ by a difference equation at the interior grid points

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = \lambda h^2 u_{i,j}, \quad 0 < i, j < N. \tag{1.43}$$

We consider the unknown variables $u_{i,j}$ as approximations of the eigenfunctions at the grid points (i,j):

$$u_{i,j} \approx u(x_i, x_j). \tag{1.44}$$

The Dirichlet boundary conditions are replaced by the equations

$$u_{i,0} = u_{i,N} = u_{0,i}, \qquad 0 < i < N. \tag{1.45}$$

At the points at the upper boundary of Ω we first take the difference equation (1.43)

$$4u_{i,N} - u_{i-1,N} - u_{i+1,N} - u_{i,N-1} - u_{i,N+1} = \lambda h^2 u_{i,N}, \quad 0 \le i \le N. \tag{1.46}$$

The value $u_{i,N+1}$ corresponds to a grid point *outside* of the domain! However the Neumann boundary conditions suggest to reflect the domain at the upper boundary and to extend the eigenfunction symmetrically beyond the boundary. This procedure leads to the equation $u_{i,N+1} = u_{i,N-1}$. Plugging this into (1.46) and multiplying the new equation by the factor $1/2$ gives

$$2u_{i,N} - \frac{1}{2}u_{i-1,N} - \frac{1}{2}u_{i+1,N} - u_{i,N-1} = \frac{1}{2}\lambda h^2 u_{i,N}, \quad 0 < i < N. \tag{1.47}$$

In summary, from (1.43) and (1.47), taking into account that (1.45) we get the matrix equation

$$\begin{pmatrix}
4 & -1 & 0 & -1 & & & & & & & & \\
-1 & 4 & -1 & 0 & -1 & & & & & & & \\
0 & -1 & 4 & 0 & 0 & -1 & & & & & & \\
-1 & 0 & 0 & 4 & -1 & 0 & -1 & & & & & \\
 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & & & & \\
 & & -1 & 0 & -1 & 4 & 0 & 0 & -1 & & & \\
 & & & -1 & 0 & 0 & 4 & -1 & 0 & -1 & & \\
 & & & & -1 & 0 & -1 & 4 & -1 & 0 & -1 & \\
 & & & & & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\
 & & & & & & -1 & 0 & 0 & 2 & -\frac{1}{2} & 0 \\
 & & & & & & & -1 & 0 & -\frac{1}{2} & 2 & -\frac{1}{2} \\
 & & & & & & & & -1 & 0 & -\frac{1}{2} & 2
\end{pmatrix}
\begin{pmatrix} u_{1,1} \\ u_{1,2} \\ u_{1,3} \\ u_{2,1} \\ u_{2,2} \\ u_{2,3} \\ u_{3,1} \\ u_{3,2} \\ u_{3,3} \\ u_{4,1} \\ u_{4,2} \\ u_{4,3} \end{pmatrix}$$

$$= \lambda h^2 \begin{pmatrix}
1 & & & & & & & & & & & \\
 & 1 & & & & & & & & & & \\
 & & 1 & & & & & & & & & \\
 & & & 1 & & & & & & & & \\
 & & & & 1 & & & & & & & \\
 & & & & & 1 & & & & & & \\
 & & & & & & 1 & & & & & \\
 & & & & & & & 1 & & & & \\
 & & & & & & & & 1 & & & \\
 & & & & & & & & & \frac{1}{2} & & \\
 & & & & & & & & & & \frac{1}{2} & \\
 & & & & & & & & & & & \frac{1}{2}
\end{pmatrix}
\begin{pmatrix} u_{1,1} \\ u_{1,2} \\ u_{1,3} \\ u_{2,1} \\ u_{2,2} \\ u_{2,3} \\ u_{3,1} \\ u_{3,2} \\ u_{3,3} \\ u_{4,1} \\ u_{4,2} \\ u_{4,3} \end{pmatrix}. \tag{1.48}$$

For arbitrary $N > 1$ we define

$$\mathbf{u}_i := \begin{pmatrix} u_{i,1} \\ u_{i,2} \\ \vdots \\ u_{i,N-1} \end{pmatrix} \in \mathbb{R}^{N-1},$$

$$T := \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{(N-1)\times(N-1)},$$

$$I := \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{(N-1)\times(N-1)}.$$

In this way we obtain from (1.43), (1.45), (1.47) the discrete eigenvalue problem

$$\begin{pmatrix} T & -I & & \\ -I & T & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & \frac{1}{2}T \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_3 \\ \mathbf{u}_4 \end{pmatrix} = \lambda h^2 \begin{pmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & & \frac{1}{2}I \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{N-1} \\ \mathbf{u}_N \end{pmatrix} \tag{1.49}$$

of size $N \times (N-1)$. This is a **matrix eigenvalue problem** of the form

$$A\mathbf{x} = \lambda M\mathbf{x}, \tag{1.50}$$

where A and M are *symmetric* and M additionally is *positive definite*. If M is the identity matrix is, we call (1.50) a *special* and otherwise a *generalized* eigenvalue problem. In these lecture notes we deal with numerical methods, to solve eigenvalue problems like these.

In the case (1.49) it is easy to obtain a special (symmetric) eigenvalue problem by a simple transformation: By left multiplication by

$$\begin{pmatrix} I & & & \\ & I & & \\ & & I & \\ & & & \sqrt{2}I \end{pmatrix}$$

we obtain from (1.49)

$$\begin{pmatrix} T & -I & & \\ -I & T & -I & \\ & -I & T & -\sqrt{2}I \\ & & -\sqrt{2}I & T \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \\ \frac{1}{\sqrt{2}}\mathbf{u}_4 \end{pmatrix} = \lambda h^2 \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \\ \frac{1}{\sqrt{2}}\mathbf{u}_4 \end{pmatrix}. \tag{1.51}$$

A property common to matrices obtained by the finite difference method are its *sparsity*. Sparse matrices have only very few nonzero elements.

In real-world applications domains often cannot be covered easily by a rectangular grid. In this situation and if boundary conditions are complicated the method of finite differences can be difficult to implement.

Because of this the finite element method is often the method of choice.

1.6.2 The finite element method (FEM)

Let $(\lambda, u) \in \mathbb{R} \times V$ be an eigenpair of problem (1.39)–(1.41). Then

$$\int_\Omega (\Delta u + \lambda u) v \, dx \, dy = 0, \quad \forall v \in V, \tag{1.52}$$

where V is vector space of bounded twice differentiable functions that satisfy the boundary conditions (1.40)–(1.41). By partial integration (Green's formula) this becomes

$$\int_\Omega \nabla u \nabla v \, dx \, dy + \int_{\partial\Omega} \alpha \, u \, v \, ds = \lambda \int_\Omega u \, v \, dx \, dy, \quad \forall v \in V, \tag{1.53}$$

or

$$a(u, v) = (u, v), \qquad \forall v \in V \tag{1.54}$$

where

$$a(u, v) = \int_\Omega \nabla u \, \nabla v \, dx \, dy + \int_{\partial\Omega} \alpha \, u \, v \, ds, \quad \text{and} \quad (u, v) = \int_\Omega u \, v \, dx \, dy.$$

We complete the space V with respect to the Sobolev norm [8, 2]

$$\sqrt{\int_\Omega (u^2 + |\nabla u|^2) \, dx \, dy}$$

to become a Hilbert space H [2, 10]. H is the space of quadratic integrable functions with quadratic integrable first derivatives that satisfy the Dirichlet boundary conditions (1.40)

$$u(x, y) = 0 \quad (x, y) \in C_1.$$

(Functions in H in general no not satisfy the so-called *natural* boundary conditions (1.41).) One can show [10] that the eigenvalue problem (1.39)–(1.41) is equivalent with the eigenvalue problem

$$\begin{array}{l} \text{Find } (\lambda, u) \in \mathbb{R} \times H \text{ such that} \\ a(u, v) = \lambda(u, v) \quad \forall v \in H. \end{array} \tag{1.55}$$

(The essential point is to show that the eigenfunctions of (1.55) are elements of V.)

The Rayleigh–Ritz–Galerkin method

In the Rayleigh–Ritz–Galerkin method one proceeds as follows: A set of linearly independent functions

$$\phi_1(x, y), \cdots, \phi_n(x, y) \in H, \tag{1.56}$$

are chosen. These functions span a *subspace* S of H. Then, problem (1.55) is solved where H is replaced by S.

$$\begin{array}{l} \text{Find } (\lambda, u) \in \mathbb{R} \times S \text{ such that} \\ a(u, v) = \lambda(u, v) \quad \forall v \in S. \end{array} \tag{1.57}$$

With the Ritz ansatz [5]

$$u = \sum_{i=1}^{n} x_i \phi_i, \tag{1.58}$$

equation (1.57) becomes

$$\begin{aligned} &\text{Find } (\lambda, \mathbf{x}) \in \mathbb{R} \times \mathbb{R}^n \text{ such that} \\ &\sum_{i=1}^{n} x_i a(\phi_i, v) = \lambda \sum_{i=1}^{n} x_i (\phi_i, v), \quad \forall v \in S. \end{aligned} \tag{1.59}$$

Eq. (1.59) must hold *for all* $v \in S$, in particular for $v = \phi_1, \cdots, \phi_n$. But since the $\phi_i, 1 \le i \le n$, form a basis of S, equation (1.59) is equivalent with

$$\sum_{i=1}^{n} x_i a(\phi_i, \phi_j) = \lambda \sum_{i=1}^{n} x_i (\phi_i, \phi_j), \quad 1 \le j \le n. \tag{1.60}$$

This is a matrix eigenvalue problem of the form

$$A\mathbf{x} = \lambda M \mathbf{x} \tag{1.61}$$

where

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}, \quad M = \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nn} \end{pmatrix} \tag{1.62}$$

with

$$a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \nabla\phi_i \, \nabla\phi_j \, dx \, dy + \int_{\partial\Omega} \alpha \, \phi_i \, \phi_j \, ds$$

and

$$m_{ij} = (\phi_i, \phi_j) = \int_\Omega \phi_i \, \phi_j \, dx \, dy.$$

The **finite element method (FEM)** is a *special case* of the Rayleigh–Ritz method. In the FEM the subspace S and in particular the basis $\{\phi_i\}$ is chosen in a particularly clever way. For simplicity we assume that the domain Ω is a simply connected domain with a polygonal boundary, c.f. Fig 1.5. (This means that the boundary is composed of straight line segments entirely.) This domain is now partitioned into triangular subdomains

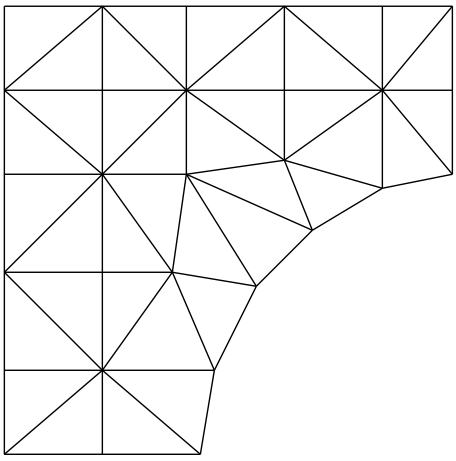


Figure 1.5: Triangulation of a domain Ω

$T_1, \cdots, T_N$, so-called *elements*, such that

$$\begin{aligned} T_i \cap T_j &= \emptyset, \qquad i \neq j, \\ \bigcup_e \overline{T_e} &= \overline{\Omega}. \end{aligned} \tag{1.63}$$

Finite element spaces for solving (1.39)–(1.41) are typically composed of functions that are *continuous* in Ω and are *polynomials* on the individual subdomains T_e. Such functions are called *piecewise polynomials*. Notice that this construction provides a subspace of the Hilbert space H but not of V, i.e., the functions in the finite element space are not very smooth and the natural boundary conditions are not satisfied.

An essential issue is the selection of the *basis* of the finite element space S. If $S_1 \subset H$ is the space of continuous, piecewise linear functions (the restriction to T_e is a polynomial of degree 1) then a function in S_1 is uniquely determined by its values at the vertices of the triangles. Let these *nodes*, except those on the boundary portion C_1, be numbered from 1 to n, see Fig. 1.6. Let the coordinates of the i-th node be (x_i, y_i). Then $\phi_i(x, y) \in S_1$ is defined by

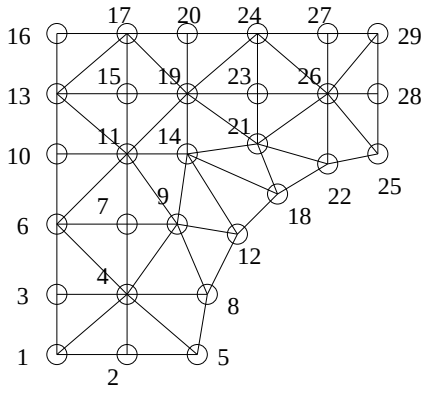


Figure 1.6: Numbering of nodes on Ω (piecewise linear polynomials)

$$\phi_i((x_j, y_j)) := \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{1.64}$$

A typical basis function ϕ_i is sketched in Figure 1.7.

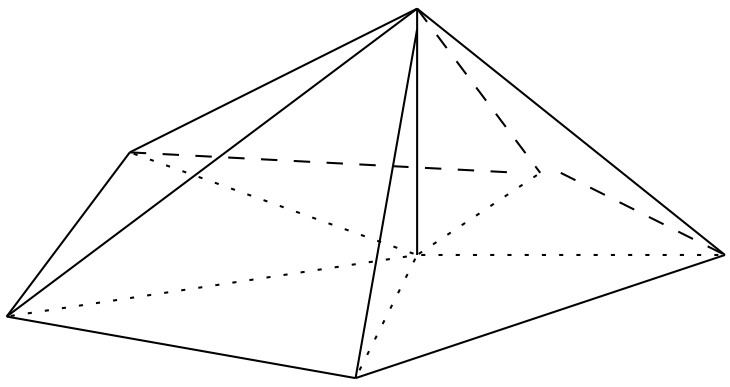


Figure 1.7: A piecewise linear basis function (or hat function)

Another often used finite element element space is $S_2 \subset H$, the space of continuous, piecewise quadratic polynomials. These functions are (or can be) uniquely determined by their values at the vertices and and edge midpoints of the triangle. The basis functions are defined according to (1.64). There are two kinds of basis functions ϕ_i now, first those that are 1 at a vertex and second those that are 1 in an edge midpoint, cf. Fig. 1.8. One

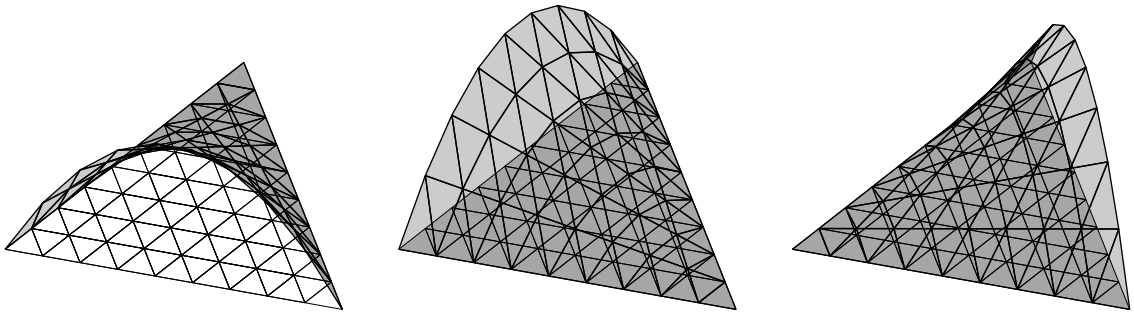


Figure 1.8: A piecewise quadratic basis function corresponding to a edge midpoint [3]

immediately sees that for most $i \neq j$

$$a(\phi_i, \phi_j) = 0, \quad (\phi_i, \phi_j) = 0. \tag{1.65}$$

Therefore the matrices A and M in (1.61) will be **sparse**. The matrix M is positive definite as

$$\mathbf{x}^T M \mathbf{x} = \sum_{i,j=1}^{N} x_i x_j m_{ij} = \sum_{i,j=1}^{N} x_i x_j (\phi_i, \phi_j) = (u, u) > 0, \quad u = \sum_{i=1}^{N} x_i \phi_i \neq 0, \tag{1.66}$$

because the ϕ_i are linearly independent and because $||u|| = \sqrt{(u,u)}$ is a norm. Similarly it is shown that

$$\mathbf{x}^T A \mathbf{x} \geq 0.$$

It is possible to have $\mathbf{x^T A x} = 0$ for a nonzero vector $\mathbf{x}$. This is the case if the constant function $u = 1$ is contained in S. This is the case if Neumann boundary conditions $\frac{\partial u}{\partial n} = 0$ are posed on the whole boundary $\partial\Omega$. Then,

$$u(x, y) = 1 = \sum_i \phi_i(x, y),$$

i.e., we have $\mathbf{x^T A x} = 0$ for $\mathbf{x} = [1, 1, \ldots, 1]$.

1.6.3 A numerical example

We want to determine the acoustic eigenfrequencies and corresponding modes in the interior of a car. This is of interest in the manufacturing of cars, since an appropriate shape of the form of the interior can suppress the often unpleasant droning of the motor. The problem is three-dimensional, but by separation of variables the problem can be reduced to two dimensions. If rigid, acoustically hard walls are assumed, the mathematical model of the problem is again the Laplace eigenvalue problem (1.23) together with Neumann boundary conditions. The domain is given in Fig. 1.9 where three finite element triangulations are shown with 87 (grid_1), 298 (grid_2), and 1095 (grid_3) vertices (nodes), respectively. The results obtained with piecewise linear polynomials are listed in Table 1.2. From the results

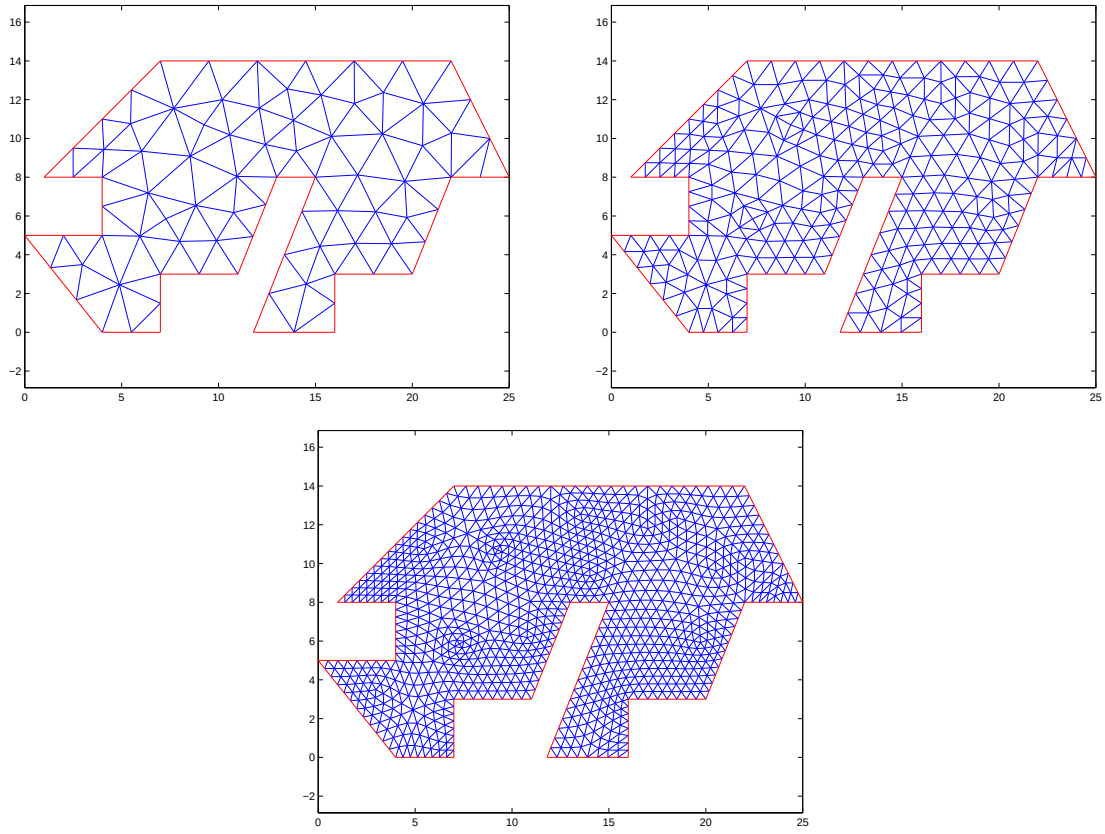


Figure 1.9: Three meshes for the car length cut

we notice the quadratic convergence rate. The smallest eigenvalue is always zero. The corresponding eigenfunction is the constant function. This function can be represented exactly by the finite element spaces, whence its value is correct (up to rounding error).

The fourth eigenfunction of the acoustic vibration problem is displayed in Fig. 1.10. The physical meaning of the function value is the difference of the pressure at a given location to the normal pressure. Large amplitudes thus means that the corresponding noise is very much noticable.

1.7 Cavity resonances in particle accelerators

The Maxwell equations in vacuum are given by

$$\begin{aligned}
\mathbf{curl}\,\mathbf{E}(\mathbf{x},t) &= -\frac{\partial \mathbf{B}}{\partial t}(\mathbf{x},t), && \text{(Faraday's law)}\\
\mathbf{curl}\,\mathbf{H}(\mathbf{x},t) &= \frac{\partial \mathbf{D}}{\partial t}(\mathbf{x},t) + \mathbf{j}(\mathbf{x},t), && \text{(Maxwell–Ampère law)}\\
\operatorname{div}\mathbf{D}(\mathbf{x},t) &= \rho(\mathbf{x},t), && \text{(Gauss's law)}\\
\operatorname{div}\mathbf{B}(\mathbf{x},t) &= 0. && \text{(Gauss's law – magnetic)}
\end{aligned}$$

where $\mathbf{E}$ is the electric field intensity, $\mathbf{D}$ is the electric flux density, $\mathbf{H}$ is the magnetic field intensity, $\mathbf{B}$ is the magnetic flux density, $\mathbf{j}$ is the electric current density, and ρ is the

Finite element method			
k	$\lambda_k(\text{grid}_1)$	$\lambda_k(\text{grid}_2)$	$\lambda_k(\text{grid}_3)$
1	0.0000	-0.0000	0.0000
2	0.0133	0.0129	0.0127
3	0.0471	0.0451	0.0444
4	0.0603	0.0576	0.0566
5	0.1229	0.1182	0.1166
6	0.1482	0.1402	0.1376
7	0.1569	0.1462	0.1427
8	0.2162	0.2044	0.2010
9	0.2984	0.2787	0.2726
10	0.3255	0.2998	0.2927

Table 1.2: Numerical solutions of acoustic vibration problem

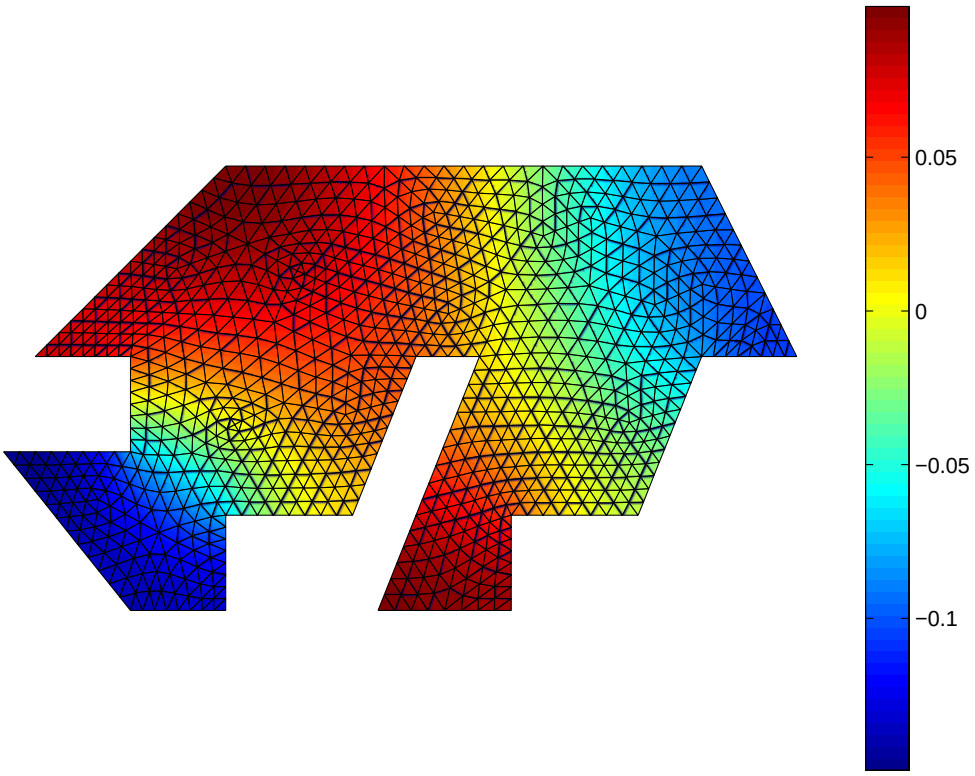


Figure 1.10: Fourth eigenmode of the acoustic vibration problem

electric charge density. Often the "optical" problem is analyzed, i.e. the situation when the cavity is not driven (cold mode), hence $\mathbf{j}$ and ρ are assumed to vanish.

Again by separating variables, i.e. assuming a *time harmonic* behavior f the fields, e.g.,

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{e}(\mathbf{x}) e^{i\omega t}$$

using the constitutive relations

$$\mathbf{D} = \epsilon \mathbf{E}, \quad \mathbf{B} = \mu \mathbf{H}, \quad \mathbf{j} = \sigma \mathbf{E},$$

one obtains after elimination of the magnetic field intensity the so called **time-harmonic Maxwell equations**

$$\begin{aligned} \mathbf{curl}\, \mu^{-1} \mathbf{curl}\, \mathbf{e}(\mathbf{x}) &= \lambda\, \epsilon\, \mathbf{e}(\mathbf{x}), && \mathbf{x} \in \Omega, \\ \operatorname{div} \epsilon\, \mathbf{e}(\mathbf{x}) &= 0, && \mathbf{x} \in \Omega, \\ \mathbf{n} \times \mathbf{e} &= 0, && \mathbf{x} \in \partial\Omega. \end{aligned} \tag{1.67}$$

Here, additionally, the cavity boundary $\partial\Omega$ is assumed to be *perfectly electrically conducting*, i.e. $\mathbf{E}(\mathbf{x}, t) \times \mathbf{n}(\mathbf{x}) = \mathbf{0}$ for $\mathbf{x} \in \partial\Omega$.

The eigenvalue problem (1.67) is a *constrained eigenvalue problem*. Only functions are taken into account that are divergence-free. This constraint is enforced by Lagrange multipliers. A weak formulation of the problem is then

Find $(\lambda, \mathbf{e}, p) \in \mathbb{R} \times H_0(\mathbf{curl}; \Omega) \times H_0^1(\Omega)$ *such that* $\mathbf{e} \neq \mathbf{0}$ *and*

$$\begin{aligned} &\text{(a)} \quad (\mu^{-1} \mathbf{curl}\, \mathbf{e}, \mathbf{curl}\, \mathbf{\Psi}) + (\mathbf{grad}\, p, \mathbf{\Psi}) = \lambda(\epsilon\, \mathbf{e}, \mathbf{\Psi}), && \forall \mathbf{\Psi} \in H_0(\mathbf{curl}; \Omega), \\ &\text{(b)} \quad (\mathbf{e}, \mathbf{grad}\, q) = 0, && \forall q \in H_0^1(\Omega). \end{aligned}$$

With the correct finite element discretization this problem turns in a matrix eigenvalue problem of the form

$$\begin{bmatrix} A & C \\ C^T & O \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \lambda \begin{bmatrix} M & O \\ O & O \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}.$$

The solution of this matrix eigenvalue problem correspond to vibrating electric fields.

1.8 Spectral clustering

This section is based on a tutorial by von Luxburg [9].

The goal of *clustering* is to group a given set of data points $x_1, \ldots, x_n$ into k clusters such that members from the same cluster are (in some sense) close to each other and members from different clusters are (in some sense) well separated from each other.

A popular approach to clustering is based on *similarity graphs*. For this purpose, we need to assume some notion of similarity $s(x_i, x_j) \geq 0$ between pairs of data points x_i and x_j. An undirected graph $G = (V, E)$ is constructed such that its vertices correspond to the data points: $V = \{x_1, \ldots, x_n\}$. Two vertices x_i, x_j are connected by an edge if the similarity s_{ij} between x_i and x_j is sufficiently large. Moreover, a weight $w_{ij} > 0$ is assigned to the edge, depending on s_{ij}. If two vertices are not connected we set $w_{ij} = 0$. The weights are collected into a *weighted adjacency matrix*

$$W = \left(w_{ij}\right)_{i,j=1}^{n}.$$

There are several possibilities to define the weights of the similarity graph associated with a set of data points and a similarity function:

fully connected graph All points with positive similarity are connected with each other and we simply set $w_{ij} = s(x_i, x_j)$. Usually, this will only result in reasonable clusters if the similarity function models locality very well. One example of such a similarity function is the Gaussian $s(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$, where $\|x_i - x_j\|$ is some distance measure (e.g., Euclidean distance) and σ is some parameter controlling how strongly locality is enforced.

k-nearest neighbors Two vertices x_i, x_j are connected if x_i is among the k-nearest neighbors of x_j or if x_j is among the k-nearest neighbors of x_i (in the sense of some distance measure). The weight of the edge between connected vertices x_i, x_j is set to the similarity function $s(x_i, x_j)$.

ϵ-neighbors Two vertices x_i, x_j are connected if their pairwise distance is smaller than ϵ for some parameter $\epsilon > 0$. In this case, the weights are usually chosen uniformly, e.g., $w_{ij} = 1$ if x_i, x_j are connected and $w_{ij} = 0$ otherwise.

Assuming that the similarity function is symmetric ($s(x_i, x_j) = s(x_j, x_i)$ for all x_i, x_j) all definitions above give rise to a symmetric weight matrix W. In practice, the choice of the most appropriate definition depends – as usual – on the application.

1.8.1 The Graph Laplacian

In the following we construct the so called *graph Laplacian*, whose spectral decomposition will later be used to determine clusters. For simplicity, we assume the weight matrix W to be symmetric. The degree of a vertex x_i is defined as

$$d_i = \sum_{j=1}^{n} w_{ij}. \tag{1.68}$$

In the case of an unweighted graph, the degree d_i amounts to the number of vertices adjacent to v_i (counting also v_i if $w_{ii} = 1$). The degree matrix is defined as

$$D = \operatorname{diag}(d_1, d_2, \ldots, d_n).$$

The *graph Laplacian* is then defined as

$$L = D - W. \tag{1.69}$$

By (1.68), the row sums of L are zero. In other words, $Le = 0$ with e the vector of all ones. This implies that 0 is an eigenvalue of L with the associated eigenvector e. Since L is symmetric all its eigenvalues are real and one can show that 0 is the smallest eigenvalue; hence L is positive semidefinite. It may easily happen that more than one eigenvalue is zero. For example, if the set of vertices can be divided into two subsets $\{x_1, \ldots, x_k\}$, $\{x_{k+1}, \ldots, x_n\}$, and vertices from one subset are not connected with vertices from the other subset, then

$$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix},$$

where L_1, L_2 are the Laplacians of the two disconnected components. Thus L has two eigenvectors $\binom{e}{0}$ and $\binom{0}{e}$ with eigenvalue 0. Of course, any linear combination of these two linearly independent eigenvectors is also an eigenvector of L.

The observation above leads to the basic idea behind spectral graph partitioning: If the vertices of the graph decompose into k connected components $V_1, \ldots, V_k$ there are k zero eigenvalues and the associated invariant subspace is spanned by the vectors

$$\chi_{V_1}, \chi_{V_2}, \ldots, \chi_{V_k}, \tag{1.70}$$

where χ_{V_j} is the indicator vector having a 1 at entry i if $x_i \in V_j$ and 0 otherwise.

1.8.2 Spectral Clustering

On a first sight, it may seem that (1.70) solves the graph clustering problem. One simply computes the eigenvectors belonging to the k zero eigenvalues of the graph Laplacian and the zero structure (1.70) of the eigenvectors can be used to determine the vertices belonging to each component. Each component gives rise to a cluster.

This tempting idea has two flaws. First, one cannot expect the eigenvectors to have the structure (1.70). Any computational method will yield an arbitrary eigenbasis, e.g., arbitrary linear combinations of $\chi_{V_1}, \chi_{V_2}, \ldots, \chi_{V_k}$. In general, the method will compute an orthonormal basis U with

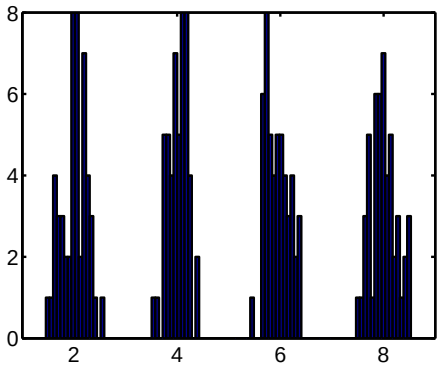
$$U = \big(v_1, \ldots, v_k\big)Q, \tag{1.71}$$

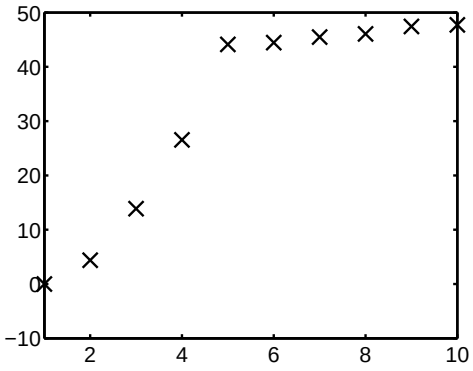
where Q is an arbitrary orthogonal $k \times k$ matrix and $v_j = \chi_{V_j}/|V_j|$ with the cardinality $|V_j|$ of V_j. Second and more importantly, the goal of graph clustering is not to detect connected components of a graph.[2] Requiring the components to be completely disconnected to each other is too strong and will usually not lead to a meaningful clustering. For example, when using a fully connected similarity graph only one eigenvalue will be zero and the corresponding eigenvector e yields one component, which is the graph itself! Hence, instead of computing an eigenbasis belonging to zero eigenvalues, one determines an eigenbasis belonging to the k smallest eigenvalues.

Example 1.1 200 real numbers are generated by superimposing samples from 4 Gaussian distributions with 4 different means:

```
m = 50; randn('state',0);
x = [2+randn(m,1)/4;4+randn(m,1)/4;6+randn(m,1)/4;8+randn(m,1)/4];
```

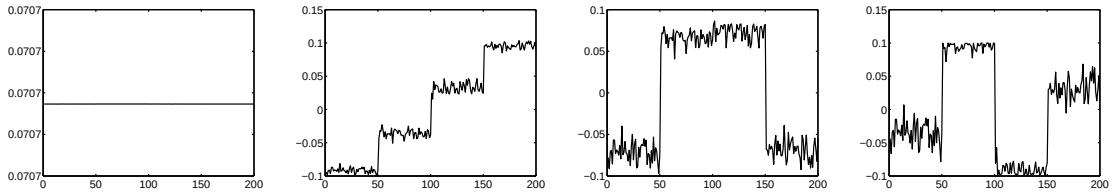
The following two figures show the histogram of the distribution of the entries of x and the eigenvalues of the graph Laplacian for the fully connected similarity graph with similarity function $s(x_i, x_j) = \exp\big(-\frac{|x_i - x_j|^2}{2}\big)$:





[2] There are more efficient algorithms for finding connected components, e.g., breadth-first and depth-first search.

As expected, one eigenvalue is (almost) exactly zero. Additionally, the four smallest eigenvalues have a clearly visible gap to the other eigenvalues. The following four figures show the entries of the 4 eigenvectors belonging to the 4 smallest eigenvalues of L:



On the one hand, it is clearly visible that the eigenvectors are well approximated by linear combinations of indicator vectors. On the other hand, none of the eigenvectors is close to an indicator vector itself and hence no immediate conclusion on the clusters is possible.

To solve the issue that the eigenbasis (1.71) may be transformed by an arbitrary orthogonal matrix, we "transpose" the basis and consider the row vectors of U:

$$U^T = \big(u_1, u_2, \ldots, u_n\big), \quad u_i \in \mathbb{R}^k.$$

If U contained indicator vectors then each of the small vectors u_i would be a unit vector e_j for some $1 \leq j \leq k$ (possibly divided by $|V_j|$). In particular, the u_i would separate very well into k different clusters. The latter property does not change if the vectors u_i undergo an orthogonal transformation Q^T. Hence, applying a clustering algorithm to $u_1, \ldots, u_n$ allows us to detect the membership of u_i independent of the orthogonal transformation. The key point is that the small vectors $u_1, \ldots, u_n$ are much better separated than the original data $x_1, \ldots, x_n$. Hence, much simpler algorithm can be used for clustering. One of the most basic algorithms is *k-means clustering*. Initially, this algorithm assigns each u_i randomly[3] to a cluster ℓ with $1 \leq \ell \leq k$ and then iteratively proceeds as follows:

1. Compute cluster centers c_ℓ as cluster means:

$$c_\ell = \sum_{i \text{ in cluster } \ell} u_i \Big/ \sum_{i \text{ in cluster } \ell} 1.$$

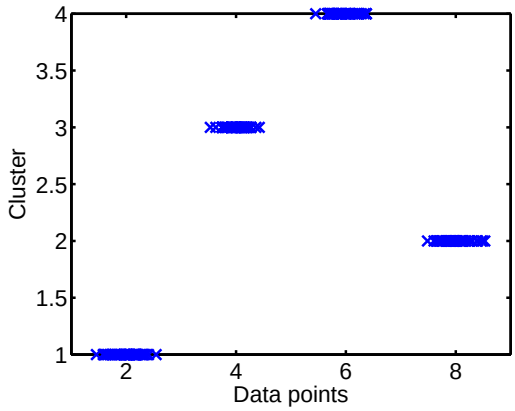
2. Assign each u_i to the cluster with the nearest cluster center.

3. Goto Step 1.

The algorithm is stopped when the assigned clusters do not change in an iteration.

Example 1.1 ctd. The k-means algorithm applied to the eigenbasis from Example 1.1 converges after 2 iterations and results in the following clustering:

[3]For unlucky choices of random assignments the k-means algorithm may end up with less than k clusters. A simple albeit dissatisfying solution is to restart k-means with a different random assignment.



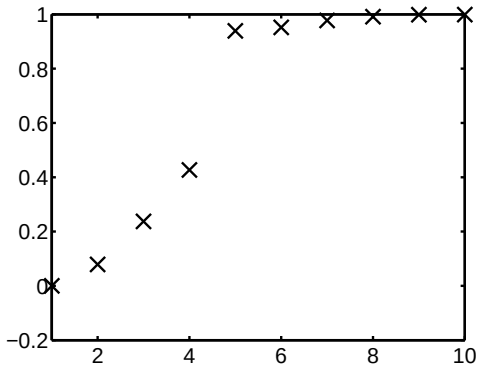
1.8.3 Normalized Graph Laplacians

It is sometimes advantageous to use a normalized Laplacian

$$D^{-1}L = I - D^{-1}W \tag{1.72}$$

instead of the standard Laplacians. Equivalently, this means that we compute the eigenvectors belonging to the smallest eigenvalues of the generalized eigenvalue problem $\lambda D - W$. Alternatively, one may also compute the eigenvalues from the symmetric matrix $D^{-1/2}WD^{-1/2}$ but the eigenvectors need to be adjusted to compensate this transformation.

Example 1.1 ctd. The eigenvalues of the normalized Laplacian for Example 1.1 are shown below:



In comparison to the eigenvalues of the standard Laplacian, the four smallest eigenvalues of the are better separated from the rest. Otherwise, the shape of the eigenvectors is similar and the resulting clustering is identical with the one obtained with the standard Laplacian.

1.9 Other Sources of Eigenvalue Problems

The selection of applications above may lead to the impression that eigenvalue problems in practice virtually always require the computation of the smallest eigenvalues of a symmetric matrix. This is *not* the case. For example, a linear stability analysis requires the

computation of all eigenvalues on or close to the imaginary axis of a nonsymmetric matrix. Computational methods for decoupling the stable/unstable parts of a dynamical system require the computation of all eigenvalues in the left and/or right half of the complex plane. The principal component analysis (PCA), which plays an important role in a large variety of applications, requires the computation of the largest eigenvalues (or rather singular values). As we will see in the following chapters, the region of eigenvalues we are interested in determines the difficulty of the eigenvalue problem to a large extent (along with the matrix order and structure). It should also guide the choice of algorithm for solving an eigenvalue problem.

Bibliography

[1] W. Ames, *Numerical Methods for Partial Differential Equations*, Academic Press, New York NY, 2nd ed., 1977.

[2] O. Axelsson and V. Barker, *Finite Element Solution of Boundary Value Problems*, Academic Press, Orlando FL, 1984.

[3] O. Chinellato, *The Complex-Symmetric Jacobi–Davidson Algorithm and its Application to the Computation of some Resonance Frequencies of Anisotropic Lossy Axisymmetric Cavities*, PhD Thesis No. 16243, ETH Zürich, 2005. (Available at URL `http://e-collection.ethbib.ethz.ch/show?type=diss&nr=16243`).

[4] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, Springer, Berlin, 1968.

[5] H. R. Schwarz, *Methode der finiten Elemente*, Teubner, Stuttgart, 3rd ed., 1991.

[6] ——, *Numerische Mathematik*, Teubner, Stuttgart, 3rd ed. ed., 1993.

[7] G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, 1986.

[8] G. Strang and G. J. Fix, *Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, 1973.

[9] U. von Luxburg, *A tutorial on spectral clustering*, Technical Report No. TR-149, Max Planck Institute for Biological Cybernetics, August 2006.

[10] H. F. Weinberger, *Variational Methods for Eigenvalue Approximation*, Regional Conference Series in Applied Mathematics 15, SIAM, Philadelphia, PA, 1974.

Chapter 2

Basics

2.1 Notation

The fields of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. Elements in $\mathbb{R}$ and $\mathbb{C}$, *scalars*, are denoted by lowercase letters, a, b, $c, \ldots$, and α, β, $\gamma, \ldots$

Vectors are denoted by boldface lowercase letters, $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}, \ldots$, and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, ... We denote the space of vectors of n *real* components by $\mathbb{R}^n$ and the space of vectors of n *complex* components by $\mathbb{C}^n$.

$$\mathbf{x} \in \mathbb{R}^n \Longleftrightarrow \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i \in \mathbb{R}. \tag{2.1}$$

We often make statements that hold for real or complex vectors or matrices. Then we write, e.g., $\mathbf{x} \in \mathbb{F}^n$.

The **inner product** of two n-vectors in $\mathbb{C}$ is defined as

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \bar{y}_i = \mathbf{y}^* \mathbf{x}, \tag{2.2}$$

that is, we require linearity in the first component and anti-linearity in the second.

$\mathbf{y}^* = (\bar{y}_1,\ \bar{y}_2, \ldots, \bar{y}_n)$ denotes conjugate transposition of complex vectors. To simplify notation we denote real transposition by an asterisk as well.

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are called **orthogonal**, $\mathbf{x} \perp \mathbf{y}$, if $\mathbf{x}^* \mathbf{y} = 0$.

The inner product (2.2) induces a **norm** in $\mathbb{F}$,

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}. \tag{2.3}$$

This norm is often called Euclidean norm or 2-norm.

The set of m-by-n **matrices** with components in the field $\mathbb{F}$ is denoted by $\mathbb{F}^{m \times n}$,

$$A \in \mathbb{F}^{m \times n} \Longleftrightarrow A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}, \quad a_{ij} \in \mathbb{F}. \tag{2.4}$$

The matrix $A^* \in \mathbb{F}^{n\times m}$,

$$A^* = \begin{pmatrix} \bar{a}_{11} & \bar{a}_{21} & \dots & \bar{a}_{m1} \\ \bar{a}_{12} & \bar{a}_{22} & \dots & \bar{a}_{m2} \\ \vdots & \vdots & & \vdots \\ \bar{a}_{1n} & \bar{a}_{2n} & \dots & \bar{a}_{nm} \end{pmatrix} \tag{2.5}$$

is the **Hermitian transpose** of A. Notice, that with this notation n-vectors can be identified with n-by-1 matrices.

The following classes of square matrices are of particular importance:

- $A \in \mathbb{F}^{n\times n}$ is called **Hermitian** if and only if $A^* = A$.
- A *real* Hermitian matrix is called **symmetric**.
- $U \in \mathbb{F}^{n\times n}$ is called **unitary** if and only if $U^{-1} = U^*$.
- *Real* unitary matrices are called **orthogonal**.

We define the norm of a matrix to be the norm induced by the vector norm (2.3),

$$\|A\| := \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|. \tag{2.6}$$

The condition number of a nonsingular matrix is defined as $\kappa(A) = \|A\|\|A^{-1}\|$. It is easy to show that if U is unitary then $\|U\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x}$. Thus the condition number of a unitary matrix is 1.

2.2 Statement of the problem

The **(standard) eigenvalue problem** is as follows.

> Given a square matrix $A \in \mathbb{F}^{n\times n}$.
> Find scalars $\lambda \in \mathbb{C}$ and vectors $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{x} \neq \mathbf{0}$, such that
>
> $$A\mathbf{x} = \lambda\mathbf{x}, \tag{2.7}$$
>
> i.e., such that
>
> $$(A - \lambda I)\mathbf{x} = \mathbf{0} \tag{2.8}$$
>
> has a nontrivial (nonzero) solution.

So, we are looking for numbers λ such that $A - \lambda I$ is *singular*.

Definition 2.1 Let the pair $(\lambda, \mathbf{x})$ be a solution of (2.7) or (2.8), respectively. Then

- λ is called an **eigenvalue** of A,
- $\mathbf{x}$ is called an **eigenvector** corresponding to λ
- $(\lambda, \mathbf{x})$ is called **eigenpair** of A.
- The set $\sigma(A)$ of *all* eigenvalues of A is called **spectrum** of A.

- The set of all eigenvectors corresponding to an eigenvalue λ together with the vector $\mathbf{0}$ form a linear subspace of $\mathbb{C}^n$ called the **eigenspace** of λ. As the eigenspace of λ is the null space of $\lambda I - A$ we denote it by $\mathcal{N}(\lambda I - A)$.
- The dimension of $\mathcal{N}(\lambda I - A)$ is called **geometric multiplicity** $g(\lambda)$ of λ.
- An eigenvalue λ is a zero of the **characteristic polynomial**

$$\chi(\lambda) := \det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0.$$

  The multiplicity of λ as a zero of χ is called the **algebraic multiplicity** $m(\lambda)$ of λ. We will later see that

$$1 \le g(\lambda) \le m(\lambda) \le n, \qquad \lambda \in \sigma(A), \quad A \in \mathbb{F}^{n\times n}.$$

Remark 2.1. A nontrivial solution solution $\mathbf{y}$ of

$$\mathbf{y}^* A = \lambda \mathbf{y}^* \tag{2.9}$$

is called **left eigenvector** corresponding to λ. A left eigenvector of A is a right eigenvector of A^*, corresponding to the eigenvalue $\bar{\lambda}$, $A^*\mathbf{y} = \bar{\lambda}\mathbf{y}$. □

Problem 2.2 Let $\mathbf{x}$ be a (right) eigenvector of A corresponding to an eigenvalue λ and let $\mathbf{y}$ be a left eigenvector of A corresponding to a *different* eigenvalue $\mu \neq \lambda$. Show that $\mathbf{x}^*\mathbf{y} = 0$.

Remark 2.2. Let A be an **upper triangular** matrix,

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix}, \qquad a_{ik} = 0 \text{ for } i > k. \tag{2.10}$$

Then we have

$$\det(\lambda I - A) = \prod_{i=1}^{n} (\lambda - a_{ii}).$$

□

Problem 2.3 Let $\lambda = a_{ii}$, $1 \le i \le n$, be an eigenvalue of A in (2.10). Can you give a corresponding eigenvector? Can you explain a situation where $g(\lambda) < m(\lambda)$?

The **(generalized) eigenvalue problem** is as follows.

Given two square matrices $A, B \in \mathbb{F}^{n\times n}$.
Find scalars $\lambda \in \mathbb{C}$ and vectors $\mathbf{x} \in \mathbb{C}$, $\mathbf{x} \neq \mathbf{0}$, such that

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{2.11}$$

or, equivalently, such that

$$(A - \lambda B)\mathbf{x} = \mathbf{0} \tag{2.12}$$

has a nontrivial solution.

Definition 2.4 Let the pair $(\lambda, \mathbf{x})$ be a solution of (2.11) or (2.12), respectively. Then

- λ is called an **eigenvalue** of A **relative to** B,
- $\mathbf{x}$ is called an **eigenvector** of A **relative to** B corresponding to λ.
- $(\lambda, \mathbf{x})$ is called an **eigenpair** of A **relative to** B,
- The set $\sigma(A; B)$ of *all* eigenvalues of (2.11) is called the **spectrum** of A **relative to** B.
- Let B be nonsingular. Then

$$A\mathbf{x} = \lambda B\mathbf{x} \Longleftrightarrow B^{-1}A\mathbf{x} = \lambda\mathbf{x} \tag{2.13}$$

- Let both A and B be Hermitian, $A = A^*$ and $B = B^*$. Let further be B positive definite and $B = LL^*$ be its Cholesky factorization. Then

$$A\mathbf{x} = \lambda B\mathbf{x} \Longleftrightarrow L^{-1}AL^{-*}\mathbf{y} = \lambda\mathbf{y}, \quad \mathbf{y} = L^*\mathbf{x}. \tag{2.14}$$

- Let A be invertible. Then $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$. That is, $0 \notin \sigma(A; B)$. Therefore,

$$A\mathbf{x} = \lambda B\mathbf{x} \Longleftrightarrow \mu\mathbf{x} = A^{-1}B\mathbf{x}, \quad \mu = \frac{1}{\lambda} \tag{2.15}$$

- *Difficult situation*: both A and B are singular.

  1. Let, e.g.,
  $$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$
  Then,
  $$A\mathbf{e}_2 = \mathbf{0} = 0 \cdot B\mathbf{e}_2 = 0 \cdot \mathbf{e}_2$$
  $$A\mathbf{e}_1 = \mathbf{e}_1 = \lambda B\mathbf{e}_1 = \lambda\mathbf{0}$$
  So 0 is an eigenvalue of A relative to B.
  As in (2.15) we may swap the roles of A and B. Then
  $$B\mathbf{e}_1 = \mathbf{0} = \mu A\mathbf{e}_1 = 0\mathbf{e}_1.$$
  So, $\mu = 0$ is an eigenvalue of B relative to A. We therefore say, informally, that $\lambda = \infty$ is an eigenvalue of A relative to B. So, $\sigma(A; B) = \{0, \infty\}$.
  2. Let
  $$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = A.$$
  Then,
  $$A\mathbf{e}_1 = 1 \cdot B\mathbf{e}_1,$$
  $$A\mathbf{e}_2 = \mathbf{0} = \lambda B\mathbf{e}_2 = \lambda\mathbf{0}, \text{ for } \mathbf{all}\ \lambda \in \mathbb{C}.$$
  Therefore, in this case, $\sigma(A; B) = \mathbb{C}$.

2.3 Similarity transformations

Definition 2.5 A matrix $A \in \mathbb{F}^{n\times n}$ is **similar** to a matrix $C \in \mathbb{F}^{n\times n}$, $A \sim C$, if and only if there is a nonsingular matrix S such that

$$S^{-1}AS = C. \tag{2.16}$$

The mapping $A \longrightarrow S^{-1}AS$ is called a **similarity transformation**.

Theorem 2.6 *Similar matrices have equal eigenvalues with equal multiplicities. If* $(\lambda, \mathbf{x})$ *is an eigenpair of* A *and* $C = S^{-1}AS$ *then* $(\lambda, S^{-1}\mathbf{x})$ *is an eigenpair of* C*.*

Proof. $A\mathbf{x} = \lambda\mathbf{x}$ and $C = S^{-1}AS$ imply that

$$CS^{-1}\mathbf{x} = S^{-1}ASS^{-1}\mathbf{x} = S^{-1}\lambda\mathbf{x}.$$

Hence, A and C have equal eigenvalues and their geometric multiplicity is not changed by the similarity transformation. From

$$\begin{aligned}\det(\lambda I - C) &= \det(\lambda S^{-1}S - S^{-1}AS)\\ &= \det(S^{-1}(\lambda I - A)S) = \det(S^{-1})\det(\lambda I - A)\det(S) = \det(\lambda I - A)\end{aligned}$$

it follows that the characteristic polynomials of A and C are equal and hence also the algebraic eigenvalue multiplicities are equal. ∎

Definition 2.7 Two matrices A and B are called **unitarily similar** if S in (2.16) is unitary. If the matrices are real the term orthogonally similar is used.

Unitary similarity transformations are very important in numerical computations. Let U be unitary. Then $\|U\| = \|U^{-1}\| = 1$, the condition number of U is therefore $\kappa(U) = 1$. Hence, if $C = U^{-1}AU$ then $C = U^*AU$ and $\|C\| = \|A\|$. In particular, if A is disturbed by δA (e.g., roundoff errors introduced when storing the entries of A in finite-precision arithmetic) then

$$U^*(A + \delta A)U = C + \delta C, \qquad \|\delta C\| = \|\delta A\|.$$

Hence, errors (perturbations) in A are not amplified by a unitary similarity transformation. This is in contrast to arbitrary similarity transformations. However, as we will see later, small eigenvalues may still suffer from large relative errors.

Another reason for the importance of unitary similarity transformations is the preservation of symmetry: If A is symmetric then $U^{-1}AU = U^*AU$ is symmetric as well.

For generalized eigenvalue problems, similarity transformations are not so crucial since we can operate with different matrices from both sides. If S and T are nonsingular

$$A\mathbf{x} = \lambda B\mathbf{x} \quad \Longleftrightarrow \quad TAS^{-1}S\mathbf{x} = \lambda TBS^{-1}S\mathbf{x}.$$

This sometimes called *equivalence transformation* of A, B. Thus, $\sigma(A; B) = \sigma(TAS^{-1}, TBS^{-1})$. Let us consider a special case: let B be invertible and let $B = LU$ be the LU-factorization of B. Then we set $S = U$ and $T = L^{-1}$ and obtain $TBU^{-1} = L^{-1}LUU^{-1} = I$. Thus, $\sigma(A; B) = \sigma(L^{-1}AU^{-1}, I) = \sigma(L^{-1}AU^{-1})$.

2.4 Schur decomposition

Theorem 2.8 ***(Schur decomposition)*** *If $A \in \mathbb{C}^{n\times n}$ then there is a unitary matrix $U \in \mathbb{C}^{n\times n}$ such that*

$$U^*AU = T \tag{2.17}$$

is upper triangular. The diagonal elements of T are the eigenvalues of A.

Proof. The proof is by induction. For $n = 1$, the theorem is obviously true.

Assume that the theorem holds for matrices of order $\leq n-1$. Let $(\lambda, \mathbf{x})$, $\|\mathbf{x}\| = 1$, be an eigenpair of A, $A\mathbf{x} = \lambda\mathbf{x}$. We construct a unitary matrix U_1 with first column $\mathbf{x}$ (e.g. the Householder reflector U_1 with $U_1\mathbf{x} = \mathbf{e}_1$). Partition $U_1 = [\mathbf{x}, \overline{U}]$. Then

$$U_1^*AU_1 = \begin{bmatrix} \mathbf{x}^*A\mathbf{x} & \mathbf{x}^*A\overline{U} \\ \overline{U}^*A\mathbf{x} & \overline{U}^*A\overline{U} \end{bmatrix} = \begin{bmatrix} \lambda & \times\cdots\times \\ \mathbf{0} & \hat{A} \end{bmatrix}$$

as $A\mathbf{x} = \lambda\mathbf{x}$ and $\overline{U}^*\mathbf{x} = \mathbf{0}$ by construction of U_1. By assumption, there exists a unitary matrix $\hat{U} \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $\hat{U}^*\hat{A}\hat{U} = \hat{T}$ is upper triangular. Setting $U := U_1(1\oplus\hat{U})$, we obtain (2.17). ∎

Notice , that this proof is not constructive as we assume the knowledge of an eigenpair $(\lambda, \mathbf{x})$. So, we cannot employ it to actually compute the Schur form. The QR algorithm is used for this purpose. We will discuss this basic algorithm in Chapter 3.

Let $U^*AU = T$ be a Schur decomposition of A with $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$. The Schur decomposition can be written as $AU = UT$. The k-th column of this equation is

$$A\mathbf{u}_k = \lambda\mathbf{u}_k + \sum_{i=1}^{k-1} t_{ik}\mathbf{u}_i, \qquad \lambda_k = t_{kk}. \tag{2.18}$$

This implies that

$$A\mathbf{u}_k \in \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}, \quad \forall k. \tag{2.19}$$

Thus, the first k **Schur vectors** $\mathbf{u}_1, \ldots, \mathbf{u}_k$ form an **invariant subspace**[1] for A. From (2.18) it is clear that the *first* Schur vector is an eigenvector of A. The other columns of U, however, are in general *not* eigenvectors of A. Notice, that the Schur decomposition is not unique. In the proof we have chosen *any* eigenvalue λ. This indicates that the eigenvalues can be arranged in any order in the diagonal of T. This also indicates that the order with which the eigenvalues appear on T's diagonal can be manipulated.

Problem 2.9 Let

$$A = \begin{bmatrix} \lambda_1 & \alpha \\ 0 & \lambda_2 \end{bmatrix}.$$

Find an orthogonal 2×2 matrix Q such that

$$Q^*AQ = \begin{bmatrix} \lambda_2 & \beta \\ 0 & \lambda_1 \end{bmatrix}.$$

Hint: the first column of Q must be the (normalized) eigenvector of A with eigenvalue λ_2.

[1] A subspace $\mathcal{V} \subset \mathbb{F}^n$ is called invariant for A if $A\mathcal{V} \subset \mathcal{V}$.

2.5 The real Schur decomposition

Real matrices can have complex eigenvalues. If complex eigenvalues exist, then they occur in *complex conjugate pairs*! That is, if λ is an eigenvalue of the real matrix A, then also $\bar{\lambda}$ is an eigenvalue of A. The following theorem indicates that complex computation can be avoided.

Theorem 2.10 ***(Real Schur decomposition)*** *If $A \in \mathbb{R}^{n\times n}$ then there is an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ such that*

$$Q^T A Q = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ & R_{22} & \cdots & R_{2m} \\ & & \ddots & \vdots \\ & & & R_{mm} \end{bmatrix} \tag{2.20}$$

is upper quasi-triangular. The diagonal blocks R_{ii} are either 1×1 or 2×2 matrices. A 1×1 block corresponds to a real eigenvalue, a 2×2 block corresponds to a pair of complex conjugate eigenvalues.

Remark 2.3. The matrix

$$\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}, \quad \alpha, \beta \in \mathbb{R},$$

has the eigenvalues $\alpha + i\beta$ and $\alpha - i\beta$. □

Proof. Let $\lambda = \alpha + i\beta$, $\beta \neq 0$, be an eigenvalue of A with eigenvector $\mathbf{x} = \mathbf{u} + i\mathbf{v}$. Then $\bar{\lambda} = \alpha - i\beta$ is an eigenvalue corresponding to $\bar{\mathbf{x}} = \mathbf{u} - i\mathbf{v}$. To see this we first observe that

$$\begin{aligned} A\mathbf{x} &= A(\mathbf{u} + i\mathbf{v}) = A\mathbf{u} + iA\mathbf{v}, \\ \lambda\mathbf{x} &= (\alpha + i\beta)(\mathbf{u} + i\mathbf{v}) = (\alpha\mathbf{u} - \beta\mathbf{v}) + i(\beta\mathbf{u} - \alpha\mathbf{v}). \end{aligned}$$

Thus,

$$\begin{aligned} A\bar{\mathbf{x}} &= A(\mathbf{u} - i\mathbf{v}) = A\mathbf{u} - iA\mathbf{v}, \\ &= (\alpha\mathbf{u} - \beta\mathbf{v}) - i(\beta\mathbf{u} + \alpha\mathbf{v}) \\ &= (\alpha - i\beta)\mathbf{u} - i(\alpha - i\beta)\mathbf{v} = (\alpha - i\beta)(\mathbf{u} - i\mathbf{v}) = \bar{\lambda}\bar{\mathbf{x}}. \end{aligned}$$

Now, the actual proof starts. Let k be the number of complex conjugate pairs. We prove the theorem by induction on k.

First we consider the case $k = 0$. In this case A has real eigenvalues and eigenvectors. It is clear that we can repeat the proof of the Schur decomposition of Theorem 2.8 in real arithmetic to get the decomposition (2.17) with $U \in \mathbb{R}^{n\times n}$ and $T \in \mathbb{R}^{n\times n}$. So, there are n diagonal blocks R_{jj} in (2.20) all of which are 1×1.

Let us now assume the the theorem is true for all matrices with fewer than k complex conjugate pairs. Then, with $\lambda = \alpha + i\beta$, $\beta \neq 0$ and $\mathbf{x} = \mathbf{u} + i\mathbf{v}$, as previously, we have

$$A[\mathbf{u}, \mathbf{v}] = [\mathbf{u}, \mathbf{v}] \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}.$$

Let $\{\mathbf{x}_1, \mathbf{x}_2\}$ be an orthonormal basis of span$([\mathbf{u}, \mathbf{v}])$. Then, since $\mathbf{u}$ and $\mathbf{v}$ are linearly independent[2], there is a nonsingular 2×2 real square matrix C with

$$[\mathbf{x}_1, \mathbf{x}_2] = [\mathbf{u}, \mathbf{v}]C.$$

[2]If u and v were linearly dependent then it follows that β must be zero.

Now,

$$A[\mathbf{x}_1, \mathbf{x}_2] = A[\mathbf{u}, \mathbf{v}]C = A[\mathbf{u}, \mathbf{v}] \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} C$$
$$= [\mathbf{x}_1, \mathbf{x}_2]C^{-1} \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} C =: [\mathbf{x}_1, \mathbf{x}_2]S.$$

S and $\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}$ are similar and therefore have equal eigenvalues. Now we construct an orthogonal matrix $[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n] =: [\mathbf{x}_1, \mathbf{x}_2, W]$. Then

$$\big[[\mathbf{x}_1, \mathbf{x}_2], W\big]^T A \big[[\mathbf{x}_1, \mathbf{x}_2], W\big] = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ W^T \end{bmatrix} \big[[\mathbf{x}_1, \mathbf{x}_2]S, AW\big] = \begin{bmatrix} S & [\mathbf{x}_1, \mathbf{x}_2]^T AW \\ O & W^T AW \end{bmatrix}.$$

The matrix $W^T AW$ has less than k complex-conjugate eigenvalue pairs. Therefore, by the induction assumption, there is an orthogonal $Q_2 \in \mathbb{R}^{(n-2)\times(n-2)}$ such that the matrix

$$Q_2^T (W^T AW) Q_2$$

is quasi-triangular. Thus, the orthogonal matrix

$$Q = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n] \begin{pmatrix} I_2 & O \\ O & Q_2 \end{pmatrix}$$

transforms A similarly to quasi-triangular form. ■

2.6 Hermitian matrices

Definition 2.11 A matrix $A \in \mathbb{F}^{n\times n}$ is *Hermitian* if

$$A = A^*. \tag{2.21}$$

The Schur decomposition for Hermitian matrices is particularly simple. We first note that A being Hermitian implies that the upper triangular Λ in the Schur decomposition $A = U\Lambda U^*$ is Hermitian and thus diagonal. In fact, because

$$\overline{\Lambda} = \Lambda^* = (U^* AU)^* = U^* A^* U = U^* AU = \Lambda,$$

each diagonal element λ_i of Λ satisfies $\overline{\lambda}_i = \lambda_i$. So, Λ has to be *real*. In summary have the following result.

Theorem 2.12 ***(Spectral theorem for Hermitian matrices)*** *Let A be Hermitian. Then there is a unitary matrix U and a real diagonal matrix Λ such that*

$$A = U\Lambda U^* = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^*. \tag{2.22}$$

The columns $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of U are eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$. They form an orthonormal basis for $\mathbb{F}^n$.

The decomposition (2.22) is called a *spectral decomposition* of A.

As the eigenvalues are real we can sort them with respect to their magnitude. We can, e.g., arrange them in ascending order such that $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$.

If $\lambda_i = \lambda_j$, then any nonzero linear combination of $\mathbf{u}_i$ and $\mathbf{u}_j$ is an eigenvector corresponding to λ_i,

$$A(\mathbf{u}_i\alpha + \mathbf{u}_j\beta) = \mathbf{u}_i\lambda_i\alpha + \mathbf{u}_j\lambda_j\beta = (\mathbf{u}_i\alpha + \mathbf{u}_j\beta)\lambda_i.$$

However, eigenvectors corresponding to *different* eigenvalues are orthogonal. Let $A\mathbf{u} = \mathbf{u}\lambda$ and $A\mathbf{v} = \mathbf{v}\mu$, $\lambda \neq \mu$. Then

$$\lambda\mathbf{u}^*\mathbf{v} = (\mathbf{u}^*A)\mathbf{v} = \mathbf{u}^*(A\mathbf{v}) = \mathbf{u}^*\mathbf{v}\mu,$$

and thus

$$(\lambda - \mu)\mathbf{u}^*\mathbf{v} = 0,$$

from which we deduce $\mathbf{u}^*\mathbf{v} = 0$ as $\lambda \neq \mu$.

In summary, the eigenvectors corresponding to a particular eigenvalue λ form a subspace, the *eigenspace* $\{\mathbf{x} \in \mathbb{F}^n, A\mathbf{x} = \lambda\mathbf{x}\} = \mathcal{N}(A - \lambda I)$. They are perpendicular to the eigenvectors corresponding to all the other eigenvalues. Therefore, the spectral decomposition (2.22) is unique up to $\pm$ signs if all the eigenvalues of A are distinct. In case of multiple eigenvalues, we are free to choose any orthonormal basis for the corresponding eigenspace.

Remark 2.4. The notion of Hermitian or symmetric has a wider background. Let $\langle \mathbf{x}, \mathbf{y} \rangle$ be an inner product on $\mathbb{F}^n$. Then a matrix A is symmetric with respect to this inner product if $\langle A\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, A\mathbf{y} \rangle$ for all vectors $\mathbf{x}$ and $\mathbf{y}$. For the ordinary Euclidean inner product $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^*\mathbf{y}$ we arrive at the element-wise Definition 2.6 if we set $\mathbf{x}$ and $\mathbf{y}$ equal to coordinate vectors.

It is important to note that all the properties of Hermitian matrices that we will derive subsequently hold similarly for matrices symmetric with respect to a certain inner product. □

Example 2.13 We consider the one-dimensional Sturm-Liouville eigenvalue problem

$$-u''(x) = \lambda u(x), \quad 0 < x < \pi, \quad u(0) = u(\pi) = 0, \tag{2.23}$$

that models the vibration of a homogeneous string of length π that is *clamped* at both ends. The eigenvalues and eigenvectors or eigenfunctions of (2.23) are

$$\lambda_k = k^2, \quad u_k(x) = \sin\, kx, \qquad k \in \mathbb{N}.$$

Let $u_i^{(n)}$ denote the approximation of an (eigen)function u at the grid point x_i,

$$u_i \approx u(x_i), \quad x_i = ih, \quad 0 \le i \le n+1, \quad h = \frac{\pi}{n+1}.$$

We approximate the second derivative of u at the *interior* grid points by

$$\frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) = \lambda u_i, \qquad 1 \le i \le n. \tag{2.24}$$

Collecting these equations and taking into account the boundary conditions, $u_0 = 0$ and $u_{n+1} = 0$, we get a (matrix) eigenvalue problem

$$T_n\mathbf{x} = \lambda\mathbf{x} \tag{2.25}$$

where

$$T_n := \frac{(n+1)^2}{\pi^2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

The matrix eigenvalue problem (2.25) can be solved explicitly [3, p.229]. Eigenvalues and eigenvectors are given by

$$\begin{aligned} \lambda_k^{(n)} &= \frac{(n+1)^2}{\pi^2}(2-2\cos\phi_k) = \frac{4(n+1)^2}{\pi^2}\sin^2\frac{k\pi}{2(n+1)}, \\ \mathbf{u}_k^{(n)} &= \left(\frac{2}{n+1}\right)^{1/2} [\sin\phi_k, \sin 2\phi_k, \ldots, \sin n\phi_k]^T, \qquad \phi_k = \frac{k\pi}{n+1}. \end{aligned} \tag{2.26}$$

Clearly, $\lambda_k^{(n)}$ converges to λ_k as $n \to \infty$. (Note that $\sin\xi \to \xi$ as $\xi \to 0$.) When we identify $\mathbf{u}_k^{(n)}$ with the piecewise linear function that takes on the values given by $\mathbf{u}_k^{(n)}$ at the grid points x_i then this function evidently converges to $\sin kx$.

Let $p(\lambda)$ be a polynomial of degree d, $p(\lambda) = \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2 + \cdots + \alpha_d\lambda^d$. As $A^j = (U\Lambda U^*)^j = U\Lambda^j U^*$ we can define a *matrix polynomial* as

$$p(A) = \sum_{j=0}^{d} \alpha_j A^j = \sum_{j=0}^{d} \alpha_j U\Lambda^j U^* = U\left(\sum_{j=0}^{d}\alpha_j\Lambda^j\right)U^*. \tag{2.27}$$

This equation shows that $p(A)$ has the same eigenvectors as the original matrix A. The eigenvalues are modified though, λ_k becomes $p(\lambda_k)$. Similarly, more complicated functions of A can be computed if the function is defined on spectrum of A.

Definition 2.14 The quotient

$$\rho(\mathbf{x}) := \frac{\mathbf{x}^*A\mathbf{x}}{\mathbf{x}^*\mathbf{x}}, \qquad \mathbf{x} \neq \mathbf{0},$$

is called the *Rayleigh quotient* of A at $\mathbf{x}$.

Notice, that $\rho(\mathbf{x}\alpha) = \rho(\mathbf{x})$, $\alpha \neq 0$. Hence, the properties of the Rayleigh quotient can be investigated by just considering its values on the unit sphere. Using the spectral decomposition $A = U\Lambda U^*$, we get

$$\mathbf{x}^*A\mathbf{x} = \mathbf{x}^*U\Lambda U^*\mathbf{x} = \sum_{i=1}^{n}\lambda_i|\mathbf{u}_i^*\mathbf{x}|^2.$$

Similarly, $\mathbf{x}^*\mathbf{x} = \sum_{i=1}^{n}|\mathbf{u}_i^*\mathbf{x}|^2$. With $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, we have

$$\lambda_1\sum_{i=1}^{n}|\mathbf{u}_i^*\mathbf{x}|^2 \le \sum_{i=1}^{n}\lambda_i|\mathbf{u}_i^*\mathbf{x}|^2 \le \lambda_n\sum_{i=1}^{n}|\mathbf{u}_i^*\mathbf{x}|^2.$$

So,

$$\lambda_1 \le \rho(\mathbf{x}) \le \lambda_n, \qquad \text{for all } \mathbf{x} \neq \mathbf{0}.$$

As

$$\rho(\mathbf{u}_k) = \lambda_k,$$

the extremal values λ_1 and λ_n are actually attained for $\mathbf{x} = \mathbf{u}_1$ and $\mathbf{x} = \mathbf{u}_n$, respectively. Thus we have proved the following theorem.

Theorem 2.15 *Let A be Hermitian. Then the Rayleigh quotient satisfies*

$$\lambda_1 = \min \rho(\mathbf{x}), \qquad \lambda_n = \max \rho(\mathbf{x}). \tag{2.28}$$

As the Rayleigh quotient is a continuous function it attains *all* values in the closed interval $[\lambda_1, \lambda_n]$.

The next theorem generalizes the above theorem to interior eigenvalues. The following theorems is attributed to Poincaré, Fischer and Pólya.

Theorem 2.16 ***(Minimum-maximum principle)*** *Let A be Hermitian. Then*

$$\lambda_p = \min_{X \in \mathbb{F}^{n\times p}, \operatorname{rank}(X)=p} \max_{\mathbf{x}\neq\mathbf{0}} \rho(X\mathbf{x}) \tag{2.29}$$

Proof. Let $U_{p-1} = [\mathbf{u}_1, \ldots, \mathbf{u}_{p-1}]$. For every X with full rank we can choose $\mathbf{x} \neq \mathbf{0}$ such that $U_{p-1}^* X\mathbf{x} = \mathbf{0}$. Then $\mathbf{0} \neq \mathbf{z} := X\mathbf{x} = \sum_{i=p}^n z_i \mathbf{u}_i$. As in the proof of the previous theorem we obtain the inequality

$$\rho(\mathbf{z}) \geq \lambda_p.$$

To prove that equality holds in (2.29) we choose $X = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$. Then

$$U_{p-1}^* X\mathbf{x} = \begin{bmatrix} 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{bmatrix} \mathbf{x} = \mathbf{0}$$

implies that $\mathbf{x} = \mathbf{e}_p$, i.e., that $\mathbf{z} = X\mathbf{x} = \mathbf{u}_p$. So, $\rho(\mathbf{z}) = \lambda_p$. ∎

An important consequence of the minimum-maximum principle is the following

Theorem 2.17 ***(Monotonicity principle)*** *Let $\mathbf{q}_1, \ldots, \mathbf{q}_p$ be normalized, mutually orthogonal vectors and $Q := [\mathbf{q}_1, \ldots, \mathbf{q}_p]$. Let $A' := Q^* A Q \in \mathbb{F}^{p\times p}$. Then the p eigenvalues $\lambda'_1 \leq \cdots \leq \lambda'_p$ of A' satisfy*

$$\lambda_k \leq \lambda'_k, \qquad 1 \leq k \leq p. \tag{2.30}$$

Proof. Let $\mathbf{w}_1, \ldots, \mathbf{w}_p \in \mathbb{F}^p$ be the eigenvectors of A',

$$A'\mathbf{w}_i = \lambda'_i \mathbf{w}_i, \qquad 1 \leq i \leq p, \tag{2.31}$$

with $\mathbf{w}^*\mathbf{w}_j = \delta_{ij}$. Then the vectors $Q\mathbf{w}_1, \ldots, Q\mathbf{w}_p$ are normalized and mutually orthogonal. Therefore, we can construct a vector

$$\mathbf{x}_0 := a_1 Q\mathbf{w}_1 + \cdots + a_k Q\mathbf{w}_k, \quad \|\mathbf{x}_0\| = 1,$$

that is orthogonal to the first $k-1$ eigenvectors of A,

$$\mathbf{x}_0^* \mathbf{x}_i = 0, \qquad 1 \leq i \leq k-1.$$

Then, with the minimum-maximum principle we have

$$\begin{aligned}\lambda_k &= \min_{\substack{\mathbf{x}\neq\mathbf{0}\\ \mathbf{x}^*\mathbf{x}_1=\cdots=\mathbf{x}^*\mathbf{x}_{k-1}=0}} R(\mathbf{x}) \le R(\mathbf{x}_0) = \mathbf{x}_0^* A\mathbf{x}_0 \\ &= \sum_{i,j=1}^{p} \bar{a}_i a_j \mathbf{w}_i^* Q^* A Q \mathbf{w}_j = \sum_{i,j=1}^{k} \bar{a}_i a_j \lambda'_i \delta_{ij} = \sum_{i=1}^{k} |a|_i^2 \lambda'_i \le \lambda'_k.\end{aligned}$$

The last inequality holds since $\|\mathbf{x}_0\| = 1$ implies $\sum_{i=1}^{k} |a|_i^2 = 1$. ■

Remark 2.5. (The proof of this statement is an exercise)

It is possible to prove the inequalities (2.30) without assuming that the $\mathbf{q}_1, \ldots, \mathbf{q}_p$ are orthonormal. But then one has to use the eigenvalues λ'_k of

$$A'\mathbf{x} = \lambda' B\mathbf{x}, \quad B' = Q^*Q,$$

instead of (2.31). □

The *trace* of a matrix $A \in \mathbb{F}^{n\times n}$ is defined to be the sum of the diagonal elements of a matrix. Matrices that are similar have equal trace. Hence, by the spectral theorem,

$$\operatorname{trace}(A) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i. \tag{2.32}$$

The following theorem is proved in a similar way as the minimum-maximum theorem.

Theorem 2.18 ***(Trace theorem)***

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = \min_{X\in\mathbb{F}^{n\times p}, X^*X=I_p} \operatorname{trace}(X^*AX) \tag{2.33}$$

2.7 Cholesky factorization

Definition 2.19 A Hermitian matrix is called **positive definite** (**positive semi-definite**) if all its eigenvalues are positive (nonnegative).

For a Hermitian positive definite matrix A, the LU decomposition can be written in a particular form reflecting the symmetry of A.

Theorem 2.20 ***(Cholesky factorization)*** *Let $A \in \mathbb{F}^{n\times n}$ be Hermitian positive definite. Then there is a* lower triangular *matrix L such that*

$$A = LL^*. \tag{2.34}$$

L is unique if we choose its diagonal elements to be positive.

Proof. We prove the theorem by giving an algorithm that computes the desired factorization.

Since A is positive definite, we have $a_{11} = \mathbf{e}_1^* A \mathbf{e}_1 > 0$. Therefore we can form the matrix

$$L_1 = \begin{bmatrix} l_{11}^{(1)} & & & \\ l_{21}^{(1)} & 1 & & \\ \vdots & & \ddots & \\ l_{n1}^{(1)} & & & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{a_{11}} & & & \\ \frac{a_{21}}{\sqrt{a_{1,1}}} & 1 & & \\ \vdots & & \ddots & \\ \frac{a_{n1}}{\sqrt{a_{1,1}}} & & & 1 \end{bmatrix}.$$

We now form the matrix

$$A_1 = L_1^{-1} A L_1^{-1*} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & a_{22} - \frac{a_{21}a_{12}}{a_{11}} & \dots & a_{2n} - \frac{a_{21}a_{1n}}{a_{11}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2} - \frac{a_{n1}a_{12}}{a_{11}} & \dots & a_{nn} - \frac{a_{n1}a_{1n}}{a_{11}} \end{bmatrix}.$$

This is the first step of the algorithm. Since positive definiteness is preserved by a congruence transformation X^*AX (see also Theorem 2.22 below), A_1 is again positive definite. Hence, we can proceed in a similar fashion factorizing $A_1(2{:}n, 2{:}n)$, etc.

Collecting $L_1, L_2, \dots$, we obtain

$$I = L_n^{-1} \cdots L_2^{-1} L_1^{-1} A (L_1^*)^{-1} (L_2^*)^{-1} \cdots (L_n^*)^{-1}$$

or

$$(L_1 L_2 \cdots L_n)(L_n^* \cdots L_2^* L_1^*) = A.$$

which is the desired result. It is easy to see that $L_1 L_2 \cdots L_n$ is a triangular matrix and that

$$L_1 L_2 \cdots L_n = \begin{bmatrix} l_{11}^{(1)} & & & & \\ l_{21}^{(1)} & l_{22}^{(2)} & & & \\ l_{31}^{(1)} & l_{32}^{(2)} & l_{33}^{(3)} & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1}^{(1)} & l_{n2}^{(2)} & l_{n3}^{(3)} & \dots & l_{nn}^{(n)} \end{bmatrix}$$

■

Remark 2.6. When working with symmetric matrices, one often stores only half of the matrix, e.g. the lower triangle consisting of all elements including and below the diagonal. The L-factor of the Cholesky factorization can overwrite this information in-place to save memory. □

Definition 2.21 The **inertia** of a Hermitian matrix is the triple (ν, ζ, π) where ν, ζ, π is the number of negative, zero, and positive eigenvalues.

Theorem 2.22 ***(Sylvester's law of inertia)*** *If $A \in \mathbb{C}^{n\times n}$ is Hermitian and $X \in \mathbb{C}^{n\times n}$ is nonsingular then A and X^*AX have the same inertia.*

Proof. The proof is given, for example, in [2]. ■

Remark 2.7. Two matrices A and B are called congruent if there is a nonsingular matrix X such that $B = X^*AX$. Thus, Sylvester's law of inertia can be stated in the following form: *The inertia is invariant under congruence transformations.* □

2.8 The singular value decomposition (SVD)

Theorem 2.23 ***(Singular value decomposition)*** *If $A \in \mathbb{C}^{m\times n}$ then there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that*

$$U^*AV = \Sigma = \begin{pmatrix} \operatorname{diag}(\sigma_1, \dots, \sigma_p) & 0 \\ 0 & 0 \end{pmatrix}, \qquad p = \min(m, n), \tag{2.35}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$.

Proof. If $A = O$, the theorem holds with $U = I_m, V = I_n$ and Σ equal to the $m \times n$ zero matrix.

We now assume that $A \neq O$. Let $\mathbf{x}$, $\|\mathbf{x}\| = 1$, be a vector that maximizes $\|A\mathbf{x}\|$ and let $A\mathbf{x} = \sigma\mathbf{y}$ where $\sigma = \|A\| = \|A\mathbf{x}\|$ and $\|\mathbf{y}\| = 1$. As $A \neq O$, $\sigma > 0$. Consider the scalar function

$$f(\alpha) := \frac{\|A(\mathbf{x} + \alpha\mathbf{y})\|^2}{\|\mathbf{x} + \alpha\mathbf{y}\|^2} = \frac{(\mathbf{x} + \alpha\mathbf{y})^* A^* A(\mathbf{x} + \alpha\mathbf{y})}{(\mathbf{x} + \alpha\mathbf{y})^*(\mathbf{x} + \alpha\mathbf{y})}$$

Because of the extremality of $A\mathbf{x}$, the derivative $f'(\alpha)$ of $f(\alpha)$ must vanish at $\alpha = 0$. This holds *for all* $\mathbf{y}$! We have

$$\frac{df}{d\alpha}(\alpha) = \frac{(\mathbf{x}^* A^* A\mathbf{y} + \bar{\alpha}\mathbf{y}^* A^* A\mathbf{y})\|\mathbf{x} + \alpha\mathbf{y}\|^2 - (\mathbf{x}^*\mathbf{y} + \bar{\alpha}\mathbf{y}^*\mathbf{y})\|A(\mathbf{x} + \alpha\mathbf{y})\|^2}{\|\mathbf{x} + \alpha\mathbf{y}\|^4}$$

Thus, we have for all $\mathbf{y}$,

$$\frac{df}{d\alpha}(\alpha)\bigg|_{\alpha=0} = \frac{\mathbf{x}^* A^* A\mathbf{y}\|\mathbf{x}\|^2 - \mathbf{x}^*\mathbf{y}\|A(\mathbf{x})\|^2}{\|\mathbf{x}\|^4} = 0.$$

As $\|\mathbf{x}\| = 1$ and $\|A\mathbf{x}\| = \sigma$, we have

$$(\mathbf{x}^* A^* A - \sigma^2\mathbf{x}^*)\mathbf{y} = (A^* A\mathbf{x} - \sigma^2\mathbf{x})^*\mathbf{y} = 0, \qquad \text{for all } \mathbf{y},$$

from which

$$A^* A\mathbf{x} = \sigma^2\mathbf{x}$$

follow. Multiplying $A\mathbf{x} = \sigma\mathbf{y}$ from the left by A^* we get $A^* A\mathbf{x} = \sigma A^*\mathbf{y} = \sigma^2\mathbf{x}$ from which

$$A^*\mathbf{y} = \sigma\mathbf{x}$$

and $AA^*\mathbf{y} = \sigma A\mathbf{x} = \sigma^2\mathbf{y}$ follows. Therefore, $\mathbf{x}$ is an eigenvector of A^*A corresponding to the eigenvalue σ^2 and $\mathbf{y}$ is an eigenvector of AA^* corresponding to the same eigenvalue.

Now, we construct a unitary matrix U_1 with first column $\mathbf{y}$ and a unitary matrix V_1 with first column $\mathbf{x}$, $U_1 = [\mathbf{y}, \bar{U}]$ and $V_1 = [\mathbf{x}, \bar{V}]$. Then

$$U_1^* A V_1 = \begin{bmatrix} \mathbf{y}^* A\mathbf{x} & \mathbf{y}^* A\overline{U} \\ \overline{U}^* A\mathbf{x} & \overline{U}^* A\overline{V} \end{bmatrix} = \begin{bmatrix} \sigma & \sigma\mathbf{x}^*\overline{U} \\ \sigma\overline{U}^*\mathbf{y} & \overline{U}^* A\overline{V} \end{bmatrix} = \begin{bmatrix} \sigma & \mathbf{0}^* \\ \mathbf{0} & \hat{A} \end{bmatrix}$$

where $\hat{A} = \overline{U}^* A\overline{V}$. ∎

The proof above is due to W. Gragg. It nicely shows the relation of the singular value decomposition with the spectral decomposition of the Hermitian matrices A^*A and AA^*,

$$A = U\Sigma V^* \quad \Longrightarrow \quad A^* A = U\Sigma^2 U^*, \qquad AA^* = V\Sigma^2 V^*. \tag{2.36}$$

Note that the proof given in [2] is shorter and may be more elegant.

The SVD of dense matrices is computed in a way that is very similar to the dense Hermitian eigenvalue problem. However, in the presence of roundoff error, it is not advisable to make use of the matrices A^*A and AA^*. Instead, let us consider the $(n+m) \times (n+m)$ Hermitian matrix

$$\begin{bmatrix} O & A \\ A^* & O \end{bmatrix}. \tag{2.37}$$

Making use of the SVD (2.35) we immediately get

$$\begin{bmatrix} O & A \\ A^* & O \end{bmatrix} = \begin{bmatrix} U & O \\ O & V \end{bmatrix} \begin{bmatrix} O & \Sigma \\ \Sigma^T & O \end{bmatrix} \begin{bmatrix} U^* & O \\ O & V^* \end{bmatrix}.$$

Now, let us assume that $m \geq n$. Then we write $U = [U_1, U_2]$ where $U_1 \in \mathbb{F}^{m\times n}$ and $\Sigma = \begin{bmatrix} \Sigma_1 \\ O \end{bmatrix}$ with $\Sigma_1 \in \mathbb{R}^{n\times n}$. Then

$$\begin{bmatrix} O & A \\ A^* & O \end{bmatrix} = \begin{bmatrix} U_1 & U_2 & O \\ O & O & V \end{bmatrix} \begin{bmatrix} O & O & \Sigma_1 \\ O & O & O \\ \Sigma_1 & O & O \end{bmatrix} \begin{bmatrix} U_1^* & O \\ U_2^* & O \\ O & V^* \end{bmatrix} = \begin{bmatrix} U_1 & O & U_2 \\ O & V & O \end{bmatrix} \begin{bmatrix} O & \Sigma_1 & O \\ \Sigma_1 & O & O \\ O & O & O \end{bmatrix} \begin{bmatrix} U_1^* & O \\ O & V^* \\ U_2^* & O \end{bmatrix}.$$

The first and third diagonal zero blocks have order n. The middle diagonal block has order $n - m$. Now we employ the fact that

$$\begin{bmatrix} 0 & \sigma \\ \sigma & 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma & 0 \\ 0 & -\sigma \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

to obtain

$$\begin{bmatrix} O & A \\ A^* & O \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}}U_1 & \frac{1}{\sqrt{2}}U_1 & U_2 \\ \frac{1}{\sqrt{2}}V & -\frac{1}{\sqrt{2}}V & O \end{bmatrix} \begin{bmatrix} \Sigma_1 & O & O \\ O & -\Sigma_1 & O \\ O & O & O \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}}U_1^* & \frac{1}{\sqrt{2}}V^* \\ \frac{1}{\sqrt{2}}U_1^* & -\frac{1}{\sqrt{2}}V^* \\ U_2^* & O \end{bmatrix}. \tag{2.38}$$

Thus, there are three ways how to treat the computation of the singular value decomposition as an eigenvalue problem. One of the two forms in (2.36) is used *implicitly* in the QR algorithm for dense matrices A, see [2],[1]. The form (2.37) is suited if A is a sparse matrix.

2.9 Projections

Definition 2.24 A matrix P that satisfies

$$P^2 = P \tag{2.39}$$

is called a **projection**.

Obviously, a projection is a square matrix. If P is a projection then $P\mathbf{x} = \mathbf{x}$ for all $\mathbf{x}$ in the range $\mathcal{R}(P)$ of P. In fact, if $\mathbf{x} \in \mathcal{R}(P)$ then $\mathbf{x} = P\mathbf{y}$ for some $\mathbf{y} \in \mathbb{F}^n$ and $P\mathbf{x} = P(P\mathbf{y}) = P^2\mathbf{y} = P\mathbf{y} = \mathbf{x}$.

Example 2.25 Let

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

The range of P is $\mathcal{R}(P) = \mathbb{F} \times \{\mathbf{0}\}$. The effect of P is depicted in Figure 2.1: All points $\mathbf{x}$ that lie on a line parallel to $\text{span}\{(2, -1)^*\}$ are mapped on the same point on the $\mathbf{x}_1$ axis. So, the projection is *along* $\text{span}\{(2, -1)^*\}$ which is the null space $\mathcal{N}(P)$ of P.

Example 2.26 Let $\mathbf{x}$ and $\mathbf{y}$ be arbitrary vectors such that $\mathbf{y}^*\mathbf{x} \neq 0$. Then

$$P = \frac{\mathbf{x}\mathbf{y}^*}{\mathbf{y}^*\mathbf{x}} \tag{2.40}$$

is a projection. Notice that the projector of the previous example can be expressed in the form (2.40).

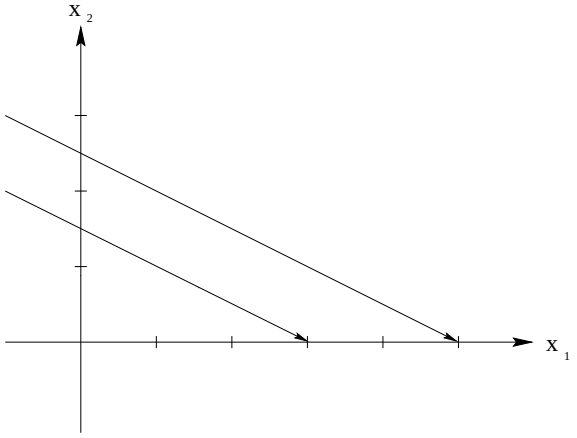


Figure 2.1: Oblique projection of example 2.9

Problem 2.27 Let $X, Y \in \mathbb{F}^{n\times p}$ such that Y^*X is nonsingular. Show that

$$P := X(Y^*X)^{-1}Y^*$$

is a projection.

If P is a projection then $I-P$ is a projection as well. In fact, $(I-P)^2 = I-2P+P^2 = I-2P+P = I-P$. If $P\mathbf{x} = \mathbf{0}$ then $(I-P)\mathbf{x} = \mathbf{x}$. Therefore, the range of $I-P$ coincides with the null space of P, $\mathcal{R}(I-P) = \mathcal{N}(P)$. It can be shown that $\mathcal{R}(P) = \mathcal{N}(P^*)^\perp$.

Notice that $\mathcal{R}(P)\cap\mathcal{R}(I-P) = \mathcal{N}(I-P)\cap\mathcal{N}(P) = \{\mathbf{0}\}$. For, if $P\mathbf{x} = \mathbf{0}$ then $(I-P)\mathbf{x} = \mathbf{x}$, which can only be zero if $\mathbf{x} = \mathbf{0}$. So, any vector $\mathbf{x}$ can be uniquely decomposed into

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2, \qquad \mathbf{x}_1 \in \mathcal{R}(P), \quad \mathbf{x}_2 \in \mathcal{R}(I-P) = \mathcal{N}(P). \tag{2.41}$$

The most interesting situation occurs if the decomposition is orthogonal, i.e., if $\mathbf{x}_1^*\mathbf{x}_2 = 0$ for all $\mathbf{x}$.

Definition 2.28 A matrix P is called an **orthogonal projection** if

$$\begin{array}{ll} \text{(i)} & P^2 = P \\ \text{(ii)} & P^* = P. \end{array} \tag{2.42}$$

Proposition 2.29 *Let P be a projection. Then the following statements are equivalent.*
(i) $P^* = P$,
(ii) $\mathcal{R}(I-P) \perp \mathcal{R}(P)$, *i.e.* $(P\mathbf{x})^*(I-P)\mathbf{y} = 0$ *for all* $\mathbf{x}, \mathbf{y}$.

Proof. (ii) follows trivially from (i) and (2.39).

Now, let us assume that (ii) holds. Then

$$\begin{aligned} \mathbf{x}^*P^*\mathbf{y} &= (P\mathbf{x})^*\mathbf{y} = (P\mathbf{x})^*(P\mathbf{y} + (I-P)\mathbf{y}) \\ &= (P\mathbf{x})^*(P\mathbf{y}) \\ &= (P\mathbf{x} + (I-P)\mathbf{x})(P\mathbf{y}) = \mathbf{x}^*(P\mathbf{y}). \end{aligned}$$

This equality holds for any $\mathbf{x}$ and $\mathbf{y}$ and thus implies (i). ∎

Example 2.30 Let $\mathbf{q}$ be an arbitrary vector of norm 1, $\|\mathbf{q}\| = \mathbf{q}^*\mathbf{q} = 1$. Then $P = \mathbf{q}\mathbf{q}^*$ is the orthogonal projection onto span$\{\mathbf{q}\}$.

Example 2.31 Let $Q \in \mathbb{F}^{n\times p}$ with $Q^*Q = I_p$. Then QQ^* is the orthogonal projector onto $\mathcal{R}(Q)$, which is the space spanned by the columns of Q.

Problem 2.32 Let $Q, Q_1 \in \mathbb{F}^{n\times p}$ with $Q^*Q = Q_1^*Q_1 = I_p$ such that $\mathcal{R}(Q) = \mathcal{R}(Q_1)$. This means that the columns of Q and Q_1, respectively, are orthonormal bases of the *same* subspace of $\mathbb{F}^n$. Show that the projector does not depend on the basis of the subspace, i.e., that $QQ^* = Q_1Q_1^*$.

Problem 2.33 Let $Q = [Q_1, Q_2]$, $Q_1 \in \mathbb{F}^{n\times p}$, $Q_2 \in \mathbb{F}^{n\times(n-p)}$ be a unitary matrix. Q_1 contains the first p columns of Q, Q_2 the last $n-p$. Show that $Q_1Q_1^* + Q_2Q_2^* = I$. Hint: Use $QQ^* = I$. Notice, that if $P = Q_1Q_1^*$ then $I - P = Q_2Q_2^*$.

Problem 2.34 What is the form of the orthogonal projection onto span$\{\mathbf{q}\}$ if the inner product is defined as $\langle \mathbf{x}, \mathbf{y}\rangle := \mathbf{y}^*M\mathbf{x}$ where M is a symmetric positive definite matrix?

2.10 Angles between vectors and subspaces

Let $\mathbf{q}_1$ and $\mathbf{q}_2$ be unit vectors, cf. Fig. 2.2. The length of the orthogonal projection of $\mathbf{q}_2$

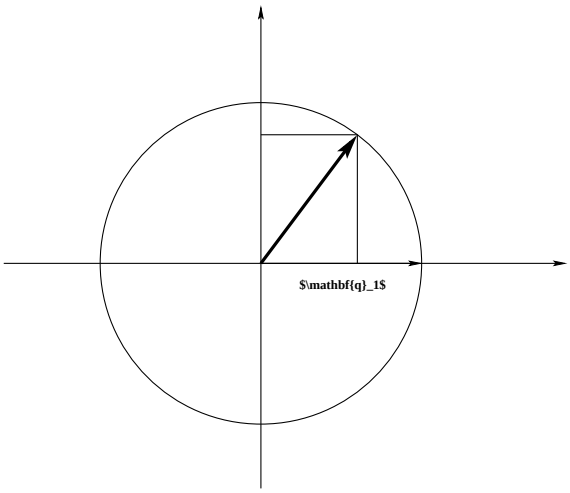


Figure 2.2: Angle between vectors $\mathbf{q}_1$ and $\mathbf{q}_2$

on span$\{\mathbf{q}_1\}$ is given by

$$c := \|\mathbf{q}_1\mathbf{q}_1{}^*\mathbf{q}_2\| = |\mathbf{q}_1{}^*\mathbf{q}_2| \le 1. \tag{2.43}$$

The length of the orthogonal projection of $\mathbf{q}_2$ on span$\{\mathbf{q}_1\}^\perp$ is

$$s := \|(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1{}^*)\mathbf{q}_2\|. \tag{2.44}$$

As $\mathbf{q}_1\mathbf{q}_1^*$ is an orthogonal projection we have by Pythagoras' formula that

$$1 = \|\mathbf{q}_2\|^2 = \|\mathbf{q}_1\mathbf{q}_1{}^*\mathbf{q}_2\|^2 + \|(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1{}^*)\mathbf{q}_2\|^2 = s^2 + c^2. \tag{2.45}$$

Alternatively, we can conclude from (2.44) that

$$\begin{aligned} s^2 &= \|(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1{}^*)\mathbf{q}_2\|^2 \\ &= \mathbf{q}_2{}^*(\mathbf{I} - \mathbf{q}_1\mathbf{q}_1{}^*)\mathbf{q}_2 \\ &= \mathbf{q}_2{}^*\mathbf{q}_2 - (\mathbf{q}_2{}^*\mathbf{q}_1)(\mathbf{q}_1{}^*\mathbf{q}_2) \\ &= 1 - c^2 \end{aligned} \tag{2.46}$$

So, there is a number, say, ϑ, $0 \leq \vartheta \leq \frac{\pi}{2}$, such that $c = \cos\vartheta$ and $s = \sin\vartheta$. We call this uniquely determined number ϑ the **angle** between the vectors $\mathbf{q}_1$ and $\mathbf{q}_2$:

$$\vartheta = \angle(\mathbf{q}_1, \mathbf{q}_2).$$

The generalization to arbitrary vectors is straightforward.

Definition 2.35 The **angle** θ between two nonzero vectors $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\vartheta = \angle(\mathbf{x}, \mathbf{y}) = \arcsin\left(\left\|\left(I - \frac{\mathbf{x}\mathbf{x}^*}{\|\mathbf{x}\|^2}\right)\frac{\mathbf{y}}{\|\mathbf{y}\|}\right\|\right) = \arccos\left(\frac{|\mathbf{x}^*\mathbf{y}|}{\|\mathbf{x}\|\|\mathbf{y}\|}\right). \tag{2.47}$$

When investigating the convergence behaviour of eigensolvers we usually show that the angle between the approximating and the desired vector tends to zero as the number of iterations increases. In fact it is more convenient to work with the sine of the angle.

In the formulae above we used the projections P and $I-P$ with $P = \mathbf{q}_1\mathbf{q}_1{}^*$. We would have arrived at the same point if we had exchanged the roles of $\mathbf{q}_1$ and $\mathbf{q}_2$. As

$$\|\mathbf{q}_1\mathbf{q}_1^*\mathbf{q}_2\| = \|\mathbf{q}_2\mathbf{q}_2^*\mathbf{q}_1\| = |\mathbf{q}_2^*\mathbf{q}_1|$$

we get

$$\|(I - \mathbf{q}_1\mathbf{q}_1^*)\mathbf{q}_2\| = \|(I - \mathbf{q}_2\mathbf{q}_2^*)\mathbf{q}_1\|.$$

This immediately leads to

Lemma 2.36 $\sin\angle(\mathbf{q}_1, \mathbf{q}_2) = \|\mathbf{q}_1\mathbf{q}_1^* - \mathbf{q}_2\mathbf{q}_2^*\|$.

Let now $Q_1 \in \mathbb{F}^{n\times p}$, $Q_2 \in \mathbb{F}^{n\times q}$ be matrices with orthonormal columns, $Q_1^*Q_1 = I_p, Q_2^*Q_2 = I_q$. Let $S_i = \mathcal{R}(Q_i)$, then S_1 and S_2 are subspaces of $\mathbb{F}^n$ of dimension p and q, respectively. We want to investigate how we can define a distance or an angle between S_1 and S_2 [2].

It is certainly straightforward to define the angle between the subspaces S_1 and S_2 to be the angle between two vectors $\mathbf{x}_1 \in S_1$ and $\mathbf{x}_2 \in S_2$. It is, however, not clear right-away how these vectors should be chosen.

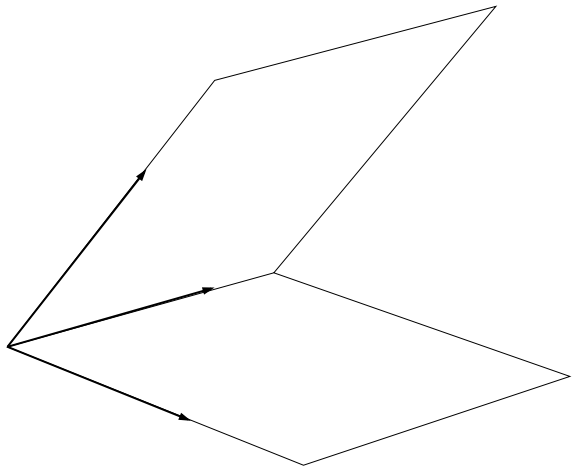


Figure 2.3: Two intersecting planes in 3-space

Let us consider the case of two 2-dimensional subspaces in $\mathbb{R}^3$, cf. Fig. (2.3). Let $S_1 = \text{span}\{\mathbf{q}_1,\ \mathbf{q}_2\}$ and $S_2 = \text{span}\{\mathbf{q}_1,\ \mathbf{q}_3\}$ where we assume that $\mathbf{q}_1^*\mathbf{q}_2 = \mathbf{q}_1^*\mathbf{q}_3 = 0$. We

might be tempted to define the angle between S_1 and S_2 as the maximal angle between any two vectors in S_1 and S_2,

$$\angle(S_1, S_2) = \max_{\substack{\mathbf{x}_1 \in S_1 \\ \mathbf{x}_2 \in S_2}} \angle(\mathbf{x}_1, \mathbf{x}_2). \tag{2.48}$$

This would give an angle of 90^o as we could chose $\mathbf{q}_1$ in S_1 and $\mathbf{q}_3$ in S_2. This angle would not change as we turn S_2 around $\mathbf{q}_1$. It would even stay the same if the two planes coincided.

What if we would take the minimum in (2.48)? This definition would be equally unsatisfactory as we could chose $\mathbf{q}_1$ in S_1 as well as in S_2 to obtain an angle of 0^o. In fact, any two 2-dimensional subspaces in 3 dimensions would have an angle of 0^o. Of course, we would like to reserve the angle of 0^o to coinciding subspaces.

A way out of this dilemma is to proceed as follows: Take any vector $\mathbf{x}_1 \in S_1$ and determine the angle between $\mathbf{x}_1$ *and its orthogonal projection* $(I - Q_2^* Q_2)\mathbf{x}_1$ on S_2. We now maximize the angle by varying $\mathbf{x}_1$ among all non-zero vectors in S_1. In the above 3-dimensional example we would obtain the angle between $\mathbf{x}_2$ and $\mathbf{x}_3$ as the angle between S_1 and S_3. Is this a reasonable definition? In particular, is it well-defined in the sense that it does not depend on how we number the two subspaces? Let us now assume that $S_1, S_2 \subset \mathbb{F}^n$ have dimensions p and q. Formally, the above procedure gives an angle ϑ with

$$\begin{aligned} \sin\vartheta &:= \max_{\substack{\mathbf{r} \in S_1 \\ \|\mathbf{r}\|=1}} \|(I_n - Q_2 Q_2^*)\mathbf{r}\| \\ &= \max_{\substack{\mathbf{a} \in \mathbb{F}^p \\ \|\mathbf{a}\|=1}} \|(I_n - Q_2 Q_2^*) Q_1 \mathbf{a}\| \qquad (2.49) \\ &= \|(I_n - Q_2 Q_2^*) Q_1\|. \end{aligned}$$

Because $I_n - Q_2 Q_2^*$ is an orthogonal projection, we get

$$\begin{aligned} \|(I_n - Q_2 Q_2^*) Q_1 \mathbf{a}\|^2 &= \mathbf{a}^* Q_1^* (I_n - Q_2 Q_2^*)(I_n - Q_2 Q_2^*) Q_1 \mathbf{a} \\ &= \mathbf{a}^* Q_1^* (I_n - Q_2 Q_2^*) Q_1 \mathbf{a} \\ &= \mathbf{a}^* (Q_1^* Q_1 - Q_1^* Q_2 Q_2^* Q_1) \mathbf{a} \qquad (2.50) \\ &= \mathbf{a}^* (I_p - (Q_1^* Q_2)(Q_2^* Q_1)) \mathbf{a} \\ &= \mathbf{a}^* (I_p - W^* W) \mathbf{a} \end{aligned}$$

where $W := Q_2^* Q_1 \in \mathbb{F}^{q \times p}$. With (2.49) we obtain

$$\begin{aligned} \sin^2\vartheta &= \max_{\|\mathbf{a}\|=1} \mathbf{a}^* (I_p - W^* W) \mathbf{a} \\ &= \text{largest eigenvalue of } I_p - W^* W \qquad (2.51) \\ &= 1 - \text{smallest eigenvalue of } W^* W. \end{aligned}$$

If we change the roles of Q_1 and Q_2 we get in a similar way

$$\sin^2\varphi = \|(I_n - Q_1 Q_1^*) Q_2\| = 1 - \text{smallest eigenvalue of } W W^*. \tag{2.52}$$

Notice, that $W^* W \in \mathbb{F}^{p \times p}$ and $W W^* \in \mathbb{F}^{q \times q}$ and that both matrices have equal rank. Thus, if W has full rank and $p < q$ then $\vartheta < \varphi = \pi/2$. However if $p = q$ then $W^* W$ and $W W^*$ have equal eigenvalues, and, thus, $\vartheta = \varphi$. In this most interesting case we have

$$\sin^2\vartheta = 1 - \lambda_{\min}(W^* W) = 1 - \sigma_{\min}^2(W),$$

where $\sigma_{\min}(W)$ is the smallest singular value of W [2, p.16].

For our purposes in the analysis of eigenvalue solvers the following definition is most appropriate.

Definition 2.37 Let $S_1, S_2 \subset \mathbb{F}^n$ be of dimensions p and q and let $Q_1 \in \mathbb{F}^{n\times p}$ and $Q_2 \in \mathbb{F}^{n\times q}$ be matrices the columns of which form orthonormal bases of S_1 and S_2, respectively, i.e. $S_i = \mathcal{R}(Q_i)$, $i = 1, 2$. Then we define the angle ϑ, $0 \le \vartheta \le \pi/2$, between S_1 and S_2 by

$$\sin\vartheta = \sin\angle(S_1, S_2) = \begin{cases} \sqrt{1 - \sigma_{\min}^2(Q_1^* Q_2)} & \text{if } p = q, \\ 1 & \text{if } p \neq q. \end{cases}$$

If $p = q$ the equations (2.49)–(2.51) imply that

$$\begin{aligned} \sin^2\vartheta &= \max_{\|\mathbf{a}\|=1} \mathbf{a}^*(I_p - W^*W)\mathbf{a} = \max_{\|\mathbf{b}\|=1} \mathbf{b}^*(I_p - WW^*)\mathbf{b} \\ &= \|(I_n - Q_2Q_2^*)Q_1\| = \|(I_n - Q_1Q_1^*)Q_2\| \\ &= \|(Q_1Q_1^* - Q_2Q_2^*)Q_1\| = \|(Q_1Q_1^* - Q_2Q_2^*)Q_2\| \end{aligned} \tag{2.53}$$

Let $\mathbf{x} \in S_1 + S_2$. Then $\mathbf{x} = \tilde{\mathbf{q}}_1 + \tilde{\mathbf{q}}_2$ with $\tilde{\mathbf{q}}_i \in S_i$. We write

$$\mathbf{x} = \tilde{\mathbf{q}}_1 + Q_1Q_1^*\tilde{\mathbf{q}}_2 + (I_n - Q_1Q_1^*)\tilde{\mathbf{q}}_2 =: \mathbf{q}_1 + \mathbf{q}_2$$

with $\mathbf{q}_1 = Q_1\mathbf{a}$ and $\mathbf{q}_2 = Q_2\mathbf{b} = (I_n - Q_1Q_1^*)Q_2\mathbf{b}$. Then

$$\begin{aligned} \|(Q_1Q_1^* - Q_2Q_2^*)\mathbf{x}\|^2 &= \|(Q_1Q_1^* - Q_2Q_2^*)(Q_1\mathbf{a} + Q_2\mathbf{b})\|^2 \\ &= \|Q_1\mathbf{a} + Q_2Q_2^*Q_1\mathbf{a} + Q_2\mathbf{b}\|^2 \\ &= \|(I_n - Q_2Q_2^*)Q_1\mathbf{a} + Q_2\mathbf{b}\|^2 \\ &= \mathbf{a}^*Q_1^*(I_n - Q_2Q_2^*)Q_1\mathbf{a} \\ &\qquad + 2\mathrm{Re}(\mathbf{a}^*Q_1^*(I_n - Q_2Q_2^*)Q_2\mathbf{b}) + \mathbf{b}^*Q_2^*Q_2\mathbf{b} \\ \sin^2\vartheta &= \max_{\|\mathbf{a}\|=1} \mathbf{a}^*Q_1^*(I_n - Q_2Q_2^*)Q_1\mathbf{a}, \\ &= \max_{\|\mathbf{a}\|=1} \mathbf{a}^*Q_1^*(Q_1Q_1^* - Q_2Q_2^*)Q_1\mathbf{a}, \\ &= \max_{\mathbf{x}\in S_1\setminus\{\mathbf{0}\}} \frac{\mathbf{x}^*(Q_1Q_1^* - Q_2Q_2^*)\mathbf{x}}{\mathbf{x}^*\mathbf{x}}. \end{aligned}$$

Thus, $\sin\vartheta$ is the maximum of the Rayleigh quotient $R(\mathbf{x})$ corresponding to $Q_1Q_1^* - Q_2Q_2^*$, that is the largest eigenvalue of $Q_1Q_1^* - Q_2Q_2^*$. As $Q_1Q_1^* - Q_2Q_2^*$ is symmetric and positive semi-definite, its largest eigenvalue equals its norm,

Lemma 2.38 $\sin\angle(S_1, S_2) = \|Q_2Q_2^* - Q_1Q_1^*\|$

Lemma 2.39 $\angle(S_1, S_2) = \angle({S_1}^\perp, {S_2}^\perp)$.

Proof. Because

$$\|Q_2Q_2^* - Q_1Q_1^*\| = \|(I - Q_2Q_2^*) - (I - Q_1Q_1^*)\|$$

the claim immediately follows from Lemma 2.38. ∎

Bibliography

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide - Release 2.0*, SIAM, Philadelphia, PA, 1994. (Software and guide are available from Netlib at URL `http://www.netlib.org/lapack/`).

[2] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 2nd ed., 1989.

[3] R. Zurmühl, *Matrizen und ihre technischen Anwendungen*, Springer, Berlin, 4th ed., 1964.