Numerical Methods for Eigenvalue Problems
Lecture Notes

Jan Brandts

December 6, 2007

Contents

Chapter 1

Preliminaries

An important concept in numerical analysis is projection on a finite dimensional subspace of a given vector space. Here we only consider orthogonal projection. A simple example of this is to consider $\mathbb{R}^2$ and its subspace $\mathcal{V} = \{(x, y) \in \mathbb{R}^2 \mid y = 0\}$, the x-axis. Intuitively, it is clear what is meant by orthogonal projection on $\mathcal{V}$: given a vector $x = (x_1, x_2)^* \in \mathbb{R}^2$, we think of the vector $P(x) = (x_1, 0)^* \in \mathcal{V}$ as its orthogonal projection on $\mathcal{V}$, because the difference $x - P(x)$ is perpendicular to $\mathcal{V}$.

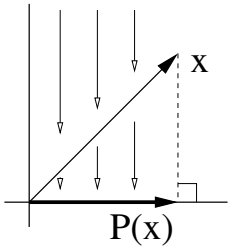

Figure 1: A vector x and its orthogonal projection on the horizontal axis.

An important fact is that $P(x)$ is the vector from $\mathcal{V}$ closest to x, when their mutual distance $x - P(x)$ is measured in the standard norm. We say that $P(x)$ is the best approximation of x in $\mathcal{V}$ with respect to this norm. This has interesting consequences in numerical analysis.

In this chapter we will show how to compute projections on subspaces of $\mathbb{R}^n$, but also on (finite dimensional) subspaces of function spaces such as $C^0(I)$. For this, we need to know what mutually orthogonal functions are. The abstract notion of inner product and its so-called derived or associated norm are hereby central concepts. They further lead to the definition of orthogonal transformations, which are important in the QR-decomposition and in the theory of canonical forms of matrices.

1.1 Norms and inner products on a vector space

Well-known examples of norms on $\mathbb{R}^n$ are the Euclidean (or standard) norm $\|\cdot\|_2$ on $\mathbb{R}^n$, and the supremum (or maximum) norm $\|\cdot\|_\infty$ on $\mathbb{R}^n$, defined by

$$\|x\|_2 = \sqrt{x^*x} = \sqrt{\sum_{j=1}^{n} x_j^2} \quad \text{and} \quad \|x\|_\infty = \sup\{|x_j| \mid j \in \{1, \ldots, n\}\}. \tag{1.1}$$

Norms can be considered as measuring the magnitude of an element of the vector space, and since the meaning of magnitude can be different in different situations, there exist many different norms. The norm axioms below express which properties any norm should have to be rightfully called a norm in the mathematical sense.

Definition 1.1.1 *A mapping $\|\cdot\| : V \to \mathbb{R}$ is a norm on V if and only if:*

- $\forall x \in V : \|x\| \geq 0$, *and* $\|x\| = 0 \Leftrightarrow x = 0$,
- $\forall \alpha \in \mathbb{R}, \forall x \in V : \|\alpha x\| = |\alpha|\|x\|$,
- $\forall x, y \in V : \|x + y\| \leq \|x\| + \|y\|$ *("triangle inequality")*

Also on other vector spaces, such as the space $C^k(I)$ of k times continuously differentiable functions on an interval I, norms and inner products exist. For example, the standard norm and the supremum norm on $C^0(I)$ are defined by

$$\|f\|_2 = \sqrt{\int_I f(x)^2 dx} \quad \text{and} \quad \|f\|_\infty = \sup\{|f(x)| \mid x \in I\}.$$

As we will show below, norms can be (but do not necessarily need to be) defined by means of inner products. Inner products additionally equip the vector space at hand with the concept of orthogonality. Inner products, like norms, are supposed to satisfy certain axioms.

Definition 1.1.2 *A mapping $(\cdot,\cdot) : V \times V \to \mathbb{R}$ is an inner product on V if and only if*

- $\forall x, y \in V : (x, y) = (y, x)$
- $\forall x, y, z \in V : (x + y, z) = (x, z) + (y, z)$
- $\forall \alpha \in \mathbb{R}, \forall x, y \in V : (\alpha x, y) = \alpha(x, y)$
- $\forall x \in V : (x, x) \geq 0$, *and* $(x, x) = 0 \Leftrightarrow x = 0$

The most important example of an inner product on $\mathbb{R}^n$ is the standard inner product defined by

$$(x, y)_{\mathbb{R}^n} = \sum_{j=1}^{n} x_j y_j. \tag{1.2}$$

This inner product is so standard that it is usually simply written as $(\cdot,\cdot)$. Recall that this inner product allows a nice and compact notation that makes use of the matrix-vector multiplication on $\mathbb{R}^n$, and which is

$$(x, y) = x^* y = (x_1, \cdots, x_n) \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{1.3}$$

On $C^0(I)$, the standard inner product is the following:

$$(f, g)_{C^0(I)} = \int_I f(x) g(x) dx.$$

Here too, the subscript is usually suppressed and we simply write $(\cdot,\cdot)$. It is no coincidence that the standard norms on $\mathbb{R}^n$ and $C^0(I)$ can be computed by means of the standard inner products on these spaces. Indeed, we have that

$$\|x\|_2 = \sqrt{(x,x)} \quad \text{and} \quad \|f\|_2 = \sqrt{(f,f)} \tag{1.4}$$

for all $x \in \mathbb{R}^n$ and all $f \in C^0(I)$. Such a relation is valid more generally. Suppose that $(\cdot,\cdot)_V$ satisfies the inner product axioms on some vector space V. Define

$$\|x\|_V := \sqrt{(x,x)_V}. \tag{1.5}$$

Then it can be verified that $\|\cdot\|_V$ satisfies the norm axioms on V. This norm is called the associated norm (also called induced norm, derived norm, corresponding norm). As we saw above, the standard norm $\|\cdot\|_2$ on $\mathbb{R}^n$ is associated to the standard inner product on $\mathbb{R}^n$, and the same holds for the standard norm and inner product on $C^0(I)$.

Remark 1.1.3 There also exist norms that are not associated to any inner product. An example is the supremum norm $\|\cdot\|_\infty$, both on $\mathbb{R}^n$ and on $C^0(I)$.

An important result on both inner products and their induced norms is the following inequality. It is frequently used in numerous applications.

Theorem 1.1.4 (Cauchy-Schwarz inequality) *Let $(\cdot,\cdot)_V$ be an inner product on V together with its induced norm $\|\cdot\|_V$. Then for each $v, w \in V$,*

$$|(v,w)_V| \leq \|v\|_V \|w\|_V. \tag{1.6}$$

Equality holds if and only if v and w are linearly dependent.

Proof. Since the inequality if trivially true if either v or w is zero, let non-zero v and w be given. Also, let $\lambda \in \mathbb{R}$ be any scalar. Then, by definition of the induced norm and using the inner product axioms, we find that

$$0 \leq \|v - \lambda w\|_V^2 = (v - \lambda w, v - \lambda w)_V = (v,v)_V - 2\lambda(v,w)_V + \lambda^2(w,w) \tag{1.7}$$

for all $\lambda \in \mathbb{R}$. In particular, it holds for

$$\lambda = \frac{(v,w)_V}{(w,w)_V}. \tag{1.8}$$

Substituting this in (1.7) shows that

$$0 \leq (v,v)_V - 2\frac{(v,w)_V}{(w,w)_V}(v,w)_V + \frac{(v,w)_V^2}{(w,w)_V^2}(w,w)_V = (v,v)_V - \frac{(v,w)^2}{(w,w)}. \tag{1.9}$$

From this we find immediately that

$$(v,w)^2 \leq (v,v)_V (w,w)_V \quad \text{hence} \quad |(v,w)_V| \leq \sqrt{(v,v)_V}\sqrt{(w,w)} = \|v\|_V \|w\|_V. \tag{1.10}$$

This proves the statement. □

Remark 1.1.5 The Cauchy-Schwarz inequality shows in particular that for all non-zero $v, w \in V$ we have that

$$-1 \leq \frac{(v,w)_V}{\|v\|_V \|w\|_V} \leq 1. \tag{1.11}$$

The expression in the middle is usually interpreted as the cosine of the angle between v and w, just as this is done in $\mathbb{R}^n$.

1.2 Orthogonal projection on a finite dimensional subspace

Inner products define what it means for two elements from the vector space to be orthogonal to each other. Notice that the following definition is consistent with Remark 1.1.5.

Definition 1.2.1 *Two elements v, w of a vector space V are said to be orthogonal to each other with respect to the inner product $(\cdot,\cdot)_V$ if and only if $(v, w)_V = 0$.*

For the standard inner product, this corresponds to our intuition of what orthogonality should be. Indeed, vectors x and y in $\mathbb{R}^3$ are orthogonal if and only if they are perpendicular to each other. But there exist other inner products on $\mathbb{R}^3$ for which this intuition is no longer valid. It may be that x and y are orthogonal even though they are not perpendicular to each other in the usual sense.

Definition 1.2.2 *Let W be a finite dimensional subspace of V. A vector $r \in V$ is orthogonal to W with respect to the inner product $(\cdot,\cdot)_V$ if and only if*

$$\forall w \in W : \quad (r, w)_V = 0. \tag{1.12}$$

Since inner products are bilinear, it is not very hard to verify that this is equivalent to demanding that r is orthogonal to each of the basis vectors of a given basis for W, as depicted below.

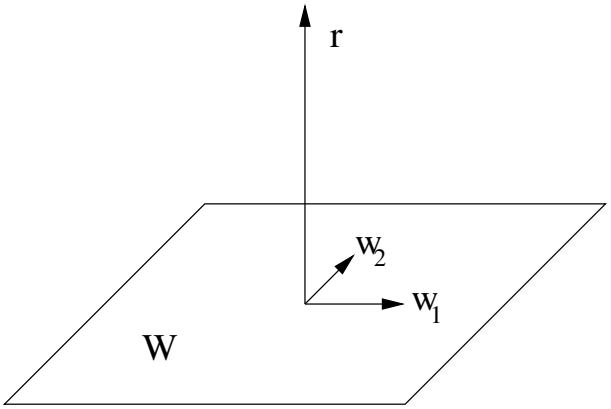

Figure 2: Orthogonality of r to W is the same as $r \perp w_1$ and $r \perp w_2$.

As a result, orthogonality of v to W can be verified by evaluation of a finite number of inner products only. Lemma 1.3.4 contains a special case of this statement.

The concept of orthogonality to a subspace is central in the definition of the orthogonal projection on a subspace.

Definition 1.2.3 *Let W be a finite dimensional subspace of V and let $v \in V$. The orthogonal projection $P_W(v)$ of v on W with respect to $(\cdot,\cdot)_V$ is the element from W such that*

$$\forall w \in W : \quad (v - P_W(v), w)_V = 0. \tag{1.13}$$

This definition is graphically illustrated in the left picture in Figure 3. At this point though, it is not yet clear if such an element exists, nor that it is unique. We will show this further on. First, we prove that the projection $P_W(v)$ of v on W with respect to $(\cdot,\cdot)_V$ is the best approximation of v in W measured in the associated norm $\|\cdot\|_V$, as depicted in the right picture in Figure 3.

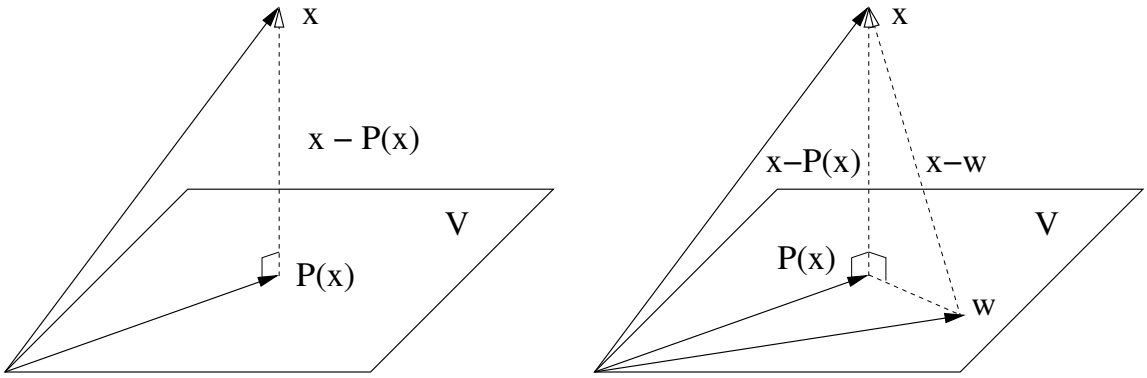

Figure 3: Illustration of Definition 1.2.3 and Theorem 1.2.4.

Theorem 1.2.4 *Let V be a vector space with subspace W and with inner product $(\cdot,\cdot)_V$ and induced norm $\|\cdot\|_V$. Denote the orthogonal projection of $v \in V$ with respect to $(\cdot,\cdot)_V$ on W by $P_W(v)$. Then for all $w \in W$ we have*

$$\|v - P_W(v)\|_V \leq \|v - w\|_V, \tag{1.14}$$

and $P_W(v)$ is called the best approximation of v within W measured in the induced norm.

Proof. Let $w \in W$ be given. Then also $w - P_W(v) \in W$, because W is a subspace of V. By definition, we find therefore,

$$(v - P_W(v), w - P_W(v))_V = 0 \tag{1.15}$$

Using this, we find, using the inner product axioms, that

$$\|v - P_W(v)\|_V^2 = (v - P_W(v), v - P_W(v))_V = (v - P_W(v), v - w + w - P_W(v))_V$$

$$= (v - P_W(v), v - w)_V. \tag{1.16}$$

Now, apply the Cauchy-Schwarz inequality (1.6) to the right-hand side. This gives

$$\|v - P_W(v)\|_V^2 \leq \|v - P_W(v)\|_V \|v - w\|_V.$$

Division by $\|v - P_W(v)\|_V$ finishes the proof. □

We will illustrate the actual computation of the projection on a subspace in four explicit situations. The simplest example is projection on a one-dimensional subspace of $\mathbb{R}^n$.

1.3 Projection on subspaces of $\mathbb{R}^n$

Since projection in $\mathbb{R}^n$ is intuitively the most clear, we first give two examples related to $\mathbb{R}^n$. The first example involves projection on a line in $\mathbb{R}^2$, the second projection on a k-dimensional subspace of $\mathbb{R}^n$.

1.3.1 Projection on a line in $\mathbb{R}^2$

As a simple example, we will compute the projection $P_v(x)$ of a given non-zero vector $x \in \mathbb{R}^2$ on the line ℓ spanned by a non-zero vector $v \in \mathbb{R}^2$, with respect to the standard inner product. Notice that $\mathcal{B} = \{v\}$ is a basis for ℓ. Now, two properties fully characterize $P_v(x)$:

- (A) $P_v(x) = \alpha v$ for some $\alpha \in \mathbb{R}$,
- (B) $x - P_v(x) \perp v$.

Property (A) expresses that the projection of x on ℓ is an element from that subspace, and thus it can be written as a yet unknown coordinate α times the basis vector v. Property (B) says that the difference between a vector and its projection is orthogonal to the space upon which is being projected. This reflects Definition 1.2.3.

We will use the compact matrix-vector notation (1.3). Especially in the upcoming section its convenience will become clear. Substituting (A) into (B) gives:

$$x - \alpha v \perp v \Leftrightarrow (x - \alpha v, v) = 0 \Leftrightarrow v^*(x - \alpha v) = 0 \Leftrightarrow \alpha = \frac{v^* x}{v^* v}. \tag{1.17}$$

Substituting this back into (A) gives that

$$P_v(x) = v\frac{v^* x}{v^* v} \quad \text{and hence} \quad P_v = \frac{v v^*}{v^* v} \tag{1.18}$$

is the 2×2 matrix that represents the projection on v.

Remark 1.3.1 Notice a subtlety in the above derivation. The expression αv with $\alpha \in \mathbb{R}$ is well-defined. But substituting for α the expression computed in (1.17) makes αv an impossible matrix-vector multiplication. However, $v\alpha$ does not suffer from this. Reason is, that if α is interpreted as a 1×1 matrix, only $v\alpha$ is a valid matrix-matrix multiplication.

Proposition 1.3.2 *For all $w \in \ell$ we have that*

$$\|x - P_v(x)\|_2 \leq \|x - w\|_2. \tag{1.19}$$

In other words, $P_v(x)$ is the vector in the line ℓ that is closest to x in the standard norm.

Proof. This is a special case of Theorem 1.2.4. □

1.3.2 Projection on a k-dimensional subspace $\mathcal{V} \subset \mathbb{R}^n$

Consider a subspace $\mathcal{V} \subset \mathbb{R}^3$ of dimension two. This subspace is a plane through the origin. Suppose that v_1 and v_2 are vectors in $\mathbb{R}^3$ that form a basis for $\mathcal{V}$. This means that for all $v \in \mathcal{V}$ there exist coordinates α_1 and α_2 such that $v = \alpha_1 v_1 + \alpha_2 v_2$. In fact, it holds that

$$\mathcal{V} = \{v \mid v = \alpha_1 v_1 + \alpha_2 v_2 \quad \text{for some } \alpha_1, \alpha_2 \in \mathbb{R}\}. \tag{1.20}$$

Another way of writing this same expression is to make use of the matrix V, which has two columns, being v_1 and v_2. We will write $V = (v_1|v_2)$ for this matrix, which has 3 rows. Write $\alpha = (\alpha_1, \alpha_2)^*$, and verify that

$$v = \alpha_1 v_1 + \alpha_2 v_2 = (v_1|v_2)\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = V\alpha. \tag{1.21}$$

The expression above makes use of a very important interpretation of the concept of *matrix-vector multiplication*. This interpretation is the first one of two, given below.

Observation 1.3.3 (Matrix-vector multiplication) Let V be a matrix with n rows and k columns. Denote the columns by $v_1, \ldots, v_k \in \mathbb{R}^n$. Let $y \in \mathbb{R}^k$, then

$$Vy = (v_1|\ldots|v_k)\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = y_1 v_1 + \cdots + y_k v_k. \tag{1.22}$$

This expresses that Vy is the specific linear combination of the columns of V with as coefficients the *entries* of the vector y. Alternatively, denoting the rows of V by $v^1, \ldots, v^n$, we see that

$$Vy = \left[\begin{array}{c} v^1 \\ \hline \vdots \\ \hline v^n \end{array}\right] y = \begin{bmatrix} v^1 y \\ \vdots \\ v^n y \end{bmatrix} \tag{1.23}$$

showing that Vy can also be seen as a collection of inner products.

Using the first interpretation of matrix-vector multiplication from the observation above, we see that $\mathcal{V}$ can alternatively be characterized as

$$\mathcal{V} = \{v \mid v = V\alpha, \quad \alpha \in \mathbb{R}^2\}. \tag{1.24}$$

With this characterization at our disposal we return to orthogonal projections. First we prove a lemma that arises from Definition 1.2.2 of orthogonality to a subspace. The lemma is formulated for the general case of a k dimensional subspace $\mathcal{V} \subset \mathbb{R}^n$.

Lemma 1.3.4 *Let $\mathcal{V} \subset \mathbb{R}^n$ be a k-dimensional subspace, and V a $k \times n$ matrix with the property that its columns $v_1, \ldots, v_k$ are a basis for $\mathcal{V}$. Then*

$$r \perp \mathcal{V} \Leftrightarrow V^* r = 0. \tag{1.25}$$

Proof. By definition (1.2.2) we have that

$$r \perp \mathcal{V} \Leftrightarrow (\forall v \in \mathcal{V} : r \perp v) \Leftrightarrow (\forall \alpha \in \mathbb{R}^k : r \perp V\alpha)$$

$$\Leftrightarrow (\forall \alpha \in \mathbb{R}^k : (V\alpha)^* r = 0) \Leftrightarrow (\forall \alpha \in \mathbb{R}^k : \alpha^* V^* r = 0) \Leftrightarrow V^* r = 0. \tag{1.26}$$

The last equivalence follows from choosing $\alpha = V^* r$ and the inner product axiom that states that only the zero vector is orthogonal to itself. □

The above lemma shows the convenience of the compact format notation $x^* y$ for the inner product between x and y. It explicitly uses the equivalences

$$(\forall j \in \{1, \ldots, n\} : v_j^* r = 0) \Leftrightarrow \left[\begin{array}{c} v_1^* \\ \hline \vdots \\ \hline v_n^* \end{array}\right] r = 0 \Leftrightarrow V^* r = 0. \tag{1.27}$$

It is now just a matter of following the same lines as in the one-dimensional example to see how the orthogonal projection $P_{\mathcal{V}}(x)$ of a vector $x \in \mathbb{R}^n$ on $\mathcal{V}$ can actually be computed in practice. We only need to have a matrix V available whose columns are a basis for $\mathcal{V}$. Then we combine the two characterizing properties:

- (A) $P_{\mathcal{V}}(x) = V\alpha$ for some $\alpha \in \mathbb{R}^k$,
- (B) $x - P_{\mathcal{V}}(x) \perp \mathcal{V}$,

by substituting (A) into (B), which leads, using the Lemma 1.3.4, to:

$$x - V\alpha \perp \mathcal{V} \Leftrightarrow V^*(x - V\alpha) = 0 \Leftrightarrow V^*V\alpha = V^*x \Leftrightarrow \alpha = (V^*V)^{-1}V^*x. \tag{1.28}$$

Substituting this back into (A) gives that

$$P_{\mathcal{V}}(x) = V(V^*V)^{-1}V^*x. \tag{1.29}$$

Proposition 1.3.5 *For all $v \in \mathcal{V}$ we have that*

$$\|x - P_{\mathcal{V}}(x)\|_2 \leq \|x - v\|_2, \tag{1.30}$$

No other element from $\mathcal{V}$ is closer to x than $P_{\mathcal{V}}(x)$ when measured in the standard norm.

Proof. This is a special case of Theorem 1.2.4. □

1.4 Projection onto finite dimensional subspaces of $C^0(I)$

Since also the space $C^0(I)$ has been equipped with an inner product, we are formally also able to compute projections on finite dimensional subspaces of $C^0(I)$. To keep things as simple as possible, here we will project on subspaces of dimension one and two only.

1.4.1 Projection on the one-dimensional subspace $\mathcal{P}^0(I)$ of $C^0(I)$

Consider the subspace $\mathcal{P}^0(I)$ of $C^0(I)$ consisting of all constant functions on $I = [a, b]$. The projection $P^0(f)$ of an arbitrary function $f \in C(I)$ on $\mathcal{P}^0(I)$ satisfies:

- (A) $P^0(f)$ is constant on I,
- (B) $f - P^0(f)$ is orthogonal to all constant functions.

It can be computed by using a basis $\mathcal{B} = \{\phi_0\}$ for the one-dimensional subspace $\mathcal{P}^0(I)$. Thus, $P^0(f) = \alpha\phi_0$ for some yet unknown coordinate $\alpha \in \mathbb{R}$. Substituting this information into (B) results in

$$f - P^0(f) \perp \mathcal{P}^0(I) \Leftrightarrow f - \alpha\phi_0 \perp \phi_0 \Leftrightarrow \alpha = \frac{(f, \phi_0)}{(\phi_0, \phi_0)}, \tag{1.31}$$

and hence,

$$P^0(f) = \frac{(f, \phi_0)}{(\phi_0, \phi_0)}\phi_0.$$

If we choose an explicit basis, we can compute some of the quantities involved. For example, choose $\phi_0 : I \to \mathbb{R} : x \mapsto 1$, then we find

$$P^0(f) = \frac{1}{b-a}\int_a^b f(x)dx, \tag{1.32}$$

which is the mean value of f over I.

Proposition 1.4.1 *For all $q \in \mathcal{P}^0(I)$ we have that*

$$\|f - P^0(f)\|_2 \leq \|f - q\|_2. \tag{1.33}$$

No $q \in \mathcal{P}^0(I)$ approximates f better than $P^0(f)$ when measured in the standard norm.

Proof. This is a special case of Theorem 1.2.4. □

1.4.2 Projection onto $\mathcal{P}^1(I)$

To make things slightly more complicated, we will now project on the two-dimensional subspace $\mathcal{P}^1(I) \subset C^0(I)$. The projection $P^1(f)$ of a function $f \in C^0(I)$ is fully characterized by:

- (A) $P^1(f) \in \mathcal{P}^1(I)$,
- (B) $f - P^1(f) \perp \mathcal{P}^1(I)$.

To compute $P^1(f)$ explicitly, we need a basis for the two-dimensional linear vector space $\mathcal{P}^1(I)$. Let ϕ_1 and ϕ_2 form a basis of $\mathcal{P}^1(I)$, then each $p \in \mathcal{P}^1(I)$, and $P^1(f)$ in particular, can be written as

$$P^1(f) = \alpha_1\phi_1 + \alpha_2\phi_2.$$

The unknown coordinates α_1, α_2 with respect to the basis ϕ_1, ϕ_2 can be computed by substitution of this expression into (B). This gives that

$$f - \alpha_1\phi_1 - \alpha_2\phi_2 \perp \mathcal{P}^1(I). \tag{1.34}$$

This is equivalent (by linearity) to demanding orthogonality to both basis functions only:

$$f - \alpha_1\phi_1 - \alpha_2\phi_2 \perp \phi_1 \quad \text{and} \quad f - \alpha_1\phi_1 - \alpha_2\phi_2 \perp \phi_2. \tag{1.35}$$

This, in turn, is equivalent to

$$(f, \phi_1) = \alpha_1(\phi_1, \phi_1) + \alpha_2(\phi_2, \phi_1) \quad \text{and} \quad (f, \phi_2) = \alpha_1(\phi_1, \phi_2) + \alpha_2(\phi_2, \phi_2), \tag{1.36}$$

and this can be written in the compact format

$$\begin{bmatrix} (\phi_1, \phi_1) & (\phi_2, \phi_1) \\ (\phi_1, \phi_2) & (\phi_2, \phi_2) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} (f, \phi_1) \\ (f, \phi_2) \end{bmatrix}. \tag{1.37}$$

This two-by-two linear system can be solved, after which the result can be plugged into $P^1(f) = \alpha_1\phi_1 + \alpha_2\phi_2$.

Again, we can compute some of the quantities involved explicitly. Choose as a basis the functions

$$\phi_1(x) = 1 \quad \text{and} \quad \phi_2(x) = x - m, \quad \text{where} \quad m = \frac{b+a}{2}. \tag{1.38}$$

This choice has as advantage that we immediately see that $(\phi_1, \phi_2) = (\phi_2, \phi_1) = 0$, because the product $\phi_1(x)\phi_2(x) = \phi_2(x)$ is odd around m and hence its integral vanishes. This gives that

$$\alpha_1 = \frac{(f, \phi_1)}{(\phi_1, \phi_1)} = \frac{1}{b-a}\int_a^b f(x)dx \quad \text{and} \quad \alpha_2 = \frac{(f, \phi_2)}{(\phi_2, \phi_2)} = \frac{12}{(b-a)^3}\int_a^b (x-m)f(x)dx, \tag{1.39}$$

and hence,

$$P^1(f) = \frac{1}{b-a}\int_a^b f(x)dx \cdot 1 + \frac{12}{(b-a)^3}\int_a^b (x-m)f(x)dx \cdot (x-m). \tag{1.40}$$

Proposition 1.4.2 *For all $\ell \in \mathcal{P}^1(I)$ we have that*

$$\|f - P^1(f)\|_2 \leq \|f - \ell\|_2. \tag{1.41}$$

No other linear function is closer to f than $P^1(f)$ when measured in the standard norm.

Proof. This is a special case of Theorem 1.2.4. □

1.5 QR-decomposition and applications

QR-decompositions of an matrix form an important concept in numerical linear algebra. They serve as a tool to solve linear systems, least-squares problems, and is an essential building block in the QR-iteration for eigenvalues. First we consider orthogonal transformations.

1.5.1 Orthogonal transformations

Here we consider the orthogonal transformations. Usually, the orthogonality to which is referred, is the orthogonality with respect to the standard inner product, although in particular situations, this may be different. For the time being, we restrict ourselves to the standard inner product.

Definition 1.5.1 (Orthogonal matrix) *A real $n \times k$ matrix Q for which the k columns are mutually orthogonal vectors of length one will be called orthogonal.* □

Recall the two interpretations of matrix-vector multiplication given in section 1.3.2. Using the second of these interpretations, we find that the orthonormality of the columns of Q is equivalent with $Q^*Q = I$. Indeed, denoting the columns of Q by $q_1, \ldots, q_k$ we see that

$$Q^*Q = \left[\begin{array}{c} q_1^* \\ \hline \vdots \\ \hline q_k^* \end{array}\right] \left[\begin{array}{c|c|c} q_1 & \ldots & q_k \end{array}\right] = \begin{bmatrix} q_1^*q_1 & \ldots & q_1^*q_k \\ \vdots & & \vdots \\ q_k^*q_1 & \ldots & q_k^*q_k \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}. \tag{1.42}$$

Notice that only in case $n = k$, or, in other words, only if Q is square, this implies that $Q^* = Q^{-1}$. In that case we have, apart from $Q^*Q = I$ also that $QQ^* = I$.

The first interpretation of matrix-vector multiplication allows us to prove the following lemma, which is characteristic for orthogonal transformations.

Proposition 1.5.2 *Let Q be an $n \times k$ orthogonal matrix. Then for all $y \in \mathbb{R}^k$ we have that*

$$\|y\| = \|Qy\|. \tag{1.43}$$

Proof. Denoting the columns of Q by $q_1, \ldots, q_k$ we see that

$$Qy = (q_1|\ldots|q_k) \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = y_1q_1 + \cdots + y_kq_k. \tag{1.44}$$

Since q_1 is orthogonal to $q_2, \ldots, q_k$, also y_1q_1 is orthogonal to $y_2q_2 + \cdots + y_kq_k$, and Pythagoras Theorem tells us that

$$\|Qy\|^2 = \|y_1q_1\|^2 + \|y_2q_2 + \cdots + y_kq_k\|^2 = y_1^2 + \|y_2q_2 + \cdots + y_kq_k\|^2. \tag{1.45}$$

This argument can be repeated for y_2q_2 and the sum $y_3q_3 + \cdots + y_kq_k$ and so on, until we reach the conclusion that

$$\|Qy\|^2 = y_1^2 + \cdots + y_k^2 = \|y\|^2, \tag{1.46}$$

which proves the statement. □

A much shorter and simpler looking proof of Proposition 1.5.2 is contained as a special case of the following, stronger result, which implies that not only lengths are preserved by orthogonal transformations, but also angles.

Proposition 1.5.3 *For all $y, z \in \mathbb{R}^k$ we have that $(y, z) = (Qy, Qz)$, where on the left we have the standard inner product on $\mathbb{R}^k$, and on the right the standard inner product on $\mathbb{R}^n$.*

Proof. By definition of the adjoint Q^* and using that $Q^*Q = I$ we find,

$$(Qy, Qz) = (Qy)^*(Qz) = y^*Q^*Qz = y^*z = (y, z), \tag{1.47}$$

where we have used that $(AB)^* = B^*A^*$ together with $Q^*Q = I$. □

1.5.2 QR-decomposition

One of the corner stones of linear algebra is the following result. Recall that a matrix R is called upper triangular if all its entries r_{ij} with $i > j$ are zero.

Theorem 1.5.4 (QR-decomposition) *Let A be a real $n \times k$ matrix. Then there exists an orthogonal matrix Q and an upper triangular matrix R such that $A = QR$.*

Proof. We use induction with respect to the number k of columns of A. If A has one column a_1 then if $\|a_1\| \neq 0$ a QR-decomposition is given by

$$a_1 = q_1 r_{11} \quad \text{where } q_1 = \frac{a_1}{\|a_1\|} \quad \text{and} \quad r_{11} = \|a_1\|. \tag{1.48}$$

In case $\|a_1\| = 0$ a QR-decomposition is given by $q_1 \cdot 0$ where q_1 with $\|q_1\| = 1$ is arbitrary.

Suppose now that A has k columns $a_1, \ldots, a_k$ and that

$$\left[\begin{array}{c|c|c} a_1 & \ldots & a_{k-1} \end{array}\right] = \left[\begin{array}{c|c|c} q_1 & \ldots & q_{k-1} \end{array}\right] \begin{bmatrix} r_{1,1} & \cdots & r_{1,k-1} \\ 0 & \ddots & \vdots \\ 0 & 0 & r_{k-1,k-1} \end{bmatrix} \tag{1.49}$$

is the QR-decomposition of the first $k-1$ columns of A. Define v_k as

$$v_k = (a_k^* q_1)q_1 + \cdots + (a_k^* q_{k-1})q_{k-1}. \tag{1.50}$$

If $a_k = v_k$ then with q_k an arbitrary vector of unit length orthogonal to $q_1, \ldots, q_{k-1}$ we have that

$$\left[\begin{array}{c|c|c} a_1 & \ldots & a_k \end{array}\right] = \left[\begin{array}{c|c|c} q_1 & \ldots & q_k \end{array}\right] \begin{bmatrix} r_{1,1} & \cdots & r_{1,k-1} & (a_k^* q_1) \\ 0 & \ddots & \vdots & \vdots \\ 0 & 0 & r_{k-1,k-1} & (a_k^* q_{k-1}) \\ 0 & \ldots & 0 & 0 \end{bmatrix}, \tag{1.51}$$

and this is clearly a QR-decomposition of A. In case $a_k \neq v_k$, let $w_k = a_k - v_k$ and $q_k = w_k/\|w_k\|$. Then $w_k \perp q_j$ for $j < k$ because

$$q_j^* w_k = q_j^*(a_k - (a_k^* q_1)q_1 + \cdots + (a_k^* q_{k-1})q_{k-1}) = q_j^* a_k - q_j^*(a_k^* q_j)q_j = 0.$$

Moreover, since $a_k = v_k + w_k = v_k + q_k\|w_k\|$ we find that

$$\left[\begin{array}{c|c|c} a_1 & \ldots & a_k \end{array}\right] = \left[\begin{array}{c|c|c} q_1 & \ldots & q_k \end{array}\right] \begin{bmatrix} r_{1,1} & \cdots & r_{1,k-1} & (a_k^* q_1) \\ 0 & \ddots & \vdots & \vdots \\ 0 & 0 & r_{k-1,k-1} & (a_k^* q_{k-1}) \\ 0 & \ldots & 0 & \|w_k\| \end{bmatrix} \tag{1.52}$$

is a QR-decomposition of A. This completes the proof. □

Remark 1.5.5 *QR-decomposition is not unique. Let D be diagonal with entries $d_{jj} = \pm 1$. Then $D^2 = I$ and $A = (QD)(DR)$, and QD is orthogonal and DR is upper triangular.* □

Remark 1.5.6 *Basically, the QR-decomposition is the result of the Gram-Schmidt orthonormalization process applied to the columns of A from left to right. If such a column is linearly dependent from the previous, this shows as a zero on the diagonal of R.* □

Notice that if $k < n$ and $A = QR$ is a QR-decomposition of A, we can do the following. Let $q \in \mathbb{R}^n$ be a unit vector orthogonal to all columns of Q. Such a vector exists because if $k < n$, the columns of Q cannot form an orthonormal basis for $\mathbb{R}^n$. Write 0_k^* for the horizontal vector of k zero entries. Then we have that

$$[Q|q] \begin{bmatrix} R \\ \hline 0_k^* \end{bmatrix} = QR + q0_k^* = QR, \tag{1.53}$$

and the product at the left consists of an orthogonal and an upper triangular matrix. Therefore, we conclude the following.

Observation 1.5.7 *If A is $n \times k$ with $k < n$, then for each $j = k, \ldots, n$ there exists a QR-decomposition of A such that Q is $n \times j$ and R is $j \times k$.* □

There are two formats for the QR-decomposition that have particular interest:

- Q is $n \times k$ and R is $k \times k$, the *thin* QR-decomposition,
- Q is $n \times n$ and R is $n \times k$, the *full* QR-decomposition.

Clearly, the computation of the thin QR-decomposition is the most economical, but in theoretical considerations it is often convenient to consider the full QR-decomposition, since it involves the square and invertible orthogonal transformation Q.

1.5.3 Example: Gram-Schmidt process for two vectors

First, we present an example of how to compute a QR-decomposition in practice. Or in other words, we illustrate the Gram-Schmidt orthogonalization process applied to two vectors.

Let A be an $n \times 2$ matrix consisting of columns a_1 and a_2. For convenience we assume that a_1 and a_2 are non-zero and linearly independent. We distinguish the following steps.

Normalization of the first column. Since $a_1 \neq 0$ we can define

$$q_1 = \frac{a_1}{\|a_1\|}. \tag{1.54}$$

This results in the QR-decomposition of the first column of A:

$$a_1 = q_1 \|a_1\|. \tag{1.55}$$

Orthogonormalization of a_2 to q_1. The vector a_2 can be decomposed in a vector in the direction of q_1 and a vector $\tilde{q}_2$ orthogonal to q_1. Indeed, define

$$\tilde{q}_2 = a_2 - q_1(q_1^* a_2), \tag{1.56}$$

then $q_1^*\tilde{q}_2 = q_1^*(a_2 - q_1(q_1^*a_2)) = 0$. Now define $q_2 = \tilde{q}_2/\|q_2\|$, which gives $q_2\|\tilde{q}_2\| = \tilde{q}_2$. Substitution in (1.56) gives $a_2 = q_1(q_1^*a_2) + q_2\|\tilde{q}_2\|$. This latter equality can be retraced by comparing the second columns in

$$[a_1|a_2] = (q_1|q_2)\begin{bmatrix} \|a_1\| & a_2^*q_1 \\ 0 & \|\tilde{q}_2\| \end{bmatrix}. \tag{1.57}$$

This is a QR-decomposition of A.

1.5.4 Application: direct solution of linear systems

A square and non-singular linear system $Ax = b$ can be solved quite efficiently if a QR-decomposition of A is available. The efficiency comes from the fact that $Q^{-1} = Q^*$ and from the comfortable way in which systems with upper triangular matrices can be solved. Indeed, multiplying both sides of $QRx = b$ by Q^* gives

$$Rx = Q^*b, \quad \text{or} \quad \begin{bmatrix} * & \cdots & * \\ 0 & \ddots & \vdots \\ 0 & 0 & * \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = Q^*\begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}. \tag{1.58}$$

Then, x_k can be solved from the last equation, and one can proceed upwards by substitution. This process, called *backward substitution*, costs only in the order of n^2 arithmetical operations. For lower triangular matrices, the solution process is called *forward substitution*.

1.5.5 Application: solving least-squares problems

Let $k \leq n$ and let A be an $n \times k$ matrix with linearly independent columns (or, in other words, A has column rank k, or is of full rank). Let $b \in \mathbb{R}^n$ be given. The least-squares problem is the problem of finding the vector $x \in \mathbb{R}^k$ for which the difference $Ax - b$ has minimal euclidean norm:

$$\text{Find } x \in \mathbb{R}^k \text{ such that for all } y \in \mathbb{R}^k,\ \|Ax - b\| \leq \|Ay - b\|. \tag{1.59}$$

Now, let $A = QR$ be the full QR-decomposition of A, then Q is square and $Q^*Q = QQ^* = I$, so we can write

$$Ay - b = QRy - QQ^*b = Q(Ry - Q^*b). \tag{1.60}$$

Therefore, Proposition 1.5.2 shows that that we only need to solve the easier problem

$$\text{Find } x \in \mathbb{R}^k \text{ such that for all } y \in \mathbb{R}^k,\ \|Rx - Q^*b\| \leq \|Ry - Q^*b\|. \tag{1.61}$$

Since R is zero below the k-th row, the last $n - k$ entries of $Ry - Q^*b$ are equal to the last $n - k$ entries of Q^*b regardless of the choice of y. By choosing y such that the first k entries of $Ry - Q^*b$ are zero, we minimize the norm of $Ry - Q^*b$. This is equivalent to solving the square linear system $Ry = Q^*b$ where Q and R form a thin QR-decomposition of A.

Therefore, in order to solve the least-squares problem, we only need to solve the first k equations of the n equations $Ry = Q^*b$. For this we only need the thin QR-decomposition of A. This nicely illustrates that the full decomposition is handy in theoretical considerations (the derivation above) while in practice, you only need to compute the thin one.

1.6 Canonical forms of matrices

Important properties of a matrix A depend strongly on whether A is real and symmetric, or complex Hermitian, possibly positive definite, normal or non-normal. The orthogonal and unitary transformations play an important role in the theory of canonical forms of matrices.

1.6.1 The Schur decomposition

For an arbitrary matrix we recall the following important classical result.

Theorem 1.6.1 (Schur decomposition) *Any $n\times n$ matrix A can be written as $A = U^*RU$, or in other words,*

$$AU = UR, \tag{1.62}$$

where U is a unitary matrix, i.e., $U^*U = I$, and R an upper triangular matrix.

Proof. By induction. For $n = 1$ the statement is trivially true. Now, suppose that for all $(n-1)\times(n-1)$ matrices the statement is true. Let $\lambda \in \mathcal{C}$ be an eigenvalue of A and $v \in \mathcal{C}^n$ a unit length eigenvector belonging to λ. Let W be unitary of dimensions $n \times (n-1)$ and such that $W^*v = 0$. Then

$$A(v|W) = (v|W) \begin{pmatrix} \lambda & w^* \\ 0 & B \end{pmatrix}. \tag{1.63}$$

By the induction hypothesis, there exists unitary Q and upper triangular R such that $BQ = QR$. Using this, we find

$$A(v|W) \begin{pmatrix} 1 & 0 \\ 0 & Q \end{pmatrix} = (v|W) \begin{pmatrix} 1 & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} \lambda & w^* \\ 0 & R \end{pmatrix}. \tag{1.64}$$

Since the product of unitary matrices is unitary, this proves the statement. □

Remark 1.6.2 Notice that R has the same eigenvalues as A. Since R is upper triangular, the eigenvalues of R are on its diagonal. Therefore, the Schur decomposition is an *eigenvalue revealing decomposition*: it shows the eigenvalues of A.□

Corollary 1.6.3 *Suppose that $A = A^*$. Then there exists a diagonal matrix Λ and a unitary matrix U such that*

$$AU = U\Lambda. \tag{1.65}$$

As a result, the columns of U are the eigenvectors of A, and the diagonal elements of Λ are the eigenvalues of A. Those eigenvalues are real numbers.

Proof. Let $U^*AU = R$ be the Schur decomposition of A. Taking the complex conjugate transpose gives that $U^*A^*U = R^*$. Since $A = A^*$, this means that $R^* = R$. Therefore, R must be diagonal, and the diagonal entries must be real. □

A matrix A for which $A^* = A$ is called *Hermitian*. The above corollary states that there exists an orthonormal basis of eigenvectors $v_1, \ldots, v_n$ of A of $\mathcal{C}^n$. With respect to this basis, A becomes real and diagonal. Real matrices A for which $A^* = A$ are called *symmetric*, and have the same properties as a Hermitian matrix. Additionally, the eigenvectors of a symmetric matrix are real. In both cases, if all eigenvalues are positive, A is called *positive definite*.

1.6.2 Normal and diagonalizable matrices

The real symmetric and complex Hermitian matrices are not the only matrices for which there exists an orthonormal basis of eigenvectors. Such matrices A are called *normal*.

Definition 1.6.4 (Normal matrices) *If there exists an orthonormal basis of the whole space consisting of eigenvectors of A, then A is called a normal matrix.*□

Remark 1.6.5 A matrix A is normal if and only if $A^*A = AA^*$. The latter characterization is usually much easier to check. We leave it as an exercise to the reader to verify that those two properties are equivalent.□

A matrix is called *non-normal* if there does not exist an orthonormal basis of eigenvectors of A. If nonetheless there still exists a basis of eigenvectors of A, then A is called *diagonalizable*.

Definition 1.6.6 (Diagonalizable matrices) *A matrix A is called diagonalizable if there exists a diagonal matrix Λ and a non-singular matrix V such that*

$$AV = V\Lambda. \tag{1.66}$$

Thus, A is diagonalizable if and only if there exists a basis of the whole space consisting of eigenvectors of A only. Matrices that cannot be diagonalized are called defective.□

Here we summarize the most important implications of the properties just defined.

- A is real symmetric $\Rightarrow$ A is Hermitian $\Rightarrow$ A is normal $\Rightarrow$ A is diagonalizable,
- A is defective $\Rightarrow$ A is non-normal $\Rightarrow$ A is non-Hermitian.

None of the above implications are an equivalence. The properties of being normal and diagonalizable are the ones to remember.

Proposition 1.6.7 *Let A be normal. Then for each $\varepsilon > 0$ there exists a non-normal matrix A_ε such that*

$$\|A - A_\varepsilon\| \leq \varepsilon. \tag{1.67}$$

In other words, the set of non-normal matrices is dense in the set of all matrices.

Proof. First of all, since all norms on finite dimensional spaces are equivalent, we need not specify the topology implied by the word *dense*. Now, let A be normal and $AU = UD$ with $U^*U = I$ and Λ diagonal. Define matrices A_ε by

$$A_\varepsilon = U_\varepsilon \Lambda U_\varepsilon^*, \quad \text{where} \quad U_\varepsilon = U \begin{pmatrix} \cos(\varepsilon) & 0 & 0 \\ \sin(\varepsilon) & 1 & 0 \\ 0 & 0 & I \end{pmatrix}. \tag{1.68}$$

Then $U_0 = U$ and $A_0 = A$, and for all ε with $0 < \varepsilon < \pi$, we have that U_ε is non-unitary. But (1.68) is a diagonalization of A_ε. Therefore, A_ε with $\varepsilon \to 0$ consists of non-normal matrices converging to A. This proves the statement. □

Proposition 1.6.8 *Let A be defective. Then for each $\varepsilon > 0$ there exists a diagonalizable matrix A_ε such that*

$$\|A - A_\varepsilon\| \leq \varepsilon. \tag{1.69}$$

In other words, the set of diagonalizable matrices is dense in the set of all matrices.

Proof. Since A is defective, it has an eigenvalue λ for which the dimension k of the eigenspace $\mathcal{V}_\lambda$ is strictly less than the multiplicity ℓ of λ, the latter being the number of times that λ occurs on the diagonal of the upper triangular factor of the Schur decomposition. It is possible to construct a parameter dependent matrix A_ε such that for all $\varepsilon > 0$, the matrix A_ε has ℓ different eigenvalues that all converge to λ for $\varepsilon \to 0$, with corresponding eigenvectors that are linearly independent but all converge to elements in $\mathcal{V}_\lambda$. We omit the explicit form of such a matrix. □

These density results are of interest if we consider the numerical approximation problem of eigenvalues of a matrix. Thinking about the effects of finite precision arithmetic, we may be tempted to state

- Every matrix A can be diagonalized numerically,
- Normal matrices have no numerical interest at all.

but both statements are not true. It turns out that it pays off to stay as closely as possible to a normal matrix, and also to stay away as far as possible from defective matrices.

1.6.3 The Singular Value Decomposition

In what follows we need the notion of *singular values*. First we recall the definition of the Singular Value Decomposition and its existence proof, together with its significance.

Theorem 1.6.9 (Singular Value Decomposition) *For each $n \times k$ matrix A there exist two unitary matrices U and V and a diagonal matrix Σ with non-negative entries such that*

$$AV = U\Sigma. \tag{1.70}$$

In the so-called thin decomposition, U is $n \times k$, and both V and Σ and $k \times k$, whereas in the full decomposition U is $n \times n$, V is $k \times k$ and Σ is $n \times k$.

Proof. We sketch a proof by induction. Assume that for all $(n-1) \times (k-1)$ matrices, the decomposition exists. Let v and u be such that $Av = \|A\|u$, and V, U such that $(v|V)$ and $(u|U)$ are unitary, then

$$A(v|V) = (u|U) \begin{pmatrix} \|A\| & w^* \\ 0 & B \end{pmatrix}.$$

Since

$$A(v|V) \begin{pmatrix} \|A\| \\ w \end{pmatrix} = (u|U) \begin{pmatrix} \|A\| & w^* \\ 0 & B \end{pmatrix} \begin{pmatrix} \|A\| \\ w \end{pmatrix} = \begin{pmatrix} \|A\|^2 + \|w\|^2 \\ Bw \end{pmatrix}$$

is a vector of size at least $\|A\|^2 + \|w\|^2$. If $w \neq 0$, the norm of A would be strictly larger than $\|A\|$, which is a contradiction. Hence $w = 0$. By the induction hypothesis, this proves the statement. □

As an application of the Singular Value Decomposition we mention the following. If $AV = U\Sigma$ and $V = (v_1|\ldots|v_k), U = (u_1|\ldots|u_n)$, then $A = U\Sigma V^*$ can be rewritten as

$$A = \sum_{j=1}^{k} \sigma_j u_j v_j^* = \sigma_1(u_1 v_1^*) + \cdots + \sigma_k(u_k v_k^*),$$

showing that A can be written as the sum of rank-one matrices, each of the factors having norm one. It can be shown that if $\sigma_k \leq \cdots \leq \sigma_1$, then each truncated expansion A_p consisting of the first p terms only, satisfies

$$\text{for all } n \times k \text{ rank } p \text{ matrices } B_p\text{:} \quad \|A - A_p\| \leq \|A - B_p\|, \tag{1.71}$$

showing that A_p is the best rank p approximation of A.

Proposition 1.6.10 *Let A be an $n \times k$ matrix and $AV = U\Sigma$ a full SVD. Then we have that*

$$A^*AV = V\Sigma^*\Sigma, \tag{1.72}$$

diagonalizes A^*A, showing that the singular values of A are the square roots of the eigenvalues of A^*A.

Proof. Since the SVD is full, both U and V are square and their inverses equal to their complex conjugate transposed. Since $AV = U\Sigma$, we find

$$V^*A^* = \Sigma^*U^* \quad \text{and thus} \quad A^*U = V\Sigma^*. \tag{1.73}$$

Therefore, $A^*AV = A^*U\Sigma = V\Sigma^*\Sigma$, and $\Sigma^*\Sigma$ is a $k \times k$ matrix with the squares of the singular values on the diagonal. □

Chapter 2

The Eigenproblem

The algebraic eigenvalue problem is one of the most challenging computational problems in the field of Numerical Linear Algebra. Given a square matrix A, it asks to find all complex numbers $\lambda \in \mathcal{C}$ for which there exists a nonzero solution $v \in \mathcal{C}$ to the equation

$$Av = v\lambda. \tag{2.1}$$

The scalars λ for which such v exists are called the *eigenvalues* of A, and the corresponding v the *eigenvectors* belonging to λ. We will often interpret λ as a one by one matrix. Clearly, if $0 \neq v \in \mathcal{C}$ is a solution of (2.1), then so are all multiples of v. So alternatively, we may look for $\lambda \in \mathcal{C}$ such that $A - \lambda I$ vanishes on a non-trivial subspace $\mathcal{V}_\lambda \subset \mathcal{C}^n$. Notice that $\mathcal{V}_\lambda$ may have a dimension that is even larger than one. Notice also that if v is an eigenvector belonging to $\lambda \in \mathcal{C}$, then λ can be computed by the following expression,

$$v^*Av = v^*v\lambda. \tag{2.2}$$

This shows that the eigenvalue problem can be formulated purely as an *eigenvector* problem, in the sense that the nonzero roots of the function

$$f : \mathcal{C}^n \to \mathcal{C}^n : v \mapsto Avv^*v - vv^*Av \tag{2.3}$$

are all eigenvectors of A.[1] Indeed, if $f(v) = 0$ for $v \neq 0$, then defining λ by (2.2) yields an eigenpair. Conversely, if $Av = v\lambda$ for $v \neq 0$ then obviously $f(v) = 0$ because of (2.2). Therefore, in spite of the fact that the eigenvalue problem is classified as a problem from Linear Algebra, the problem itself is non-linear; the function f is a quadratic matrix function. Finding its roots can, in principle, be attempted by using for instance Newton's Method, and in fact, there exists a wide range of literature on this topic. However, as it will turn out here, there are other attractive alternatives.

2.1 Perturbation theory

Before we turn to computational methods for eigenvalue problems, it is important to know more about the specific properties of the problem that we aim to deal with. For instance, it is of interest to know what happens to the eigendata of A if we slightly change its entries. This will be the topic of this section on perturbation theory. First we motivate its relevance.

[1] Notice that under the assumption that $\|v\| = 1$, the function f slightly simplifies to $f(v) = (I - vv^*)Av$.

2.1.1 Motivation to study perturbation theory

The following beautiful but nonetheless trivial observation is at the heart of many important results in eigenvalue approximation.

Definition 2.1.1 (Eigenvalue residual) *Let A be an $n \times n$ matrix, v an arbitrary vector of unit length, and μ a given scalar. The residual r for the pair (v, μ) is defined as*

$$r = Av - v\mu. \tag{2.4}$$

Proposition 2.1.2 *Let A be an $n \times n$ matrix, v an arbitrary vector of unit length, and μ a given scalar. Then (v, μ) is an eigenpair of a $\hat{A} = A + E$ where $E = -rv^*$.*

Proof. We have that

$$\hat{A}v = (A + E)v = (A - rv^*)v = Av - r = Av - (Av - v\mu) = v\mu, \tag{2.5}$$

showing that (v, μ) is an eigenpair of $\hat{A}$. □

Notice that the perturbation $E = -rv^*$ in Proposition 2.1.2 has a very special structure, with corresponding special properties. It is a so-called *rank-one* matrix, satisfying:

- Ew is a multiple of r, the scalar factor being the inner product v^*w.
- The norm $\|E\|$ equals $\|r\|$. This is because $\|Ev\| = \|r\|\|v\|$ and,

$$\forall w, \quad \|Ew\| = \|rv^*w\| = \|r\||v^*w| \leq \|r\|\|v\|\|w\| = \|r\|\|w\|. \tag{2.6}$$

  Therefore,

$$\|E\| = \sup_{w \neq 0} \frac{\|Ew\|}{\|w\|} = \|r\|. \tag{2.7}$$

This motivates the study of the differences between the eigendata of A and matrices $A + E$, where $\|E\| \leq \varepsilon$ for some small $\varepsilon > 0$. This is because eigenvalue algorithms usually terminate as soon as the residual is smaller than some given tolerance ε. The computed eigenvalues are then eigenvalues of $A - rv^*$ instead of A.

A motivation to study in particular the eigenvalues of $A + E$ where $E = -rv^*$ is given in the following proposition. It states that there exists no perturbation F of A smaller than $E = -rv^*$ having (μ, v) as an eigenpair.

Proposition 2.1.3 *There exists no matrix F with $\|F\| < \|r\|$ such that $(A + F)v = v\mu$.*

Proof. Clearly, $(A + F)v = v\mu$ implies

$$\|Fv\| = \|Av - \mu v\| = \|r\|, \quad \text{hence} \quad \|F\| = \sup_{w \neq 0} \frac{\|Fw\|}{\|w\|} \geq \frac{\|Fv\|}{\|v\|} = \|r\|. \tag{2.8}$$

In fact, $E = -rv^*$ is the smallest perturbation of A having (v, μ) as eigenpair. □

With the above in mind, it can be seen that there exist two relevant but very different questions in eigenvalue approximation. The first one is to develop algorithms that make sure that they produce approximations with small residuals. If this is done, it depends on the properties of the matrix A whether this results in accurate eigenvalues.

2.1.2 Classical perturbation bounds

The perturbation properties of a matrix A depends strongly on whether A is real and symmetric, or complex Hermitian, possibly positive definite, normal or non-normal. The diagonalizable and non-normal matrices are the subject of two classical theorems in perturbation theory. In these theorems, the change in the eigenvalues of A in the complex plane under perturbations E is measured in terms of the distance of A to being normal or defective.

Theorem 2.1.4 (Bauer-Fike) *Suppose that $AV = V\Lambda$ with V nonsingular and Λ diagonal. Let μ be an eigenvalue of $A + E$. Then there exists an eigenvalue λ of A such that*

$$|\lambda - \mu| \leq \kappa(V)\|E\|, \quad \text{where} \quad \kappa(V) = \|V\|\|V^{-1}\|. \tag{2.9}$$

Proof. We may assume that μ is not an eigenvalue of A. Since $(A + E - \mu I)$ is singular, so are $V^{-1}(A - \mu I + E)V$ and $(\Lambda - \mu I)(I + (\Lambda - \mu I)^{-1}V^{-1}EV)$ and $I + (\Lambda - \mu I)^{-1}V^{-1}EV$. Each singular matrix $I + F$ satisfies $\|F\| \geq 1$ because there exists an x with $Fx = x$. Therefore, by multiplicativity of $\|\cdot\|$ we find,

$$1 \leq \|(\Lambda - \mu I)^{-1}V^{-1}EV\| \leq \kappa(V)\|(\Lambda - \mu I)^{-1}\|\|E\|, \tag{2.10}$$

and the norm of a diagonal matrix equals its entry furthest away from zero. □

Corollary 2.1.5 *Let A be a normal matrix, and μ an eigenvalue of $A + E$ with $\|E\| \leq \varepsilon$. Then there exists an eigenvalue λ of A such that*

$$|\lambda - \mu| \leq \varepsilon. \tag{2.11}$$

Proof. Follows immediately since $\kappa(V) = 1$. □

In order to prove the counterpart of the Bauer-Fike theorem for non-normal matrices, we first formulate a technical though useful lemma.

Lemma 2.1.6 *Let D and N be $n \times n$ matrices. Suppose D is nonsingular and diagonal and N strictly upper triangular. Let p be such that $N^p = 0$ but $N^{p-1} \neq 0$. Then $(DN)^p = 0$ and*

$$(D - N)^{-1} = \sum_{j=0}^{p-1} (D^{-1}N)^j D^{-1}. \tag{2.12}$$

Proof. Recall the formula $(I + F + \cdots + F^k)(I - F) = I - F^{k+1}$, valid for arbitrary F. Applying this with $F = D^{-1}N$ and $k = p - 1$ proves the statement. □

Theorem 2.1.7 (Henrici) *Let $AQ = QR$ be a Schur decomposition of A. Write $R = \Lambda + N$, with Λ diagonal and N strictly upper triangular, and let p be the smallest integer such that $N^p = 0$. Then for each eigenvalue μ of $A + E$ there exists an eigenvalue λ of A such that*

$$|\lambda - \mu| \leq \max(\theta, \theta^{\frac{1}{p}}), \quad \text{where} \quad \theta = \|E\| \sum_{k=0}^{p-1} \|N\|^k. \tag{2.13}$$

The number $\nu(A) = \|N\|$ is called A's departure from normality.

Proof. Since $A+E-\mu I$ is singular, and assuming that μ is not an eigenvalue of A we find that $(R-\mu I)(I+(R-\mu I)^{-1}Q^*EQ)$ and $I+(R-\mu I)^{-1}Q^*EQ$ are singular. Therefore,

$$1 \leq \|(R-\mu I)^{-1}Q^*EQ\| \leq \|(R-\mu I)^{-1}\|\|Q^*EQ\| = \|(\Lambda-\mu I+N)^{-1}\|\|E\|. \tag{2.14}$$

Apply Lemma 2.1.6 to the term $(\Lambda-\mu I+N)^{-1}$. This results in

$$\|(\Lambda-\mu I+N)^{-1}\| \leq \sum_{k=0}^{p-1}\|N\|^k\|(\Lambda-\mu I)^{-1}\|^{k+1} \leq \theta\max(\|(\Lambda-\mu I)^{-1}\|, \|(\Lambda-\mu I)^{-1}\|^p).$$

Using (2.14), and distinguishing the cases $\|(\Lambda-\mu I)^{-1}\|$ being either smaller or larger than one, we find the statement. □

The Bauer-Fike and the Henrici Theorems show that if a matrix A is non-normal or close to defective, the eigenvalues $A+E$ can differ significantly from the ones of A, even if $\|E\|$is small. Therefore, a small eigenvalue residual does not guarantee a good eigenvalue approximations.

2.2 Pseudo-eigenvalues

The perturbation theorems in the previous section produce upper bounds for the distance that eigenvalues may differ from their perturbed counterparts. Here we consider the exact subset $\Lambda_\varepsilon(A)$ of the complex plane around eigenvalues of A that eigenvalues can reach under perturbations of a given size $\varepsilon > 0$. Such a subset is called the ε-pseudospectrum of A.

2.2.1 Pseudo-eigenvalues and elementary properties

The eigenvalues of a matrix are points z in the complex plane where $A-zI$ is singular. It may be interesting to consider particular sections in the complex plane close to these singularities, which inspires the following definition.

Definition 2.2.1 *The ε-pseudospectrum $\Lambda_\varepsilon(A)$ of a A is the set of points z in the complex plane such that z is an eigenvalue of $A+E$ for some perturbation E with $\|E\| \leq \varepsilon$.*

This defines the exact bounds for how far an eigenvalue of A can go under perturbations E. It turns out that there are two useful equivalent characterizations of the ε-pseudospectrum. Both do not involve perturbations.

Proposition 2.2.2 *$\Lambda_\varepsilon(A)$ is the set of points $z \in \mathcal{C}$ for which*

- *$\sigma \leq \varepsilon$, where σ is the smallest singular value of $A-zI$,*
- *$\|(A-zI)^{-1}\| \geq \varepsilon^{-1}$.*

Proof. Assuming that z is not an eigenvalue of A, we get that $A+E-zI$ is singular, hence $(A-zI)(I+(A-zI)^{-1}E)$ and $(I+(A-zI)^{-1}E)$ are singular. Therefore,

$$1 \leq \|(A-zI)^{-1}E\| \leq \|(A-zI)^{-1}\|\|E\| \leq \varepsilon\|(A-zI)^{-1}\|, \tag{2.15}$$

which shows that the definition implies $\|(A-zI)^{-1}\| \geq \varepsilon^{-1}$. Conversely, assume the latter, then there exists a vector v for which

$$\|(A-zI)^{-1}v\| \geq \varepsilon^{-1}\|v\|. \tag{2.16}$$

Writing v as $v = (A - zI)w$ and multiplying both sides of the inequality by ε shows that there exists a vector w such that

$$\|(A - zI)w\| \leq \varepsilon \|w\|. \tag{2.17}$$

Hence, the pair $(u = w/\|w\|, z)$ approximates an eigenpair of A with residual r and $\|r\| \leq \varepsilon$. According to Proposition 2.1.2 we find that z is an eigenvalue of $A + E$ with $E = -ru^*$ and $\|E\| = \|r\| \leq \varepsilon$. This proves the equivalence of Definition 2.2.1 with the second characterization above. The equivalence between the first and the second is standard. □

A fairly trivial observation on the appearance of the set $\Lambda_\varepsilon(A)$ is formulated as follows. For this purpose we write $\mathcal{B}_\varepsilon(z)$ for the closed disc in the complex plane around z with radius ε.

Proposition 2.2.3 *For all A and $\varepsilon \geq 0$ we have that*

$$\bigcup_{\lambda \in \Lambda_0(A)} \mathcal{B}_\varepsilon(\lambda) \subset \Lambda_\varepsilon(A), \tag{2.18}$$

whereas equality holds if A is normal.

Proof. Let $z \in \mathcal{B}_\varepsilon(\lambda)$ be given, then with $E = (z - \lambda)I$ we have that $\|E\| = |z - \lambda| \leq \varepsilon$ and $A + E$ has eigenvalue z. If A is normal, Corollary 2.1.5 proves the reverse inclusion. □

Therefore, if A is normal, the pseudo-spectrum of a matrix is not interesting, for the simple reason that every set $\Lambda_\varepsilon(A)$ is completely determined by the magnitude of ε and $\lambda_0(A)$. For non-normal matrices however, the pseudospectrum $\Lambda_\varepsilon(A)$ need not be the union of discs.

Theorem 2.2.4 *Let (v, μ) be an approximate eigenpair of A, and assume that $\|v\| = 1$. Then $\mu \in \Lambda_\varepsilon(A)$ with $\varepsilon = \|r\|$.*

Proof. According to Proposition 2.1.2 we have that μ is an eigenvalue of $A + E$ with $E = -rv^*$, where $r = Av - v\mu$. Since $\|E\| = \|r\|$, we find that $\mu \in \Lambda_\varepsilon(A)$ with $\varepsilon = \|r\|$. □

This theorem, simple as it is, shows us relevant information on eigenvalue approximations obtained by some numerical method, regardless of the method employed. We end with a last observation, which involves optimization of the approximate eigenvalue once the approximate eigenvector has been fixed.

Theorem 2.2.5 *Let (v, μ) be an approximate eigenpair of A, and assume that $\|v\| = 1$. Let θ be the positive acute angle between r and v, then*

$$\hat{\mu} = v^* A v \in \Lambda_{\sin(\theta)\|r\|}(A). \tag{2.19}$$

Proof. By definition, $\hat{\mu}$ is such that $Av - v\hat{\mu} \perp v$, which means that $\hat{r} = Av - v\hat{\mu}$ is strictly smaller in size than $\|r\|$. Closer inspection shows that

$$\|\hat{r}\| = \sin(\theta)\|r\|. \tag{2.20}$$

The statement is now a simple corollary from Theorem 2.2.4. □

Chapter 3

Small Eigenproblems

We will discuss methods to compute eigenvalues of a moderately sized matrix by means of the QR-iteration. The Power Method will be derived as a by-product of the QR-iteration. This method computes only a single eigenpair of a given matrix, and is usually only employed if A is too large to apply the QR-iteration efficiently. Its convergence properties give much insight in the convergence of the QR-iteration, which will not be proved here.

3.1 The QR-iteration

The solution of small and medium size eigenvalue problems can be done using the QR-decomposition in an iterative fashion, which will result in the QR-iteration. First however we consider QR-decomposition of upper Hessenberg matrices.

3.1.1 QR-factorization of an upper Hessenberg matrix

There exists an important class of matrices that is almost upper triangular in the following sense, and for which the costs of computing a QR-decomposition is relatively low: upper Hessenberg matrices.

Definition 3.1.1 *A matrix H is called upper Hessenberg if it is zero below the diagonal that is directly below the main diagonal.*

For $n \times n$ upper Hessenberg matrices $H = (h_{ij})$, the QR-factorization can be computed far more cheaply then for general matrices. This is done by the following iterative procedure:

- Set $j = 1$, and repeat until $j = n - 1$,
- Let QR be the QR-factorization of the $2 \times n$ matrix $H(j : j + 1, :)$,
- Store the 2×2 matrix $Q_j = Q$ and overwrite $H(j : j + 1, :)$ by $R(j : j + 1, :)$.

To be precise, suppose that T is $j \times j$ and upper triangular, that E is $j \times (n - j)$, and K is $(n - j) \times (n - j)$ upper Hessenberg. Consider the $n \times n$ matrices

$$H = \left[\begin{array}{c|c} T & E \\ \hline 0 & K \end{array}\right], \quad \text{and} \quad Q_j = \left[\begin{array}{c|c|c} I_j & 0 & 0 \\ \hline 0 & Q^* & 0 \\ \hline 0 & 0 & I_{n-j-2} \end{array}\right] \tag{3.1}$$

where I_k is the $k \times k$ identity matrix. The 2×2 diagonal block Q^* of Q_j is the adjoint of the orthogonal factor Q of the QR-factorization of the top two rows of K. Notice that Q is already almost determined by the top 2×1 part of K, because it should be valid that

$$Q \begin{bmatrix} k_{11} \\ k_{21} \end{bmatrix} = \begin{bmatrix} \pm\sqrt{k_{11}^2 + k_{21}^2} \\ 0 \end{bmatrix}. \tag{3.2}$$

With this definition of Q_j, the product $Q_j H$ has the same block structure as H, but the upper triangular part has become one row and column larger in size.

Remark 3.1.2 Each of the matrices Q can be chosen as a rotation in the plane and therefore be represented by a single real number: the angle of rotation.

Proposition 3.1.3 *The computation of a QR-decomposition of an $n \times n$ upper Hessenberg matrix H can be done in $\mathcal{O}(n^2)$ flops.*

Proof. The computation of each 2×2 matrix Q costs only $\mathcal{O}(1)$ flops. Application of Q_j costs $\mathcal{O}(n-j)$ operations, since it acts on two rows with each only $n-j$ non-zero elements. Summing from $j = 1$ to $n-1$ gives the statement □

Exercise 3.1.4 Write a Matlab code that takes a possibly non-square upper Hessenberg matrix H as input, and that computes its QR-factorization according to the above strategy. The output should solely consist of the angles of rotation.

Exercise 3.1.5 Prove that the QR-factorization of a Hermitian tridiagonal matrix can be performed in $\mathcal{O}(n)$ flops. Write a Matlab code that implements this.

3.1.2 The basic QR-iteration

A very simple idea to compute approximation of eigenvalues of a matrix would be to do a *fixed point iteration* on the Schur factorization $AQ = QR$ as follows:

$$\text{Start with } Q_0 = I, \quad \text{and iterate} \quad AQ_n = Q_{n+1}R_{n+1}, \tag{3.3}$$

where the right-hand side $Q_{n+1}R_{n+1}$ is obtained by computing a QR-decomposition of the left-hand side AQ_n.

Notice that each fixed point of this iteration generates a Schur decomposition. This simple fact forms the heart of one of the most successful eigenvalue algorithms to compute the Schur decomposition matrix, the *QR-iteration*, which is a more advanced and efficient implementation of this idea.

Definition 3.1.6 (Basic QR-iteration) *The basic QR-iteration is a fixed point iteration applied to the Schur decomposition, using QR-decomposition at every iteration step.*

To be more precise, the QR-algorithm is usually presented as follows:

$$\text{Start with } A = \hat{Q}_1\hat{R}_1, \quad \text{iterate} \quad \hat{Q}_{n+1}\hat{R}_{n+1} = \hat{R}_n\hat{Q}_n. \tag{3.4}$$

In this form, the intuition behind it is less clear, but the algorithm is more robust and cheaper. Some manipulations give that (3.4) actually produces $\hat{Q}_{n+1}$ and $\hat{R}_{n+1}$ such that

$$A\hat{Q}_1 \cdots \hat{Q}_n = \hat{Q}_1 \cdots \hat{Q}_n\hat{Q}_{n+1}\hat{R}_{n+1}, \tag{3.5}$$

which shows that the transformation Q_{n+1} from (3.3) is generated as a product of transformations $\hat{Q}_j$, and that the upper triangular matrices are in principle equal for both iterations.

Exercise 3.1.7 Show that this is true, by comparing the first few iterates of both sequences. What is meant by "in principle"?

Another favourable aspect of formulation (3.4) is that the right hand side $\hat{R}_n\hat{Q}_n$ is spectrally equivalent to the original matrix A, which follows immediately from $\hat{R}_1\hat{Q}_1 = \hat{Q}_1^* A\hat{Q}_1$ and an induction argument. Moreover, if the algorithm converges, we have that $\hat{Q}_j \to I$, so $\hat{R}_j\hat{Q}_j \to R$, the triangular QR-factor of A.

Exercise 3.1.8 Apply the basic QR iteration (in either form) to a real matrix having complex conjugate eigenvalues and study the orthogonal and upper triangular matrices during the iteration.

3.1.3 Computational considerations

Last, but surely not least, we have the following properties, which are of great practical importance.

Proposition 3.1.9 *If H is an upper Hessenberg matrix, and $H = QR$ a QR-decomposition, then both Q and the reversed product RQ are also upper Hessenberg. The statement also holds with upper Hessenberg replaced by Hermitian tridiagonal.*

Therefore, if A is upper Hessenberg, each iteration step of (3.4) costs only $\mathcal{O}(n^2)$ flops instead of the usual $\mathcal{O}(n^3)$ for general matrices. In case A is Hermitian tridiagonal, each iteration costs even less, only $\mathcal{O}(n)$ flops. This is a considerable difference compared to the general case. However, notice that if at the end of the iteration the matrix Q is explicitly needed, it will cost $\mathcal{O}(n^2)$ flops to compute it from the rotation angles, even in the Hermitian setting. This can be explained by the fact that Q is a full matrix: it contains the eigenvector approximations as columns.

Remark 3.1.10 Consider iteration (3.3). If A is upper Hessenberg, then so is Q_1. However, AQ_n is not upper Hessenberg anymore, it has one more non-trivial subdiagonal. This becomes worse in every step. For theoretical analysis, this is of course not relevant.

The bottom line is that it always pays off to transform A to Hermitian tridiagonal form before starting the QR-iteration, or if this is not possible, to upper Hessenberg form.

Proposition 3.1.11 *For each square matrix A, there exists a matrix V such that $V^{-1}AV = H$ is upper Hessenberg. If $A = A^*$, the result H is tridiagonal. V can even be chosen orthogonal, and can be computed in $\mathcal{O}(n^3)$ flops by a direct (non-iterative) method.*

Proof. Compute the Arnoldi factorization $AV_k = V_{k+1}H_{k+1,k}$ until $k = n$. This produces an orthonormal basis V_n for $K^n(A, v_0) = \mathbb{R}^n$. By construction, A is upper Hessenberg with respect to this basis, and even tridiagonal is $A = A^*$. □.

We will now study the convergence of the QR iteration. We start doing this by considering only the first column of the orthogonal matrices Q_j.

3.2 The Power iteration

The Power iteration, or Power method, is a well-known method to compute the dominant eigenvector and its corresponding eigenvalue of a matrix for which it is computationally unattractive to perform the QR-iteration. The iteration is often explained and analyzed without any reference to the QR-iteration, because for some simple though instructive cases, convergence can easily be proved. Here, alternatively, we choose to emphasize this connection.

Proposition 3.2.1 *Consider the basic QR-iteration (3.3). The first j columns $Q_{k+1}^{(j)}$ of Q_{k+1} can be alternatively computed from partial iteration*

$$\text{Start with the first } j \text{ columns of } I =: Q_0^{(j)}, \quad \text{and iterate} \quad AQ_n^{(j)} = Q_{n+1}^{(j)} R_{n+1}^{(j)}, \tag{3.6}$$

where the right-hand side is computed by QR-factorization. Then also $R_{n+1}^{(j)}$ is the top left $j \times j$ part of R_{n+1}.

Proof. This follows from the simple fact that the first j columns of Q_{k+1} can be computed from the first j columns of AQ_k, and to compute the first j columns of AQ_k we only need the first j columns of Q_k. The statement now follows by induction. □

Even though this statement was not so difficult to prove, another directly related question is more difficult: if it converges, then whereto does it converge? Obviously, we then have an $n \times j$ orthogonal matrix $Q^{(j)}$ and a $j \times j$ upper triangular $R^{(j)}$ for which

$$AQ^{(j)} = Q^{(j)} R^{(j)}. \tag{3.7}$$

From this we see that the column span $\mathcal{V}$ of $Q^{(j)}$ is an invariant subspace for A, since it is mapped into itself by A. Consequently, the eigenvalues of $R^{(j)}$ must be eigenvalues of A. But which ones? We will answer this question for the simple case $j = 1$. The resulting iteration is called the Power iteration, and can be written as

$$\text{Start with a unit vector } q_0, \quad \text{and iterate} \quad Aq_k = q_{k+1} r_{k+1}, \tag{3.8}$$

in which the QR-decomposition is nothing more than scaling Aq_k to unit length. We will now prove convergence of the Power method in the most simple case, that of a real symmetric matrix A.

Theorem 3.2.2 *Let A be real symmetric and suppose that $AQ = Q\Lambda$, where $Q = (v_1|\dots|v_n)$ is orthogonal and Λ diagonal with diagonal entries $\lambda_1, \dots, \lambda_n$. Assume that $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Then, if $q_0^* v_1 \neq 0$ we get*

$$|\sin(\theta_k)| \leq |\tan(\theta_0)| \left|\frac{\lambda_2}{\lambda_1}\right|^k, \tag{3.9}$$

where $\theta_k \in [0, \pi/2[$ is the angle between q_k and v_1, and

$$||r_{k+1}| - |\lambda_1|| \leq ||\lambda_1| - |\lambda_n|| \tan^2(\theta_0) \left(\frac{\lambda_2}{\lambda_1}\right)^{2k}, \tag{3.10}$$

showing convergence of the Power iteration to the dominant eigenpair (v_1, λ_1).

Proof. By definition, q_k is a multiple of $A^k q_0$,which implies that

$$q_k = \frac{A^k q_0}{\|A^k q_0\|}. \tag{3.11}$$

Using this, it is not hard to see that

$$\sin^2(\theta_k) = 1 - \cos^2(\theta_k) = 1 - \left(v_1^T q_k\right)^2 = 1 - \left(\frac{v_1^* A^k q_0}{\|A^k q_0\|}\right)^2. \tag{3.12}$$

Since $q_0 = QQ^* q_0$ we have that $q_0 = Qy$ with $y = Q^* q_0$ and $\|y\| = 1$. This is nothing else than writing q_0 as a linear combination of the eigenvectors of A, as

$$q_0 = Qy = (v_1|\dots|v_n)\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = y_1 v_1 + \cdots + y_n v_n. \tag{3.13}$$

The eigendecomposition $AQ = Q\Lambda$ helps us find $A^k q_0 = Q\Lambda^k y$, which implies that $v_1^* A^k q_0 = y_1 \lambda_1^k$, and substituting this in(3.12) we conclude that

$$\sin^2(\theta_k) = 1 - \left(\frac{y_1 \lambda_1^k}{\|\Lambda^k y\|}\right)^2. \tag{3.14}$$

The following facts hold:

$$\|\Lambda^k y\|^2 = y_1^2 \lambda_1^{2k} + \sum_{j=2}^{n} y_j \lambda_j^{2k} \quad \text{and} \quad 0 < y_1^2 \lambda_1^{2k} \le \|\Lambda^k y\|^2 \quad \text{and} \quad \left|\frac{\lambda_j}{\lambda_1}\right| \le \left|\frac{\lambda_2}{\lambda_1}\right| \quad \text{for } j \ge 2.$$

These can be employed to bound

$$\sin^2(\theta_k) \le \frac{1}{y_1^2} \sum_{j=2}^{n} y_j^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2k} \le \frac{1}{y_1^2} \sum_{j=2}^{n} y_j^2 \left(\frac{\lambda_2}{\lambda_1}\right)^{2k} = \frac{1 - y_1^2}{y_1^2} \left(\frac{\lambda_2}{\lambda_1}\right)^{2k}. \tag{3.15}$$

Since $\cos^2(\theta_0) = y_1^2$, this proves the first statement. To prove the second statement we first observe that

$$r_{k+1}^2 = \|Aq_k\|^2 = \left(\frac{\|A^{k+1} q_0\|}{\|A^k q_0\|}\right), \tag{3.16}$$

where we made use of (3.11). Substituting $A^j q_0 = \Lambda^j y$ gives that

$$\left|\lambda_1^2 - r_{k+1}^2\right| = \left|\frac{\lambda_1^2 y^* \Lambda^{2k} y - y^* \Lambda^{2k+2} y}{y^* \Lambda^{2k} y}\right| = \left|\frac{y^* \Lambda^{2k}(\lambda_1^2 I - \Lambda^2) y}{y^* \Lambda^{2k} y}\right| \le \frac{\|\Lambda^{2k}(\lambda_1^2 I - \Lambda^2)\|}{y_1^2 \lambda_1^{2k}}. \tag{3.17}$$

Since $\Lambda^{2k}(\lambda_1^2 I - \Lambda^2)$ is diagonal we can easily compute its norm as the maximum over the entries $\lambda_j^{2k}(\lambda_1^2 - \lambda_j^2)$, which is bounded by $\lambda_2^{2k}(\lambda_1^2 - \lambda_n^2)$, and hence,

$$\left|\lambda_1^2 - r_{k+1}^2\right| \le \tan^2(\theta_0) \left|\frac{\lambda_2}{\lambda_1}\right|^{2k} (\lambda_1^2 - \lambda_n^2). \tag{3.18}$$

Dividing left and right hand sides by $|\lambda_1| + |r_k|$ results in

$$||\lambda_1| - |r_{k+1}|| \le \tan^2(\theta_0) \left|\frac{\lambda_2}{\lambda_1}\right|^{2k} (|\lambda_1| - |\lambda_n|), \tag{3.19}$$

since $|\lambda_1| + |r_{k+1}| \ge |\lambda_1| + |\lambda_n|$. This finishes the proof. □

Remark 3.2.3 Instead of studying the convergence of $|r_k|$ it is also possible to consider $\mu_k = q_k^* A q_k$ as approximation of the eigenvalue. This has as advantage that its sign will be correct.

Exercise 3.2.4 With $\mu_k = q_k^* A q_k$, prove along the lines of Theorem 3.2.2 that

$$|\mu_k - \lambda_1| \leq |\lambda_1 - \lambda_n| \tan^2(\theta_0) \left(\frac{\lambda_2}{\lambda_1}\right)^{2k}. \tag{3.20}$$

From the proof above, it becomes clear that symmetry of A is merely handy and not relevant for the proof. A closer analysis shows that for each diagonalizable matrix having an eigenvalue that is strictly separated from the others in magnitude, convergence to this eigenvalue takes place. So generically, the Power iteration produces approximations of the eigenvalue that is largest in size, and the speed of convergence depends on the relative magnitude of the second largest eigenvalue. This results immediately in the following corollary.

Corollary 3.2.5 *Let A be real symmetric. If the Power iteration (3.8) is applied to the matrix $(A - \mu I)^{-1}$ for some shift $\mu \in \mathbb{R}$, then the convergence takes place to the eigenvalue of A that is closest to the μ.*

Heuristically, it becomes also clear how the convergence of the QR-iteration can be explained. Consider for example the partial iteration (3.6) with $j = 2$. The first column of $Q_k^{(2)}$ converges to the eigenvector of A belonging to the dominant eigenvalue. The second column is kept orthogonal to this eigenvector approximation at all times. In the limit, it is the matrix A restricted to the orthogonal complement of v_1 that is being iteratively applied to the second column. The dominant eigenvalue of this restricted matrix is λ_2, hence convergence takes place to the eigenvector v_2 of A. Continuing like this, we would expect that the columns of Q_k in the full QR-iteration converge to the eigenvectors $v_1, \ldots, v_n$ respectively, at least, in case all eigenvalues are strictly separated in magnitude.

3.2.1 The QR-iteration with shifts

It would go to far to prove the convergence of the QR-algorithm it its most general form, but we will give some more intuition.

Proposition 3.2.6 *Let H be unreduced upper Hessenberg, which means that there are no zero elements on the first subdiagonal. If H is singular, and $H = QR$ is a QR-deomposition of H, then the bottom row of Q^*HQ is zero.*

Proof. Since H is unreduced, its singularity can only be caused by the last column being a linear combination of the others. This means that the column span of H is the column span if its first $n-1$ columns. Hence, if $H = QR$ is a QR-decomposition, then the last column of Q is not in the column span of H and with $R = (r_{ij})$ we must have that $r_{kk} = 0$. Therefore, $RQ = Q^*HQ$ has bottom row zero. □

Exercise 3.2.7 Show that all eigenvalues of an unreduced upper Hessenberg matrix have eigenspaces of dimension one.

The *QR-algorithm with shifts* exploits this idea to accelerate the original QR-algorithm.

Definition 3.2.8 (QR-iteration with shifts) *Let A be given. Transform A to upper Hessenberg form H_1. Then,*

$$\text{Start with } \hat{Q}_1\hat{R}_1 = H_1 - \mu_1 I, \quad \text{iterate} \quad \left\{ \begin{array}{ll} H_{n+1} & := \hat{R}_n\hat{Q}_n + \mu_n I \\ \hat{Q}_{n+1}\hat{R}_{n+1} & := H_{n+1} - \mu_{n+1} I \end{array} \right. , \tag{3.21}$$

where the $\mu_j \in \mathcal{C}$ can be chosen arbitrarily in each iteration step.

The intuition is, that each upper Hessenberg matrix H_n has the same eigenvalues as the original matrix H. If $H_n - \mu_n I$ is singular, then the last row of H_n equals μe_n^*, showing explicitly that H has an eigenvalue equal to μ. We could then proceed the QR-iteration with the $(n-1) \times (n-1)$ upper-left block of H, which is again an upper Hessenberg matrix, and which contains the remaining eigenvalues of H.

Based on a continuity argument, we might expect that if we use a shift that is close to an eigenvalue, then for the bottom row $e_n^* H_n$ of H_n we have

$$e_n^* H_n = \mu e_n^* + h^*, \tag{3.22}$$

where $h = \alpha e_{n-1} + \beta e_n$ has a relative small norm compared to $|\mu|$. This would make $\hat{\mu} = \mu + \beta$ an approximation of an eigenvalue, whereas based on the Gershgorin circle theorem, the size of α would indicate *how good* this approximation is. It may be a better approximation than μ is, in which case we could continue the QR-algorithm with next shift equal to $\hat{\mu}$.

Remark 3.2.9 As soon as a sub-diagonal element at some position $(j+1, j)$ of H_n is very small, the problem may be split in two by replacing this small element by zero, and applying separate QR-iterations the $j \times j$ and $(k-j-1) \times (k-j-1)$ diagonal blocks of H_n.

3.2.2 Real Schur Decomposition

If A is a real matrix with complex eigenvalues, then it is clearly impossible to iterate towards a Schur decomposition. Every matrix H_n that is encountered, will have real entries as long as real shifts are used. In that case, convergence will take place to the so-called *Real Schur Decomposition*

Theorem 3.2.10 (Real Schur Decomposition) *Each real $n \times n$ matrix A can be written as $A = QRQ^*$, where Q is real orthogonal and R is real quasi upper triangular, by which we mean that it is the sum of a real upper triangular matrix and a block diagonal matrix with block size at most two by two.*

Proof. Similar to the proof of the Schur Decomposition, with some extra technicalities. □

Remark 3.2.11 The eigenvalues of the two-by-two blocks of a quasi upper triangular matrix are exactly the complex conjugate eigenpairs of A. Writing $\lambda = \gamma + \mu i$ for one of the conjugates and $v = y + iz$ for a corresponding eigenvector, the real invariant subspace belonging to the conjugate pair is spanned by y and z and

$$A(y|z) = (y|z) \begin{bmatrix} \gamma & \mu \\ -\mu & \gamma \end{bmatrix}. \tag{3.23}$$

Like the complex Schur form, the real Schur form is not unique, in the sense that the diagonal elements and blocks can appear in any prescribed order.

Chapter 4

Subspace Methods

4.1 Selection

Let A be a non-singular $n \times n$ matrix, and $\mathcal{V} \subset \mathcal{C}^n$ a subspace with $\dim(\mathcal{V}) = k < n$. Let $\mathcal{W} = A\mathcal{V}$. Given a full rank matrix V with column span $\mathcal{V}$, set $W = AV$. We will investigate systematic approaches to find approximate eigenpairs of the eigenvalue problem $Az = z\lambda$ using nothing but the information contained in V and W. Notice that this information costs k matrix-vector multiplications with the matrix A.

We will discuss two strategies. The first one is demanding orthogonality of the eigenvalue residual to $\mathcal{V}$, whereas the second demands orthogonality to $\mathcal{W}$.

4.1.1 $\mathcal{V}$-orthogonal residuals: Ritz values and Ritz vectors

The first strategy is to find $v \in \mathcal{V}$ and $\theta \in \mathcal{C}$ such that the eigenvalue residual $r = Av - v\theta$ is orthogonal to $\mathcal{V}$. Recall that each $v \in \mathcal{V}$ can be written as $v = Vy$ for some $y \in \mathcal{C}^k$, in fact,

$$Vy = (v_1|\ldots|v_k)\begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = y_1 v_1 + \ldots y_k v_k.$$

Hence, demanding that $r \perp \mathcal{V}$ is equivalent to $r \perp Vy$ for all $y \in \mathcal{C}^k$, which can also be expressed as $r^*Vy = y^*V^*r = 0$ for all $y \in \mathcal{C}^k$. Since $V^*r \in \mathcal{C}^k$, this shows that

$$r \perp \mathcal{V} \Leftrightarrow V^*r = 0.$$

Combining the above, our strategy boils down to solving (y, θ) from $V^*(AVy - y\theta) = 0$, or, using the notation $W = AV$, from

$$V^*Wy = V^*Vy\theta \tag{4.1}$$

The pairs (Vy, θ) that result from the generalized eigenvalue problem (4.1) can be interpreted as approximate eigenpairs.

Definition 4.1.1 (Ritz data) *The approximate eigenpairs (Vy, θ) are called Ritz pairs of A in $\mathcal{V}$, consisting of Ritz values and Ritz vectors.*

Notice that the problem to solve in (4.1) is a generalized eigenvalue problem of size $k \times k$, and that in accordance with our aims, only the matrices V and W are needed to produce the approximations. In case $k << n$, the computational costs are low compared to those for the original problem.

4.1.2 $\mathcal{W}$-orthogonal residuals: Harmonic Ritz values and vectors

A second strategy is given in by the following symmetry-argument. If A is invertible, then the eigenvalue problem $Az = \lambda z$ can be formulated equivalently as $A^{-1}z = z\lambda^{-1}$. Since $V = A^{-1}W$, we may as well look for $w \in \mathcal{W}$ and scalars $\mu \in \mathcal{C}$ such that $A^{-1}w - w\mu \perp \mathcal{W}$. Since w can be written as Wy for some $y \in \mathcal{C}^k$, the orthogonality constraint can be rewritten as follows,

$$W^*Vy = W^*Wy\mu. \tag{4.2}$$

The pairs (Wy, μ) that result from the generalized eigenvalue problem (4.2) clearly are Ritz pairs for A^{-1} in $\mathcal{W}$.

Interestingly, the above is equivalent to the following strategy: find $v \in \mathcal{V}$ and $\mu \in \mathcal{C}$ such that $Av - v\mu^{-1} \perp \mathcal{W}$. Indeed, substituting $v = Vy$ and writing $I = A^{-1}A$ this transforms to $AVy - A^{-1}AVy\mu^{-1} \perp \mathcal{W}$. Multiplying by μ, substituting $W = AV$ and writing $w = Wy$, this leads us back to $w\mu - A^{-1}w \perp \mathcal{W}$. This shows that there are two equivalent interpretations:

- replace $A, \lambda, W = AV$ from the previous section by $A^{-1}, \lambda^{-1}, V = A^{-1}W$,
- instead of finding $v \in \mathcal{V}$ with $r \perp \mathcal{V}$, we look for $v \in \mathcal{V}$ with $r \perp \mathcal{W}$.

In the latter case however, it is the pair (Vy, μ) that is considered as an approximation of an eigenpair of A.

Definition 4.1.2 (Harmonic Ritz data) *The approximate eigenpairs (Vy, μ^{-1}) are called Harmonic Ritz pairs, consisting of Harmonic Ritz values and Harmonic Ritz vectors.*

The above does not depend on the basis V that is chosen for $\mathcal{V}$ and as long as $W = AV$, the approximations remain unchanged. Now, notice that:

- If $W^*W = I$, then (4.2) reduces to $W^*Vy = y\mu$.
- If $V^*V = I$, then (4.1) reduces to $V^*Wy = y\theta$.

It seems that μ and θ are each others complex conjugates, but this is due to a hidden abuse of notation. This abuse is the result of the fact that due to $W = AV$, it is not possible to assume that *both* V and W are orthogonal matrices.

4.1.3 Optimality properties for eigenvalue approximation

In case $A^* = A$ is positive definite, there are some optimality properties that we can prove. They are based on the following result.

Lemma 4.1.3 *Let $A = A^*$ be positive definite, with eigenvalues $\lambda_n \leq \cdots \leq \lambda_1$. If for all $0 \neq v \in \mathcal{V} \subset \mathbb{R}^n$ we have $\|Av\| > \lambda_k\|v\|$, then* $\dim(\mathcal{V}) < k$.

Proof. Let p be the largest subscript with $\lambda_k < \lambda_p$. Denote the span of the eigenvectors corresponding to $\lambda_1, \ldots, \lambda_p$ by $\mathcal{U}_p$. Since $\dim(\mathcal{U}_p) = p < k$, each $\mathcal{V} \subset \mathbb{R}^n$ with $\dim(\mathcal{V}) \geq k$

contains a vector $0 \neq v \perp \mathcal{U}_p$. For this vector v we clearly have $\|Av\| \leq \lambda_k \|v\|$. This proves the statement. □

Consider the $\mathcal{V}$-orthogonal residual selection principle. Assume that $V^*V = I$. Then the eigenpairs (y_i, θ_i) of V^*AV induce the $\mathcal{V}$-orthogonal residual approximations (Vy_i, θ_i) of the eigenpairs of A and

$$AVy_j = \theta_j Vy_j + r_j \quad \text{and} \quad r_j \perp \mathcal{V}. \tag{4.3}$$

Theorem 4.1.4 *The approximate eigenvalues $\theta_k \leq \cdots \leq \theta_1$ obtained by the $\mathcal{V}$-orthogonal residual approach satisfy $\theta_i \leq \lambda_i$ for all $i = 1, \ldots, k$.*

Proof. If $j \geq 1$ is such that $\theta_j > \lambda_j$, then $\lambda_j < \theta_j \leq \cdots \leq \theta_1$ and the span $\mathcal{V}_j$ of $Vy_1, \ldots, Vy_j$ is a j-dimensional subspace of $\mathbb{R}^n$. Using (4.3), we have for all $v \in \mathcal{V}_j$ that

$$Av = A \sum_{i=1}^{j} Vy_i \alpha_i = \sum_{i=1}^{j} (\theta_i Vy_i + r_i)\alpha_i = \sum_{i=1}^{j} \theta_i Vy_i \alpha_i + \sum_{i=1}^{j} r_i \alpha_i. \tag{4.4}$$

The summands at the right are mutually orthogonal, hence by Pythagoras' theorem we find

$$\|Av\|^2 \geq \sum_{i=1}^{j} \theta_i^2 \alpha_i^2 \geq \theta_j^2 \sum_{i=1}^{j} \alpha_i^2 = \theta_j^2 \|v\|^2 > \lambda_j^2 \|v\|^2. \tag{4.5}$$

Lemma 4.1.3 shows that $\dim(\mathcal{V}_j) < j$ which is a contradiction. Hence, there exists no $j \geq 1$ with $\theta_j > \lambda_j$. □

In words, Theorem 4.1.4 states that it is not possible to have j approximate eigenvalues that are all strictly larger than λ_j. We will now prove, by means of a trick that is not unusual in settings like this, that it is also not possible to have j approximate eigenvalues that are all strictly smaller than λ_{n-j+1}.

Lemma 4.1.5 *If $AVy_j = \theta_j Vy_j + r_j$ with $r_j \perp \mathcal{V}$ arise from $\mathcal{V}$-orthogonal aproximation of the eigendata of A, then the eigendata of $\alpha I - A$ give the $\mathcal{V}$-orthogonal approximations*

$$(\alpha I - A)Vy_j = (\alpha I - \theta_j)Vy_j - r_j \quad \text{and} \quad r_j \perp \mathcal{V}. \tag{4.6}$$

Proof. By inspection. □

Theorem 4.1.6 *The approximate eigenvalues $\theta_k \leq \cdots \leq \theta_1$ obtained by the $\mathcal{V}$-orthogonal residual approach satisfy $\lambda_{n+1-i} \leq \theta_{k+1-i}$ for all $i = 1, \ldots, k$.*

Proof. Let $\alpha > \lambda_1$ and define $H = \alpha I - A$. Then H is positive definite with eigenvalues $\alpha - \lambda_1 \leq \cdots \leq \alpha - \lambda_n$. Applying the $\mathcal{V}$-orthogonal residual selection strategy gives, according to Lemma 4.1.5 approximate eigenvalues $\alpha - \theta_1 \leq \cdots \leq \alpha - \theta_k$. Theorem 4.1.4 can be applied, which shows that

$$\forall j = 1, \ldots, k, \quad \alpha - \theta_{k+1-j} \leq \alpha - \lambda_{n+1-j}, \tag{4.7}$$

from which the statement follows immediately. □

The reason for introducing this little trick is, that even though Lemma 4.1.3 can be adjusted in a straightforward way, the proof of Theorem 4.1.4 cannot. It uses that Av is larger than its projection on $\mathcal{V}$, and here the word larger cannot be replaced by smaller in the reversed setting.

Corollary 4.1.7 *The approximate eigenvalue θ_j lies in $[\lambda_{n+1-j}, \lambda_j]$. In particular, if $\dim(\mathcal{V}) = n-1$, we have $\lambda_k \leq \theta_{k-1} \leq \lambda_{k-1}$ for all $k = 1, \ldots, k-1$.*

To conclude this section, we notice that for the $\mathcal{W}$-orthogonal selection principle, the similar statements can be derived for the inverses of the eigenvalues of A, using the interpretation of this method given in Section 4.1.2.

4.2 Expansion

We have now arrived at the second ingredient of subspace methods, which is expansion of the subspace. In the Section 5.2 we have investigated two selection principles from a given subspace of fixed dimension, and it seems reasonable to consider the situation that we are not satisfied with the approximations at hand, and we wish to improve them.

4.2.1 Arbitrary expansion

The most general setting is the one in which we expand the space $\mathcal{V}$ in the direction given by a unit vector v. Assuming that V has a basis for $\mathcal{V}$ as columns, the matrix $V_+ = (V|v)$ then spans the expanded space, which we will call $\mathcal{V}_+$. To be able to apply either of the methods from Section 5.2, we also need to compute $w = Av$, after which $W_+ = (W|w)$ spans the expanded space $\mathcal{W}_+$.

The matrices that are needed to select approximations from $\mathcal{V}_+$ according to either $\mathcal{V}_+$-orthogonal or $\mathcal{W}_+$-orthogonal residuals, are

$$V_+^* V_+ = (V|v)^*(V|v) = \left(\begin{array}{c|c} V^*V & V^*v \\ \hline v^*V & v^*v \end{array}\right), \tag{4.8}$$

which can be computed from V^*V by evaluating only V^*v and v^*v, its $\mathcal{W}_+$ counterpart (which we will not write out) and the matrix $V_+^* W_+$ given by

$$V_+^* W_+ = (V|v)^*(W|w) = \left(\begin{array}{c|c} V^*W & V^*w \\ \hline v^*W & v^*w \end{array}\right), \tag{4.9}$$

which can be formed from V^*W by computing additionally the vectors V^*w, v^*W and the scalar v^*w.

Proposition 4.2.1 *Let $A = A^*$ be positive definite, and let $\theta_k \leq \cdots \leq \theta_1$ be the $\mathcal{V}$-orthogonal residual approximations of the eigenvalues $\lambda_n \leq \cdots \leq \lambda_1$ of A, and $\theta_{k+1}^+ \leq \cdots \leq \theta_1^+$ the ones in $\mathcal{V}_+$. Then*

$$\theta_{k+1}^+ \leq \theta_k \leq \theta_k^+ \leq \cdots \leq \theta_2^+ \leq \theta_1 \leq \theta_1^+. \tag{4.10}$$

Proof. Assume that $V_+^* V_+ = I$. Write $M_+ = V_+^* A V_+$, and let $\mathcal{U}$ be the k-dimensional subspace of $\mathbb{R}^{k+1}$ of all vectors having last component equal to zero. Let U be the first k columns of the $(k+1) \times (k+1)$ identity matrix, then U has column span the subspace $\mathcal{U}$. Applying the $\mathcal{U}$-orthogonal residual selection to the matrix M gives

$$U^* M U = U^* V_+^* A V_+ U = V^* A V, \tag{4.11}$$

showing that the eigenvalues θ_j of V^*AV are the $\mathcal{U}$-orthogonal residual approximations of the eigenvalues θ_j^+ of M. Corollary 4.1.7 now proves the statement. □

Theorem 4.2.2 *Let $A = A^*$ be positive definite, and $\mathcal{V}_1 \subset \cdots \subset \mathcal{V}_{n-1} \subset \mathcal{V}_n = \mathbb{R}^n$ be a nested sequence of subspaces with $\dim(\mathcal{V}_j) = j$. Write $\theta_k^{(k)} \leq \cdots \leq \theta_1^{(k)}$ for the k eigenvalue approximations from $\mathcal{V}_k$. Then we have,*

$$\forall j, \quad \theta_j^{(j)} \leq \theta_j^{(j+1)} \leq \cdots \leq \theta_j^{(n)} = \lambda_j, \tag{4.12}$$

expressing that the sequence of j-th largest approximate eigenvalue converges monotonely to the j-th largest exact eigenvalue. Conversely,

$$\forall j, \quad \lambda_j = \theta_j^{(n)} \leq \theta_{j-1}^{(n-1)} \leq \cdots \leq \theta_1^{(n-j+1)}, \tag{4.13}$$

expressing that the sequence of j-th smallest approximate eigenvalue converges monotonely to the j-th smallest exact eigenvalue.

Proof. The proof of the two statements is an immediate consequence of Proposition 4.2.1, and can easiest be read from the following schematic expression:

$$\begin{array}{ccccccccc} & & & & \theta_1^{(1)} & & & & \\ & & & \theta_2^{(2)} & & \theta_1^{(2)} & & & \\ & & \theta_3^{(3)} & & \theta_2^{(3)} & & \theta_1^{(3)} & & \\ & \theta_4^{(4)} & & \theta_3^{(4)} & & \theta_2^{(4)} & & \theta_1^{(4)} & \\ \theta_5^{(5)} & & \theta_4^{(5)} & & \theta_3^{(5)} & & \theta_2^{(5)} & & \theta_1^{(5)} \end{array} \tag{4.14}$$

The approximate eigenvalues on line k lie between the ones on line $k+1$, which expresses Proposition 4.2.1. The convergence statements can now be formulated as

- The j-th left diagonal with positive slope decreases monotonely to λ_{n+1-j},
- The j-th right diagonal with negative slope increases monotonely to λ_j.

This clarifies the statements. □

4.2.2 Resulting algorithms for the eigenvalue problem

Consider the $\mathcal{V}$-orthogonal residual approach. By symmetry of formulation, everything said about the $\mathcal{V}$-orthogonal residual method can be easily translated in corresponding statements for the $\mathcal{W}$-orthogonal method.

In the $\mathcal{V}$-orthogonal residual approach, it is possible to force the approximating problem to be a standard eigenvalue problem by making sure that the matrix V is orthogonal at all times. This is what is sketched in the algorithm below.

In this general framework, we observe the following. The space $\mathcal{V}$ is expanded in some arbitrary direction v. This vector is orthonormalized against $\mathcal{V}$, and its image under A is stored in w. The matrix M is being updated as in (4.9), which we copy here again as

$$M_+ = (V|v)^*(W|w) = \left(\begin{array}{c|c} M & V^*w \\ \hline v^*W & v^*w \end{array}\right). \tag{4.15}$$

The eigenpairs of M_+ induce new eigenvalue approximations resulting from the $\mathcal{V}$-orthogonal residual approach applied to the expanded space $\mathcal{V}_+$. In the algorithm below, we assume diagonalizability of M with diagonalization $MY = Y\Lambda$. The resulting approximations for the eigenvectors are VY, and the residuals corresponding to those are the columns of the matrix $R = AVY - Y\Lambda$.

Remark 4.2.3 It is also possible to work with Schur Decomposition of the matrix M. In that case, we could replace the relevant lines by $MQ = QU$ and $R = AVQ - VQU$. The columns of R would then indicate the distance to a Schur Decomposition.

$V = v, \|v\| = 1, W = AV, M = W^*V, r = W - VM;$
for $j = 1, \ldots, n-1$ **do**
  $v = $ *expansion vector of* $\mathcal{V}$ *into* $\mathcal{V}_+$;
  $v = v - V(V^*v)$;
  $\gamma = \sqrt{v^*v}$;
  $v = \gamma^{-1}v$;
  $w = Av$;
  $M = [M, V^*w; v^*W, v^*w]$;
  $MY = Y\Lambda$;
  $R = AVY - VY\Lambda$;
  $V = (\, V|v\,)$;
  $W = (W|w)$;
end

By definition of $\mathcal{V}$ orthogonal methods, each residual r is orthogonal to $\mathcal{V}$. Therefore, expanding $\mathcal{V}$ with one of the residuals r we may skip the orthogonalization. Doing so from the beginning, this defines the sequence of subspaces uniquely once the initial vector $v = v_1$ has been chosen. We will prove this in the next section, and study the resulting algorithm in detail. It is called the *Arnoldi Method*, and reduces to the *Lanczos Method* if $A = A^*$.

4.3 The Arnoldi Method

The Arnoldi Method arises if in the $\mathcal{V}$-orthogonal residual method for the eigenvalue problem, we choose a current residual as expansion vector for $\mathcal{V}$. Since all residuals are multiples from each other if we expand like this from the beginning, this gives a unique way to build a sequence $(\mathcal{V}_j)_j$ of subspaces once the initial vector v_1 has been fixed.

4.3.1 Analyzing the first steps of the algorithm

In order to get a good understanding of the algorithm that results from our choice, we will first perform a few steps of it, right from the beginning. Afterwards, we will derive some important properties.

First step. Let an initial vector v_1 with $\|v_1\|$ be given, then $\mathcal{V}_1$ is the column span of the matrix V_1 having v_1 as single column. We set $w_1 = Av_1$ and define $\mathcal{W}_1$ as the column span of the matrix W_1 having w_1 as single column. Then we compute the first $\mathcal{V}$-orthogonal residual approximation as the number θ for which $r_1 = w_1 - \theta v_1 \perp \mathcal{V}$, which gives $\theta_1^{(1)} = v_1^* w_1$ as approximating eigenvalue, which is the one at the top (4.14) of the triangle, and v_1 as approximating eigenvector.

Second step. By definition, $r_1 \perp \mathcal{V}_1$, and according to our discussion above, we expand $\mathcal{V}_1$ in the direction of r_1. For this purpose, we define $v_2 = r_1/\|r_1\|$, and accordingly we set $V_2 = (V_1|v_2)$. Notice that due to

$$\|r_1\|v_2 = r_1 = Av_1 - v_1\theta_1^{(1)}, \tag{4.16}$$

we have the schematic relation

$$W_1 = AV_1 = V_2 \begin{pmatrix} * \\ * \end{pmatrix}. \tag{4.17}$$

We compute $w_2 = Av_2$ and set $W_2 = (W_1|w_2)$. This gives the following approximate eigenproblem to solve:

$$V_2^* W_2 y = y\theta, \tag{4.18}$$

where $V_2^*W_2$ is a 2×2 matrix. This results in two approximate eigenvalues $\theta_2^{(2)}$ and $\theta_1^{(2)}$, and corresponding approximate eigenvectors $V_2(y_1|y_2)$ with residuals

$$R = W_2(y_1|y_2) - V_2(y_1|y_2) \begin{pmatrix} \theta_1^{(2)} & 0 \\ 0 & \theta_1^{(2)} \end{pmatrix}. \tag{4.19}$$

A important observation is now that both residuals are multiples of one vector. This is because they are linear combinations of the columns of V_2 and W_2. Orthogonalization to $\mathcal{V}_2$ includes, according to (4.17), orthogonalization to w_1. Hence, at most one direction remains in which they can point.

Third step. The next step is to expand $\mathcal{V}_2$ in the direction of the residuals. As we just argued, we can just select one, say r, and define

$$\|r\|v_3 = r = Av_2 - V_2V_2^*Av_2. \tag{4.20}$$

This shows again that $w_2 = Av_2$ is a linear combination of $v_1, \ldots, v_3$, and we can expand the relation (4.17) as follows,

$$W_2 = AV_2 = V_3 \begin{pmatrix} * & * \\ * & * \\ 0 & * \end{pmatrix}. \tag{4.21}$$

The residuals are linear combinations of the columns of V_3 and W_3, and orthogonal to V_3 which includes, according to (4.21), the column space of $\mathcal{W}_2$. So again, there is only one direction in which the residual can point.

4.3.2 Arnoldi factorization and uniqueness properties

The algorithm that we have just studied is the Arnoldi method for approximation of eigenvalues. Its characteristics are the orthonormal basis V_k for the space $\mathcal{V}_k$ and the fact that W_k and V_k are related by means of multiplication with an upper Hessenberg matrix $H_{k+1,k}$ as follows,

$$W_k = AV_k = V_{k+1}H_{k+1,k}. \tag{4.22}$$

It is of importance to realize that the subspaces $\mathcal{V}_k$ are equal to the Krylov subspace $\mathcal{K}^k(A, v_1)$, which is an analogy with what happened in the linear system setting. Throughout this section we assume that Krylov subspaces $\mathcal{K}^j(\cdot, \cdot)$ have dimension j.

Proposition 4.3.1 *The column span $\mathcal{V}_k$ of the matrices V_k is uniquely defined by the initial vector v_1 and the matrix A and equal to $\mathcal{K}^k(A, v_1)$, where*

$$\mathcal{K}^k(A, v_1) = \text{span}\{v_1, Av_1, \ldots, A^{k-1}v_1\}.$$

Proof. By induction. For $k = 1$ the statement is trivially true. Now, suppose that V_k spans $\mathcal{K}^k(A, v_1)$. Then according to (4.22) we have that $A\mathcal{K}^k(A, v_1) \subset V_{k+1}$. Since the span of v_1 is also included in the column span van V_{k+1}, this proves the statement □

Definition 4.3.2 (Flag) *We say that an $n \times k$ matrix V is a flag for $\mathcal{K}^k(A, v_1)$ if for all $j \in \{1, \ldots, k\}$, the first j columns of V span $\mathcal{K}^j(A, v_1)$.*

By our assumption that Krylov spaces have full dimension, it is clear that if V and W are flags for the same Krylov space, then $V = WR$ with R upper triangular and non-singular. An immediate consequence is the following.

Corollary 4.3.3 *Modulo the sign of its columns, the matrix V_k produced by the Arnoldi method is the only $n \times k$ orthogonal flag for $\mathcal{K}^k(A, v_1)$.*

Proof. If Q and U are orthogonal flags for $\mathcal{K}^k(A, v_1)$ then $Q = UR$ with R upper triangular and orthogonal. Hence, R is diagonal with entries equal to plus and minus one only. □

Definition 4.3.4 (Arnoldi factorization) *The relation $AV = (V|v)H$, where $(V|v)$ is an orthogonal $n \times (k + 1)$ flag for $\mathcal{K}^k(A, Ve_1)$ is called an Arnoldi factorization.*

An important result is that if $AV = (V|v)H$ holds for orthogonal matrices V and $(V|v)$ and H is upper Hessenberg, then it is an Arnoldi factorization, and hence $(V|v)$ spans $\mathcal{K}^{k+1}(A, Ve_1)$. It is formulated in Theorem 4.3.6 below.

Lemma 4.3.5 *If $(V|v)$ is $n \times (k + 1)$ non-singular and $AV = (V|v)H$ with H upper Hessenberg, then $(V|v)$ is a flag for $\mathcal{K}^{k+1}(A, v_1)$.*

Proof. For $j = 1$ the statement is trivial. Assume that $(v_1|\ldots|v_j)$ is a flag for $\mathcal{K}^j(A, v_1)$. By (1) we find that $v_{j+1} \notin \mathcal{K}^j(A, v_1)$ and by (2) that $v_{j+1} \in A\mathcal{K}^j(A, v_1) \subset \mathcal{K}^{j+1}(A, v_1)$. Hence, $(v_1|\ldots|v_{j+1})$ is a flag for $\mathcal{K}^{j+1}(A, v_1)$. Induction completes the proof. □

Theorem 4.3.6 *If $(V|v)$ is orthogonal and $AV = (V|v)H$ with H upper Hessenberg, then $AV = (V|v)H$ is an Arnoldi factorization.*

Proof. Follows immediately from Corollary 4.3.3 and Lemma 4.3.5. □

4.3.3 Alternative implementation of the Arnoldi method

Returning to the approximate eigenvalue problem, we find that the approximate eigenvalues, the Ritz values, are the eigenvalues of

$$V_k^* W_k = V_k^* V_{k+1} H_{k+1,k} = H_{k,k}, \tag{4.23}$$

where $H_{k,k}$ consists of the top k rows of $H_{k+1,k}$. We recall the following facts,

- Each matrix $H_{j,j-1}$ is the top left block of $H_{j+1,j}$,
- Applying the QR-iteration to upper Hessenberg matrices is relatively inexpensive.
- If $H_{k,k}$ is unreduced, each eigenspace has dimension one.

We can now present an updated version of the algorithm. In this version, we refrain from storing the matrix W_k, since according to (4.22) it is sufficient to store $H_{k+1,k}$ and V_{k+1}. Moreover, instead of expanding with the residual, we expand by Av_k. By Proposition 4.3.1, this does not change the subspaces $\mathcal{V}_k$. It requires the orthogonalization of Av_k against the columns of V_k, but it is not necessary anymore to compute the residuals.

Even if we do not compute the residuals, we can still find out for which k the matrix $H_{k,k}$ has eigenvalues with small residuals. Indeed, since

$$r_k = AV_k y_k - V_k y_k \theta = h_{k+1,k} v_{k+1} e_k^* y_k, \tag{4.24}$$

and using that $\|y_k\| = \|e_k\| = 1$, we see that

$$\|r_k\| = |h_{k+1,k} e_k^* y_k| \leq |h_{k+1,k}|, \tag{4.25}$$

where $h_{k+1,k}$ the entry bottom right of $H_{k+1,k}$. Therefore, the computation of the approximate eigenvalues can be delayed until this entry is small enough.

$V = v, \|v\| = 1, H = [], p = \infty;$
while $\gamma > \varepsilon$ **do**
    $v = Av;$
    $h = V^* v;$
    $v = v - Vh;$
    $\gamma = \sqrt{v^* v};$
    $H = [H, h; 0, \gamma];$
    $V = (V|v/\gamma);$
end

Finally, if $A = A^*$, we know on beforehand that $H_{k,k}$ is tridiagonal. Hence, Av_k only needs to be orthogonalized to v_{k-1} and v_{k-2}. The resulting method is the *Lanczos Method*

Unfortunately, in the Arnoldi method, the amount of computational work increases with the dimension of the subspace. Moreover, it may not give approximations of the eigenvalues in which you are interested. To cure this, we introduce the concept of the *implicit restart* in the next section.

To conclude, notice the remarkable fact that the few lines of code above comprise such a richness of mathematical ideas and algorithmic subtleties.

4.4 Implicit restart of the Arnoldi method

In the previous section, we have derived the Arnoldi method. The Arnoldi method is the method that arises if $\mathcal{V}$-orthogonal residual selection is applied to the nested sequence of subspaces $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \ldots$, where each $\mathcal{V}_{j+1}$ is defined by expansion of $\mathcal{V}_j$ with the unique direction in which all residuals $r_j \perp \mathcal{V}_j$ point. This expansion can be implemented in two mathematically equivalent ways:

- Compute the Ritz pairs and their residuals for every k and expand with one of the residuals $r_j, j \in \{1, \ldots, k\}$. Orthogonalization of r_j to $\mathcal{V}_k$ is then for free.
- Expand $\mathcal{V}_k$ with Av_k. Then orthogonalization is not for free, but it saves the computation of the Ritz pairs. The size of the residual is bounded by $|h_{k+1,k}|$.

Both implementations give, in exact arithmetic, the same Ritz pairs. Each has its advantages and disadvantages.

4.4.1 The influence of the initial vector

Clearly, the Ritz pairs that are produced in step k of the Arnoldi method depend only on the start vector v_1. In the general case, it is not possible to give a full analysis of their approximation properties. If however $A = A^*$, the convergence theory is well understood, and combines the theory in Section 4.1.3 with the particular choice $\mathcal{V}_k = \mathcal{K}^k(A, v_1)$. We refer to [4] and the references therein for an excellent treatment of the sometimes rather technical material.

A few simple statements can nevertheless be proved for the general non-hermitian case. For instance, if v_1 lies in an invariant subspace for A, then the corresponding exact eigenvalues are found.

Proposition 4.4.1 *Let $A(q_1|\ldots|q_n) = (q_1|\ldots|q_n)R$ be a Schur decomposition of A and write $Q_k = (q_1|\ldots|q_k)$. Then, if*

$$v_1 = Q_k y \quad \text{for some } y \in \mathbb{R}^k, \tag{4.26}$$

then there exists an $m \leq k$ such that the Arnoldi method terminates by division by zero after m steps, producing m eigenvalues of A.

Proof. Since $v_1 \in \text{colspan}(Q_k)$, which is an invariant subspace for A of dimension k, also $\mathcal{K}^k(A, v_1)$ is an invariant subspace, of some dimension $m \leq k$. Applying the $\mathcal{V}$-orthogonal residual selection with an m-dimensional invariant subspace gives m Ritz pairs that are equal to exact eigenpairs of A. □

By a continuity argument, one may expect that if the start vector v_1 is close to an invariant subspace, the eigenvalues beloning to this invariant subspace will be approximated sooner and better by the Arnoldi method than other eigenvalues of A.

Unfortunately, this will only become visible after an investment of computational effort. An important question is what to do if the Arnoldi method seems to produce approximations of eigenvalues in which there is no interest.

4.4.2 Restarting the algorithm with a different start vector

Suppose that we have done $k > p$ steps of the Arnoldi method with start vector v_1, resulting in

$$AV_k = V_{k+1}H_{k+1,k}. \tag{4.27}$$

This gives us k Ritz values, the eigenvalues of $H_{k,k}$. The idea is to divide those into two groups: those we find uninteresting, say $\mu_1, \cdots, \mu_\ell$, because they are relatively far away from

some target $\tau \in \mathcal{C}$, and the remaining ones $\mu_{\ell+1}, \cdots, \mu_k$. Compute

$$\hat{v_1} = \frac{\tilde{v_1}}{\|\tilde{v_1}\|}, \quad \text{where} \quad \tilde{v_1} = \Pi_{j=1}^{\ell}(A - \mu_j I)v_1, \tag{4.28}$$

and start the algorithm again but now with $\hat{v_1}$ as startvector. The hope is that $\hat{v}_1$ will be less close to the invariant subspace belonging to the eigenvalues that are approximated by $\mu_1, \cdots, \mu_\ell$, and closer to the one belonging to $\mu_{\ell+1}, \cdots, \mu_k$.

For ease of presentation, we consider the case $\ell = 1$. The other cases can be covered simply by repeating the below arguments ℓ times. We will write $\mu = \mu_1$.

Observation 4.4.2 *Using the Arnoldi factorization (4.27) we have that*

$$\hat{v}_1 = Av_1 - \mu v_1 = AV_k e_1 - \mu v_1 = V_{k+1}H_{k+1,k}e_1 - \mu v_1 = (h_{11} - \mu)v_1 + h_{21}v_2. \tag{4.29}$$

Therefore, to compute $\hat{v}_1$, no matrix multiplications with A are needed.

Now notice that since $\hat{v_1} \in K^1(A, v_1)$, we have that that

$$K^k(A, \hat{v_1}) \subset K^{k+1}(A, v_1). \tag{4.30}$$

Because of this, it is possible to recover this subspace, together with an orthonormal flag for it without computing any additional matrix vector products with A.

Theorem 4.4.3 *Let $H_{k+1,k} - \mu I_{k+1,k} = Q_{k+1,k}R_{k,k}$ denote the QR-decomposition of its left-hand side. Define W_j as the first j columns of W_k, where*

$$W_k = V_{k+1}Q_{k+1,k}. \tag{4.31}$$

Then W_k is the orthogonal flag for $\mathcal{K}^k(A, \hat{v}_1)$. Moreover,

$$AW_{k-1} = W_k\hat{H}_{k,k-1} \quad \text{where} \quad \hat{H}_{k,k-1} = R_{k,k}Q_{k,k-1} + \mu I_{k,k-1}, \tag{4.32}$$

is the corresponding Arnoldi factorization.

Proof. Because $Q_{k+1,k}$ is upper Hessenberg, so is $\hat{H}_{k,k-1}$. Furthermore, since $Q_{k+1,k}$ is upper Hessenberg, we also have the somewhat surprising equality

$$I_{k+1,k}Q_{k,k-1} = Q_{k+1,k}I_{k,k-1}. \tag{4.33}$$

Using this, it can easily be verified using the various definitions above, that

$$\begin{aligned} AW_{k-1} = AV_kQ_{k,k-1} = V_{k+1}H_{k+1,k}Q_{k,k-1} = V_{k+1}(\mu I_{k+1,k} + Q_{k+1,k}R_{k,k})Q_{k,k-1} \\ = V_{k+1}Q_{k+1,k}(\mu I_{k,k-1} + R_{k,k}Q_{k,k-1}) = W_k\hat{H}_{k,k-1}. \end{aligned} \tag{4.34}$$

Since W_k is orthogonal, Theorem 4.3.6 yields that (4.32) is an Arnoldi decomposition. The equality

$$W_k e_1 = V_k Q_{k+1,k}\frac{R_{k,k}e_1}{r_{11}} = \frac{1}{r_{11}}V_k(H_k - \mu I)e_1 = \frac{1}{r_{11}}(AV_k - \mu V_k)e_1 = \hat{v}_1 \tag{4.35}$$

shows that W_k is the orthogonal flag for $\mathcal{K}^k(A, \hat{v}_1)$. □

The conclusion is that without performing any computations in which the matrix A is involved, it is possible to recover the Arnoldi factorization (4.32) from (4.27).

Chapter 5

Recent Developments

The formulation of eigenproblems as generalized algebraic Riccati equations removes the non-uniqueness problem of eigenvectors. This basic idea gave birth to the Jacobi-Davidson (JD) method of Sleijpen and Van der Vorst (1996). JD converges quadratically when the current iterate is close enough to the solution that one targets for. Unfortunately, it may take quite some effort to get close enough to this solution. Here we present a remedy for this. Instead of linearizing the Riccati equation (which is done in JD) and replacing the linearization by a low-dimensional linear system, we propose to replace the Riccati equation by a low-dimensional Riccati equation and to solve it exactly. The performance of the resulting *Riccati algorithm* compares favorable to JD while the extra costs per iteration compared to JD are in fact negligible.

5.1 Introduction

The standard eigenvalue problem $Ax = \lambda x$ for possibly non-Hermitian matrices A is one of the basic building blocks in computational sciences. Solution methods, which are necessarily iterative, range from the QR algorithm (if all eigenvalues are wanted) via the Arnoldi [1] method to Jacobi-Davidson [6] for large problems from which only few eigenpairs are needed. As a matter of fact, both the Arnoldi method and the Jacobi-Davidson method can be derived as Ritz-Galerkin projection methods that use subspaces of growing dimension, and in which the expansion of the subspace is governed by adding approximations of solution(s) of a generalized algebraic Riccati equation. We will show this in Section 5.2, and then discuss in Section 5.3 a third, very natural method based on the same idea. This method was briefly introduced in [2]. In Section 5.4, convincing numerical evidence of the success of the new approach is given, using matrices from the Matrix Market test-collection.

5.2 Projection on expanding subspaces

A straightforward tool to tackle the eigenvalue problem $Ax = \lambda x$ in $\mathcal{R}^n$ is to project it on a k dimensional subspace $\mathcal{V}$ of $\mathcal{R}^n$ with $k << n$. By this we mean the following. Assume that V is an $n \times k$ matrix of which the columns span $\mathcal{V}$. Approximations of eigenpairs of A can be found in $\mathcal{V}$ by computing vectors $v \in \mathcal{V}$ such that, instead of Av itself, the *orthogonal projection* $P_{\mathcal{V}}Av$ of Av onto $\mathcal{V}$ is a (scalar) multiple of v. The condition $v \in \mathcal{V}$ can be expressed as $\exists y \in \mathcal{R}^k : v = Vy$, whereas the condition $w \perp \mathcal{V}$ translates to $V^*w = 0$, i.e., w is orthogonal

to each of the columns of V and hence to their span $\mathcal{V}$. So, the problem to solve, usually called the *projected problem*, becomes

$$\text{find } y \in \mathcal{R}^k \text{ and } \mu \in \mathcal{R} \text{ such that } V^*(AVy - \mu Vy) = 0. \tag{5.1}$$

Note that both V^*AV and V^*V are small $k \times k$ matrices, and that when the columns of V are an *orthonormal basis* for $\mathcal{V}$, then V^*V is the identity matrix, hence (5.1) a small standard eigenvalue problem, which can be solved routinely by for instance the QR algorithm. The couples (μ, Vy) can be interpreted as *approximate eigenpairs* for A, and are usually called *Ritz-pairs*. If the *residuals* $r = AVy - \mu Vy$, which can easily be computed *a posteriori*, are not small enough according to the wishes of the user, expansion of the subspace, followed by solving a new projected problem, may lead to improvement.

5.2.1 Algorithm of the general framework

The following pseudo-code presents a general framework for this approach. In each execution of the while-loop, the subspace dimension increases by one, in some direction q, to be specified below.

input: A, V, ε;
$W = AV$;
$M = V^*W$; *initial projected matrix; note that* $M = V^*AV$
$r = s =$ *residual of projected problem*;
while $\|r\| > \varepsilon\|s\|$
  q = *expansion vector for* $\mathcal{V}$;
  v = *with span* $(V|v)$ *equal to span* $(V|q)$,
    and $(V|v)$ *is orthonormal;*
  $w = Av$;
  $M = \left(\begin{array}{c|c} M & V^*w \\ \hline v^*W & v^*w \end{array}\right)$ *efficient implementation of projection*
    $M = (V|v)^*(W|w)$ *using previous* M;
  $V = (V|v)$ *expansion of* $\mathcal{V}$;
  $W = (W|w)$ *expansion of* $\mathcal{W}$;
  r = *residual of the new problem*;
end (while)

5.2.2 Expansion strategies

We will now discuss expansion strategies, i.e., how to choose q such that projection on the expanded space may yield better approximate eigenpairs. For this, suppose that for the eigenvalue problem $Ax = \lambda x$ we have an approximate eigenvector v with $\|v\| = 1$ that was obtained by projection of the problem on a subspace $\mathcal{V}$. Consider the affine variety

$$v^\perp = \{x + v | x^*v = 0\}. \tag{5.2}$$

Generally, there will be n points v_j in $v^\perp$ corresponding to eigenvectors of the matrix A, which are the intersections of lines (eigenvector directions) through the origin, with the affine variety $v^\perp$. Each of those points, obviously, can be written as $v_j = v + q_j$, with $q_j^*v = 0$. Writing $\mu = v^*Av$ and $r = Av - \mu v$, it is not hard to show that the vectors q_j are the roots of the following generalized algebraic Riccati equation in q,

$$q^*v = 0 \quad \text{and} \quad (I - vv^*)Aq - q\mu = q(v^*A)q - r. \tag{5.3}$$

If we intend to use approximations to solutions q_j of this equation to expand the subspace $\mathcal{V}$, we should find a balance between the effort spent on computing such approximations, and what we get back in terms of improved approximations from the bigger space. Note that one of the crudest (hence cheapest) approximations would result from replacing A by the identity I, which gives $\hat{q} = -(1-\mu)^{-1}r$ as approximate root q. The resulting algorithm is in fact the Arnoldi method. Note that the orthogonalization step in the while loop above becomes then superfluous, making the method even less numerically expensive.

5.2.3 The Jacobi-Davidson Method

The Jacobi-Davidson method [6] results, when the Riccati equation (5.3) is linearized around $q = 0$. The linearized equation

$$\hat{q}^* v = 0 \quad \text{and} \quad (I - vv^*)A\hat{q} - \hat{q}\mu = -r \tag{5.4}$$

is, in turn, usually only solved approximately. The approximate solution of (5.4) is then used to expand the subspace $\mathcal{V}$, and a new approximate eigenvector v is extracted from the expanded space by projection, and the process repeated. Equation (5.4) can be approximated by projection as well, say on a ℓ-dimensional subspace $\mathcal{U}$ with and $r \in \mathcal{U}$ and $\mathcal{U} \perp v$. Note that the latter requirement assures that an approximation in $v^\perp$ is found. If U is a matrix whose orthonormal columns span $\mathcal{U}$, then the projected equation would be

$$U^* A U \hat{z} - \hat{z}\mu = -U^* r. \tag{5.5}$$

For $\ell = 1$, this gives the Arnoldi method again. Using higher values for ℓ results in a structurally different method, whereas solving (5.4) exactly is the "full" Jacobi-Davidson method of which can be proved that asymptotically, it converges quadratically. Therefore, much attention has been paid to finding good preconditioners to solve (5.4) in high precision with little effort.

5.2.4 Stagnation due to unjust linearization

Perhaps the most important observation is, that due to the linearization, the effort in solving (5.4) accurately is only worthwhile if the quadratic term in (5.3) could really be neglected, which is only the case if there is a solution q of (5.3) with $\|q\|$ small enough, i.e., close enough to zero. Thus, v needs to be a rather good approximation of an eigenvector. It has been quantified in [3] that this is the case if

$$\sigma^2 - 12\|r\|\|v^* A\| > 0, \tag{5.6}$$

where σ is the smallest singular value of A projected on $v^\perp$. It has moreover been observed in experiments that this condition seems necessary, and also that it can be very restrictive. This explains why the Jacobi-Davidson method shows an initial phase of no structural residual reduction, before it plunges into the quadratically convergence region. Especially if the start-vector for Jacobi-Davidson is chosen badly (for instance, randomly), this initial phase can be very long and hence very expensive.

5.3 Curing the stagnation: The Riccati method

The phase in the Jacobi-Davidson method in which the current eigenvector approximation is still outside the quadratic convergence region can be significantly reduced, as was firstly observed in Section 4.2 of [2]. There we proposed to project (5.3) directly on an ℓ-dimensional subspace $\mathcal{U}$, and to observe that if ℓ is moderate, we can refrain from linearizing the resulting ℓ-dimensional nonlinear equation, and compute *all* its roots instead.

5.3.1 Main idea

Consider the Jacobi-Davidson method. Suppose that in solving (5.4), projection on a subspace $\mathcal{U}$ with $r \in \mathcal{U}$ and $\mathcal{U} \perp v$ is used, as suggested below equation (5.4). As before, let U contain an orthonormal basis for $\mathcal{U}$. Then the projected linearized equation reads as

$$U^*AU\hat{z} - \mu\hat{z} = -U^*r, \tag{5.7}$$

after which $U\hat{z}$ is the approximation of $\hat{q}$ with which $\mathcal{V}$ is going to be expanded. Alternatively, we could also have used the subspace $\mathcal{U}$ to approximate the Riccati equation (5.3) directly, without linearization, yielding the ℓ-dimensional projected Riccati equation

$$U^*AUz - \mu z = z(v^*AU)z - U^*r. \tag{5.8}$$

In fact, (5.7) is the linearization around $z = 0$ of (5.8), so it seems as if no progress has been made, apart from the fact that linearization and projection commute. There is, however, an important difference, which is that there is no need anymore to linearize the low-dimensional Riccati equation. It can be solved exactly by realizing that it is equivalent to the $(\ell+1)\times(\ell+1)$ eigenvalue problem

$$\left(\begin{array}{c|c} \mu & v^*AU \\ \hline U^*r & U^*AU \end{array}\right)\left(\begin{array}{c} 1 \\ z \end{array}\right) = (\mu + v^*AUz)\left(\begin{array}{c} 1 \\ z \end{array}\right). \tag{5.9}$$

The gain in comparison to Jacobi-Davidson is, that instead of obtaining an approximation of a unique correction $\hat{q}$ of (5.4), we get a small number $\ell+1$ approximations $\tilde{q}_j$ of the solutions q of (5.3). This gives extra freedom in deciding how to expand the space $\mathcal{V}$, for example, by computing the Ritz values corresponding to the vectors $v + \tilde{q}_j, j \in \{1, \ldots, \ell+1\}$ and using the $\tilde{q}_j$ that gives a Ritz value closest to a given target to expand $\mathcal{V}$. We will call the resulting method the *Riccati method* if $\mathcal{U}$ is chosen as the ℓ-dimensional Krylov subspace for r and $(I - vv^*)A$. Note that the presence of the projection $(I - vv^*)$ assures that this Krylov subspace $\mathcal{U}$ will be orthogonal to v.

5.3.2 Discussion

For moderately small $\ell << n$, the costs for solving the eigenvalue problem (5.9) are negligible compared to the ℓ matrix-vector multiplications with A that are needed to construct the projected matrix U^*AU, which is needed in both JD (5.7) and Riccati (5.9). This shows that the Riccati method is only slightly more expensive per iteration than JD. By one iteration we mean expansion of $\mathcal{V}$ with a new vector q.

If v is very close to an eigenvector, then there exists a solution q of (5.3) with small norm. If there exists a solution z of (5.8) such that Uz is very close to q, then the solution $\hat{z}$ of

(5.7) will yield an accurate approximation $U\hat{z}$ of q, since the quadratic term in z will then be negligible. Hence, JD would give a good approximation of this eigenvector. In the Riccati method, we would have the *option* to expand the space in (almost) the same way as in JD, resulting in similarly good approximations. In this case, the advantage of Riccati is, that if we are not interested in this particular eigenvector but in another one, we can refrain from expansion with the JD correction and point towards a different direction. This shows that it is unlikely that JD will outperform Riccati: if the JD correction $\hat{z}$ is good, then one of the possible corrections in the Riccati method will be very close to $\hat{z}$.

In case v is *not* a good approximation of the eigenvector in which one is interested, the Jacobi-Davidson correction $\hat{q}$ may not make sense in the same way as a step of the Newton method may not make sense far away from the objective solution. Since in the Riccati method there are other possible expansion vectors (also called corrections) to choose from, this may lead to substantial improvement.

In solving practical problems with the Jacobi-Davidson method, it is not unusual that the expansion vector $\hat{q}$ is computed from (5.4) by using a few iterations of a Krylov subspace method like Conjugate Gradients or GMRES. We have just argued that the Krylov subspace built in those methods could better be used to project the Riccati equation (5.3) on, resulting in a small eigenvalue problem (5.9). We will now support this claim by numerical experiments with matrices from the Matrix Market test collection.

For information on the matrices from Matrix Market, see the webpage

http://math.nist.gov/MatrixMarket/index.html

5.4 Numerical experiments

We will now list some results in comparing Jacobi-Davidson with the Riccati method. In order to keep things as simple as possible and to illustrate the main idea, we did not include sophisticated features that could enhance both methods equally well. Therefore, we did not consider restart techniques, harmonic projections, and preconditioning.

For each matrix, we selected a random start vector. This start vector was used for each of the experiments with the same matrix. JD and Riccati were compared in two aspects: cpu-time and number of iteration steps needed to reduce the initial residual by a factor 10^{10}. Since the absolute numbers are not important for the comparison, we list, in the tabular below, the relative numbers only. As an example, the number 0.31 that is listed for the matrix plat1919 for $\ell = 5$ and belonging to cpu-time, indicates that Riccati needed only 0.31 of the cpu-time that JD needed to attain the same residual reduction.

5.5 Conclusions

As appears from the numerical experiments, the plain Riccati method almost always outperforms the plain Jacobi-Davidson method, and in many cases by a large factor. When Riccati loses, the difference is almost negligible. This suggests to incorporate the Riccati idea in existing JD codes which use Krylov subspaces to solve the Jacobi-Davidson correction equation.

As suggestions for further research one could think of recursiveness of this idea, since it is

in principle a nested subspace method with inner and outer loop, i.e. the space $\mathcal{V}$ is the space built in the outer loop, whereas for each expansion of $\mathcal{V}$ in one direction, as space $\mathcal{U}$ is constructed. It may be of interest to develop general theory for nested subspace methods for eigenvalue problems, which do not take into account that $\mathcal{U}$ is a Krylov subspace. Other choices may be of interest as well.

For readers more interested in the theoretical aspects of the method, and in its adaptation as a block method for the computation of invariant subspaces we refer to [2]. In contrast to [2], the underlying paper presents clear heuristics for its success and is therefore particularly suitable for people from computational science, computer science, and physics.

matrix name:	size	ℓ	5	10	20
sherman4	1104	cpu	0.68	0.10	0.05
real unsymm.		its	0.71	0.29	0.11
nnc1374	1374	cpu	0.55	0.28	0.12
real unsymm.		its	0.80	0.52	0.22
plat1919	1919	cpu	0.31	0.08	0.02
symm. indef.		its	0.53	0.17	0.06
utm3060	3060	cpu	1.07	0.81	0.43
real unsymm.		its	0.97	0.68	0.37
lshp3466	3466	cpu	0.79	1.09	1.07
symm. indef.		its	0.90	0.92	0.76
bcsstm24	3562	cpu	0.26	0.14	0.06
symm. posdef.		its	0.44	0.22	0.10
rw5151	5151	cpu	0.99	1.03	0.82
real unsymm.		its	1.00	1.00	0.81
cry10000	10000	cpu	0.34	0.10	0.05
real unsymm.		its	0.44	0.16	0.07
memplus	17758	cpu	0.17	0.19	0.08
real unsymm.		its	0.33	0.23	0.10
af23560	23560	cpu	0.83	0.49	0.91
real unsymm.		its	0.89	0.60	0.90
bcsstk32	44609	cpu	1.28	0.90	0.25
symm. indef.		its	0.65	0.39	0.13

Bibliography

[1] W.E. Arnoldi (1951). The principle of minimized iteration in the solution of the matrix eigenvalue problem, *Quart. Appl. Math.*,9:17–29.

[2] J.H. Brandts (2003). The Riccati method for eigenvalues and invariant subspaces of matrices with inexpensive action. *Linear Algebra and its Applications*, 358:333-363.

[3] J. Demmel (1987). Three methods for refining estimates of invariant subspaces, *Computing*, 38:43–57.

[4] B.N. Parlett (1998). *The Symmetric Eigenvalue Problem*. Classics in Applied Mathematics 20. SIAM, Philadelphia. Firstly appeared in 1980.

[5] W. Ritz (1908). Über eine neue Methode zu Lösung gewisser Variationsprobleme der mathematischen Physik. *J. reine angew. Math.*, 135:1–61.

[6] G.L.G. Sleijpen and H.A. van der Vorst (1996). Jacobi-Davidson iteration method for linear eigenvalue problems, *SIAM J. Matrix Anal. Applic.*, 17:401–425.