December 3, 2014

SPECTRAL PROPERTIES OF SELF-ADJOINT MATRICES

RODICA D. COSTIN

Contents

1. Review

1.1. **The spectrum of a matrix.** If L is a linear transformation *on a finite dimensional vector space the set of its eigenvalues $\sigma(L)$ is called the* **spectrum** *of L.*

Note that: 1. the spectrum $\sigma(L)$ contains no information on the multiplicity of each eigenvalue;

2. $\lambda \in \sigma(L)$ if and only if $L - \lambda I$ is not invertible.

Remark: It will be seen that for linear transformations (*linear operators*) in infinite dimensional vector spaces the spectrum of L is defined using property 2. above, and it may contain more numbers than just eigenvalues of L.

1.2. **Brief overview of previous results.**

Let F denote the scalar field $\mathbb{R}$ or $\mathbb{C}$.

1. A matrix M is called *diagonalizable* if it is similar to a diagonal matrix: exists an invertible matrix S so that $S^{-1}MS = \Lambda$=diagonal. The diagonal entries of Λ are precisely the eigenvalues of M and the columns of S are eigenvectors of M.

2. *An $n \times n$ matrix is diagonalizable if and only if it has n linearly independent eigenvectors.*

3. If $T : F^n \to F^n$ is the linear transformation given by $T\mathbf{x} = M\mathbf{x}$ then M is the matrix associated to T in the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of F^n, while $S^{-1}MS$ is the matrix associated to T in the basis $S\mathbf{e}_1, \ldots, S\mathbf{e}_n$ of F^n (recall that $S\mathbf{e}_j$ is the column j of S, eigenvector of M).

4. Assume that the n-dimensional matrix M is diagonalizable, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be linearly independent eigenvectors. Let $S = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$. Then $S^{-1}MS = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

$S^{-1}MS$ is the matrix of the linear transformation $F^n \to F^n$ given by $\mathbf{x} \mapsto M\mathbf{x}$ in the basis of F^n consisting of the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of M.

5. *Eigenvectors corresponding to different eigenvalues are linearly independent.*

As a consequence, an $n \times n$ matrix with n distinct eigenvalues is diagonalizable.

6. More generally, a matrix M is diagonalizable if and only if for every eigenvalue λ the eigenspace $V_\lambda = \{\mathbf{v} \mid M\mathbf{v} = \lambda\mathbf{v}\}$ has dimension equal to the multiplicity of λ.

7. If the matrix M is not diagonalizable, then there exists an invertible matrix S (whose columns are eigenvectors or generalized eigenvectors of M) so that $S^{-1}MS = J$=Jordan normal form: a block diagonal matrix, consisting of Jordan blocks which have a repeated eigenvalue on the diagonal and 1 above the diagonal.

8. If $J_p(\lambda)$ is a Jordan $p \times p$ block, with λ on the diagonal, then any power $J_p(\lambda)^k$ is an upper triangular matrix, with λ^k on the diagonal.

9. Let $q(t)$ be a polynomial.

If M is diagonalizable by S: $S^{-1}MS = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ then $q(M)$ is also diagonalizable by S and $S^{-1}q(M)S = \Lambda = \text{diag}(q(\lambda_1), \dots, q(\lambda_n))$.

If M is brought to a Jordan normal form by S: $S^{-1}MS = J$ then $q(M)$ is brought to an upper triangular form by S, having $q(\lambda_j)$ on the diagonal.

As a consequence:

Theorem 1. The spectral mapping theorem. *The eigenvalues of $q(M)$ are precisely $q(\lambda_1), \dots, q(\lambda_n)$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of M.*

Similar results hold for more general functions, when $q(t) = \sum_{k=0}^{\infty} c_k t^k$ and the series has radius of convergence strictly greater than $\max_j |\lambda_j|$, for example, for $q(t) = \exp(t)$.

1.3. **More on similar matrices.** Recall that similar matrices have the same eigenvalues.

Here is an additional result[1], to complete the picture (it is not proved here):

Theorem 2. *Two matrices are similar: $S^{-1}MS = N$ if and only if M and N have the same eigenvalues, and the dimensions of their corresponding eigenspaces are equal:* $\dim V_{\lambda_j}^{[M]} = \dim V_{\lambda_j}^{[N]}$ *for all j.*

2. Self-adjoint matrices

2.1. **Definitions.**

Definition 3. *Let $(V, \langle , \rangle)$ be an inner product space. The linear transformation $L : V \to V$ is called* **self-adjoint** *if $L^* = L$, that is, if*

$$\langle L\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, L^*\mathbf{y}\rangle \quad \textit{for all } \mathbf{x}, \mathbf{y} \in V$$

Recall that the matrix M of a linear transformation L with respect to orthonormal bases is related to the matrix M^* of L^* by $M^* = \overline{M^T}$ ($=\overline{M}^T$). Note that

$$(M^*)^* = M, \quad (MN)^* = N^*M^*$$

[1]See P.D. Lax, *Linear Algebra and Its Applications*, Wiley, 2007.

Recall that if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is an orthonormal basis of V then the inner product is the usual dot product of coordinates (also called the Euclidian inner product):

$$\text{if } \mathbf{x} = \sum_{k=0}^{n} x_k \mathbf{u}_k, \ \mathbf{y} = \sum_{k=0}^{n} y_k \mathbf{u}_k \ \text{ then } \ \langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=0}^{n} \overline{x_k} \mathbf{y}_k$$

So it suffices (for a while) to assume $V = \mathbb{R}^n$ or $V = \mathbb{C}^n$ equipped with the Euclidian inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^{n} x_j y_j \quad \text{on } \mathbb{R}^n \tag{1}$$

and respectively[2]

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^{n} \overline{x_j} y_j \quad \text{on } \mathbb{C}^n \tag{2}$$

For a unitary treatment we write $V = F^n$ and use the inner product (2). Of course for $F = \mathbb{R}$ the inner product (2) is just (1).

We will often use interchangeably the expressions "the matrix M" and "the linear transformation $\mathbf{x} \mapsto M\mathbf{x}$".

Definition 4. *A matrix A is called* **self-adjoint** *if $A = A^*$.*

Note that only square matrices can be self-adjoint, and that $A = A^*$ means, entrywise, that $A_{ij} = \overline{A_{ji}}$ (elements which are positioned symmetrically with respect to the diagonal are complex conjugates of each other).

When it is needed (or just desired) to distinguish between matrices with real coefficients and those with complex coefficients (perhaps not all real), the following terms are used:

Definition 5. *A self-adjoint matrix with real entries is called* **symmetric**.
A self-adjoint matrix with complex entries is called **Hermitian**.

Note that a symmetric matrix A satisfies $A^T = A$, hence its entries are symmetric with respect to the diagonal.

Notations.
$\mathcal{M}_n$ denotes the set of $n \times n$ matrices with entries in $\mathbb{C}$.
$\mathcal{M}_n(\mathbb{R})$ denotes the set of $n \times n$ matrices with entries in $\mathbb{R}$.

2.2. **Self-adjoint matrices are diagonalizable I.** We start with a few special properties of self-adjoint matrices.

Proposition 6. *If $A \in \mathcal{M}_n$ is a self-adjoint matrix: $A = A^*$, then*

$$\langle \mathbf{x}, A\mathbf{x} \rangle \in \mathbb{R} \quad \textit{for all } \mathbf{x} \in \mathbb{C}^n \tag{3}$$

[2]Some texts use conjugation in the second argument, rather than in the first one. Make sure you know the convention used in the text you are reading.

Proof:

$$\langle \mathbf{x}, A\mathbf{x}\rangle = \langle A^*\mathbf{x}, \mathbf{x}\rangle = \langle A\mathbf{x}, \mathbf{x}\rangle = \overline{\langle \mathbf{x}, A\mathbf{x}\rangle}$$

hence (3). □

Proposition 7. *If $A \in \mathcal{M}_n$ is a self-adjoint matrix: $A = A^*$, then all its eigenvalues are real: $\sigma(A) \subset \mathbb{R}$.*

Proof:
Let $\lambda \in \sigma(A)$. Then there is $\mathbf{v} \in \mathbb{C}^n,\ \mathbf{v} \neq \mathbf{0}$ so that $A\mathbf{v} = \lambda\mathbf{v}$. Then on one hand

$$\langle \mathbf{v}, A\mathbf{v}\rangle = \langle \mathbf{v}, \lambda\mathbf{v}\rangle = \lambda\langle \mathbf{v}, \mathbf{v}\rangle = \lambda\|\mathbf{v}\|^2$$

and on the other hand

$$\langle \mathbf{v}, A\mathbf{v}\rangle = \langle A^*\mathbf{v}, \mathbf{v}\rangle = \langle A\mathbf{v}, \mathbf{v}\rangle = \langle \lambda\mathbf{v}, \mathbf{v}\rangle = \overline{\lambda}\langle \mathbf{v}, \mathbf{v}\rangle = \overline{\lambda}\|\mathbf{v}\|^2$$

therefore $\lambda\|\mathbf{v}\|^2 = \overline{\lambda}\|\mathbf{v}\|^2$ and since $\mathbf{v} \neq \mathbf{0}$ then $\lambda = \overline{\lambda}$ hence $\lambda \in \mathbb{R}$. □

If a matrix is symmetric, not only its eigenvalues are real, but its eigenvectors as well:

Proposition 8. *If A is a symmetric matrix then all its eigenvectors are real.*

Indeed, the eigenvalues are real by Proposition 7. Then the eigenvectors are real since they are solutions linear systems with real coefficients (which can be solved using $+, -, \times, \div$, operations that performed with real numbers do yield real numbers (as opposed to solving polynomials equations, which may have nonreal solutions). □

For self-adjoint matrices, eigenvectors corresponding to distinct eigenvalues are not only linearly independent, they are even orthogonal:

Proposition 9. *If $A \in \mathcal{M}_n$ is a self-adjoint matrix: $A = A^*$, then eigenvectors corresponding to distinct eigenvalues are orthogonal: if $\lambda_{1,2} \in \sigma(A)$ and $A\mathbf{v}_1 = \lambda_1\mathbf{v}_1,\ A\mathbf{v}_2 = \lambda_2\mathbf{v}_2$ ($\mathbf{v}_{1,2} \neq \mathbf{0}$) then*

$$\lambda_1 \neq \lambda_2 \Longrightarrow\ \mathbf{v}_1 \perp \mathbf{v}_2$$

Proof:
On one hand

$$\langle \mathbf{v}_1, A\mathbf{v}_2\rangle = \langle \mathbf{v}_1, \lambda_2\mathbf{v}_2\rangle = \lambda_2\langle \mathbf{v}_1, \mathbf{v}_2\rangle$$

and on the other hand

$$\langle \mathbf{v}_1, A\mathbf{v}_2\rangle = \langle A^*\mathbf{v}_1, \mathbf{v}_2\rangle = \langle A\mathbf{v}_1, \mathbf{v}_2\rangle = \langle \lambda_1\mathbf{v}_1, \mathbf{v}_2\rangle = \lambda_1\langle \mathbf{v}_1, \mathbf{v}_2\rangle$$

(since the eigenvalues are real, by Proposition 7). Therefore $\lambda_2\langle \mathbf{v}_1, \mathbf{v}_2\rangle = \lambda_1\langle \mathbf{v}_1, \mathbf{v}_2\rangle$ and since $\lambda_1 \neq \lambda_2$ then $\langle \mathbf{v}_1, \mathbf{v}_2\rangle = 0$. □

As a consequence of Proposition 9: *if $A = A^*$ and all the eigenvalues of A are distinct, then the n independent eigenvectors form an orthogonal set.* We can normalize the eigenvectors, to be unit vectors, and then the eigenvectors form an orthonormal set, hence the matrix S which by conjugation diagonalizes A is a unitary matrix:*there is U unitary so that U^*AU =diagonal.*

In fact this is true for general self-adjoint matrices, as stead in Theorem 15. Its proof is included in §2.5 and requires establishing additional results, which are important in themselves.

2.3. **Further properties of unitary matrices.**

Proposition 10. *Every eigenvalue of a unitary matrix U has absolute value 1: $\sigma(U) \subset S^1 = \{z \in \mathbb{C} \,|\, |z| = 1\}$.*

Proof:
Let λ be an eigenvalue of U: $U\mathbf{v} = \lambda\mathbf{v}$ for some $\mathbf{v} \neq \mathbf{0}$. Then $\|U\mathbf{v}\| = \|\lambda\mathbf{v}\|$ and since U is an isometry then $\|\mathbf{v}\| = \|\lambda\mathbf{v}\|$ which implies $\|\mathbf{v}\| = |\lambda|\,\|\mathbf{v}\|$ which implies $|\lambda| = 1$ since $\mathbf{v} \neq \mathbf{0}$. □

Exercises.
1. Show that the product of two unitary matrices is also a unitary matrix.
2. Show that the determinant of a unitary matrix has absolute value 1. What is the determinant of an orthogonal matrix?

Proposition 11. *Eigenvectors of a unitary matrix U corresponding to different eigenvalues are orthogonal.*

Proof:
Let $U\mathbf{v}_1 = \lambda_1\mathbf{v}_1$ and $U\mathbf{v}_2 = \lambda_2\mathbf{v}_2$ ($\mathbf{v}_j \neq \mathbf{0}$). Since U preserves angles, $\langle U\mathbf{v}_1, U\mathbf{v}_2\rangle = \langle \mathbf{v}_1, \mathbf{v}_2\rangle$ which implies $\langle \lambda_1\mathbf{v}_1, \lambda_2\mathbf{v}_2\rangle = \langle \mathbf{v}_1, \mathbf{v}_2\rangle$ therefore $\overline{\lambda_1}\lambda_2\langle \mathbf{v}_1, \mathbf{v}_2\rangle = \langle \mathbf{v}_1, \mathbf{v}_2\rangle$. For $\lambda_1 \neq \lambda_2$ we have $\overline{\lambda_1}\lambda_2 \neq 1$ (using Proposition 10) therefore $\langle \mathbf{v}_1, \mathbf{v}_2\rangle = 0$. □

We will see a bit later that *unitary matrices are diagonalizable.*

Example. Consider the matrix

$$Q = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

(its action on $\mathbb{R}^3$ is a renumbering of the axes: $\mathbf{e}_j \mapsto \mathbf{e}_{j-1}$ cyclically). The columns of Q form an orthonormal set, therefore Q is a unitary matrix. Since its entries are real numbers, then Q is an orthogonal matrix.

The characteristic polynomial of Q is easily found to be $1 - \lambda^3$, therefore its eigenvalues are the three cubic roots of 1, namely $1, (-1 \pm i\sqrt{3})/2$. The eigenvector corresponding to 1 is $(1, 1, 1)^T$. In the plane orthogonal to this

eigenvector is spanned by the other two eiegnevctors (rather, the real and imaginary parts, if we choose to work in $\mathbb{R}^3$) the action of Q is a rotation seen geometrically, and algebraically by the presence of the two complex eigenvalues.

2.4. **Triangularization by conjugation using a unitary matrix.** Diagonalization of matrices is not always possible, and even when it is, it is computationally expensive. However, matrices can be easily brought to a triangular form, which suffices for many applications:

Theorem 12. *Given any square matrix $M \in \mathcal{M}_n$ there is an n dimensional unitary matrix U so that $U^*MU = T$=upper triangular.*

Of course, the diagonal elements of T are the eigenvalues of M.

Proof:

The matrix U is constructed in successive steps.

1^o. Choose an eigenvalue λ_1 of M and a corresponding *unit* eigenvector $\mathbf{u}_1$. Then complete $\mathbf{u}_1$ to an orthonormal basis of $\mathbb{C}^n$: $\mathbf{u}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ (recall that this is always possible!). Let U_1 be the unitary matrix

$$U_1 = [\mathbf{u}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] = [\mathbf{u}_1 \,\big|\, X_1]$$

The goal is to simplify M by replacing it with $U_1^*MU_1$. Note that

(4)
$$U_1^*MU_1 = \left[\begin{array}{c} \mathbf{u}_1^* \\ - \\ X_1^* \end{array}\right] M[\mathbf{u}_1 \,\big|\, X_1] = \left[\begin{array}{c} \mathbf{u}_1^* \\ - \\ X_1^* \end{array}\right] [M\mathbf{u}_1 \,\big|\, MX_1] = \left[\begin{array}{c} \mathbf{u}_1^* \\ - \\ X_1^* \end{array}\right] [\lambda_1\mathbf{u}_1 \,\big|\, *]$$

$$= \left[\begin{array}{c|c} \lambda_1\mathbf{u}_1^*\mathbf{u}_1 & * \\ - & - \\ \lambda_1 X_1^*\mathbf{u}_1 & * \end{array}\right] = \left[\begin{array}{c|c} \lambda_1 & * \\ - & - \\ 0 & * \end{array}\right] \equiv \tilde{M}_1$$

where we used the fact that the vector $\mathbf{u}_1$ is a unit vector, and that the columns of X_1 are orthogonal to $\mathbf{u}_1$.

Denote by M_1 the $(n-1)\times(n-1)$ lower right submatrix of $\tilde{M}_1$:

$$\tilde{M}_1 = \left[\begin{array}{c|c} \lambda_1 & * \\ - & - \\ 0 & M_1 \end{array}\right]$$

Note that $Sp(\mathbf{e}_2, \ldots, \mathbf{e}_n) \equiv \mathbb{C}^{n-1}$ is an invariant space for $U_1^*MU_1$ and $U_1^*MU_1$ acts on $\mathbb{C}^{n-1}$as multiplication by M_1.

Note also that $\sigma(M) = \sigma(\tilde{M}_1)$ (conjugate matrices have the same eigenvalues) and that $\sigma(\tilde{M}_1) = \{\lambda_1\} \cup \sigma(M_1)$.

2^o. We repeat the first step for the $n-1$ dimensional matrix M_1: let λ_2 be an eigenvalue of M_1, and $\mathbf{u}_2 \in \mathbb{C}^{n-1}$ a unit eigenvector, and complete it

to an orthonormal basis $\mathbf{u}_2, \mathbf{y}_2, \ldots, \mathbf{y}_n$ of $\mathbb{C}^{n-1}$. If $U_2 = [\mathbf{u}_2, \mathbf{y}_2, \ldots, \mathbf{y}_n]$ then

$$U_2^* M_1 U_2 = \left[\begin{array}{c|c} \lambda_2 & * \\ - & - \\ 0 & M_2 \end{array}\right] \tag{5}$$

where $M_2 \in \mathcal{M}_{n-2}$.

We can extend U_2 to an n dimensional unitary matrix by

$$\tilde{U_2} = \left[\begin{array}{c|c} 1 & 0 \\ - & - \\ 0 & U_2 \end{array}\right]$$

and it is easy to check that

$$\tilde{U_2}^* \tilde{M_1} \tilde{U_2} = \left[\begin{array}{cc|c} \lambda_1 & * & * \\ 0 & \lambda_2 & * \\ - & - & - \\ 0 & 0 & M_3 \end{array}\right] \equiv \tilde{M_2}$$

Note that the matrix $U_1\tilde{U_2}$ (which is unitary) conjugates M to $\tilde{M_2}$.

$3^o \ldots n^o$ Continuing this procedure we obtain the unitary $U = U_1\tilde{U_2}\ldots\tilde{U_n}$ which conjugates M to an upper triangular matrix. □

2.5. **All self-adjoint matrices are diagonalizable II.** Let A be a self-adjoint matrix: $A = A^*$. By Theorem 12 there is a unitary matrix U so that $U^*AU = T =$ upper triangular. The matrix U^*AU is self-adjoint, since $(U^*AU)^* = U^*A^*(U^*)^* = U^*AU$ and a triangular matrix which is self-adjoint must be diagonal! We proved Theorem 15:

Any self-adjoint matrix A is diagonalizable and there is U unitary so that $U^*AU =$diagonal.

2.6. **Normal matrices.** We saw that any self-adjoint matrix is diagonalizable, has a complete set of orthonormal eigenvectors, and its diagonal form is real (by Proposition 7). It is natural to inquire: what are the matrices which are diagonalizable also having a complete set of orthonormal eigenvectors, but having possible nonreal eigenvalues?

It is easy to see that such matrices have special properties. For example, if $N = U\Lambda U^*$ for some unitary U and diagonal Λ, then, by taking the adjoint, $N^* = U\overline{\Lambda}U^*$ and it is easy to see that N commutes with its adjoint:

$$NN^* = N^*N \tag{6}$$

It turns out that condition (6) suffices to ensure that a matrix is diagonalizable by a unitary. Indeed, we know that any matrix can be conjugated to an upper triangular form by a unitary: $U^*MU = T$ as in Theorem 2.4; therefore also $U^*M^*U = T^*$. If M satisfies $MM^* = M^*M$, then also $TT^* = T^*T$; a simple calculation shows that this can only happen if T is, in fact, diagonal.

For example, in the 2-dimensional case:

$$\text{for } T = \begin{bmatrix} \lambda_1 & \alpha \\ 0 & \lambda_2 \end{bmatrix} \quad \text{then } T^* = \begin{bmatrix} \overline{\lambda_1} & 0 \\ \overline{\alpha} & \overline{\lambda_2} \end{bmatrix}$$

and therefore

$$TT^* = \begin{bmatrix} |\lambda_1|^2 + |\alpha|^2 & \alpha\overline{\lambda_2} \\ \overline{\alpha}\lambda_2 & |\lambda_2|^2 \end{bmatrix} \text{ and } T^*T = \begin{bmatrix} |\lambda_1|^2 & \alpha\overline{\lambda_1} \\ \overline{\alpha}\lambda_1 & |\lambda_2|^2 + |\alpha|^2 \end{bmatrix}$$

and $TT^* = T^*T$ if and only if $\alpha = 0$.

Exercise. Show that an upper triangular matrix T commutes with its adjoint if and only if T is diagonal.

Definition 13. *A matrix N which commutes with its adjoint, $NN^* = N^*N$, is called* **normal**.

We proved:

Theorem 14. The spectral theorem for normal matrices
*A square matrix N can be diagonalized by a unitary matrix: $U^*NU =$*diagonal *for some unitary U, if and only if N is normal: $N^*N = NN^*$.*

In particular:

Theorem 15. The spectral theorem for self-adjoint matrices
*A self-adjoint matrix $A = A^*$ can be diagonalized by a unitary matrix: $U^*AU =$*real diagonal*, for some unitary U.*

And in the more particular case:

Theorem 16. The spectral theorem for symmetric matrices
*A symmetric matrix $A = A^T \in \mathcal{M}_n(\mathbb{R})$ can be diagonalized by an orthogonal matrix: $Q^TAQ =$*real diagonal*, for some orthogonal matrix Q.*

Exercise. True or False? "A normal matrix with real eigenvalues is self-adjoint."

Note. *Unitary matrices are normal, hence are diagonalizable by a unitary.*

2.7. **Generic matrices (or: "beware of roundoff errors").** Choosing the entries of a square matrix M at random, *it is almost certain* that M has distinct eigenvalues and a complete set of eigenvectors which is not orthogonal. In other words, M is almost certainly diagonalizable, but not by a unitary conjugation, rather by conjugation with an S which requires a deformation of the space: modifications of angles, stretching of lengths.

Why is that?

I. Generic matrices have no repeated eigenvalues, hence are diagonalizable.

Indeed, of all matrices in $\mathcal{M}_n$ (a vector space of dimension n^2) the set of matrices with repeated eigenvalues form a surface of lower dimension since their entries satisfy the condition that the characteristic polynomial and

its derivative have a common zero (two polynomial equations with n^2+1 unknowns).

To illustrate, consider 2 dimensional real matrices. A matrix

$$M \in \mathcal{M}_2(\mathbb{R}), \quad M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{7}$$

has its characteristic polynomial $p(\lambda) = \lambda^2 - (a+d)\,\lambda + (ad - bc)$. Then some λ is not a simple eigenvalue if and only if λ satisfies

$$p(\lambda) = 0 \ \text{ and } \ p'(\lambda) = 0$$

Equation $p'(\lambda) = 0$ implies $\lambda = (a+d)/2$ which substituted in $p(\lambda) = 0$ gives

$$(a-d)^2 + 4\,bc = 0 \tag{8}$$

which is the condition that a two dimensional matrix has multiple eigenvalues: one equation in the four dimensional space of the entries (a, b, c, d); its solution is a three dimensional surface.

II. Among diagonalizable matrices, those diagonalizable by a unitary matrix form a set of lower dimension, due to the conditions that eigenvectors be orthogonal.

To illustrate, consider 2 dimensional matrices (7) with distinct eigenvalues. The eigenvalues are

$$\lambda_\pm = \frac{a+d}{2} \pm \frac{1}{2}\sqrt{(a-d)^2 + 4\,bc}$$

and the eigenvectors are

$$\mathbf{v}_\pm = \begin{bmatrix} b \\ \lambda_\pm - a \end{bmatrix} \text{ if } b \neq 0, \text{ and } \ \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \ \mathbf{v}_2 = \begin{bmatrix} a-d \\ c \end{bmatrix} \text{ if } b = 0$$

which are orthogonal if and only if $b = c$, a 3-dimensional subspace in the four dimensional space of the parameters (a, b, c, d).

Why study normal and self-adjoint transformations? Problems which come from applications often have symmetries (coming from conservation laws which systems have, or are approximated to have) and normal or self-adjoint matrices often appear. We will see that in infinite dimensions, there are important linear transformations which are, or are reducible to, self-adjoint ones; differentiation is one of them.

2.8. **Anti-self-adjoint (skew-symmetric, skew-Hermitian) matrices.**

Definition 17.
A matrix M satisfying $M^* = -M$ is called **anti-self-adjoint**.
In particular:
A matrix $B \in \mathcal{M}_n(\mathbb{R})$ so that $B^T = -B$ is called **skew-symmetric**.
A matrix $K \in \mathcal{M}_n(\mathbb{C})$ so that $K^H = -K$ is called **skew-Hermitian**.

Note that:
1. Anti-self-adjoint matrices are normal.
2. If $K^* = -K$ then $A = \pm iK$ is a self-adjoint matrix.
Therefore, eigenvalues of anti-self-adjoint matrices are purely imaginary.

Exercises.
1. Show that a skew-symmetric matrix of odd dimension has determinant zero.
2. Show that if λ is an eigenvalue of a skew-symmetric matrix, then $-\lambda$ is also an eigenvalue.
3. Show that if K is skew-Hermitian then e^K is unitary. What kind of matrix is e^K when K is skew-symmetric?
4. Consider $\mathbf{u}(t)$ a solution of the linear differential equation $\mathbf{u}'(t) = M\mathbf{u}(t)$ where M is a skew-symmetric matrix. Show that $\|\mathbf{u}(t)\| = \|\mathbf{u}(0)\|$ for any t.
5. For $A, B \in \mathcal{M}_n(\mathbb{R})$ define their commutator $[A, B] = AB - BA$. Show that $[A, B]$ is skew-symmetric.
6. If N is normal, show that $N = A + K$ where A is self-adjoint and K is anti-self-adjoint. Hint: let $A = \frac{1}{2}(N + N^*)$ and $K = \frac{1}{2}(N - N^*)$.

Note that skew-symmetric matrices have the diagonal entries zero, since if $B^T = -B$ then this imples that $B_{jj} = -B_{jj}$ hence $B_{jj} = 0$. For general anti-self-adjoint matrices $K^* = -K$ implies for the diagonal entries that $K_{jj} = -\overline{K_{jj}}$ hence if $K_{jj} = a_j + ib_j$ (with a_j, b_j real) then $a_j + ib_j = -(a_j - ib_j)$ hence the diagonal entries are purely imaginary (or zero).

2.9. **Application to linear differential equations.** *The Schrödinger equation* has the form

$$-i\frac{d\mathbf{y}}{dt} = A\mathbf{y}, \quad A \in \mathcal{M}_n(\mathbb{R}) \tag{9}$$

where $A = A^*$. (In the proper Schrödinger equation A is more general than a matrix, it is a linear transformation in infinite dimensions -an operator, usually $A = \Delta - V(\mathbf{x})$.)

Equation (9) is, of course, the same as $\mathbf{y}' = iA\mathbf{y}$ with iA an anti-self-adjoint matrix. The general solution is $\mathbf{y}(t) = e^{itA}\mathbf{y}_0$.

It turns out that the evolution of a Schrödinger equation preserves the norm of vectors: $\|\mathbf{y}(t)\| = \|\mathbf{y}(0)\|$ for all t.

First note the derivative of an inner product obeys the product rule:

$$\frac{d}{dt}\langle \mathbf{x}(t), \mathbf{y}(t)\rangle = \langle \frac{d}{dt}\mathbf{x}(t), \mathbf{y}(t)\rangle + \langle \mathbf{x}(t), \frac{d}{dt}\mathbf{y}(t)\rangle$$

because the inner product is a sum of products: $\langle \mathbf{x}(t), \mathbf{y}(t)\rangle = \sum_{j=1}^{n} \overline{x_j(t)} y_j(t)$.

To see that the evolution preserves the norm calculate

$$\frac{d}{dt}\|\mathbf{y}(t)\|^2 = \frac{d}{dt}\langle \mathbf{y}(t), \mathbf{y}(t)\rangle = \langle \frac{d}{dt}\mathbf{y}(t), \mathbf{y}(t)\rangle + \langle \mathbf{y}(t), \frac{d}{dt}\mathbf{y}(t)\rangle$$

$= \langle iA\mathbf{y}(t), \mathbf{y}(t)\rangle + \langle \mathbf{y}(t), iA\mathbf{y}(t)\rangle = \langle iA\mathbf{y}(t), \mathbf{y}(t)\rangle + \langle (iA)^*\mathbf{y}(t), \mathbf{y}(t)\rangle =$

$= \langle iA\mathbf{y}(t), \mathbf{y}(t)\rangle + \langle (-iA^*)\mathbf{y}(t), \mathbf{y}(t)\rangle = \langle iA\mathbf{y}(t), \mathbf{y}(t)\rangle + \langle (-iA)\mathbf{y}(t), \mathbf{y}(t)\rangle = 0$

and therefore the function $\|\mathbf{y}(t)\|$ is constant:

$$\|\mathbf{y}(t)\| = \|e^{itA}\mathbf{y}_0\| = \|\mathbf{y}_0\| \text{ for all } \mathbf{y}_0 \text{ and all } t$$

which means that the matrix e^{itA} (called *propagator*) is unitary.

The evolution of a system with an anti-self-adjoint matrix is unitary (preserves the norm).

2.10. **Diagonalization of unitary matrices.** Noting that unitary matrices are normal, we re-obtain Proposition 11 for free. Also, Proposition 2.3 becomes obvious, since if U is unitary and $S^*US = \Lambda$ =diagonal with S unitary, then $UU^* = I$ implies $(S\Lambda S^*)(S\Lambda S^*)^* = I$ therefore $\Lambda\Lambda^* = I$ so all $\lambda_j\overline{\lambda_j} = 1$ hence all $|\lambda_j| = 1$.

3. Quadratic forms and Positive definite matrices

3.1. **Quadratic forms.** *Example:* a quadratic form in $\mathbb{R}^2$ is a function $q : \mathbb{R}^2 \to \mathbb{R}$ of the form

$$q(x_1, x_2) = ax_1^2 + 2bx_1x_2 + cx_2^2$$

The function q can be written using matrices and the usual inner product as

$$q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle, \quad \text{where } A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

Note that the matrix A is symmetric. But the same quadratic form can be written using many other matrices which are not symmetric:

$$q(\mathbf{x}) = ax_1^2 + 2bx_1x_2 + cx_2^2 = \langle \mathbf{x}, C\mathbf{x} \rangle, \quad \text{where } C = \begin{bmatrix} a & b+d \\ b-d & c \end{bmatrix}$$

Definition 18. *A* **quadratic form in** $\mathbb{R}^n$ *is a function* $q : \mathbb{R}^n \to \mathbb{R}$ *of the form*

$$q(\mathbf{x}) = \sum_{i,j=1}^{n} A_{ij} x_i x_j = \langle \mathbf{x}, A\mathbf{x} \rangle$$

The matrix A can be assumed symmetric. Indeed, if $q(\mathbf{x}) = \langle \mathbf{x}, C\mathbf{x} \rangle$ is a quadratic form defined by an arbitrary (square) matrix C, then C can be replaced by $A = \frac{1}{2}(C + C^T)$ which is symmetric (entrywise, $A_{ij} = \frac{1}{2}(C_{ij} + C_{ji})$), and gives *the same* quadratic form:

$$\langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, \frac{1}{2}(C+C^T)\mathbf{x} \rangle = \frac{1}{2}\left(\langle \mathbf{x}, C\mathbf{x} \rangle + \langle \mathbf{x}, C^T\mathbf{x} \rangle \right) = \frac{1}{2}\left(\langle \mathbf{x}, C\mathbf{x} \rangle + \langle C\mathbf{x}, \mathbf{x} \rangle \right)$$

$$= \frac{1}{2}\left(\langle \mathbf{x}, C\mathbf{x} \rangle + \langle \mathbf{x}, C\mathbf{x} \rangle \right) = \langle \mathbf{x}, C\mathbf{x} \rangle$$

3.1.1. *Diagonalization by orthogonal matrices.* Since $A \in \mathcal{M}_n(\mathbb{R})$ is symmetric, then there exists an orthogonal matrix Q so that $Q^T A Q = \Lambda$ with $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$, the eigenvalues λ_j of A being real. Then

$$q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, Q\Lambda Q^T\mathbf{x} \rangle = \langle Q^T\mathbf{x}, \Lambda Q^T\mathbf{x} \rangle$$

In the new coordinates $\mathbf{y}$ given by $\mathbf{y} = Q^T\mathbf{x}$ the quadratic form has completed squares:

$$q(\mathbf{x}) = q(Q\mathbf{y}) = \langle \mathbf{y}, \Lambda\mathbf{y} \rangle = \sum_{j=1}^{n} \lambda_j y_j^2 \tag{10}$$

The columns of $Q = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ ($\mathbf{u}_j$ are eigenvectors of A) are called *a diagonalizing basis* of the quadratic form, their spanned subspaces $\mathbb{R}\mathbf{u}_j$ are called **principal axes** of the quadratic form, and formula (10) is called **reduction to principal axes** (or *diagonal form* of the quadratic function).

Using (10) for $\mathbf{y} = \mathbf{e}_j$ it follows that

$$q(\mathbf{u}_j) = \lambda_j$$

Examples of quadratic form in $\mathbb{R}^2$, and their level sets.

1. Consider the quadratic form

$$q(x_1, x_2) = 9x_1^2 + 2\sqrt{3}x_1x_2 + 11x_2^2$$

Note that

$$q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x} \rangle \quad \text{where } A = \begin{bmatrix} 9 & \sqrt{3} \\ \sqrt{3} & 11 \end{bmatrix}$$

The matrix A has the eigenvalue $\lambda_1 = 12$ with eigenvector $\mathbf{u}_1 = (\frac{1}{2}, \frac{\sqrt{3}}{2})^T$ and the eigenvalue $\lambda_2 = 8$ with eigenvector $\mathbf{u}_2 = (-\frac{\sqrt{3}}{2}, \frac{1}{2})^T$. The orthogonal matrix Q diagonalizing A is

$$Q = \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{bmatrix} \quad \text{and } Q^T A Q = \begin{bmatrix} 12 & 0 \\ 0 & 8 \end{bmatrix}$$

The level curves $q(\mathbf{x}) = k$ are ellipses (for $k > 0$) with principal axes in the directions of the two eigenvectors. For $k = 0$ the level set is a point, $\mathbf{0}$ (a degenerate ellipse).

The point $\mathbf{x} = \mathbf{0}$ is a minimum of the function $q(\mathbf{x})$.

1'. Consider

$$q_n(x_1, x_2) = -9x_1^2 - 2\sqrt{3}x_1x_2 - 11x_2^2$$

Of course, $q_n = -q$, hence q_n has the same eigenvectors as q in Example 1., and eigenvalues of opposite sign: -12 and -8.

The level curves $q_n(\mathbf{x}) = k$ are ellipses (for $k < 0$) with principal axes in the directions of the two eigenvectors. The point $\mathbf{x} = \mathbf{0}$ is a maximum of the function $q_n(\mathbf{x})$.

2. Consider the quadratic form:

$$q_s(x_1, x_2) = -3x_1^2 + 10\sqrt{3}x_1x_2 + 7x_2^2$$

with

$$q_s(\mathbf{x}) = \langle \mathbf{x}, A_s\mathbf{x} \rangle \quad \text{where } A_s = \begin{bmatrix} -3 & 5\sqrt{3} \\ 5\sqrt{3} & 7 \end{bmatrix}$$

The matrix A_s has the eigenvalues $\lambda_1 = 12$ and $\lambda_2 = -8$ with the same eigenvectors as in Example 1. The diagonal form is

$$q_s(\mathbf{x}) = q_s(Q\mathbf{y}) = 12y_1^2 - 8y_1^2$$

The level sets $q(\mathbf{x}) = k$ are hyperbolas (for $k \neq 0$) with asymptotes in the directions of the two eigenvectors. For $k = 0$ the level set is formed of two lines (a degenerate hyperbola).

The point $\mathbf{x} = \mathbf{0}$ is a minimum in the direction of $(y_1, 0)$ and a maximum in the direction of $(0, y_2)$: it is a saddle point of the function $q_s(\mathbf{x})$.

3. As an example of a quadratic form with one zero eigenvalue consider

$$q_d(x_1, x_2) = 3x_1^2 + 6\sqrt{3}x_1x_2 + 9x_2^2$$

with

$$q_d(\mathbf{x}) = \langle \mathbf{x}, A_d\mathbf{x}\rangle \text{ where } A_d = \begin{bmatrix} 3 & 3\sqrt{3} \\ 3\sqrt{3} & 9 \end{bmatrix}$$

where A_d has the eigenvalues $\lambda_1 = 12$ and $\lambda_2 = 0$ (and the same eigenvectors as above). The diagonal form is

$$q_d(\mathbf{x}) = q_d(Q\mathbf{y}) = 12y_1^2$$

The level sets $q_d(\mathbf{x}) = k$ are pairs of lines.

In general, the level curves of quadratic forms are *quadrics*, whose nature depends on the signs of the eigenvalues of A.

3.1.2. *Diagonalization using a non-unitary matrix.* How much more can we simplify a normal form (10) of a quadratic form? Can we replace the λ's with, say, all 1?

Change the coordinates even further: in (10) substitute $\mathbf{y} = D\mathbf{v}$ where D is the diagonal matrix with entries D_{jj}; (10) becomes

$$q(\mathbf{x}) = q(Q\mathbf{y}) = q(QD\mathbf{v}) = \sum_{j=1}^{n} \lambda_j D_{jj}^2 v_j^2 \tag{11}$$

which can be most simplified by choosing

$$D_{jj} = \begin{cases} 1/\sqrt{\lambda_j} & \text{if } \lambda_j > 0 \\ 1/\sqrt{-\lambda_j} & \text{if } \lambda_j < 0 \\ 0 & \text{if } \lambda_j = 0 \end{cases}$$

in which case and (11) becomes sum/difference of squares

$$q(\mathbf{x}) = q(Q\mathbf{y}) = q(QD\mathbf{v}) = \sum_{j=1}^{n} \text{sign}(\lambda_j) v_j^2 \tag{12}$$

where

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t > 0 \\ -1 & \text{if } t < 0 \\ 0 & \text{if } t = 0 \end{cases}$$

The linear change of coordinates $\mathbf{x} \mapsto \mathbf{v}$ after which the quadratic form that the simple form of (12) is $\mathbf{x} = QD\mathbf{v}$. The matrix A is diagonalized by QD, a matrix which preserves orthogonality, but stretches or compresses lengths.

Theorem 20 shows that this all that can be done: the modulus of any $\lambda_j \neq 0$ can be changed (in appropriate coordinates), but not its sign.

3.1.3. *A quadratic form after a change of basis.* What matrices represent the same quadratic form, only in different coordinates (i.e. in a different basis) of $\mathbb{R}^n$? Let $q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x}\rangle$ be the quadratic form associated to the symmetric matrix A using the usual inner product in $\mathbb{R}^n$. Let S be an invertible matrix, and let $\mathbf{x} = S\mathbf{y}$ be a change of coordinates in $\mathbb{R}^n$. Then

$$q(\mathbf{x}) = q(S\mathbf{y}) = \langle S\mathbf{y}, AS\mathbf{y}\rangle = \langle \mathbf{y}, S^T AS\mathbf{y}\rangle$$

and in the new coordinates $\mathbf{y}$ the quadratic form is associated to the symmetric matrix S^TAS.

Definition 19. *Two symmetric matrices A and B linked by the transformation $B = S^TAS$ for some nonsingular matrix S are called* **congruent**.

Congruent matrices represent the same quadratic form in different bases.

The quadratic form $q(\mathbf{x})$ is called **diagonalized** in the basis consisting of the columns of S if $q(S\mathbf{y}) = \sum_j a_j y_j^2$.

We saw that the numbers a_j can be changed by changing the basis. What cannot be changed is their sign (except for a rearrangement):

Theorem 20. Sylvester's law of inertia for quadratic forms

Let $q(\mathbf{x})$ be a quadratic form with real coefficients.

The number of coefficients of a given sign does not depend on a particular choice of diagonalizing basis.

In other words, congruent symmetric matrices have the same number of positive eigenvalues, the same number of negative eigenvalues, and the same number of zero eigenvalues.

The number of positive eigenvalues of the symmetric matrix A defining a quadratic form defines its character, as we will explore on a few examples.

Examples.

1. Quadratic forms in $\mathbb{R}^2$:

(i) Forms with both eigenvalues positive have level curves ellipses (for $k > 0$), and have a minimum at the point $\mathbf{x} = \mathbf{0}$.

(ii) Forms with both eigenvalues negative have level curves in the shape of ellipses (for $k < 0$), and have a maximum at the point $\mathbf{x} = \mathbf{0}$.

(iii) Forms with one positive and one negative eigenvalue have level curves in the shape of hyperbolas (for $k \neq 0$), and the point $\mathbf{x} = \mathbf{0}$ is a saddle point (in some directions it is a minimum, while in other directions it is a maximum). For $k = 0$ the level curves are two lines.

2. Quadratic forms in $\mathbb{R}^n$:

(i) Forms with all eigenvalues positive have a minimum at the point $\mathbf{x} = \mathbf{0}$.

(ii) Forms with all eigenvalues negative have a maximum at $\mathbf{x} = \mathbf{0}$.

(iii) Forms with some positive eigenvalues and the other negative, have a saddle point at $\mathbf{x} = \mathbf{0}$.

3. The character of a surface in $\mathbb{R}^3$ given by an equation $ax^2+by^2+cz^2 = 1$ depends only on the signs of the constants a, b, c: if all are positive, then it is an ellipsoid; if two are positive and one is negative, then it is a hyperboloid of one sheet; if one positive and two are negative, then it is a hyperboloid of two sheets.

3.2. **Critical points of functions of several variables.** Let $f(\mathbf{x})$ be a function of n real variables defined on a domain in $\mathbb{R}^n$.

Recall that the first derivatives of f can be organized as the gradient vector

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)^T$$

and its second derivatives can be organized as a matrix, the Hessian,

$$Hf = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{i,j=1,\dots,n}$$

which is a symmetric matrix if the mixed derivatives are continuous, and therefore equal.

Recall the form of the Taylor series of a function $f(\mathbf{x})$ of n variables:

$$\sum_{\mathbf{k} \in \mathbb{N}^n} \frac{1}{\mathbf{k}!} \frac{\partial^{\mathbf{k}} f}{\partial \mathbf{x}^{\mathbf{k}}}(\mathbf{a}) \, (\mathbf{x} - \mathbf{a})^{\mathbf{k}}$$

where the following notations are used

$$\mathbf{k}! = \prod_{j=1}^n (k_j)! \,, \quad \frac{\partial^{\mathbf{k}}}{\partial \mathbf{x}^{\mathbf{k}}} = \prod_{j=1}^n \frac{\partial^{k_j}}{\partial x_j^{k_j}} \,, \quad \mathbf{v}^{\mathbf{k}} = \prod_{j=1}^n v_j^{k_j}$$

Retaining the first few terms of the Taylor series of f at $\mathbf{x} = \mathbf{a}$ we obtain the quadratic approximation

(13)

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{1}{1!} \sum_{j=1}^n \frac{\partial f}{\partial x_j}(\mathbf{a})(x_j - a_j) + \frac{1}{2!} \sum_{i,j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a})(x_i - a_i)(x_j - a_j)$$

or, more compactly,

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{1}{1!} \langle \nabla f(\mathbf{a}), (\mathbf{x} - \mathbf{a}) \rangle + \frac{1}{2!} \langle \mathbf{x} - \mathbf{a}, Hf(\mathbf{a})(\mathbf{x} - \mathbf{a}) \rangle$$

The point $\mathbf{x} = \mathbf{a}$ is called a *critical point* if $\nabla f(\mathbf{a}) = \mathbf{0}$. At a critical point the Taylor approximation (13) becomes

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{1}{2!} \langle \mathbf{x} - \mathbf{a}, Hf(\mathbf{a})(\mathbf{x} - \mathbf{a}) \rangle$$

In real analysis it is proved that if the Hessian matrix $Hf(\mathbf{a})$ has all its eigenvalues nonzero, then the character of the critical point $\mathbf{x} = \mathbf{a}$ of $f(\mathbf{x})$ is the same as for the Hessian quadratic form $\langle \mathbf{x} - \mathbf{a}, Hf(\mathbf{a})(\mathbf{x} - \mathbf{a}) \rangle$ (they both have minimum, or a maximum, or a saddle point).

It is then important to have practical criteria to decide when quadratic forms have a minimum, a maximum, or a saddle point. This brings us to the topic of next section, positive definite matrices.

3.3. **Positive definite matrices.** In this section matrices are not necessarily real. But the notions of positive/negative definite matrices is used only for self-adjoint matrices.

Definition 21. *A self-adjoint matrix $A = A^*$ is called* **positive definite** *if*

$$\langle \mathbf{x}, A\mathbf{x}\rangle > 0 \quad \textit{for all } \mathbf{x} \neq \mathbf{0}$$

Note that $\langle \mathbf{x}, A\mathbf{x}\rangle \in \mathbb{R}$ since A is self-adjoint.

Exercise. Find which of following matrices are positive definite

$$A_1 = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix},\ A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix},\ A_3 = \begin{bmatrix} -2 & 0 \\ 0 & 3 \end{bmatrix}.$$

The following theorem lists equivalent characterization of positive matrices:

Theorem 22. *Let $A = A^*$ be an $n\times n$ matrix. Then following are equivalent:*

(i) $\langle \mathbf{x}, A\mathbf{x}\rangle > 0$ for all $\mathbf{x} \neq \mathbf{0}$;

(ii) all the eigenvalues of A are positive;

(iii) the function $\mathbb{C}^n \times \mathbb{C}^n \to \mathbb{R}$ defined by $(\mathbf{x}, \mathbf{y}) \to \langle \mathbf{x}, \mathbf{y}\rangle_A$ where

$$\langle \mathbf{x}, \mathbf{y}\rangle_A = \langle \mathbf{x}, A\mathbf{y}\rangle$$

is an inner product on $\mathbb{C}^n$ (in fact, every inner product on $\mathbb{C}^n$ has this form);

(iv) there are n linearly independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{C}^n$ so that

$$A_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j\rangle, \quad i, j = 1, \ldots, n$$

in other words, $A = M^*M$ where $M = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ is an invertible matrix, and if A is a real matrix, then M is a real matrix and $A = M^TM$. (In fact, the same is true for M a rectangular $m \times n$ matrix, with $m \geq n$ and of maximal rank);

(v) all the upper left $k \times k$ ($k = 1, 2, \ldots, n$) submatrices

$$[A_{1,1}],\ [A_{i,j}]_{i,j=1,2}, \ldots, [A_{i,j}]_{i,j=1,\ldots,k}, \ldots, [A_{i,j}]_{i,j=1,\ldots,n}$$

have positive determinants.

Proof:

The equivalence of (i) and (ii) is immediate, since if U is a unitary matrix so that $U^*AU = \Lambda$ =diagonal then

$$\langle \mathbf{x}, A\mathbf{x}\rangle = \langle \mathbf{x}, U\Lambda U^*\mathbf{x}\rangle = \langle U^*\mathbf{x}, \Lambda U^*\mathbf{x}\rangle = \sum_{j=1}^{n} \lambda_j \left| \left(U^*\mathbf{x}\right)_j \right|^2 \tag{14}$$

If all $\lambda_j > 0$ then clearly $\langle \mathbf{x}, A\mathbf{x}\rangle > 0$. Conversley, if all $\langle \mathbf{x}, A\mathbf{x}\rangle > 0$ then for each $\mathbf{x} = \mathbf{u}_k$ (kth column of U) then $U\mathbf{e}_k = \mathbf{u}_k$ hence $U^*\mathbf{u}_k = \mathbf{e}_k$ and (14) implies $\lambda_k > 0$.

The equivalence of (i) and (iii) is immediate.

(i) implies (iv): using $A = U\Lambda U^*$ and since $\lambda_j > 0$ we can define the radical $\sqrt{\Lambda}$ =diag$(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. Then $A = U\sqrt{\Lambda}\sqrt{\Lambda}U^* = M^*M$ where

$M = U\sqrt{\Lambda}U^*$ is a positive definite matrix (since U is invertible and $\sqrt{\Lambda}$ has positive eigenvalues, hence it is also invertible). If A is a real matrix, then U is real (an orthogonal matrix) by Proposition **??**.

There are many other matrices M, for example $M = \sqrt{\Lambda}U^*$. Rectangular matrices M can be found too.

(iv) implies (i): a matrix M^*M is self-adjoint since $(M^*M)^* = M^*(M^*)^* = M^*M$ and it is positive definite since $\langle \mathbf{x}, M^*M\mathbf{x}\rangle = \langle M\mathbf{x}, M\mathbf{x}\rangle = \|M\mathbf{x}\| > 0$ for $\mathbf{x} \neq \mathbf{0}$.

The equivalence of (i) and (v) is not proved here[3]. □

3.4. **Negative definite, semidefinite and indefinite matrices.**

Definition 23. *A self-adjoint matrix $A = A^*$ is called* **negative definite** *if*

$$\langle \mathbf{x}, A\mathbf{x}\rangle < 0 \quad \textit{for all } \mathbf{x} \neq \mathbf{0}$$

Of course, A is positive definite if and only if $-A$ is negative definite, so the properties of negative definite matrices can be easily deduced from those of positive definite ones.

As a word of caution: in Theorem 22 part (v) becomes: *A is negative definite is equivalent to*
(v_-) the upper left submatrices of odd dimension have negative determinant, while those of even dimension have positive determinant.

Definition 24. *A self-adjoint matrix $A = A^*$ is called* **positive semidefinite** *if*

$$\langle \mathbf{x}, A\mathbf{x}\rangle \geq 0 \quad \textit{for all } \mathbf{x}$$

The analogue of Theorem 22 is, quite obviously:

Theorem 25. *Let $A = A^*$ be an $n\times n$ matrix. Then following are equivalent:*
(i') $\langle \mathbf{x}, A\mathbf{x}\rangle \geq 0$ *for all* $\mathbf{x}$*;*
(ii') all the eigenvalues of A are ≥ 0;
[(iii) is obviously not true if A has zero eigenvalues.]
(iv') there are n vectors (possibly dependent) $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{C}^n$ so that

$$A_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j\rangle, \quad i, j = 1, \ldots, n$$

in other words, $A = M^*M$ where $M = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ (possibly not of maximal rank);
(v') all the upper left $k \times k$ ($k = 1, 2, \ldots, n$) submatrices

$$[A_{1,1}],\ [A_{i,j}]_{i,j=1,2}, \ldots, [A_{i,j}]_{i,j=1,\ldots,k}, \ldots, [A_{i,j}]_{i,j=1,\ldots,n}$$

have nonnegative determinants.

Definition 26. *A self-adjoint matrix $A = A^*$ is called* **negative semidefinite** *if*

$$\langle \mathbf{x}, A\mathbf{x}\rangle \leq 0 \quad \textit{for all } \mathbf{x}$$

[3]For a proof see Strang's book.

Of course, A is positive semidefinite if and only if $-A$ is negative semidefinite, and an analogue of Theorem 25 can be obtained.

Definition 27. *A self-adjoint matrix which is neither positive- nor negative-semidefinite is called* **indefinite**.

Indefinite matrices have both positive and negative eigenvalues.

3.5. **Applications to critical points of several variables.**

3.5.1. *Functions of two variables.* We can (re)obtain the second derivative test for functions of two variables: if $f(x, y)$ has a critical point at (a, b), then its nature can be seen by looking at the Hessian matrix, §3.2,

$$Hf(a,b) = \begin{bmatrix} f_{xx}(a,b) & f_{xy}(a,b) \\ f_{yx}(a,b) & f_{yy}(a,b) \end{bmatrix}$$

if its determinant is not zero (meaning that no eigenvalue is zero).

If the mixed derivatives are equal, then $Hf(a, b)$ is a symmetric matrix. Using Theorem 22 (v), and its analogue for negative definite matrices we obtain the second derivative test:

• If $Hf(a, b)$ is positive definite, that is, if

$$f_{xx}(a,b) > 0, \quad f_{xx}(a,b)f_{yy}(a,b) - f_{xy}(a,b)^2 > 0$$

then (a, b) is a minimum.

• If $Hf(a, b)$ is negative definite, that is, if

$$f_{xx}(a,b) < 0, \quad f_{xx}(a,b)f_{yy}(a,b) - f_{xy}(a,b)^2 > 0$$

then (a, b) is a maximum.

• If $Hf(a, b)$ has one positive and one negative eigenvalue, that is, if its determinant is negative,

$$f_{xx}(a,b)f_{yy}(a,b) - f_{xy}(a,b)^2 < 0$$

then (a, b) is a saddle point.

3.5.2. *Functions of n variables.* If $f(\mathbf{x})$ is a function of n real variables, let $\mathbf{x} = \mathbf{a}$ be a critical point: $\nabla f(\mathbf{a}) = \mathbf{0}$. As discussed in §3.2, **if the Hessian matrix $Hf(\mathbf{a})$ has all eigenvalues nonzero**, then the nature of the critical point of f is the same as the nature of the point for the Hessian quadratic form $\langle \mathbf{x} - \mathbf{a}, Hf(\mathbf{a})(\mathbf{x} - \mathbf{a})\rangle$:

(i) $\mathbf{a}$ is a point of (local) minimum of f if and only if $\mathbf{y} = \mathbf{0}$ is a minimum for $q(\mathbf{y}) = \langle \mathbf{y}, Hf(\mathbf{a})\mathbf{y}\rangle$, hence if and only if $Hf(\mathbf{a})$ is positive definite;

(ii) $\mathbf{a}$ is a point of (local) maximum of f if and only if $\mathbf{y} = \mathbf{0}$ is a maximum for $q(\mathbf{y}) = \langle \mathbf{y}, Hf(\mathbf{a})\mathbf{y}\rangle$, hence if and only if $Hf(\mathbf{a})$ is negative definite;

(iii) $\mathbf{a}$ is a saddle point of f if and only if $\mathbf{0}$ is a saddle point for $q(\mathbf{y}) = \langle \mathbf{y}, Hf(\mathbf{a})\mathbf{y}\rangle$, hence if and only if $Hf(\mathbf{a})$ is indefinite.

3.6. **Application to differential equations: Lyapunov functions.** Consider a linear system of differential equation of first order

$$\frac{d\mathbf{y}}{dt} = M\mathbf{y}, \quad M \in \mathcal{M}_n(\mathbb{R}) \tag{15}$$

Recall that $\mathbf{0}$ is a stationary point, in the sense that $\mathbf{y}(t) = \mathbf{0}$ is a solution of 15), and if all the eigenvalues of M have negative real parts, then this stationary point is asymptotically stable.

Another method for proving stability (for linear, or nonlinear equations) is by finding a Lyapunov function. Consider a (possibly nonlinear) system

$$\frac{d\mathbf{y}}{dt} = \mathbf{F}(\mathbf{y}), \quad \mathbf{y} \in \mathbb{R}^n \tag{16}$$

with an equilibrium point $\mathbf{y}_0$: $F(\mathbf{y}_0) = \mathbf{0}$.

Definition 28. *A function $L : \mathbb{R}^n \to \mathbb{R}$ is called a Lyapunov function for the system (16) and the equilibrium point $\mathbf{0}$ if:*
(i) $L(\mathbf{y}_0) = 0$,
(ii) $L(\mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{y}_0$, and
(iii) L decreases along the solutions:

$$\frac{d}{dt} L(\mathbf{y}(t)) < 0 \tag{17}$$

for all $\mathbf{y}(t)$ solution of (16) stating close enough to $\mathbf{y}_0$.

Note that it is not needed that we actually know the solutions $\mathbf{y}(t)$ since

$$\frac{d}{dt} L(\mathbf{y}(t)) = \sum_{j=1}^{n} \frac{\partial L}{\partial y_j}(\mathbf{y}(t)) y_j'(t) = \langle \nabla L(\mathbf{y}(t)), F(\mathbf{y}(t)) \rangle$$

Then if

$$\langle \nabla L(\mathbf{x}), F(x) \rangle < 0 \ \text{ for all } \mathbf{x} \neq \mathbf{y}_0, \mathbf{x} \text{ close to } \mathbf{y}_0 \tag{18}$$

then (iii) is satisfied.

It is proved in the theory of differential equation that if a Lyapunov function exists, then $\mathbf{y}_0$ is stable.

Consider a linear system (15). A good candidate for a Lyapunov function[4] is the function $\|\mathbf{x}\|$. Intuitively, if $\|\mathbf{y}(t)\|$ decreases in time, then trajectories keep on approaching $\mathbf{0}$.

It is better to consider the norm squared instead, since it is easier to differentiate: let $L(\mathbf{x}) = \|\mathbf{x}\|^2$. Calculating as in §2.9 it is found that

$$\frac{d}{dt} \|\mathbf{y}(t)\|^2 = \langle (M + M^T)\mathbf{y}(t), \mathbf{y}(t) \rangle$$

[4]"There is no general method to construct or find a Lyapunov-candidate-function which proves the stability of an equilibrium, and the inability to find a Lyapunov function is inconclusive with respect to stability, which means, that not finding a Lyapunov function does not mean that the system is unstable." From Wikipedia.

Note that the matrix $M + M^T$ is symmetric. If $M + M^T$ is negative definite, then $\langle (M + M^T)\mathbf{y}(t), \mathbf{y}(t)\rangle < 0$ for all $\mathbf{y}(t) \neq \mathbf{0}$ which shows that $L(\mathbf{x}) = \|\mathbf{x}\|^2$ is a Lyapunov function and that $\mathbf{0}$ is asymptotically stable.

We used a Lyapunov function to show that if the symmetric matrix $M + M^T$ is negative definite then M has eigenvalues with negative real part. Can we find a simpler proof?

Proposition 29. *If the self-adjoint matrix $M + M^*$ negative definite then M has eigenvalues with negative real part.*

Proof:

Let λ be an eigenvalue of M and $\mathbf{v}$ a corresponding eigenvector. Then $\langle \mathbf{v}, (M + M^*)\mathbf{v}\rangle < 0$. On the other hand

$$\langle \mathbf{v}, (M + M^*)\mathbf{v}\rangle = \langle \mathbf{v}, M\mathbf{v}\rangle + \langle \mathbf{v}, M^*\mathbf{v}\rangle = \langle \mathbf{v}, M\mathbf{v}\rangle + \langle M\mathbf{v}, \mathbf{v}\rangle$$
$$= \langle \mathbf{v}, \lambda\mathbf{v}\rangle + \langle \lambda\mathbf{v}, \mathbf{v}\rangle = (\lambda + \overline{\lambda})\|\mathbf{v}\|^2 = 2\Re\lambda\, \|\mathbf{v}\|^2$$

hence $\Re\lambda < 0$. □

The converse is not true. For example the matrix

$$M = \begin{bmatrix} -1 & 4 \\ 0 & -1 \end{bmatrix}$$

has eigenvalues $-1, -1$, while

$$M + M^* = \begin{bmatrix} -2 & 4 \\ 4 & -2 \end{bmatrix}$$

has eigenvalues $2, -6$.

3.7. **Solving linear systems by minimization.** A scalar equation $Ax = b$ can be solved by minimization: its solution coincides with the point $x = x_m$ where the parabola $p(x) = \frac{1}{2}Ax^2 - bx$ has a minimum (assuming $A > 0$). To generalize this minimization solution to higher dimensions, note that the derivative of p is $p'(x) = Ax - b$, whose critical point is the solution of the linear equation $Ax = b$, and the second derivative is $p''(x) = A > 0$, which ensures that the critical point is a minimum. The idea is then to construct a function $p(\mathbf{x})$ whose gradient is $A\mathbf{x} - \mathbf{b}$, and Hessian is A:

Theorem 30. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a positive definite matrix, and $\mathbf{b} \in \mathbb{R}^n$.*
The quadratic form

$$p(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle - \langle \mathbf{x}, \mathbf{b}\rangle$$

has a global minimum.

The minimum is attained at a point $\mathbf{x}_m$ satisfying $A\mathbf{x}_m = \mathbf{b}$ and the minimum value is $p(\mathbf{x}_m) = -\frac{1}{2}\langle A^{-1}\mathbf{b}, \mathbf{b}\rangle$.

Proof:
The gradient of $p(\mathbf{x})$ equals $A\mathbf{x} - \mathbf{b}$ since from

$$p(\mathbf{x}) = p(x_1, \dots, x_n) = \frac{1}{2}\sum_{i,j=1,\dots,n} A_{ij}x_i x_j - \sum_{i=1,\dots,n} b_i x_i$$

we get, using that A is symmetric,

$$\frac{\partial p}{\partial x_k} = \frac{1}{2}\sum_{j=1,\dots,n} A_{kj}x_j + \frac{1}{2}\sum_{i=1,\dots,n} A_{ik}x_i - b_k = \sum_{j=1,\dots,n} A_{kj}x_j + b_k = (A\mathbf{x})_k - b_k$$

and the Hessian of $p(\mathbf{x})$ is A since $\frac{\partial^2 p}{\partial x_l \partial x_k} = A_{lk}$.

[Here is a more compact way to do this calculation:

$$\frac{\partial p}{\partial x_k} = \frac{1}{2}\langle \frac{\partial}{\partial x_k}\mathbf{x}, A\mathbf{x}\rangle + \frac{1}{2}\langle \mathbf{x}, A\frac{\partial}{\partial x_k}\mathbf{x}\rangle - \langle \frac{\partial}{\partial x_k}\mathbf{x}, \mathbf{b}\rangle$$

$$= \frac{1}{2}\langle \mathbf{e}_k, A\mathbf{x}\rangle + \frac{1}{2}\langle \mathbf{x}, A\mathbf{e}_k\rangle - \langle \mathbf{e}_k, \mathbf{b}\rangle = \langle \mathbf{e}_k, A\mathbf{x}\rangle - b_k = (A\mathbf{x})_k - b_k$$

and the Hessian of $p(\mathbf{x})$ is A since $\frac{\partial^2 p}{\partial x_j \partial x_k} = \frac{\partial}{\partial x_j}\langle \mathbf{e}_k, A\mathbf{x}\rangle = \langle \mathbf{e}_k, A\frac{\partial}{\partial x_j}\mathbf{x}\rangle = \langle \mathbf{e}_k, A\mathbf{e}_j\rangle = A_{jk}$.]

Note that the linear system $A\mathbf{x} = \mathbf{b}$ has a unique solution $\mathbf{x} = \mathbf{x}_m = A^{-1}\mathbf{b}$ (since A has no zero eigenvalues).

Write the Taylor expansion of p at $\mathbf{x} = \mathbf{x}_m$, which stops at the second order terms, since p is a polynomial of degree 2, hence derivatives of higher order vanish:

$$p(\mathbf{x}) = p(\mathbf{x}_m) + \langle \nabla p(\mathbf{x}_m), (\mathbf{x} - \mathbf{x}_m)\rangle + \frac{1}{2}\langle (\mathbf{x} - \mathbf{x}_m), (Hp)(\mathbf{x}_m)\,(\mathbf{x} - \mathbf{x}_m)\rangle$$

$$= p(\mathbf{x}_m) + \langle (A\mathbf{x}_m - \mathbf{b}), (\mathbf{x} - \mathbf{x}_m)\rangle + \frac{1}{2}\langle (\mathbf{x} - \mathbf{x}_m), A\,(\mathbf{x} - \mathbf{x}_m)\rangle$$

$$= p(\mathbf{x}_m) + \frac{1}{2}\langle (\mathbf{x}-\mathbf{x}_m), A\,(\mathbf{x}-\mathbf{x}_m)\rangle \quad > \; p(\mathbf{x}_m) \text{ for all } \mathbf{x} \neq \mathbf{x}_m$$

where the last inequality holds since A is positive definite; this means that $p(\mathbf{x}_m)$ is an absolute minimum. Furthermore

$$p(\mathbf{x}_{min}) = \frac{1}{2}\langle \mathbf{x}_{min}, A\mathbf{x}_{min}\rangle - \langle \mathbf{x}_{min}, \mathbf{b}\rangle = -\frac{1}{2}\langle \mathbf{x}_{min}, \mathbf{b}\rangle = -\frac{1}{2}\langle A^{-1}\mathbf{b}, \mathbf{b}\rangle$$

□

3.8. **Generalized eigenvalue problems.** The following type of problem appears in applications (for example they arise in discretizations of continuous problems):

Problem: given two $n \times n$ matrices A and B find the numbers λ so that

$$A\mathbf{v} = \lambda B\mathbf{v} \quad \text{for some } \mathbf{v} \neq \mathbf{0} \tag{19}$$

Are there such numbers λ, and if so, how many?

Clearly, such numbers λ must satisfy the equation $\det(A-\lambda B) = 0$, which is a polynomial in λ of degree at most n.

Since $\det(A - \lambda B) = \lambda^n \det(\frac{1}{\lambda}A - B)$ wee that the coefficient of λ^n is $\det(-B)$. Therefore, if B is not invertible, the degree of $\det(A - \lambda B)$ is less than n.

On the other hand, if B is invertible, then (19) is equivalent to $B^{-1}A\mathbf{v} = \lambda\mathbf{v}$ so λ and $\mathbf{v}$ are the eigenvalues and eigenvectors of the matrix $B^{-1}A$.

For real matrices, A symmetric and B positive definite, the eigenvalues/vectors of the generalized problem have special properties, which are derived below.

If B is positive definite and real, then $B = M^T M$ for an invertible real matrix M, by Theorem 22 (iv). Equation (19) is $A\mathbf{v} = \lambda M^T M\mathbf{v}$ therefore $(M^T)^{-1}A\mathbf{v} = \lambda M\mathbf{v}$, where denoting $M\mathbf{v} = \mathbf{y}$ and $C = M^{-1}$ the equation becomes

$$(C^T AC)\mathbf{y} = \lambda\mathbf{y} \tag{20}$$

which is an eigenvalue problem for the symmetric matrix $C^T AC$: there are n real eigenvalues, and n orthonormal eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. Going back through the substitution, let $\mathbf{v}_j$'s be so that $\mathbf{u}_j = M\mathbf{v}_j$, we find the eigenvectors of the generalized problem (19) as $\mathbf{v}_1, \ldots, \mathbf{v}_n$ where

$$\delta_{ij} = \langle \mathbf{u}_j, \mathbf{u}_i\rangle = \langle M\mathbf{v}_j, M\mathbf{v}_i\rangle = \langle \mathbf{v}_j, M^T M\mathbf{v}_i\rangle = \langle \mathbf{v}_j, B\mathbf{v}_i\rangle$$

and therefore

$$\langle \mathbf{v}_j, B\mathbf{v}_i\rangle = \delta_{ij}, \quad \text{for } i, j = 1, \ldots, n \tag{21}$$

meaning that the $\mathbf{v}_j$ are B-orthonormal, i.e. orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle_B$, see Theorem 22 (iii).

Also

$$\langle \mathbf{v}_j, A\mathbf{v}_i\rangle = \langle \mathbf{v}_j, \lambda_i B\mathbf{v}_i\rangle = \lambda_i \langle \mathbf{v}_j, B\mathbf{v}_i\rangle = \lambda_i\, \delta_{ij}$$

so

$$\langle \mathbf{v}_j, A\mathbf{v}_i\rangle = \delta_{ij}\lambda_i, \quad \text{for } i,j = 1,\dots,n \tag{22}$$

meaning that the vectors $\mathbf{v}_1,\dots,\mathbf{v}_n$ are also A-orthogonal.

In matrix notation, if $S = [\mathbf{v}_1,\dots,\mathbf{v}_n]$ is the matrix with columns the generalized eigenvectors, then (22) is $S^TAS = \Lambda$, and (21) is $S^TBS = I$.

The matrices A and B are simultaneously diagonalized by a congruence transformation! We proved:

Theorem 31. *Consider the real matrices: A symmetric and B positive definite. Then the quadratic forms*

$$q(\mathbf{x}) = \langle \mathbf{x}, A\mathbf{x}\rangle,\ r(\mathbf{x}) = \langle \mathbf{x}, B\mathbf{x}\rangle$$

are simultaneously diagonalizable.

More precisely, there exists an invertible matrix S so that $S^TAS = \Lambda$ and $S^TBS = I$ and if $\mathbf{x} = S\mathbf{y}$ then

$$q(\mathbf{x}) = q(S\mathbf{y}) = \langle \mathbf{y}, \Lambda\mathbf{y}\rangle, \quad r(\mathbf{x}) = r(S\mathbf{y}) = \langle \mathbf{y}, \mathbf{y}\rangle$$

The diagonal matrix Λ consists of the generalized eigenvalues solutions of the problem $A\mathbf{v} = \lambda B\mathbf{v}$ and the columns of S are its generalized eigenvectors.

4. The Rayleigh's principle. The minimax theorem for the eigenvalues of a self-adjoint matrix

Eigenvalues of self-adjoint matrices are easy to calculate. This section shows how this is done using a minimization, or maximization procedure.

4.1. The Rayleigh's quotient.

Definition 32. *Let $A = A^*$ be a self-adjoint matrix.* **The Rayleigh's quotient** *is the function*

$$R(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\|\mathbf{x}\|^2}, \quad \text{for } \mathbf{x} \neq \mathbf{0}$$

Note that

$$R(\mathbf{x}) = \langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, A\frac{\mathbf{x}}{\|\mathbf{x}\|} \rangle = \langle \mathbf{u}, A\mathbf{u} \rangle \quad \text{where } \mathbf{u} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$$

so in fact, it suffices to define the Rayleigh's quotient on unit vectors.

The set of unit vectors in $\mathbb{R}^n$ (or in $\mathbb{C}^n$), is called the $n-1$ dimensional sphere in $\mathbb{R}^n$ (or in $\mathbb{C}^n$):

$$S_F^{n-1} = \{\mathbf{u} \in F^n \,|\, \|\mathbf{u}\| = 1\}$$

For example, the sphere in $\mathbb{R}^2$ is the unit circle (it is a curve, it has dimension 1), the sphere in $\mathbb{R}^3$ is the unit sphere (it is a surface, it has dimension 2); for higher dimensions we need to use our imagination.

4.2. Extrema of the Rayleigh's quotient.

4.2.1. *Closed sets, bounded sets, compact sets.* You probably know very well the extreme value theorem for continuous function on the real line:

Theorem 33. The extreme value theorem in dimension one.
A functions $f(x)$ which is continuous on a closed and bounded interval $[a, b]$ has a maximum value (and a minimum value) on $[a, b]$.

To formulate an analogue of this theorem in higher dimensions we need to specify what we mean by a *closed* set and by a *bounded* set.

Definition 34. *A set S is called* **closed** *if it contains all its limit points: if a sequence of points in S, $\{\mathbf{x}_k\}_k \subset S$ converges, $\lim_{k\to\infty} = \mathbf{x}$, then the limit $\mathbf{x}$ is also in S.*

For example, the intervals $[2, 6]$ and $[2, +\infty)$ are closed in $\mathbb{R}$, but $[2, 6)$ is not closed. The closed unit disk $\{\mathbf{x} \in \mathbb{R}^2 \,|\, \|\mathbf{x}\| \leq 1\}$ is closed in $\mathbb{R}^2$, but the punctured disk $\{\mathbf{x} \in \mathbb{R}^2 \,|\, 0 < \|\mathbf{x}\| \leq 1\}$ or the open disk $\{\mathbf{x} \in \mathbb{R}^2 \,|\, \|\mathbf{x}\| < 1\}$ are not closed sets.

Definition 35. *A set S is called* **bounded** *if there is a number larger than all the lengths of the vectors in S: there is $M > 0$ so that $\|\mathbf{x}\| \leq M$ for all $\mathbf{x} \in S$.*

For example, the intervals $[2,6]$ and $[2,6)$ are bounded in $\mathbb{R}$, but $[2,+\infty)$ is not. The square $\{\mathbf{x}\in\mathbb{R}^2\,|\,|x_1|<1, \text{ and } |x_2|<1\}$ is bounded in $\mathbb{R}^2$, but the vertical strip $\{\mathbf{x}\in\mathbb{R}^2\,|\,|x_1|<1\}$ is not.

Theorem 36. The extreme value theorem in finite dimensions.
A functions $f(x)$ which is continuous on a closed and bounded set S in $\mathbb{R}^n$ or $\mathbb{C}^n$ has a maximum value (and a minimum value) on S.

In infinite dimensions Theorem 36 is not true in this form. A more stringent condition on the set S is needed to ensure existence of extreme values of continuous functions on S (the set must be *compact*).

It is intuitively clear (and rigorously proved in mathematical analysis) that any sphere in F^n is a closed and bounded set.

4.2.2. *Minimum and maximum of the Rayleigh's quotient.* The Rayleigh's quotient calculated on unit vectors is a quadratic polynomial, and therefore it is a continuous function on the unit sphere, and therefore

Proposition 37. *The Rayleigh's quotient has a maximum and a minimum.*

What happens if A is not self-adjoint? Recall that the quadratic form $\langle\mathbf{x},A\mathbf{x}\rangle$ has the same value if we replace A by its self-adjoint part, $\frac{1}{2}(A+A^*)$, therefore the Rayleigh's quotient of A is the same as the Rayleigh's quotient of the self-adjoint part of A (information about A is lost).

The extreme values of the Rayleigh's quotient are linked to the eigenvalues of the self-adjoint matrix A. To see this, diagonalize the quadratic form $\langle\mathbf{x},A\mathbf{x}\rangle$: consider a unitary matrix U which diagonalizes A:

$$U^*AU=\Lambda=diag(\lambda_1,\dots,\lambda_n)$$

In the new coordinates $\mathbf{y}=U^*\mathbf{x}$ we have

$$\langle\mathbf{x},A\mathbf{x}\rangle=\langle\mathbf{x},U\Lambda U^*\mathbf{x}\rangle=\langle U^*\mathbf{x},\Lambda U^*\mathbf{x}\rangle=\langle\mathbf{y},\Lambda\mathbf{y}\rangle=\sum_{j=1}^n\lambda_j|y_j|^2$$

which together with $\|\mathbf{x}\|=\|U\mathbf{y}\|=\|\mathbf{y}\|$ give

$$R(\mathbf{x})=R(U\mathbf{y})=\frac{\sum_{j=1}^n\lambda_j|y_j|^2}{\|\mathbf{y}\|^2}=\sum_{j=1}^n\lambda_j\frac{|y_j|^2}{\|\mathbf{y}\|^2}\equiv R_U(\mathbf{y}) \tag{23}$$

Since A is self-adjoint, its eigenvalues λ_j are real; assume them ordered in an increasing sequence:

$$\lambda_1\le\lambda_2\le\ldots\le\lambda_n \tag{24}$$

Then clearly

$$\sum_{j=1}^n\lambda_j|y_j|^2\le\lambda_n\sum_{j=1}^n|y_j|^2=\lambda_n\|\mathbf{y}\|^2$$

and

$$\sum_{j=1}^{n} \lambda_j |y_j|^2 \geq \lambda_1 \sum_{j=1}^{n} |y_j|^2 = \lambda_1 \|\mathbf{y}\|^2$$

therefore

$$\lambda_1 \leq R(\mathbf{x}) \leq \lambda_n \quad \text{for all } \mathbf{x} \neq \mathbf{0}$$

Equalities are attained since $R_U(\mathbf{e}_1) = \lambda_1$ and $R_U(\mathbf{e}_n) = \lambda_n$. Going to coordinates $\mathbf{x}$ minimum is attained for $\mathbf{x} = U\mathbf{e}_1 = \mathbf{u}_1$ = eigenvector corresponding to λ_1 since $R(\mathbf{u}_1) = R_U(\mathbf{e}_1) = \lambda_1$, and for $\mathbf{x} = U\mathbf{e}_n = \mathbf{u}_n$ = eigenvector corresponding to λ_n, maximum is attained since $R(\mathbf{u}_n) = R_U(\mathbf{e}_n) = \lambda_n$. This proves:

Theorem 38. *If A is a self-adjoint matrix then*

$$\max \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\|\mathbf{x}\|^2} = \lambda_n \textit{ the max eigenvalue of } A, \textit{ attained for } \mathbf{x} = \mathbf{u}_n$$

and

$$\min \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\|\mathbf{x}\|^2} = \lambda_1 \textit{ the min eigenvalue of } A, \textit{ attained for } \mathbf{x} = \mathbf{u}_1$$

As an important consequence in numerical calculations: the maximum eigenvalue of A can be found by solving a maximization problem, and the minimum eigenvalue - by a minimization problem.

4.3. **The minimax principle.** Reducing the dimension of A we can find all the eigenvalues, one by one. Consider the eigenvalues (24) of A and the corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ which form an orthonormal basis: $F^n = \oplus_{j=1}^{n} F\mathbf{u}_j$.

We saw that $\max R(\mathbf{x}) = \lambda_n = R(\mathbf{u}_n)$. The subspace $Sp(\mathbf{u}_n)$ and its orthogonal complement $Sp(\mathbf{u}_n)^\perp = \oplus_{j=1}^{n-1} Sp(\mathbf{u}_j)$ are invariant under A.

Consider A as a linear transformation of the $n-1$ dimensional vector space $Sp(\mathbf{u}_n)^\perp$ to itself: its eigenvalues are $\lambda_1, \ldots, \lambda_{n-1}$, the largest being λ_{n-1}. We reduced the dimension!

Using Theorem 38 for A as a linear transformation on the vector space $Sp(\mathbf{u}_n)^\perp$ it follows that

$$\max_{\mathbf{x} \in Sp(\mathbf{u}_n)^\perp} R(\mathbf{x}) = \lambda_{n-1} \quad \text{is attained for } \mathbf{x} = \mathbf{u}_{n-1} \tag{25}$$

The statement $\mathbf{x} \in Sp(\mathbf{u}_n)^\perp$ can be formulated as the constraint $\langle \mathbf{x}, \mathbf{u}_n \rangle = 0$:

$$\max_{\mathbf{x}: \langle \mathbf{x}, \mathbf{u}_n \rangle = 0} R(\mathbf{x}) = \lambda_{n-1}$$

We can do even better: we can *obtain* λ_{n-1} *without knowing* $\mathbf{u}_n$ *or* λ_n. To achieve this, subject $\mathbf{x}$ to *any* constraint: $\langle \mathbf{x}, \mathbf{z} \rangle = 0$ for some $\mathbf{z} \neq \mathbf{0}$.

It is easier to see what happens in coordinates $\mathbf{y} = U^*\mathbf{x}$ in which A is diagonal. The constraint $\langle \mathbf{x}, \mathbf{z} \rangle = 0$ is equivalent to $\langle \mathbf{y}, \mathbf{w} \rangle = 0$ where $\mathbf{w} = U\mathbf{z}$ is some nonzero vector.

Step I. We have

$$\max_{\mathbf{y}:\,\langle \mathbf{y},\mathbf{w}\rangle=0} R_U(\mathbf{y}) \geq \lambda_{n-1} \qquad \text{for all } \mathbf{w}\neq \mathbf{0} \tag{26}$$

since there is some nonzero vector $\tilde{\mathbf{y}}$ belonging to both the $n-1$ dimensional subspace $\{\mathbf{y}:\,\langle \mathbf{y},\mathbf{w}\rangle=0\}$ and the two dimensional subspace $F\mathbf{e}_{n-1}\oplus F\mathbf{e}_n$. (Such a vector is easy to find: $\tilde{\mathbf{y}}=(0,\dots,0,y_{n-1},y_n)^T$ with $\langle \tilde{\mathbf{y}},\mathbf{w}\rangle=0$; if $w_n\neq 0$ take $y_{n-1}=1$ and $y_n=-w_{n-1}/w_n$, and if $w_n=0$ take $y_{n-1}=0,\ y_n=1$). Using formula (23)

$$R_U(\tilde{\mathbf{y}}) = \frac{\lambda_{n-1}|y_{n-1}|^2+\lambda_n|y_n|^2}{|y_{n-1}|^2+|y_n|^2} \geq \frac{\lambda_{n-1}|y_{n-1}|^2+\lambda_{n-1}|y_n|^2}{|y_{n-1}|^2+|y_n|^2} = \lambda_{n-1} \tag{27}$$

proving (26).

Step II. Inequality (26) implies that

$$\min_{\mathbf{w}\neq\mathbf{0}}\ \max_{\mathbf{y}:\,\langle \mathbf{y},\mathbf{w}\rangle=0} R_U(\mathbf{y}) \geq \lambda_{n-1} \tag{28}$$

Step III. We now show that equality is attained in (28) for special $\mathbf{w}$.

For $\mathbf{w}=\mathbf{e}_n$ we have, by (25),

$$\max_{\mathbf{y}:\,\langle \mathbf{y},\mathbf{e}_n\rangle=0} R_U(\mathbf{y}) = \lambda_{n-1} \ \text{ attained for } \mathbf{y}=\mathbf{e}_{n-1}$$

hence in (28) there is equality

$$\min_{\mathbf{w}\neq\mathbf{0}}\ \max_{\mathbf{y}:\,\langle \mathbf{y},\mathbf{w}\rangle=0} R_U(\mathbf{y}) = \lambda_{n-1}$$

In a similar way it is shown that λ_{n-2} is obtained by a minimum-maximum process, but with two constraints:

$$\min_{\mathbf{w}_1,\mathbf{w}_2\neq\mathbf{0}}\ \max_{\substack{\langle \mathbf{y},\mathbf{w}_1\rangle=0\\ \langle \mathbf{y},\mathbf{w}_2\rangle=0}} R_U(\mathbf{y}) = \lambda_{n-2} \tag{29}$$

Indeed, consider a nonzero vector $\tilde{\mathbf{y}}=(0,\dots,0,y_{n-2},y_{n-1},y_n)^T$ satisfying $\langle \tilde{\mathbf{y}},\mathbf{w}_1\rangle=0$ and $\langle \tilde{\mathbf{y}},\mathbf{w}_2\rangle=0$. Then in formula (23)

$$R_U(\tilde{\mathbf{y}}) = \frac{\lambda_{n-2}|y_{n-2}|^2+\lambda_{n-1}|y_{n-1}|^2+\lambda_n|y_n|^2}{|y_{n-2}|^2+|y_{n-1}|^2+|y_n|^2}$$

$$\geq \frac{\lambda_{n-2}|y_{n-2}|^2+\lambda_{n-2}|y_{n-1}|^2+\lambda_{n-2}|y_n|^2}{|y_{n-2}|^2+|y_{n-1}|^2+|y_n|^2} = \lambda_{n-2}$$

which shows that

$$\max_{\substack{\langle \mathbf{y},\mathbf{w}_1\rangle=0\\ \langle \mathbf{y},\mathbf{w}_2\rangle=0}} R_U(\mathbf{y}) \geq \lambda_{n-2} \tag{30}$$

Since for $\mathbf{w}_1=\mathbf{e}_n$ and $\mathbf{w}_2=\mathbf{e}_{n-1}$ we have equality in (30), and this implies (29).

Step by step, adding one extra constraint, the minimax procedure yields the next largest eigenvalue.

Going back to the variable $\mathbf{x}$ it is found that:

Theorem 39. The minimax principle

Let A be a self-adjoint matrix, with its eigenvalues numbered in an increasing sequence:

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

corresponding to the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Then its Rayleigh's quotient

$$R(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x} \rangle}{\|\mathbf{x}\|^2}$$

satisfies

$$\max_{\mathbf{x} \neq 0} R(\mathbf{x}) = \lambda_n$$

$$\min_{\mathbf{z} \neq \mathbf{0}} \max_{\langle \mathbf{x}, \mathbf{z} \rangle = 0} R(\mathbf{x}) = \lambda_{n-1}$$

$$\min_{\mathbf{z}_1, \mathbf{z}_2 \neq \mathbf{0}} \max_{\substack{\langle \mathbf{x}, \mathbf{z}_1 \rangle = 0 \\ \langle \mathbf{x}, \mathbf{z}_2 \rangle = 0}} R(\mathbf{x}) = \lambda_{n-2}$$

and in general

$$\min_{\mathbf{z}_1, \ldots \mathbf{z}_k \neq \mathbf{0}} \max_{\substack{\langle \mathbf{x}, \mathbf{z}_1 \rangle = 0 \\ \vdots \\ \langle \mathbf{x}, \mathbf{z}_k \rangle = 0}} R(\mathbf{x}) = \lambda_{n-k}, \qquad k = 1, 2, \ldots, n-1$$

Remark. Sometimes the minimax principle is formulated as

$$\min_{V_j} \max_{\mathbf{x} \in V_j} R(\mathbf{x}) = \lambda_j, \quad j = 1, \ldots, n$$

where V_j denotes an arbitrary subspace of dimension j.

The two formulations are equivalent since the set

$$V_{n-k} = \{\mathbf{x} \,|\, \langle \mathbf{x}, \mathbf{z}_1 \rangle = 0, \ldots, \langle \mathbf{x}, \mathbf{z}_k \rangle = 0\}$$

is a vector space of dimension $n - k$ if $\mathbf{z}_1, \ldots, \mathbf{z}_k$ are linearly independent.

A similar construction starting with the lowest eigenvalue produces:

Theorem 40. The maximin principle

Under the assumptions of Theorem 39

$$\min_{\mathbf{x} \neq 0} R(\mathbf{x}) = \lambda_1$$

$$\max_{\mathbf{z} \neq \mathbf{0}} \min_{\langle \mathbf{x}, \mathbf{z} \rangle = 0} R(\mathbf{x}) = \lambda_2$$

and in general

$$\max_{\mathbf{z}_1,\dots\mathbf{z}_k\neq\mathbf{0}} \quad \min_{\substack{\langle \mathbf{x},\mathbf{z}_1\rangle = 0\\ \vdots \\ \langle \mathbf{x},\mathbf{z}_k\rangle = 0}} \quad R(\mathbf{x}) = \lambda_{k+1}, \qquad k = 1,2,\dots,n-1$$

4.4. **The minimax principle for the generalized eigenvalue problem.** Suppose $\lambda_1 \le \lambda_1 \le \ldots \le \lambda_n$ are eigenvalues for the problem

$$A\mathbf{v} = \lambda B\mathbf{v}, \quad A \text{ symmetric, } B \text{ positive definite} \tag{31}$$

It was seen in §3.8 that if $S = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ is the matrix whose columns are the generalized eigenvectors of the problem (31), then both matrices A and B are diagonalized using a congruence transformation: $S^T A S = \Lambda$ and $S^T B S = I$.

Defining

$$R(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{\langle \mathbf{x}, B\mathbf{x}\rangle}$$

it is found that in coordinates $\mathbf{x} = S\mathbf{y}$:

$$R(\mathbf{x}) = R(S\mathbf{y}) = \frac{\langle S\mathbf{y}, AS\mathbf{y}\rangle}{\langle S\mathbf{y}, BS\mathbf{y}\rangle} = \frac{\langle \mathbf{y}, S^T AS\mathbf{y}\rangle}{\langle \mathbf{y}, S^T BS\mathbf{y}\rangle} = \frac{\lambda_1 y_1^2 + \ldots + \lambda_n y_n^2}{y_1^2 + \ldots + y_n^2}$$

and therefore

$$\max R(\mathbf{x}) = \lambda_n, \quad \min R(\mathbf{x}) = \lambda_1$$

5. Singular Value Decomposition

5.1. **Rectangular matrices.**

For rectangular matrices M the notions of eigenvalue/vector cannot be defined. However, the products MM^* and/or M^*M (which are square, even self-adjoint, and even positive semi-definite matrices) carry a lot of information about M:

Proposition 41. *Let M be an $m \times n$ matrix. Then*

$$\mathcal{N}(M^*M) = \mathcal{N}(M) \tag{32}$$

$$\mathcal{R}(MM^*) = \mathcal{R}(M) \tag{33}$$

Proof. To show (32), let $\mathbf{x} \in \mathcal{N}(M^*M)$; then $M^*M\mathbf{x} = \mathbf{0}$, so that $0 = \langle M^*M\mathbf{x}, \mathbf{x}\rangle = \langle M\mathbf{x}, M\mathbf{x}\rangle$ which implies $M\mathbf{x} = \mathbf{0}$, showing that $\mathcal{N}(M^*M) \subset \mathcal{N}(M)$. The converse inclusion is immediate.

To show (33), note that (32), used for M interchanged with M^* implies that $\mathcal{N}(MM^*) = \mathcal{N}(M^*)$, hence $\mathcal{N}(MM^*)^\perp = \mathcal{N}(M^*)^\perp$, which is exactly (33) (recall that for any linear transformation L we have $\mathcal{N}(L^*)^\perp = \mathcal{R}(L)$). □

Moreover, MM^* and M^*M have the same nonzero eigenvalues:

Proposition 42. *Let M be an $m \times n$ matrix. The matrices MM^* and M^*M are positive semi-definite. Moreover, they have the same nonzero eigenvalues (with the same multiplicity).*

More precisely, let $\lambda_1, \dots, \lambda_r$ be the positive eigenvalues. If $M^*M\mathbf{v}_j = \lambda_j\mathbf{v}_j$ with $\mathbf{v}_1, \dots, \mathbf{v}_r$ an orthonormal set, then $MM^*\mathbf{u}_j = \lambda_j\mathbf{u}_j$ for $\mathbf{u}_j = \frac{1}{\sqrt{\lambda_j}}M\mathbf{v}_j$ and $\mathbf{u}_1, \dots, \mathbf{u}_r$ is an orthonormal set.

Proof. MM^* and M^*M obviously self-adjoint; they are positive semi-definite since $\langle \mathbf{x}, M^*M\mathbf{x}\rangle = \langle M\mathbf{x}, M\mathbf{x}\rangle \geq 0$ and $\langle \mathbf{x}, MM^*\mathbf{x}\rangle = \langle M^*\mathbf{x}, M^*\mathbf{x}\rangle \geq 0$.

Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ be an orthonormal set of eigenvectors of M^*M, the first r corresponding to nonzero eigenvalues: $M^*M\mathbf{v}_j = \lambda_j\mathbf{v}_j$ with $\lambda_j > 0$, for $j = 1, \dots, r$ and $M^*M\mathbf{v}_j = \mathbf{0}$ for $j > r$.

Applying M we discover that $MM^*M\mathbf{v}_j = \lambda_j M\mathbf{v}_j$ with $\lambda_j > 0$, for $j = 1, \dots, r$ and $MM^*M\mathbf{v}_j = \mathbf{0}$ for $j > r$ which would mean that $M\mathbf{v}_j$ are eigenvectors to MM^* corresponding to the eigenvalue λ_j provided we ensure that $M\mathbf{v}_j \neq \mathbf{0}$. This is true for $j \leq r$ by (32).

Also, all $M\mathbf{v}_1, \dots, M\mathbf{v}_r$ are mutually orthogonal, since $\langle M\mathbf{v}_j, M\mathbf{v}_i\rangle = \langle \mathbf{v}_j, M^*M\mathbf{v}_i\rangle = \lambda_i\delta_{ij}$ so $M\mathbf{v}_j \perp M\mathbf{v}_i$ for all $i \neq j \leq r$, and $\|M\mathbf{v}_j\|^2 = \lambda_j$. Therefore, all the nonzero eigenvalues of M^*M are also eigenvalues for MM^*, with corresponding orthonormal eigenvectors $\mathbf{u}_j = \frac{1}{\sqrt{\lambda_j}}M\mathbf{v}_j$, $j = 1, \dots, r$.

The same argument can be applied replacing M by M^*, showing that indeed, MM^* and M^*M have the same nonzero eigenvalues and with the same multiplicity. □

5.2. **The SVD theorem.** We are going to bring any $m \times n$ matrix M to a (rectangular) diagonal form by writing $M = U\Sigma V^*$ where Σ is a diagonal $m \times n$ matrix, and U and V are unitary (of obvious dimensions). The diagonal elements σ_j of Σ are called **the singular values of** M.

The SVD has a myriad applications in filtering, image reconstruction, image compression, statistics, to name just a few.

Theorem 43. Singular Value Decomposition

Let M be an $m \times n$ matrix. Then

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

where:

- *U is a unitary matrix whose columns are eigenvectors of MM^**
- *V is a unitary matrix whose columns are eigenvectors of M^*M*
- *Σ is an $m \times n$ diagonal matrix*

More precisely:

∘ *if $U = [\mathbf{u}_1, \dots, \mathbf{u}_r, \mathbf{u}_{r+1}, \dots, \mathbf{u}_m]$ and $V = [\mathbf{v}_1, \dots, \mathbf{v}_r, \mathbf{v}_{r+1}, \dots, \mathbf{v}_n]$ then for $j = 1, \dots r$ the vectors $\mathbf{u}_j$ and $\mathbf{v}_j$ correspond to the eigenvalue $\lambda_j \neq 0$ while all the others correspond to the eigenvalue 0.*

∘ *The diagonal matrix Σ has $\Sigma_{jj} = \sigma_j = \sqrt{\lambda_j}$ for $j = 1, \dots, r$, and all other elements are 0.*

∘ *Also, $\mathbf{u}_j = \frac{1}{\sigma_j} M\mathbf{v}_j$ for $j = 1, \dots, r$.*

Remarks. 1. $M^*M = V\Lambda_n V^*$ and $MM^* = U\Lambda_m U^*$ where $\Lambda_{m,n}$ are diagonal matrices with entries $\lambda_1, \dots, \lambda_r$ and 0 everywhere else.

2. The singular values are preferred be listed in decreasing order $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ for reasons coming from applications, see §5.5.

Proof of Theorem 43.

Let $\mathbf{v}_1, \dots, \mathbf{v}_r$ and $\mathbf{u}_1, \dots, \mathbf{u}_r$ be as in Proposition 42; $\mathbf{u}_{r+1}, \dots, \mathbf{u}_m$ and $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$correspond to the eigenvalue 0.

Calculating

$$U^*MV = \begin{bmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_m^* \end{bmatrix} M \left[\mathbf{v}_1, \dots, \mathbf{v}_n\right] = \begin{bmatrix} \mathbf{u}_1^* \\ \vdots \\ \mathbf{u}_m^* \end{bmatrix} \left[M\mathbf{v}_1, \dots, M\mathbf{v}_n\right] = \Sigma$$

where Σ is a matrix with elements $\Sigma_{ij} = \mathbf{u}_i^* M\mathbf{v}_j$.

For $j > r$ we have $M^*M\mathbf{v}_j = \mathbf{0}$, hence by (41) also $M\mathbf{v}_j = \mathbf{0}$, hence $\Sigma_{ij} = 0$, while for $j \leq r$ we have $\mathbf{u}_i^* M\mathbf{v}_j = \mathbf{u}_i^*(\sqrt{\lambda_j})\mathbf{u}_j = \sqrt{\lambda_j}\delta_{ij}$, showing that Σ is the diagonal matrix stated. □

5.3. **Examples and applications of SVD.**

Example 1. How does the SVD look like for a square, diagonal matrix? Say

$$M = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} \tag{34}$$

In this case

$$MM^* = \begin{bmatrix} |a_1|^2 & 0 \\ 0 & |a_2|^2 \end{bmatrix} = M^*M$$

therefore $\sigma_j = |a_j|$, $V = I$, and $\mathbf{u}_j = \frac{1}{\sigma_j} M\mathbf{e}_j = \frac{a_j}{|a_j|}\mathbf{e}_j$.

By the polar decomposition of complex numbers, write $a_j = |a_j|e^{i\theta_j}$ then $\mathbf{u}_j = e^{i\theta_j}\mathbf{e}_j$ and the SVD is

$$\begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} = \begin{bmatrix} e^{i\theta_1} & 0 \\ 0 & e^{i\theta_2} \end{bmatrix} \begin{bmatrix} |a_1| & 0 \\ 0 & |a_2| \end{bmatrix}$$

which is called the polar decomposition of the matrix (34).

In general:

Proposition 44. Polar decomposition of square matrices.

Every square matrix M can be decomposed as $M = US$ with U unitary and S positive semidefinite.

Proof.

Writing the SDV of the matrix $M = U\Sigma V^* = (UV^*)(V\Sigma V^*)$ which is the polar decomposition since UV^* is a unitary matrix and $V\Sigma V^*$ is a self-adjoint matrix with non-negative eigenvalues. □

Example 2. A rectangular diagonal matrix, say

$$\begin{bmatrix} a_1 & 0 \\ 0 & a_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} e^{i\theta_1} & 0 & 0 \\ 0 & e^{i\theta_2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} |a_1| & 0 \\ 0 & |a_2| \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Example 3. A column matrix

$$\begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \sqrt{5} \\ 0 \end{bmatrix} [1]$$

Example 4. An orthogonal matrix Q is its own SVD since $QQ^* = Q^*Q = I$ hence $V = I$, $\sigma_j = 1$ and $U = Q$.

5.4. **The matrix of an oblique projection.** Recall that a square matrix P is a projection if $P^2 = P$; then P projects onto $R = \mathcal{R}(P)$, parallel to $N = \mathcal{N}(P)$.

For given complementary subspaces R and N a simple formula for the matrix of P can be obtained from the singular value decomposition.

Since the eigenvalues of P can only be 0 or 1, then all $\sigma_j = 1$.

The vectors $\mathbf{v}_1, \dots, \mathbf{v}_r$ in Theorem 43 form an orthonormal basis for $\mathcal{R}(P^*P) = \mathcal{N}(P^*P)^\perp = N^\perp$, and $\mathbf{u}_1 = P\mathbf{v}_1, \dots, \mathbf{u}_r = P\mathbf{v}_r$ form an orthonormal basis for $\mathcal{R}(PP^*) = \mathcal{R}(P) = R$. Split the matrices U, V into blocks, the first one containing the first r columns: $U = [U_A\ U']$, $V = [V_B\ V']$, and since Σ has its upper-left $r \times r$ diagonal sub matrix equal to the identity, the SDV of P becomes

$$P = U\Sigma V^* = [U_A\ U'] \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_B^* \\ 0 \end{bmatrix} = U_A V_B^*$$

Note that $V_B^* U_A = I$. Indeed, the elements of this matrix are $(V_B^* U_A)_{i,j} = \langle \mathbf{v}_i, \mathbf{u}_j \rangle = \frac{1}{\sigma_j} \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \frac{1}{\sigma_j} \delta_{ij}$ and for projections $\sigma_j = 1$.

Let $\mathbf{y}_1, \dots, \mathbf{y}_r$ be any basis of R; then $B = [\mathbf{y}_1, \dots, \mathbf{y}_r] = V_B S$ for some invertible $r \times r$ matrix S. Similarly, if $\mathbf{x}_1, \dots, \mathbf{x}_r$ is any basis of $N^\perp$, then $A = [\mathbf{x}_1, \dots, \mathbf{x}_r] = U_A T$ for some invertible $r \times r$ matrix T. Then $A(B^*A)^{-1}B^*$ is the matrix of P since

$$A(B^*A)^{-1}B^* = U_A T (S^* V_B^* U_A T)^{-1} S^* V_B^* = U_A (V_B^* U_A)^{-1} V_B^* = U_A I V_B^* = P$$

5.5. **Low-rank approximations, image compression.** Suppose an $m \times n$ matrix M is to be approximated by a matrix X of same dimensions, but lower rank k. If $M = U\Sigma V^*$ with singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k \geq \dots \geq \sigma_r$, let $X = U\Sigma_k V^*$ where Σ_k has the same singular values $\sigma_1, \dots, \sigma_k$ and 0 everywhere else. Then the sum of the squares of the singular values of $M - X$ is minimum among all matrices $m \times n$ of rank k (in the sense that the Frobenius norm of $M - X$ is minimum).

This low rank approximations are used in image compression, noise filtering and many other applications.

6. Pseudoinverse

There are many ways to define a matrix which behaves, in some sense, like the inverse of a matrix which is not invertible. This section describes the Moore-Penrose pseudoinverse.

Finding the "best fit" solution (in the least square sense) to a possibly overdetermined linear system $M\mathbf{x} = \mathbf{b}$ yields a vector $\mathbf{x}^+$ which depends linearly on $\mathbf{b}$, hence there is a matrix M^+ so that $\mathbf{x} = M^+\mathbf{b}$; this is the Moore-Penrose pseudoinverse of M.

Recall the construction of this solution.

Step I. If $M\mathbf{x} = \mathbf{b}$ is overdetermined (i.e. has no solutions) this is because $\mathbf{b} \notin \mathcal{R}(M)$. Then find $\overline{\mathbf{x}}$ so that $\|M\overline{\mathbf{x}} - \mathbf{b}\|$ is minimum. This happens if $M\overline{\mathbf{x}} = P\mathbf{b}$ where $P\mathbf{b}$ is the orthogonal projection of $\mathbf{b}$ on $\mathcal{R}(M)$.

Step II. Now $M\overline{\mathbf{x}} = P\mathbf{b}$ is solvable. The solution is not unique if $\mathcal{N}(M)$ is not $\{\mathbf{0}\}$, in which case, if $\mathbf{x}_p$ is a solution, then all vectors in $\mathbf{x}_p + \mathcal{N}(M)$ are solutions.

Choosing among them the solution of minimal length. Since $F^n = \mathcal{N}(M) \oplus \mathcal{N}(M)^\perp$ and $\mathcal{N}(M)^\perp = \mathcal{R}(M^*)$, any $\mathbf{x} \in F^n$ can be uniquely written as $\mathbf{x} = \mathbf{x}_N + \mathbf{x}_R$ with $\mathbf{x}_N \in \mathcal{N}(M)$ and $\mathbf{x}_R \in \mathcal{R}(M^*)$. Since $\|\mathbf{x}\|^2 = \|\mathbf{x}_N\|^2 + \|\mathbf{x}_R\|^2$ (by the Pythagorean theorem) the solution of minimal length will be the unique solution $\mathbf{x}^+ \in \mathbf{x}_p + \mathcal{N}(M)$ which belongs to $\mathcal{R}(M^*)$. (It exists: since $\|\mathbf{x}_p + \mathbf{w}\|$ is the distance between $-\mathbf{x}_p$ and $\mathbf{w} \in \mathcal{N}(M)$ it is minimum when $\mathbf{w}$ is the orthogonal projection of $-\mathbf{x}_p$ on $\mathcal{N}(M)$.)

Then M^+ is defined by $M^+\mathbf{x} = \mathbf{x}^+$ for all $\mathbf{x}$.

Example. Solve $\Sigma\mathbf{x} = \mathbf{b}$ for

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Clearly $\mathcal{R}(\Sigma) = \{\mathbf{y} \in \mathbb{R}^3 | y_3 = 0\}$ hence $P\mathbf{b} = P(b_1, b_2, b_3)^T = (b_1, b_2, 0)^T$. Then $\Sigma\overline{x} = P\mathbf{b}$ has the solutions $\overline{x}$ with $\overline{x}_j = b_j/\sigma_j$ for $j = 1, 2$ and $\overline{x}_{3,4}$ arbitrary, which has minimal norm for $\overline{x}_{3,4} = 0$. We obtained

$$\mathbf{x}^+ = \begin{bmatrix} b_1/\sigma_1 \\ b_2/\sigma_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/\sigma_1 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix} \equiv \Sigma^+\mathbf{b}$$

For a general $m \times n$ diagonal matrix Σ with singular values $\Sigma_{jj} = \sigma_j$ similar arguments show that its pseudoinverse Σ^+ is an $n \times m$ diagonal matrix with singular values $\Sigma^+_{jj} = 1/\sigma_j$.

For a general $m \times n$ matrix M with singular value decomposition $M = U\Sigma V^*$, solving $M\mathbf{x} = \mathbf{b}$ is equivalent to solving $\Sigma\mathbf{y} = U^*\mathbf{b}$ where $\mathbf{y} = V^*\mathbf{x}$. This that the optimal solution $\mathbf{y}^+ = \Sigma^+U^*\mathbf{b}$, therefore (since U preserves distances) $\mathbf{x}^+ = V\Sigma^+U^*\mathbf{b}$. We proved

Theorem 45. *The pseudoinverse of a matrix M with singular value decomposition $M = U\Sigma V^*$ is $M^+ = V\Sigma^+U^*$.*

The pseudoinverse has many properties similar to those of an inverse. The following statements are left as exercises.

1. If M is invertible, then $M^{-1} = M^+$.
2. $MM^+M = M$ and $M^+MM^+ = M^+$ (though MM^+ and M^+M are not necessarily the identity).
3. MM^+ and M^+M are orthogonal projectors.
4. The operator $^+$ commutes with complex conjugation and transposition.
5. $(\lambda M)^+ = \frac{1}{\lambda}M^+$

6. If λ is a scalar (think $M = [\lambda]$) then λ^+ equals 0 if $\lambda = 0$ and $1/\lambda$ if $\lambda \neq 0$.

7. The pseudoinverse of a vector $\mathbf{x}$ is $\mathbf{x}^+ = \frac{\mathbf{x}^*}{\|\mathbf{x}\|^2}$ if $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{0}^T$ if $\mathbf{x} = \mathbf{0}$.