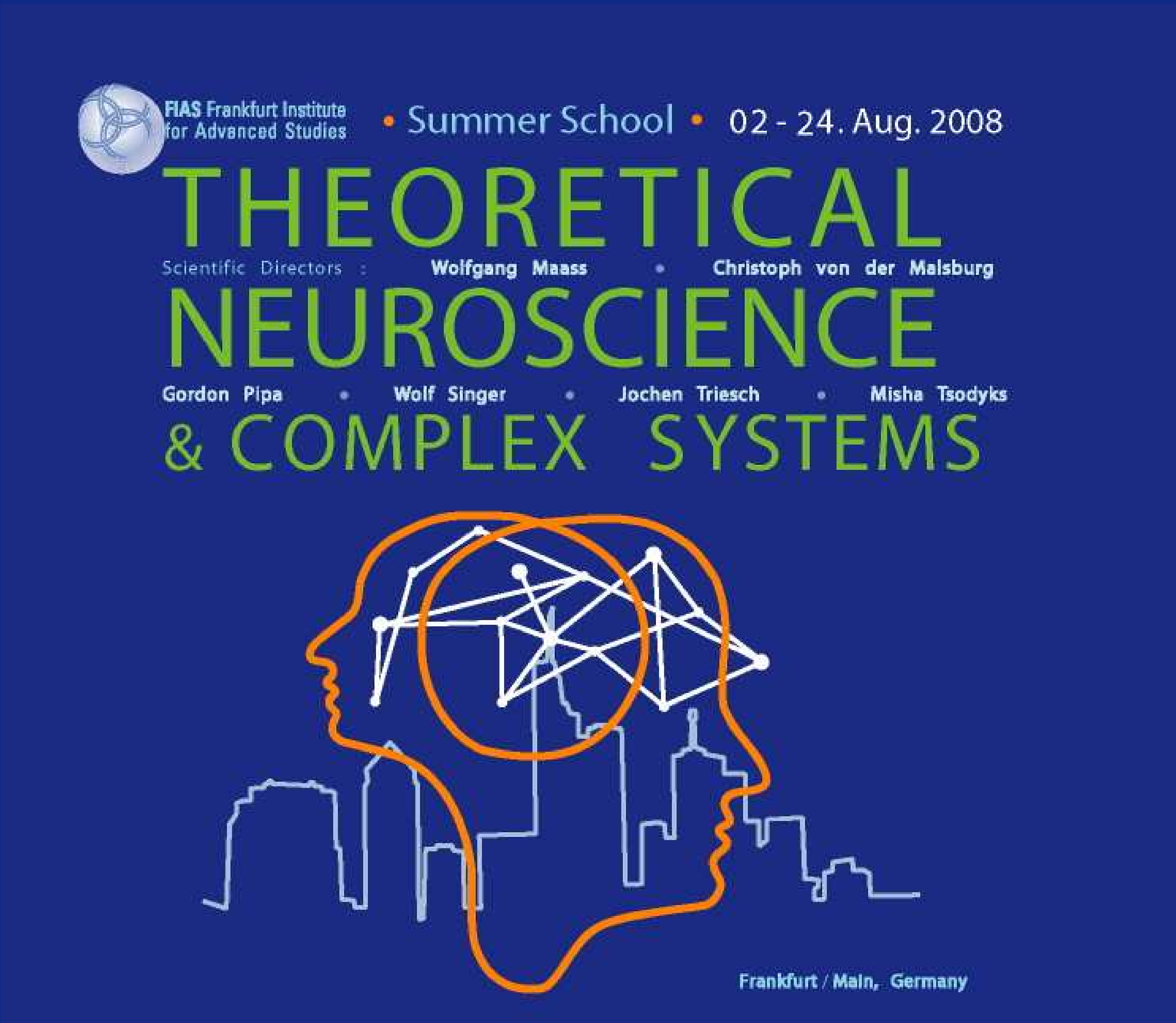


Pre-School
Linear Algebra

Cornelius Weber

Matrix Product
The elements of a matrix $C = AB$ are obtained as:

$$c_{mn} = \sum_q^Q a_{mq} b_{qn}$$

which is written in matrix notation as (small capital letters denote number of columns and rows):

$$\begin{pmatrix} & & {}^{Q} \\ & A & \\ {}_{M} & & \end{pmatrix} \begin{pmatrix} & B & {}^{N} \\ {}_{Q} & & \end{pmatrix} = \begin{pmatrix} & & {}^{N} \\ & C & \\ {}_{M} & & \end{pmatrix}$$

Note that the number of columns of A must equal the number of rows of B.

Special Cases

- Dot product: $M = 1$, $N = 1$. Then $A = \vec{a}^{\mathrm{T}}$, $B = \vec{b}$ are vectors in $\mathbb{R}^Q$:

$$AB = \vec{a}^{\mathrm{T}}\vec{b} = (a_1, \ldots, a_Q) \begin{pmatrix} b_1 \\ \vdots \\ b_Q \end{pmatrix} = c, \quad \text{(scalar)}$$

- Tensor product: $Q = 1$. Then $A = \vec{a}$ is a vector in $\mathbb{R}^M$ and $B = \vec{b}^{\mathrm{T}}$ is a vector in $\mathbb{R}^N$:

$$AB = \vec{a}\vec{b}^{\mathrm{T}} = \begin{pmatrix} a_1 \\ \vdots \\ a_M \end{pmatrix} (b_1, \ldots, b_N) = \begin{pmatrix} & & {}^{N} \\ & C & \\ {}_{M} & & \end{pmatrix}$$

- Regarding the scalar a as a 1×1-matrix, the following order is correct:

$$\vec{c} = \vec{b}a \qquad \text{or} \qquad \vec{c}^T = a\vec{b}^T$$

- The product

$$A\vec{x} = \left(\begin{pmatrix} \vec{a}_1 \end{pmatrix} \cdots \begin{pmatrix} \vec{a}_M \end{pmatrix} \right) \begin{pmatrix} x_1 \\ . \\ x_M \end{pmatrix} = \begin{pmatrix} \vec{a}_1 \end{pmatrix} x_1 + \ldots + \begin{pmatrix} \vec{a}_1 \end{pmatrix} x_M$$

  is a linear combination of the column-vectors $\vec{a}_m$ of A.
  Important is the equation $\quad A\vec{x} = 0$.
  If $\vec{x} \neq 0$, then the column vectors of A must be linearly dependent ($det(A) = 0$).

Simple Matrix Rules

$$A + B = B + A$$

$$AB \neq BA, \quad \text{in general}$$

$$(AB)^{\mathrm{T}} = B^{\mathrm{T}} A^{\mathrm{T}} \quad \text{(mind the order!)}$$

Determinant
The idea is to characterize a square matrix by a real number. So one defines a function $det : M(n \times n) \to \mathbb{R}$ from an $n \times n$-matrix to a real number. The following three axioms lead to a unique definition of such a function:

D1
$det(\mathbb{1}) = 1$
(det is normalized.)

D2
if A has two similar rows, then $det(A) = 0$
(det is alternating.)

D3
a) if for three row vectors a, b, c holds $a + b = c$, then:

$$det \begin{pmatrix} \vdots \\ c \\ \vdots \end{pmatrix} = det \begin{pmatrix} \vdots \\ a \\ \vdots \end{pmatrix} + det \begin{pmatrix} \vdots \\ b \\ \vdots \end{pmatrix}$$

b) if for two row vectors a, b holds $a = \lambda b$, then:

$$det \begin{pmatrix} \vdots \\ a \\ \vdots \end{pmatrix} = \lambda \, det \begin{pmatrix} \vdots \\ b \\ \vdots \end{pmatrix}$$

(det is linear in every row.)

The following follows from these axioms:

- if the matrix has a row a where $a = (0, \ldots, 0)$, then $det(A) = 0$.
  This follows from axiom 3.
- if we can construct B from A by adding the jth row to the ith row, $j \neq i$, then $det(B) = det(A)$.
  To see this, partition B according to axiom 3 into two matrices, then one will have two similar rows and hence $det = 0$ according to axiom 2.
- if A is diagonal, then $detA = a_{11} \cdot a_{22} \cdot \ldots \cdot a_{nn}$.
  To see this, start with the identity matrix (axiom 1) and apply operations of axiom 3.
- $det(A) = 0 \Leftrightarrow$ the row vectors (also column vectors) are linearly dependent.
  (Has to do with axiom 2: via elementary row operations, bring the matrix into block diagonal form. Then $det(A)$ is the product over the pivot elements. It holds: vectors linearly independent $\Leftrightarrow$ pivot elements $\neq 0$.)

Furthermore:

- $det \begin{pmatrix} a \\ \vdots \\ b \end{pmatrix} = -\, det \begin{pmatrix} b \\ \vdots \\ a \end{pmatrix}$ (det is anti-symmetric)
- $det(AB) = det(A) \cdot det(B)$
- but: $det(A + B) \neq det(A) + det(B)$, in general.
- A invertible $\Leftrightarrow det(A) \neq 0$

Application to a homogeneous system of equations

$$A\vec{x} = 0$$

has a non-trivial solution ($\vec{x} \neq 0$) only if $det(A) = 0$.

Reason: consider $det(A) \neq 0 \Leftrightarrow$ (column) vectors are linearly independent. Hence, there is *no* linear combination of column vectors (that's what is done by multiplying A from the right with a vector $\vec{x}$), which fulfills $A\vec{x} = 0$ (except the trivial solution $\vec{x} = 0$).
So in order to allow for a non-trivial solution of $\vec{x}$, it has to be $det(A) = 0$. The solution $\vec{x}$ will not be unique, because $c\,\vec{x}$ will also be a solution.

Application to an inhomogeneous system of equations

$$A\vec{x} = \vec{b}$$

has a *unique* solution iff $det(A) \neq 0$.

Reason: $det(A) \neq 0 \Leftrightarrow$ vectors are linearly independent, and form basis vectors of the vector space. Each point $\vec{b}$ can be reached by unique coefficients $\vec{x}$ of these basis vectors.

Geometrical interpretation

The determinant of a matrix is the area of the parallelogram spanned by its row (column-) vectors (neglecting the determinant's sign).
To see this, one transforms the parallelogram via two kinds of operations into a cube. The first kind makes a rectangular box out of the parallelogram while keeping the volume constant. This is done by adding λ times the jth row vector $\vec{v}_j$ to the ith row vector $\vec{v}_i$, $j \neq i$. That does not change the determinant (see above), and also not the volume (see Figure).

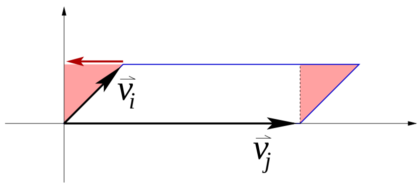


Figure 1: Varying $\vec{v}_i$ by a portion of $\vec{v}_j$ shifts the shaded volume of the parallelogram.

Now our parallelogram has right angles. The second kind of operations scales each vector to unity length. The determinant – just as the volume – scales linear with the scaling of each of these vectors. These scaling factors tell us about the original volume.
Resulting is a cube with volume 1. We might have to rotate it to align with the axis, which neither changes the volume, nor the determinant since a rotation matrix has determinant 1.

Another interpretation: if one applies the matrix to the points of a cube, then the determinant gives the distortion of the volume. A determinant which is smaller than 0 means that the volume has changed its orientation (the order of the axes permutes by an odd number).

Calculation of the determinant (Leibniz' rule)

$$det\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \stackrel{\mathbf{D3}}{=} \sum_{i_1=1}^{n} a_{1i_1} \cdot det\begin{pmatrix} e_{i_1} \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \stackrel{\mathbf{D1}}{=} \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_n=1}^{n} a_{1i_1} \cdot a_{2i_2} \cdot \ldots \cdot a_{ni_n} \cdot det\begin{pmatrix} e_{i_1} \\ \vdots \\ e_{i_n} \end{pmatrix}$$

The first step uses **D3**a) by constructing a row from unity vectors, as well as **D3**b) by multiplying with the corresponding value.

Example of Leibniz' rule

$$\begin{aligned} det\begin{pmatrix} a & b \\ c & d \end{pmatrix} &= a\, det\begin{pmatrix} 1 & 0 \\ c & d \end{pmatrix} + b\, det\begin{pmatrix} 0 & 1 \\ c & d \end{pmatrix} \\ &= a\,c\, det\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} + a\,d\, det\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\,c\, det\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + b\,d\, det\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \\ &= a\,d\; - \; b\,c \end{aligned}$$

Source: Gerd Fischer Lineare Algebra P. 142 (in german language)

The Eigenvalue Equation

Let V be a M-dimensional vector space $\mathbb{R}$. Let A: $V \to V$ be a linear function (A can be expressed as a $M \times M$-matrix). A real number λ_m is an eigenvalue of A, if there exists a vector $\vec{v}_m \in V$ with $\vec{v}_m \neq 0$, so the eigenvalue equation holds:

$$A\vec{v}_m = \lambda_m \vec{v}_m \tag{1}$$

Hence, the $\vec{v}_m$ are those vectors, which are extended/compressed, but **not** rotated, by the function A. If $\vec{v}_m$ is an eigenvector, then $c\,\vec{v}_m$ is also an eigenvector. The eigenvalue equation for M eigenvalues and eigenvectors in matrix notation is:

$$AU = U\Lambda \tag{2}$$

or

$$\begin{pmatrix} & & \\ & A & \\ & & \end{pmatrix} \left(\begin{pmatrix} \vec{v}_1 \\ \cdot \\ \cdot \end{pmatrix} \cdots \begin{pmatrix} \vec{v}_M \\ \cdot \\ \cdot \end{pmatrix} \right) = \left(\begin{pmatrix} \vec{v}_1 \\ \cdot \\ \cdot \end{pmatrix} \cdots \begin{pmatrix} \vec{v}_M \\ \cdot \\ \cdot \end{pmatrix} \right) \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{pmatrix}$$

Λ is a diagonal matrix, with the eigenvalues on the diagonal. U is a matrix, with the eigenvectors as its columns. Mind the order on both sides of the equation!

In Eq. 2 we assume that there are M eigenvectors. This sounds reasonable, because in an M-dimensional vector space there are maximally M linear independent vectors.

In the following, we will address the following questions:

- How to compute eigenvalues & eigenvectors?
- What is an inverse matrix?
- How to invert a matrix?

Let $U^{-1}U = \mathbb{1}_M$ (here U^{-1} is the inverse matrix of U) then we can write:

$$U^{-1}AU = \Lambda \qquad \Leftrightarrow \qquad AU = U\Lambda \qquad \Leftrightarrow \qquad A = U\Lambda U^{-1}.$$

Left: Eq. 2 left-multiplied by U^{-1} Right: Eq. 2 right-multiplied by U^{-1}.

Relevance

We have: $\vec{y} = A\vec{x}$ $\leftarrow$ the elements of $\vec{x}$ are combined in a complicated matter; an element of $\tilde{y}$ is made from many elements of $\tilde{x}$

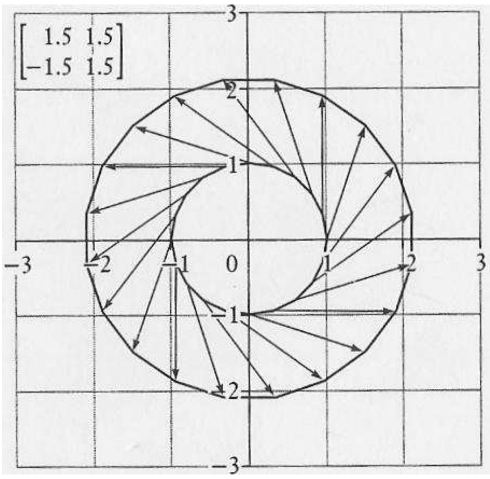
We want: $\tilde{y} = \Lambda\,\tilde{x}$ $\leftarrow$ each element of $\tilde{x}$ is multiplied by an eigenvalue; an element of $\tilde{y}$ is made from only one element of $\tilde{x}$

How to get there: $\vec{y} = U\Lambda U^{-1}\,\vec{x}$

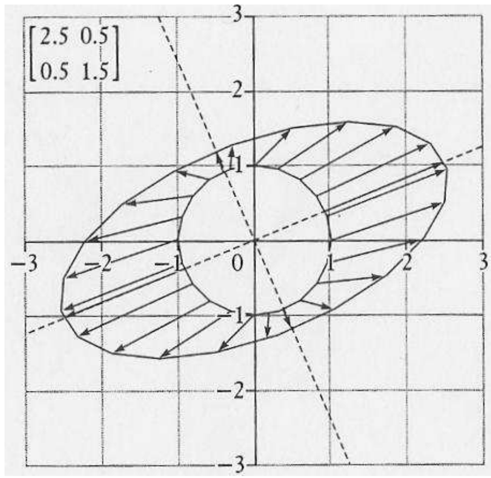
Left-multiply by U^{-1}: $U^{-1}\vec{y} = \Lambda U^{-1}\,\vec{x}$

Define: $\tilde{y} := U^{-1}\vec{y}$ and $\tilde{x} := U^{-1}\,\vec{x}$

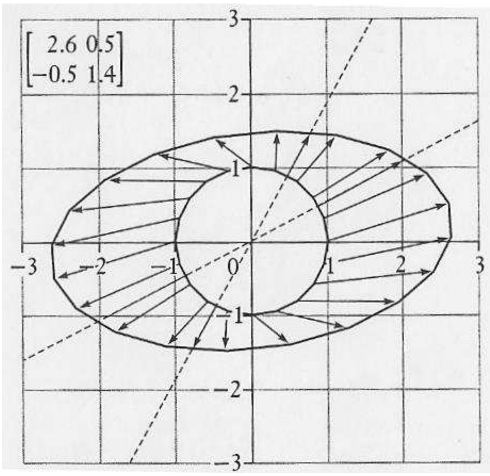
Back-transform: $\vec{y} = U\tilde{y}$ and $\vec{x} = U\,\tilde{x}$



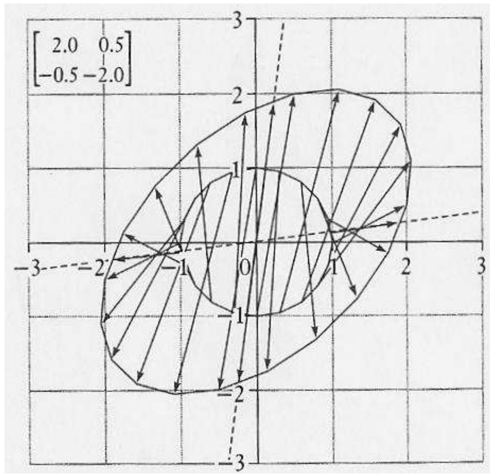
rotation matrix
no eigenvectors



symmetric matrix
orthogonal eigenvectors



two positive eigenvalues



one positive, one negative eigenvalue

Figure 1: Effect of 2×2-Matrices on sample vectors.

Source: "Consider the Lowly 2x2 Matrix" by Alexander Kujath: http://www2.in.tu-clausthal.de/~hormann/teaching/ProSemSS05/PS.CG.24.05.2005.a.ppt

Orthonormal: M vectors $\in \mathbb{R}^M$ are orthonormal if

$$\vec{v}^i \cdot \vec{v}^j = \delta_{ij}$$

In matrix notation:

$$\begin{pmatrix} (\ldots \vec{v}^1 \ldots) \\ \cdot \\ (\ldots \vec{v}^M \ldots) \end{pmatrix} \left(\begin{pmatrix} \vec{v}^1 \\ \cdot \\ \cdot \end{pmatrix} \cdots \begin{pmatrix} \vec{v}^2 \\ \cdot \\ \cdot \end{pmatrix} \right) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

or

$$SS^{\mathrm{T}} = \mathbb{1}_M \tag{3}$$

which is equivalent to (by left-multiplying with S^{-1})

$$S^{\mathrm{T}} = S^{-1}$$

The column vectors of S^{T} form an orthonormal base of $\mathbb{R}^M$. Hence, the eigenvectors would span an orthonormal base, if U (on the previous page) is orthonormal. Note that orthogonal eigenvectors may always be scaled to be orthonormal.

Source: Gerd Fischer Lineare Algebra P. 200 (in german language)

When are the eigenvectors orthogonal? We will show that of any symmetric matrix, eigenvectors belonging to different eigenvalues are orthogonal.
Let $A\vec{v}_j = \lambda_j \vec{v}_j$ and $A\vec{v}_i = \lambda_i \vec{v}_i$ (i.e. $\vec{v}_j$ is eigenvector belonging to eigenvalue λ_j and $\vec{v}_i$ is eigenvector belonging to eigenvalue λ_i). Then:

$$\lambda_j \vec{v}_i^{\mathrm{T}} \vec{v}_j = \vec{v}_i^{\mathrm{T}} \lambda_j \vec{v}_j = \vec{v}_i^{\mathrm{T}} A \vec{v}_j = (A^{\mathrm{T}} \vec{v}_i)^{\mathrm{T}} \vec{v}_j \overset{\text{symm}}{=} (A\vec{v}_i)^{\mathrm{T}} \vec{v}_j = \lambda_i \vec{v}_i^{\mathrm{T}} \vec{v}_j$$

If $\lambda_j \neq \lambda_i$ then it has to be $\vec{v}_i^{\mathrm{T}} \vec{v}_j = 0$.

Source: Gerd Fischer Lineare Algebra P. 208 (in german language)

How find eigenvalues & eigenvectors? The eigenvalue equation (1) is equivalent to:

$$(A - \lambda_m \mathbb{1})\vec{v}_m = 0 \tag{4}$$

This system of equations for $\vec{v}_m$ has non-trivial solutions if

$$\det(A - \lambda_m \mathbb{1}) = 0 \tag{5}$$

Computing this determinant results in a polynomial in λ_m, the *characteristic polynomial* of A. Its solutions are the eigenvalues $\{\lambda_m\}$ of A.
The eigenvectors $\vec{v}_m$ are then found by inserting the found λ_m into equation (4). Since the rows of this matrix are linearly dependent (det $= 0$), the solutions are not unique. One obtains only a relation which fixes the direction but not the length of the eigenvectors.

When are the eigenvalues real? As an example, let us look only at the characteristic polynomial of a 2×2-matrix.

$$\det \begin{pmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{pmatrix}$$

$$= \lambda^2 - a_{11}\lambda - a_{22}\lambda + a_{11}a_{22} - a_{21}a_{12} = \lambda^2 + (-a_{11} - a_{22})\lambda + a_{11}a_{22} - a_{21}a_{12}$$

which we set to zero to obtain the eigenvalues. We apply the quadratic formula

$$x_{1/2} = -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} - c}$$

where we identify $b = (-a_{11} - a_{22})$ and $c = a_{11}a_{22} - a_{21}a_{12}$. The eigenvalues are *both* real if: $c \leq \frac{b^2}{4}$, i.e.

$$a_{11}a_{22} - a_{21}a_{12} \leq \frac{(-a_{11} - a_{22})^2}{4} = \frac{a_{11}^2 + a_{22}^2 + 2a_{11}a_{22}}{4}$$

$$0 \leq \frac{a_{11}^2 + a_{22}^2 + 2a_{11}a_{22} - 4a_{11}a_{22} + 4a_{21}a_{12}}{4}$$

$$0 \leq a_{11}^2 + a_{22}^2 - 2a_{11}a_{22} + 4a_{21}a_{12}$$

$$0 \leq (a_{11} - a_{22})^2 + 4a_{21}a_{12}$$

This is in particular true if the off-diagonal elements have the same sign.
In general on can show: A symmetric matrix has real eigenvalues.

When are all eigenvalues positive? If the matrix is positive definite – later.

A matrix that has no eigenvectors A rotation matrix in $\mathbb{R}^2$ is given by

$$A = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

The characteristic polynomial P_A is

$$\lambda^2 - 2\cos\alpha \cdot \lambda + \cos^2\alpha + \sin^2\alpha = \lambda^2 - 2\cos\alpha \cdot \lambda + 1$$

We set this to zero and apply the quadratic formula to get the eigenvalues:

$$x_{1/2} = -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} - c} \quad \longrightarrow \quad \lambda_{1/2} = \cos\alpha \pm \sqrt{\cos^2\alpha - 1} \overset{\sin^2 + \cos^2 = 1}{=} \cos\alpha \pm \sqrt{-\sin^2\alpha}$$

The term under the square root is non-negative only for $\alpha = 0,\ \pi$, etc. Then the matrix is similar to the identity matrix.
(Defining the complex number $i = \sqrt{-1}$ we have $\lambda_{1/2} = \cos\alpha \pm i\sin\alpha$.)

Source: Gerd Fischer Lineare Algebra P. 169 (in german language)

Derivation of the quadratic formula

$$\begin{aligned}
x^2 + bx + c &= 0 \\
x^2 + bx &= -c \\
x^2 + bx + \frac{b^2}{4} &= \frac{b^2}{4} - c \quad \text{this step is motivated by the following} \\
(x + \frac{b}{2})^2 &= \frac{b^2}{4} - c \\
x_{1,2} + \frac{b}{2} &= \pm\sqrt{\frac{b^2}{4} - c} \\
x_{1,2} &= -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} - c}
\end{aligned}$$

What is an inverse matrix?
An $M \times M$ matrix A is invertible, if there exists an *inverse matrix* A^{-1} with

$$A\,A^{-1} = \mathbb{1}_M$$

Since $A^{-1} = A^{-1}\mathbb{1}_M = A^{-1}(AA^{-1}) = (A^{-1}A)A^{-1}$ it will also be

$$A^{-1}A = \mathbb{1}_M$$

How to invert a matrix?
1. Method: For a given matrix S search the inverse unknown matrix $X = S^{-1}$. Ansatz:

$$SX = \mathbb{1}$$

For each of the M column vectors of X this is a system of equations with M unknowns (the components of the column vector) and M equations (the rows of S). The solution of each of these systems of equations results in a column of X.

Source: Bronstein Taschenbuch der Mathematik P. 206 (in german language)

2. Method: With **elementary matrices** B one can transform every invertible matrix S into the identity matrix $\mathbb{1}$. In the following, apply these to S (left side) so that it becomes $\mathbb{1}$, and apply the same operations to $\mathbb{1}$ (right side) so that it becomes the inverse S^{-1} of S:

$$B_n \cdot \ldots \cdot B_1 \cdot S = \mathbb{1} \qquad \Leftrightarrow \qquad B_n \cdot \ldots \cdot B_1 \cdot \mathbb{1} = S^{-1}$$

Source: Gerd Fischer Lineare Algebra P. 94 (in german language)

What are elementary matrices? Every invertible matrix S can be expressed as a product of elementary matrices. These are square matrices which differ from $\mathbb{1}$-matrix in only few entries. They perform – from left multiplied – elementary row operations:
(i) multiplying a row by λ,
(ii) adding λ times the j-th row to the i-th row and
(iii) interchanging two rows.

A^{-1} is symmetric if A is symmetric, written also as $A^{-1} = (A^{-1})^{\mathrm{T}}$. Reason:

$$\mathbb{1}_M = A^{-1}A = (A^{-1}A)^{\mathrm{T}} = A^{\mathrm{T}}(A^{-1})^{\mathrm{T}} = A\underbrace{(A^{-1})^{\mathrm{T}}}_{\text{must be } A^{-1}}$$

Source: Christopher M. Bishop Neural Networks for Pattern Recognition Appendix A

A and A^{-1} have the same eigenvectors; the eigenvalues are reciprocal. Reason:

$$\begin{aligned} A\vec{v} &= \lambda\vec{v} && \big|\ \text{left-multiply with } A^{-1} \\ \vec{v} &= \lambda A^{-1}\vec{v} && \big|\ \text{divide by } \lambda \\ \lambda^{-1}\vec{v} &= A^{-1}\vec{v} && \leftarrow \text{eigenvalue equation for the inverse matrix} \end{aligned}$$

Source: Christopher M. Bishop Neural Networks for Pattern Recognition Appendix A

Taylor Series

Assume one can write a function $f : \mathbb{R} \rightarrow \mathbb{R}$ as a polynomial:

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_n x^n$$

The Taylor formula states, if this polynomial series is valid in an interval, the following: The function value at any point x within this intervall can be computed if one knows all derivatives of a function at (another) point x_0 which is also in this interval.

$$f(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)^2 + a_3(x-x_0)^3 + \ldots + a_n(x-x_0)^n$$

(For $x_0 = 0$ this yields the topmost equation.)
Now we need the coefficients $\{a_k\}$. We can obtain these be comparing the coefficients above to the values obtained from deriving the function at x_0. The derivatives at x_0 are:

$$\begin{array}{lllllll} f'(x) & = & a_1 + 2a_2(x-x_0) & + 3a_3(x-x_0)^2 & + \ldots + & na_n(x-x_0)^{n-1} \\ f^{(2)}(x) & = & 2a_2 & + 6a_3(x-x_0) & + \ldots + & n(n-1)a_n(x-x_0)^{n-2} \\ f^{(3)}(x) & = & & 6a_3 & + \ldots + & n(n-1)(n-2)a_n(x-x_0)^{n-3} \end{array}$$

At our expansion point $x = x_0$ we have:

$$\begin{array}{lll} f(x_0) = a_0 & \Leftrightarrow & a_0 = f(x_0) \\ f'(x_0) = a_1 & \Leftrightarrow & a_1 = f'(x_0) \\ f^{(2)}(x_0) = 2a_2 & \Leftrightarrow & a_2 = \frac{f^{(2)}(x_0)}{2} \\ f^{(3)}(x_0) = 6a_3 & \Leftrightarrow & a_3 = \frac{f^{(3)}(x_0)}{3!} \\ f^{(n)}(x_0) = n!\,a_n & \Leftrightarrow & a_n = \frac{f^{(n)}(x_0)}{n!} \end{array}$$

With these coefficients a_k the Taylor series can be written as:

$$f(x) = \sum_{k=0}^{\infty} a_k (x-x_0)^k = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

Examples (for the last one we need $i^2 = -1$):

$$\begin{array}{lll} e^x & = & 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 + \ldots = \sum_k^\infty \frac{x^k}{k!} \\ \cos x & = & 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \ldots = \sum_k^\infty \frac{x^{2k}}{(2k)!}(-1)^k \\ \sin x & = & x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 + \ldots = \sum_k^\infty \frac{x^{2k+1}}{(2k+1)!}(-1)^k \\ & & \\ e^{ix} & = & 1 + ix - \frac{1}{2!}x^2 - i\frac{1}{3!}x^3 + \frac{1}{4!}x^4 + i\frac{1}{5!}x^5 - \ldots = \sum_k^\infty \frac{ix^k}{k!} \\ & = & \cos x + i \sin x \end{array}$$

The latter is the *Euler Formula*.

From the Euler Formula follows:
(i) $e^{ix} + e^{-ix} = \cos x + i\sin x + \cos x - i \sin x = 2\cos x$
(ii) $e^{ix} - e^{-ix} = \cos x + i \sin x - (\cos x - i \sin x) = 2i \sin x$
(iii) $\cos^2 x + sin^2 x = \frac{e^{2ix}+e^{-2ix}+2-e^{2ix}-e^{-2ix}+2}{4} = 1$
(iv) $\cos(\varphi+\psi) + i\sin(\varphi+\psi) = e^{i(\varphi+\psi)} = e^{i\varphi}e^{i\psi} = (\cos\varphi + i\sin\varphi)(\cos\psi + i \sin\psi) = \cos\varphi\cos\psi - \sin\varphi\sin\psi + i(\sin\varphi\cos\psi + cos\varphi\sin\psi)$
From these series one can furthermore see:

$$\cosh x := \frac{e^x + e^{-x}}{2} = \sum_k^\infty \frac{x^{2k}}{(2k)!} = \cos ix \quad \text{or} \quad \cos x = \cosh ix$$

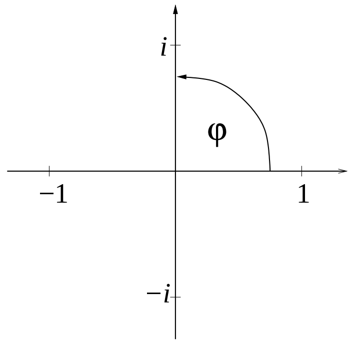
$$\sinh x := \frac{e^x - e^{-x}}{2} = \sum_k^\infty \frac{x^{2k+1}}{(2k+1)!} = -i \sin ix \quad \text{or} \quad \sin x = -i \sinh ix$$

$$\cosh^2 x - \sinh^2 x = \frac{e^{2x}+e^{-2x}+2}{4} - \frac{e^{2x}+e^{-2x}-2}{4} = 1$$

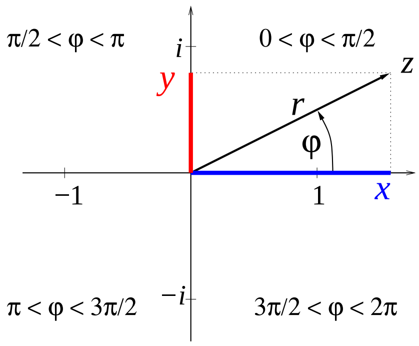
Complex Numbers

Problem: we cannot find a solution to a negative square root. The solution is to define:

$$i := \sqrt{-1}$$



Squaring i means doubling its angle. So multiplications are somehow defined by rotations. Then e.g. $(-i)^2 = -1$, but $-i \cdot i = 1$.
Real numbers are a subset of complex numbers, where the angle is a multiple of π.
Complex numbers are isomorph to $\mathbb{R}^2$.



$$\begin{aligned} z &= x + iy \\ &= r\cos\varphi + ir\sin\varphi \;=\; r\,e^{i\varphi} \end{aligned}$$

The latter is the Euler formula (verified by comparing Taylor series of $\cos\varphi$, $\sin\varphi$ and $e^{i\varphi}$).
Addition of complex numbers ...

$$\begin{aligned} z_1 + z_2 &= \overbrace{x_1 + x_2 + i(y_1 + y_2)} \\ &= r_1 e^{i\varphi_1} + r_2 e^{i\varphi 2} \end{aligned}$$

... is vector addition (which we can better see with the Cartesian coordinates).
Multiplication of complex numbers ...

$$\begin{aligned} z_1 \cdot z_2 &= x_1x_2 + ix_1y_2 + ix_2y_1 - 1 \cdot y_1y_2 \;=\; x_1x_2 - y_1y_2 + 2ix_1y_2 \\ &= r_1e^{i\varphi_1} \cdot r_2e^{i\varphi 2} \;=\; \underbrace{r_1r_2\, e^{i(\varphi_1+\varphi_2)}} \end{aligned}$$

... multiplies the absolute values and adds the angles (better seen in the polar coordinates).

Revisiting a matrix without (real) eigenvectors

$$A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

The characteristic polynomial P_A is

$$det \begin{pmatrix} a-\lambda & b \\ -b & a-\lambda \end{pmatrix} = \lambda^2 - 2a\,\lambda + a^2 + b^2 \stackrel{!}{=} 0$$

Applying the quadratic formula

$$\lambda_{1/2} \;=\; a \pm \sqrt{-b^2} \;=\; \underbrace{a \pm ib} \;=\; r\,(\cos\varphi \,+\, i\,\sin\varphi) \;=\; r\,e^{\pm i\varphi}$$

Indeed, in the original matrix we could have already expressed (a, b) as $(r\cos\varphi, r\sin\varphi)$ for some suitable (r, φ). Hence, we have a “rotation matrix” that also scales the vector by r. We insert the given eigenvalues (in the notation of the under-braced term) into the eigenvalue equation to yield the eigenvectors. For $\lambda_1 = a + ib$:

$$\begin{aligned} av_1 - av_1 - ibv_1 + bv_2 &= 0 \qquad \Rightarrow v_2 = iv_1 \\ -bv_1 + av_2 - av_2 - ibv_2 &= 0 \end{aligned}$$

For $\lambda_2 = a - ib$:

$$\begin{aligned} av_1 - av_1 + ibv_1 + bv_2 &= 0 \qquad \Rightarrow v_2 = -iv_1 \\ -bv_1 + av_2 - av_2 + ibv_2 &= 0 \end{aligned}$$

The two found eigenvectors are

$$\begin{pmatrix} 1 \\ i \end{pmatrix}, \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

Note that each of these vectors has 4 entries (2 of which are zero here), so vectors are in a 4-dimensional space. Any complex multiples of these vectors are also eigenvectors.

Interpretation of the eigenvalue equation in complex space:

- each component of the eigenvector is scaled by the absolute value of the eigenvalue
- the direction of each component is turned by the angle argument of the eigenvalue

Hence, the direction of the eigenvectors does not remain constant in complex space.

Potentiation Method

Finding eigenvalues (and eigenvectors) can be difficult since there is no formula for solving a (characteristic) polynomial of higher than 3rd order. A simple way to find the eigenvector belonging to the largest eigenvalue of a symmetric matrix A is the following iterative procedure.
First, choose an initial vector $\vec{v}$. Then repeat:

1. apply $\quad A\,\vec{v}$
2. normalize $\vec{v}$ (e.g. l^2 or l^∞ norm)

until converged. The eigenvalue is then trivially obtained.

Source: http://alf.math.uni-rostock.de/ kurt/wiwi/slides/wiwi-15.pdf

Correlation Based Learning

Let us consider a linear neuron which computes its activation y from an input $\vec{x}$ as

$$y = \sum_j w_j x_j$$

The Hebbian Rule for the change Δw_{ki} of a weight (synaptic efficacy) between a neuron i with activity x_i of an input layer and a neuron k with activity y_k of the output layer is

$$\Delta w_{ki} = \epsilon \cdot x_i \cdot y_k = \epsilon \cdot x_i \cdot \sum_j w_{kj} x_j$$

where ϵ is a small learning step size. Let us express the input activations $\{x_i\}$ as a vector $\vec{x}$:

$$\Delta \vec{w}_k = \epsilon \vec{x} y_k = \epsilon \vec{x}(\vec{w}_k^{\mathrm{T}} \cdot \vec{x}) = \epsilon \vec{x}\vec{x}_k^{\mathrm{T}} \vec{w} \overset{averaged}{=} \epsilon \langle \vec{x}\vec{x}_k^{\mathrm{T}} \rangle \vec{w} =: \epsilon C \vec{w}$$

C is the correlation matrix of the input patterns (stimuli) in the input layer.

First order derivatives in higher dimensions

f: $\mathbb{R} \mapsto \mathbb{R}$

The derivative of a function f: $\mathbb{R} \mapsto \mathbb{R}$ is the slope of the tangent. This tangent is expressed in the Taylor series

$$f(x_0 + \xi) \;=\; f(x_0) \;+\; a\xi \;+\; \ldots$$

by the first term (after the constant), namely the linear function $a \cdot \xi$. There, $a = \frac{df(x)}{dx}$ is the first order derivative of f.

f: $\mathbb{R}^N \mapsto \mathbb{R}$

Let f: $\mathbb{R}^N \mapsto \mathbb{R}$, then the derivative is given by a tangent (hyper-) plane: along each of the N coordinate axes of the input space there may be a different slope. We can write these values into a vector, the *gradient* $\vec{a}$ of f. The variation of the function value around the point of expansion $\vec{x}_0$ depends on the direction of the vector of variation $\vec{\xi}$. This is expressed by the dot product $\vec{a} \cdot \vec{\xi} = \vec{a}^{\mathrm{T}}\vec{\xi} = \sum_{i=1}^{N} a_i \xi_i$:

$$f(\vec{x}_0 + \vec{\xi}) \;=\; f(\vec{x}_0) \;+\; \vec{a}^{\mathrm{T}}\vec{\xi} \;+\; \ldots \qquad (i = 1, \ldots, N)$$

f: $\mathbb{R}^N \mapsto \mathbb{R}^M$

Let f: $\mathbb{R}^N \mapsto \mathbb{R}^M$, so for each of the M components of the function value (a vector) we have such an equation:

$$f_j(\vec{x}_0 + \vec{\xi}) \;=\; f_j(\vec{x}_0) \;+\; \vec{a}_j^{\mathrm{T}}\vec{\xi} \;+\; \ldots \qquad (j = 1, \ldots, M)$$

We write these M equations below each other to form one matrix equation

$$\vec{f}(\vec{x}_0 + \vec{\xi}) \;=\; \vec{f}(\vec{x}_0) \;+\; A\vec{\xi} \;+\; \ldots$$

So the first order term of a Taylor series in a higher dimensional vector space is represented by a matrix, since a matrix represents a linear operator.

The $M \times N$-matrix A is called *Differential* or *Jacobi-Matrix*. An entry a_{ji} is the partial derivative of the j-th component of the function value w.r.t. the i-th component of the input vector:

$$a_{ji} = \frac{\partial f_j}{\partial x_i}(\vec{x}_0) = \lim_{(\vec{\xi})_i \to 0} \frac{f_j(\vec{x}_0 + (\vec{\xi})_i) - f_j(\vec{x}_0)}{(\vec{\xi})_i}$$

2nd order derivative

f: $\mathbb{R}^N \mapsto \mathbb{R}^1$

Let f: $\mathbb{R}^N \mapsto \mathbb{R}^1$, then we can express the Taylor series up to 2nd order easily:

$$\begin{aligned} f(\vec{x}_0+\vec{\xi}) &= f(\vec{x}_0) + \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}\xi_i + \frac{1}{2}\sum_i^N \sum_j^N \frac{\partial^2 f}{\partial x_i \partial x_j}\xi_i\xi_j + \ldots \\ &= f(\vec{x}_0) + \vec{\nabla} f \cdot \vec{\xi} + \frac{1}{2}\vec{\xi}^{\mathrm{T}} H \vec{\xi} + \ldots \end{aligned}$$

The first order derivative yields the gradient vector $\vec{\nabla} f$. The second order derivative is represented by the *Hesse matrix*. The *Hesse Matrix* is symmetric, because $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.

Explanation for f: $\mathbb{R}^2 \mapsto \mathbb{R}^1$

Let f: $\mathbb{R}^2 \mapsto \mathbb{R}^1$, then the change of a function value along the direction of a vector $\vec{\xi}$ is composed of its change along each of the vector's components, ξ_1 and ξ_2:

$$f(x_1^0+\xi_1, x_2^0+\xi_2) = f(x_1^0, x_2^0) + \underbrace{\left.\frac{\partial f}{\partial x_1}\right|_{(x_1^0,x_2^0)}}_{g(x_1^0,x_2^0)} \xi_1 + \underbrace{\left.\frac{\partial f}{\partial x_2}\right|_{(x_1^0,x_2^0)}}_{h(x_1^0,x_2^0)} \xi_2 \qquad (1)$$

To get 2nd-order precision we must consider that the derivatives (here defined as g and h) themselves change along $\vec{\xi}$:

$$\begin{aligned} \Delta g &:= g(\vec{x}^0+\vec{\xi}) - g(\vec{x}^0) = \frac{\partial g}{\partial x_1}\xi_1 + \frac{\partial g}{\partial x_2}\xi_2 \\ \Delta h &:= h(\vec{x}^0+\vec{\xi}) - h(\vec{x}^0) = \frac{\partial h}{\partial x_1}\xi_1 + \frac{\partial h}{\partial x_2}\xi_2 \end{aligned}$$

This expresses how much the derivatives have changed along the whole extent of $\vec{\xi}$. The contributions to the change of the function value f are only the mean of these changes: $g + \frac{1}{2}\Delta g$ and $h + \frac{1}{2}\Delta h$.

So in Eq. 1 we have to replace $g \to g + \frac{1}{2}\Delta g$ and $h \to h + \frac{1}{2}\Delta h$. We obtain the additional terms:

$$\frac{1}{2}\Delta g\,\xi_1 + \frac{1}{2}\Delta h\,\xi_2 = \frac{1}{2}\Big((\frac{\partial^2 f}{\partial x_1^2}\xi_1 + \frac{\partial^2 f}{\partial x_2 \partial x_1}\xi_2)\xi_1 + (\frac{\partial^2 f}{\partial x_1 \partial x_2}\xi_1 + \frac{\partial^2 f}{\partial x_2^2}\xi_2)\xi_2 \Big)$$

which is in matrix notation (using $\frac{\partial^2 f}{\partial x_2 \partial x_1} = \frac{\partial^2 f}{\partial x_1 \partial x_2}$):

$$\frac{1}{2}\ (\xi_1\ \xi_2) \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} =: \frac{1}{2}\vec{\xi}^{\mathrm{T}} H \vec{\xi}$$

with the Hesse matrix H.

Positive definite

Definition: A symmetric matrix A is positive definite if $\vec{x}^{\mathrm{T}} A\vec{x} > 0 \quad \forall \vec{x}$ with $\vec{x} \neq 0$.
Theorem: A positive definite $\Leftrightarrow$ all eigenvalues > 0.
Proof:
"$\Rightarrow$" for the eigenvectors we have:

$$\vec{v}^{\mathrm{T}} A\vec{v} = \vec{v}^{\mathrm{T}}\lambda\vec{v} = \lambda \underbrace{\vec{v}^{\mathrm{T}}\vec{v}}_{>0} \stackrel{!}{>} 0 \;\Rightarrow \lambda > 0$$

"$\Leftarrow$" express the vector $\vec{x}$ as a linear combination of eigenvectors: $\vec{x} = S^{-1}\vec{\alpha}$. ($S^{-1}$ is the matrix, where the column vectors are the eigenvectors of A. These are orthogonal – orthonormal if we wish – in case of a symmetric matrix, if the eigenvalues are different; hence such a linear combination can always be made.) So we can write

$$\vec{x}^{\mathrm{T}} A\vec{x} = (S^{-1}\vec{\alpha})^{\mathrm{T}} A S^{-1}\vec{\alpha} = \vec{\alpha}^{\mathrm{T}} {S^{-1}}^{\mathrm{T}} A S^{-1}\vec{\alpha} = \vec{\alpha}^{\mathrm{T}} \underbrace{{S^{-1}}^{\mathrm{T}} S^{-1}}_{\mathbb{1}} \Lambda\vec{\alpha} = \sum_i \lambda_i \alpha_i^2 > 0$$

The under-braced term assumes orthonormal eigenvectors. Λ is the diagonal matrix of the eigenvalues.

When are matrices positive definite?

- If $\vec{y} = A\vec{x}$ then $\vec{x}^{\mathrm{T}} A\vec{x} = \vec{x}^{\mathrm{T}}\vec{y}$ is a dot product of $\vec{x}$ with $\vec{y}$. This being positive means that the angle between both vectors is less than 90°. Thus, a positive definite matrix does not turn a vector by more than 90°.

- A, B positive definite $\Rightarrow$ the matrix $A + B$ also. Reason:

$$\vec{x}^{\mathrm{T}}(A + B)\vec{x} \;=\; \underbrace{\vec{x}^{\mathrm{T}} A\vec{x}}_{>0} \;+\; \underbrace{\vec{x}^{\mathrm{T}} B\vec{x}}_{>0}$$

- A positive definite $\Rightarrow$ A^r positive definite, where $r \in \mathbb{R}$.
  Reason: eigenvalues λ_i are positive $\Rightarrow$ λ_i^r positive.

- Let R be any matrix with linearly independent columns (not necessarily square matrix). Then $R^{\mathrm{T}} R$ is positive definite.
  Reason: $\vec{x}^{\mathrm{T}} R^{\mathrm{T}} R\vec{x} = (R\vec{x})^{\mathrm{T}}(R\vec{x}) = \|R\vec{x}\|^2$
  This is always positive. The term $R\vec{x}$ which is a linear combination of the columns of R cannot become zero for $\vec{x} \neq 0$, because the columns are linearly independent.

- R as above, further let A be positive definite. $\Leftrightarrow$ $R^{\mathrm{T}} AR$ is positive definite.

Source: Gilbert Strang Introduction to Applied Mathematics P. 63, 34/35

Relevance
Let A be a symmetric 2×2-matrix.

$$f(\vec{x}) := \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2$$

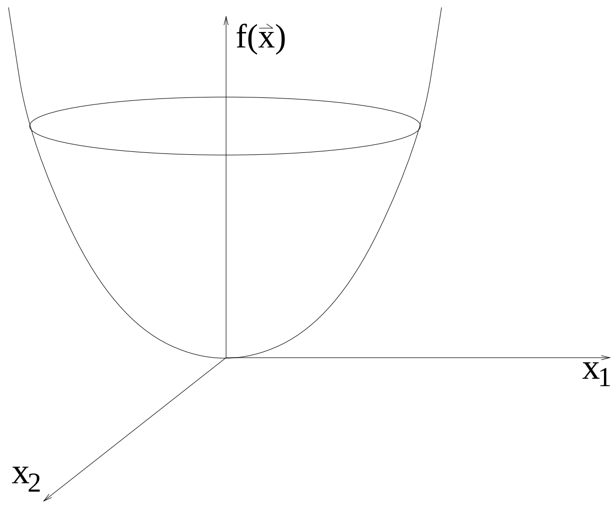
This might be the 2nd order term of a Taylor series. We have

- $f(0) = 0$
- $\frac{\partial f}{\partial x_1} = 2a_{11}x_1 + 2a_{12}x_2$

  $\frac{\partial f}{\partial x_2} = 2a_{22}x_2 + 2a_{12}x_1$

  So the first derivative at the origin is zero: $\frac{\partial f}{\partial x_{1,2}}\Big|_{\vec{x}=0} = 0$
- $\frac{\partial^2 f}{\partial x_1^2} = 2a_{11}, \quad \frac{\partial^2 f}{\partial x_2^2} = 2a_{22}, \quad \frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1} = 2a_{12}$

If A is positive definite, then $f(\vec{x}) > 0$ for all $\vec{x} \neq 0$. So $f(\vec{x})$ looks like a parabola centred at the origin.



Source: Gilbert Strang Introduction to Applied Mathematics P. 17-19

Def: $(\text{Hesse}f)(\vec{x})_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$, Hesse matrix, symmetric ("2nd derivative matrix").

Theorem: Let $f : V \mapsto \mathbb{R}$ be 2x continuously differentiable and $\vec{\nabla} f(x_o) = 0$. Then:
If $(\text{Hesse}f)(\vec{x})$ positive (negative) definite, then f has in x_o an isolated minimum (maximum).

Source: Otto Forster Analysis 2 P. 61/62