Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät

Deep Neural Networks

Chapter 2: Linear Algebra for Deep Learning

2.1 Scalars, Vectors, Matrices, Tensors

- Scalars: a single number, e.g. x_1
- Vectors: $\boldsymbol{x}$ is an array of numbers, written in ***Euclid bold italics.***

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

Special conventions:
If $S = \{1, 3{,}6\}$ and $n = 6$ then

$$\boldsymbol{x}_S = \begin{bmatrix} x_1 \\ x_3 \\ x_6 \end{bmatrix} \in \mathbb{R}^3 \qquad \boldsymbol{x}_{-S} = \begin{bmatrix} x_2 \\ x_4 \\ x_5 \end{bmatrix} \in \mathbb{R}^3 \qquad \boldsymbol{x}_{-1} = \begin{bmatrix} x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n-1}$$

- Matrices: a 2D array of numbers.

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \in \mathbb{R}^{3\times 2}$$

This matrix has 3 rows and 2 columns.

- $A_{i,:}$ denotes the row vector with row number i.
- $A_{:,j}$ denotes the column vector with column number j.
- $f(\boldsymbol{A})_{i,j}$ gives the element (i, j) of the matrix computed by applying the function f to A.
- Tensors: an n-dimensional array of numbers with n>2. If **A** is a 3D Tensor, then $\mathbf{A}_{i,j,k}$ denotes an element at coordinates (i, j, k).

Matrix Transpose

- An important matrix operation is the transpose. It is the mirror image of the matrix on the diagonal line.
  It generates an

$$\text{If } \boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \in \mathbb{R}^{3\times 2}, \text{ then } \boldsymbol{A}^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix} \in \mathbb{R}^{2\times 3}$$

- In general it is defined $\left(\boldsymbol{A}^T\right)_{i,j} = A_{j,i}$.
- To turn a row vector $\boldsymbol{x}$ into a column vector $\boldsymbol{y}$ we write $\boldsymbol{y} = \boldsymbol{x}^T$.
- We can add two matrices, if they have the same shape, by adding their corresponding elements:

$$\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}, \text{ where } C_{i,j} = A_{i,j} + B_{i,j}$$

2.2 Multiplying Matrices and Vectors

- The matrix product of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$ is a third matrix $\boldsymbol{C} \in \mathbb{R}^{m \times p}$. $\boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{A}\boldsymbol{B}$
  Note that the number of columns of A must match the number of rows of B.
$$C_{i,j} = \sum_k A_{i,k} B_{k,j}$$
- The element-wise multiplication of two matrices A and B of identical dimensions is called the Hadamard product
$$\boldsymbol{D} = \boldsymbol{A} \odot \boldsymbol{B}$$
- The dot product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same dimension is the matrix $\boldsymbol{x}^T \boldsymbol{y}$.

Matrix Product Properties

- Matrix multiplication is distributive:

$$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{A}\boldsymbol{B}+\boldsymbol{A}\boldsymbol{C}$$

- It is also associative:

$$\boldsymbol{A}(\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{A}\boldsymbol{B})\boldsymbol{C}$$

- It is <u>not</u> commutative:

$$\boldsymbol{A}\boldsymbol{B} \neq \boldsymbol{B}\boldsymbol{A} \text{ in most cases}$$

- The dot product between two vectors is commutative:

$$\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x}$$

- The transpose of a matrix has a simple form:

$$(\boldsymbol{A}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A}^T$$

Linear Equation Systems

- The standard form of a linear equation system is
$$\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b}$$
where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a known matrix, $\boldsymbol{b} \in \mathbb{R}^{m}$ is a known vector, and $\boldsymbol{x} \in \mathbb{R}^{n}$ is a vector of unknown variables for which we want to know the values.
- This is a short form of

$$\begin{gathered} A_{1,1}x_1 + \ldots + A_{1,n}x_n = b_1 \\ \ldots \\ A_{m,1}x_1 + \ldots + A_{m,n}x_n = b_m \end{gathered} \qquad \text{or of} \qquad \begin{gathered} A_{1,:}x = b_1 \\ \ldots \\ A_{m,:}x = b_m \end{gathered}$$

- There are several methods to solve these linear equation systems, like Gaussian elimination, LU decomposition, Cholesky decomposition, iterative methods.

2.3 Identity and Inverse Matrices

- Matrix inversion is a tool to analytically solve the matrix equation
- The identity matrix is a matrix that does not change any vector when we multiply it with the matrix

$$\forall \boldsymbol{x} \in \mathbb{R}^{n \times n}, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}$$

- It consists of ones in the diagonal and zeroes everywhere else.

$$\boldsymbol{I}_n = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

- The matrix inverse of $\boldsymbol{A}$ is denoted as $\boldsymbol{A}^{-1}$ and is defined as the matrix such that

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$$

- We can now solve the matrix equation 2.1 with the following steps: $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$

Inverse Matrices

$$\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$$

$$\boldsymbol{A}^{-1}\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

$$\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

- This derivation requires that $\boldsymbol{A}^{-1}$ exists, which is not always the case.
- Usually a linear system of equations is solved more efficiently with other methods than computing $\boldsymbol{A}^{-1}$.

2.4 Linear Dependence and Span

- For $\boldsymbol{A}^{-1}$ to exist, the linear equation system $\boldsymbol{A}\ \boldsymbol{x} = \boldsymbol{b}$ must have exactly one solution for every value of $\boldsymbol{b}$. The other alternatives are no solution or infinitely many solutions for a particular $\boldsymbol{b}$.
- If both $\boldsymbol{x}$ and $\boldsymbol{y}$ are solutions of $\boldsymbol{A}\ \boldsymbol{x} = \boldsymbol{b}$, then
$$\boldsymbol{z} = \alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}$$
is also a solution for any real α.
- The column vectors of $\boldsymbol{A}$ specify basis vectors pointing in different directions from the origin.
$$\boldsymbol{A}\boldsymbol{x} = \sum_i x_i \boldsymbol{A}_{:,i}$$
- A linear combination of a set of vectors $\left\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)}\right\}$ is
$$\sum_i c_i \boldsymbol{v}^{(i)}$$

Span

- In this case the x_i correspond to the c_i, and the column vectors $\boldsymbol{A}_{:,i}$ correspond to the vectors $\boldsymbol{v}_i$.
- The span (dt.: lineare Hülle) of a set of vectors is the set of all linear combinations of these vectors.
- Determining whether $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ has a solution amounts to testing whether $\boldsymbol{b}$ is in the span of the columns of $\boldsymbol{A}$. This span is also called the column space, or the range of $\boldsymbol{A}$.
- In order for the system $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ to have a solution for all values of $\boldsymbol{b} \in \mathbb{R}^m$, the column space of $\boldsymbol{A}$ must be all $\mathbb{R}^m$. Therefore $\boldsymbol{A}$ must have at least m columns, i.e. $n \geq m$.

Linear Independence

- A set of vectors is linearly independent, if no vector in the set is a linear combination of the other vectors.
- So for the column space of the matrix $\boldsymbol{A}$ to encompass all of $\mathbb{R}^m$, the matrix must contain at least m linearly independent columns.
- This ensures that $\boldsymbol{A}$ has a solution for all $\boldsymbol{b} \in \mathbb{R}^m$.
- For the matrix $\boldsymbol{A}$ to have an inverse we also need to ensure that $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ has at most one solution for each $\boldsymbol{b}$. To do so the matrix may have at most m columns.
- Together, this means that $\boldsymbol{A}$ must be square, i.e. $m = n$, and that all columns are independent.
- A square matrix with linearly dependent columns is called singular.

2.5 Norms

- If we need to measure the size of a vector we can use a norm.
- The L^p norm is defined as $\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$ for $p \in \mathbb{R}, p \geq 1$.
- Norms map vectors to non-negative values.
- A norm is a function satisfying the following 3 properties:
  - $f(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{x} = 0$
  - $f(\boldsymbol{x}+\boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ the triangle inequality
  - $\forall \alpha \in \mathbb{R}, f(\alpha\boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$
- The most frequent norm is the L^2 norm (Euclidian norm)

$$\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}\| = \sqrt{\sum_i x_i^{\,2}}$$

2.5 Norms

- The squared L^2 norm $\|\boldsymbol{x}\|^2 = \sum_i x_i^{\,2}$ is more convenient to deal with, especially with derivatives, as
$$\frac{\partial}{\partial x_j}\|\boldsymbol{x}\|^2 = \frac{\partial}{\partial x_j}\sum_i x_i^{\,2} = 2x_j$$
- If the difference between small values and zero is important, we use the L^1 norm:
$$\|\boldsymbol{x}\|_1 = \sum_i |x_i|$$
This norm adds up each small distance from 0.
- Another norm is the L^∞ norm, also called max norm.
$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|$$
It computes the absolute value of the element of the largest magnitude in the vector:

Norms

- If we want to measure the size of a matrix, we can do this with the Frobenius norm:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$$

which is analogous to the L^2 norm of a vector.

- The dot product can be written in terms of norms:

$$\boldsymbol{x}^\mathrm{T}\boldsymbol{y} = \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \cos\theta$$

However, the dot product itself is not a norm.

2.6 Special Kinds of Matrices and Vectors

- Diagonal matrices have nonzero entries only in their main diagonal. $\boldsymbol{D}$ is diagonal iff $D_{i,j} = 0$ for all $i \neq j$.
- $diag(\boldsymbol{v})$ denotes a square diagonal matrix whose diagonal entries are given by the vector $\boldsymbol{v}$.
- Multiplying a vector $\boldsymbol{x}$ with a diagonal matrix $diag(\boldsymbol{v})$ is done by simply scaling each element x_i by v_i:
$$diag(\boldsymbol{v}) \cdot \boldsymbol{x} = \boldsymbol{v} \odot \boldsymbol{x}$$
- Inverting a square diagonal matrix is also efficient
$$diag(\boldsymbol{v})^{-1} = diag([1/v_1, \cdots, 1/v_n]^T)$$
The inverse exists only if every v_i is nonzero.
- It is possible to construct non-square (rectangular) diagonal matrices, which have trailing columns or rows with 0s. The product $\boldsymbol{D} \cdot \boldsymbol{x}$ is computed similarly.

Symmetric Matrices

- Symmetric matrices are any matrix that is equal to its transpose:

$$\boldsymbol{A} = \boldsymbol{A}^T$$

- For example, distance matrices are symmetric. If A_{ij} is the distance from point i to point j, then $A_{i,j} = A_{j,i}$, because the distance function is symmetric.
- A unit vector is a vector with unit norm:

$$\|\boldsymbol{x}\|_2 = 1$$

- Vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal to each other, if $\boldsymbol{x}^T \boldsymbol{y} = 0$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal and have nonzero norm, they have an angle of 90° to each other.

Orthogonal Matrices

- An orthogonal matrix is a square matrix whose rows are mutually orthonormal (orthogonal and normal) and whose columns are mutually orthonormal:

$$\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$$

- This implies that

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^T$$

- Orthogonal matrices are of interest because their inverse is very cheap to compute.
- Note that for a matrix to be orthogonal, the columns and rows need to be mutually orthonormal, not only orthogonal.

2.7 Eigendecomposition

- In eigendecomposition we decompose a matrix into a set of eigenvectors and eigenvalues.
- An eigenvector of a square matrix $\boldsymbol{A}$ is a nonzero vector $\boldsymbol{v}$ which has the property that multiplication by $\boldsymbol{A}$ only alters the scale of v.

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

- The scalar λ is called the eigenvalue of $\boldsymbol{v}$.
- If $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{A}$, then so is any rescaled vector $s \cdot \boldsymbol{v}$ for $s \in \mathbb{R}, s \neq 0$. Also, $s \cdot \boldsymbol{v}$ still has the same eigenvalue.
- Suppose that a matrix $\boldsymbol{A}$ has n linearly independent eigenvectors $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(n)}\}$ with eigenvalues $\{\lambda_1, \cdots, \lambda_n\}$.

2.7 Eigendecomposition

- We now concatenate all eigenvectors to form a matrix $\boldsymbol{V}$ with one eigenvector per column: $\boldsymbol{V} = [\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(n)}]$.
  We also concatenate all eigenvalues to form a vector $\boldsymbol{\lambda}$:
  $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_n]$.
- Then the eigendecomposition of $\boldsymbol{A}$ is given by
  $$\boldsymbol{A} = \boldsymbol{V} \cdot diag(\boldsymbol{\lambda}) \cdot \boldsymbol{V}^{-1}$$
- We often want to decompose matrices into their eigenvectors and eigenvalues.
- Not every matrix may be decomposed into eigenvectors and eigenvalues. Sometimes the decomposition involves complex numbers.

Eigendecomposition

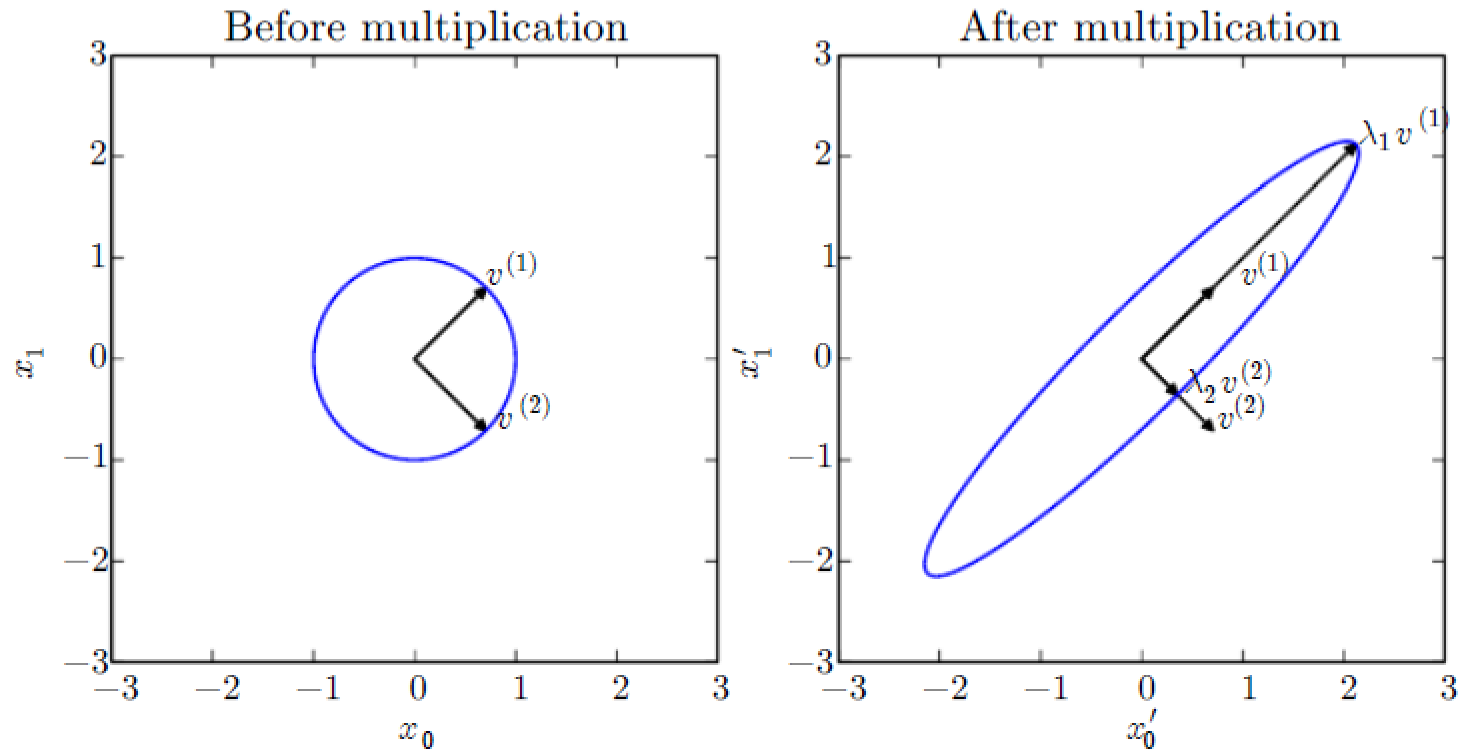
- Every real symmetric matrix $\boldsymbol{A}$ can be decomposed into matrices of real-valued eigenvectors and eigenvalues:

$$\boldsymbol{A} = \boldsymbol{Q} \cdot \Lambda \cdot \boldsymbol{Q}^T$$

  where $\boldsymbol{Q}$ is an orthogonal matrix composed of eigenvectors of $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix. The eigenvalue $\Lambda_{i,i}$ is associated with column i of $\boldsymbol{Q}$. Because $\boldsymbol{Q}$ is an orthogonal matrix, we can think of $\boldsymbol{A}$ as scaling the space by λ_i in the direction $\boldsymbol{v}^{(i)}$.

- The eigendecomposition of a real symmetric matrix may not be unique. If any two or more eigenvectors share the same eigenvalue, then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue, and we could choose a $\boldsymbol{Q}$ with these eigenvectors instead.

Effect of Eigenvectors & Eigenvalues



- Assume a matrix $\boldsymbol{A}$ with eigenvectors $\boldsymbol{v}^{(1)}$ and $\boldsymbol{v}^{(2)}$ with eigenvalues λ_1 and λ_2. Left: the set of all unit vectors $u \in \mathbb{R}^2$. Right: the set of all points $\boldsymbol{A}\boldsymbol{u}$. We see that $\boldsymbol{A}$ scales the space in direction $\boldsymbol{v}^{(i)}$ by eigenvalue λ_i.

Eigendecomposition

- A matrix is singular iff any of the eigenvalues are zero.
- A matrix is positive definite if all eigenvalues are positive.
- A matrix is positive semidefinite if all eigenvalues are positive or zero.
- A matrix is negative definite if all eigenvalues are negative.
- A matrix is negative semidefinite if all eigenvalues are negative or zero.
- Positive semidefinite matrices assure that $\forall \boldsymbol{x},\ \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$
- Positive definite matrices additionally guarantee that

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0$$

2.8 Singular Value Decomposition

- The singular value decomposition is a way to factorize a matrix into singular vectors and singular values.
- Every real matrix has a singular value decomposition but not every real matrix has an eigenvalue decomposition.
- The eigendecomposition decomposes a matrix $\boldsymbol{A}$ into a Matrix $\boldsymbol{V}$ of eigenvectors and a vector λ of eigenvalues:

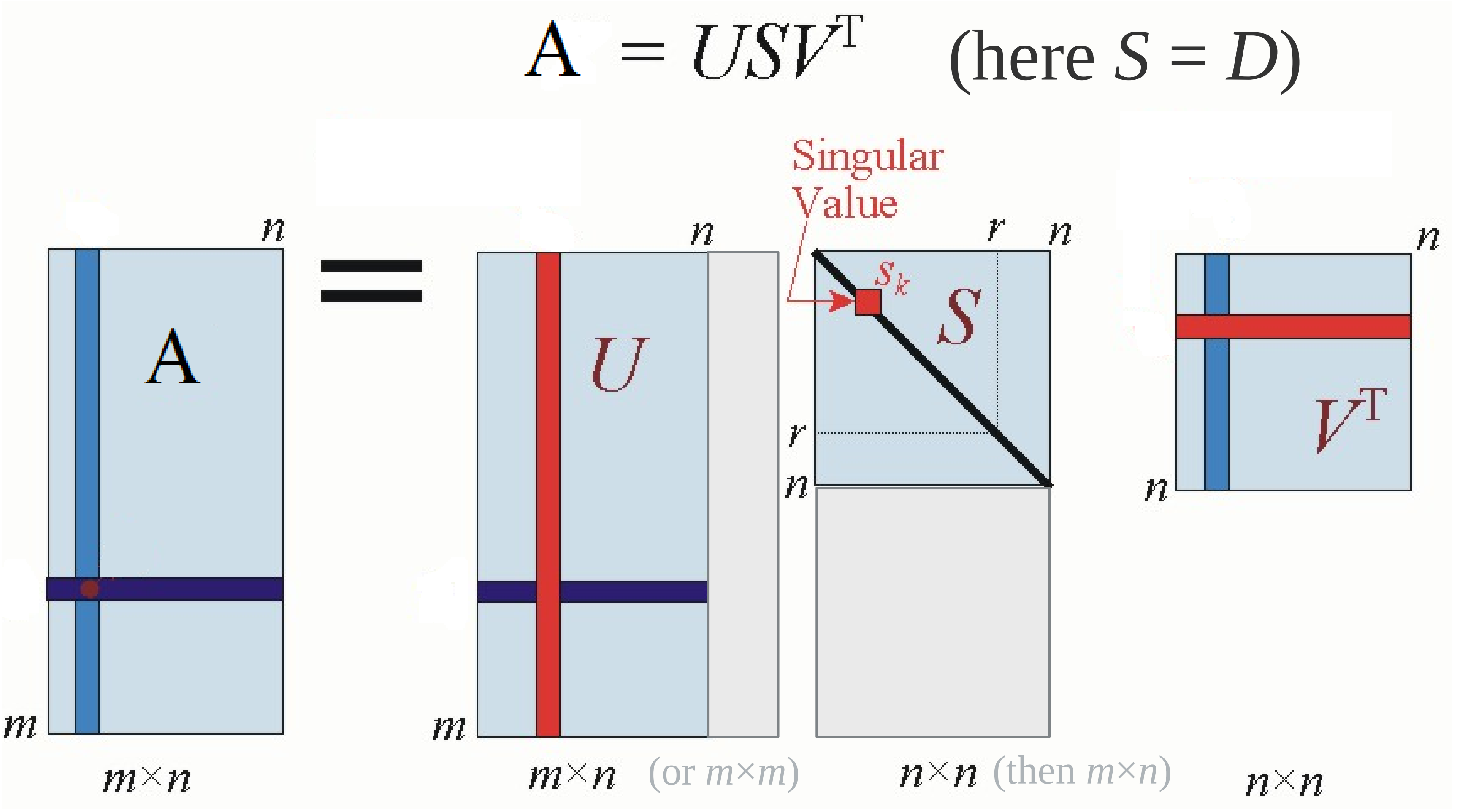
$$\boldsymbol{A} = \boldsymbol{V} \cdot diag(\lambda) \cdot \boldsymbol{V}^{-1}$$

- The singular value decomposition decomposes $\boldsymbol{A}$ into three matrices:

$$\boldsymbol{A} = \boldsymbol{U} \cdot \boldsymbol{D} \cdot \boldsymbol{V}^{T} = \boldsymbol{U} \cdot \Sigma \cdot \boldsymbol{V}^{T}$$

- If $\boldsymbol{A}$ is an $m \times n$ matrix then $\boldsymbol{U}$ is an $m \times m$ matrix and $\boldsymbol{V}$ is an $n \times n$ matrix. $\boldsymbol{U}$, $\boldsymbol{D}$ and $\boldsymbol{V}$ have a special structure.

Singular Value Decomposition

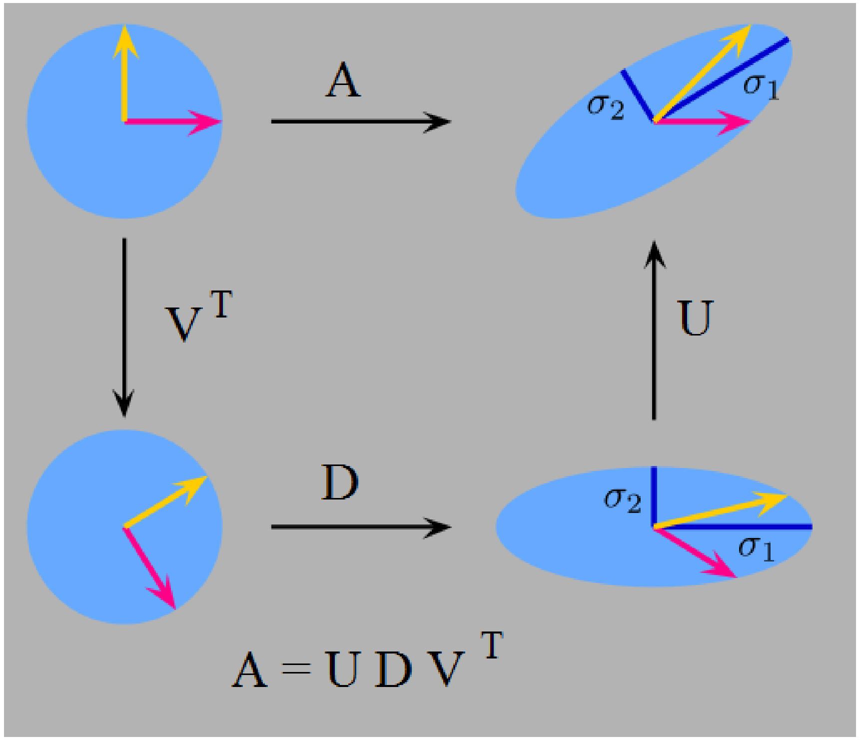


Singular Value Decomposition

- $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, $\boldsymbol{D}$ is a diagonal matrix, but is not necessarily square.
- The elements along the diagonal of $\boldsymbol{D}$ are known as the singular values of the matrix $\boldsymbol{A}$.
- The columns of $\boldsymbol{U}$ are known as the left-singular vectors.
- The columns of $\boldsymbol{V}$ are known as right-singular vectors.
- The left-singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$.
- The right-sing. vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$.
- The non-zero singular values of $\boldsymbol{A}$ are the square roots of the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$.
- We can use the singular value decomposition (SVD) to generalize matrix inversion to non-square matrices.

Singular value decomposition

- First, we see the unit disc with the two unit vectors. The matrix $\boldsymbol{A}$ distorts the disk to an ellipse.
- The SVD decomposes $\boldsymbol{A}$ into 3 transformations: an initial rotation $\boldsymbol{V}^{\mathrm{T}}$, a scaling $\boldsymbol{D}$ along the coordinate axes, and a final rotation $\boldsymbol{U}$.
- The lengths σ_1 and σ_2 of the semi-axes of the ellipse are the singular values of $\boldsymbol{A}$, namely $D_{1,1}$ and $D_{2,2}$.



https://de.wikipedia.org/wiki/Datei:
Singular-Value-Decomposition.svg

2.9 The Moore-Penrose Pseudoinverse

- Matrix inversion is only defined for square matrices.
- Suppose we want to make a left-inverse $\boldsymbol{B}$ of a matrix $\boldsymbol{A}$, so that we can solve a linear equation
$$\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b}$$
by left-multiplying each side to obtain
$$\boldsymbol{x} = \boldsymbol{B} \cdot \boldsymbol{y}$$
- If $\boldsymbol{A}$ is taller than wide, there may be no solution.
- If $\boldsymbol{A}$ is wider than tall, there could be multiple solutions.
- The Moore-Penrose pseudoinverse is defined as
$$\boldsymbol{A}^{+} = \lim_{\alpha \to 0} (\boldsymbol{A}^T \boldsymbol{A} + \alpha \boldsymbol{I})^{-1} \boldsymbol{A}^T$$
yielding in the limit
$$\boldsymbol{A}^{+} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T$$

The Moore-Penrose Pseudoinverse

- Practical algorithms for the pseudoinverse are often based on the formula
$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{D}^+\boldsymbol{U}^T$$
where $\boldsymbol{U}$, $\boldsymbol{D}$ and $\boldsymbol{V}$ are the singular value decomposition of $\boldsymbol{A}$, and the pseudoinverse $\boldsymbol{D}^+$ of a diagonal matrix $\boldsymbol{D}$ is obtained by taking the reciprocal of its non-zero elements, then taking the transpose of this matrix.
- When $\boldsymbol{A}$ has more columns than rows, the pseudo-inverse provides the solution $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{y}$ with minimal Euclidean norm $\|\boldsymbol{x}\|_2$ among all possible solutions.
- When $\boldsymbol{A}$ has more rows than columns, the pseudo-inverse gives us the $\boldsymbol{x}$ for which $\boldsymbol{A}\boldsymbol{x}$ is as close as possible to $\boldsymbol{y}$ in terms of Euclidean norm $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2$.

2.10 The Trace Operator

- The trace operator gives the sum of all of the diagonal entries of a matrix:
$$Tr(\boldsymbol{A}) = \sum_i A_{i,i}$$
- The trace operator provides an alternative way of writing the Frobenius norm of a matrix:
$$\|\boldsymbol{A}\|_F = \sqrt{Tr(\boldsymbol{A}\boldsymbol{A}^T)}$$
- The trace operator is invariant to the transpose operator
$$Tr(\boldsymbol{A}) = Tr(\boldsymbol{A}^T)$$
- The trace of a square matrix composed of many factors is also invariant to cyclic permutation, if the shapes of the corresponding matrices are suitable:
$$Tr(\boldsymbol{ABC}) = Tr(\boldsymbol{CAB}) = Tr(\boldsymbol{BCA})$$

The Trace Operator

- This invariance to cyclic permutation holds even if the resulting product has a different shape.
  For example, for $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times m}$, we have
  $$Tr(\boldsymbol{AB}) = Tr(\boldsymbol{BA})$$
  even though $\boldsymbol{AB} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{BA} \in \mathbb{R}^{n \times n}$.
- A scalar is its own trace: $Tr(a) = a$.

2.11 The Determinant

- The determinant of a square matrix, denoted $\det(\boldsymbol{A})$, is equal to the product of all the eigenvalues of the matrix.
- The absolute value of the determinant can be thought of as a measure of how much the multiplication by the matrix expands or contracts space.
- If the determinant is 0, then space is contracted completely along at least one dimension, causing it to lose all of its volume.
- If the determinant is 1, then the transformation preserves volume.

$$\det\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = a_{11}a_{22} - a_{12}a_{21}$$

$$\det\left(\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\right) = 4 - 6 = -2$$

$$\det\left(\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}\right) = 4 - 4 = 0$$