Deep Learning

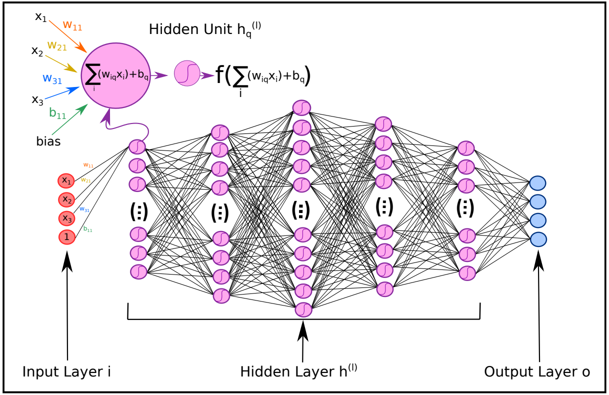
Summer Semester 2024

Monday, April 8, 2024

Prof. Dr.-Ing. Christian Bergler | OTH Amberg-Weiden

Multi-Layer Perceptron

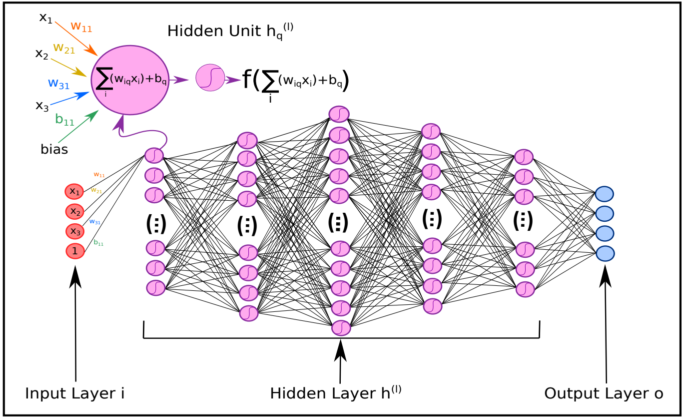
Hidden Neuron



- Neural networks model a functional representation
- Input $x = [x_1, x_2, x_3, 1] \xrightarrow{\texttt{f}(\texttt{x},\theta)}$ Output $y = [y_1, y_2, y_3, y_4]$
- Single layer-specific hidden neuron $h_q^{(l)} = \texttt{f}(\sum_i w_{iq}x_i + b_q)$

Source: Christian Bergler, Dissertation "Deep Learning Applied To Animal Linguistics", Figure 10.2, 2023

Multi-Layer Perceptron
Hidden Neuron



- $h_1^{(1)} =$

Source: Christian Bergler, Dissertation "Deep Learning Applied To Animal Linguistics", Figure 10.2, 2023

Ostbayerische Technische Hochschule Amberg-Weiden

Hidden Layer as Vector-Matrix-Product

$$h^{(l)} = \mathtt{f}\left(\begin{bmatrix} w_{11} & \cdots & w_{1M} \\ w_{21} & \cdots & w_{2M} \\ \vdots & \vdots & \vdots \\ w_{N1} & \cdots & w_{NM} \\ b_1 & \cdots & b_M \end{bmatrix}^{T(l)}_{N+1\times M} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \\ 1 \end{bmatrix}^{(l-1)}_{N+1\times 1} \right) = \begin{bmatrix} \mathtt{f}(w_{11}x_1 + \cdots + w_{N1}x_N + b_1) \\ \mathtt{f}(w_{12}x_1 + \cdots + w_{N2}x_N + b_2) \\ \vdots \\ \mathtt{f}(w_{1M}x_1 + \cdots + w_{NM}x_N + b_M) \end{bmatrix}^{(l)}_{M\times 1}$$

- $h^{(l)} = \mathtt{f}(z^{(l)}) = \mathtt{f}(W^{T(l)}x + b^{(l)})$
- $N =$ Number of inputs
- $M =$ Number of hidden neurons
- $l =$ Layer

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Continuity

Let $D \subseteq \mathbb{R}$ be an interval and $f : D \to \mathbb{R}$ a function. Then f is called continuous in $\bar{x} \in D$ if for all sequences (x_n) in D with $\lim\limits_{n\to\infty} x_n = \bar{x}$ it holds that

$$\lim_{n\to\infty} f(x_n) = f(\bar{x})$$

Furthermore, f is called continuous (in D) if f is continuous at all points $\bar{x} \in D$.

Alternative Specification:

f is called continuous in $\bar{x} \in D$ if $\lim_{x\to\bar{x}} = f(\bar{x})$ is valid.

Example Discontinuous Function:

The function $f : \mathbb{R} \to \mathbb{R}$ with

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \;, \\ 0 & \text{otherwise} \end{cases}$$

is discontinuous in $\bar{x} = 0$.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Derivation

Let $f : \mathbb{R} \to \mathbb{R}$ be a function and $x \in \mathbb{R}$. We say that f is differentiable at the point $x \in \mathbb{R}$ if the limit value

$$f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

exists. The limit value is called the **derivative** of f at the point x, or $f'(x)$ for short. The function f is called **differentiable** if it is differentiable at all points. The function $f' : x \mapsto f'(x)$ is then called **derivative function** of f.

Exercise: Calculate the derivative of $f(x) = x^2$ at the point x

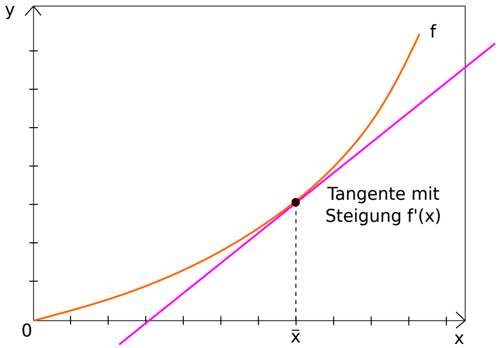
Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Additional Notes

Additional Notes

Mathematical Foundation

(Geometric) Interpretation Derivative



- The derivative at a point $\bar{x}$ indicates the gradient (slope) of a function at this point

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Ostbayerische Technische Hochschule Amberg-Weiden

Rule Set

Let $f, g : \mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R}$ and f, g be differentiable at the point x Then

(i) $f + g$ is differentiable at the point x with

$$(f+g)'(x) = f'(x) + g'(x)$$

(ii) $f \cdot g$ is differentiable at the point x with

$$(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$$

(iii) $\frac{f}{g}$ differentiable at the point x with

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

- Of particular importance in the field of “Deep Learning” is the Chain Rule, which is used in a process called backpropagation

Rule Set

Let $f, g : \mathbb{R} \to \mathbb{R}$ be functions and let g be differentiable at the point $x \in \mathbb{R}$ and f at the point $y := g(x) \in \mathbb{R}$. Then the concatenation $f \circ g$ is differentiable at the point x and the following applies

$$(f \circ g)'(x) = f'(y) \cdot g'(x) = f'(g(x)) \cdot g'(x)$$

Example: Calculate $f'(x)$ for $f(x) = \sin(x)^2$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

- The derivative can be used to specify an optimality criterion for differentiable functions

Criterion

If $f : \mathbb{R} \rightarrow \mathbb{R}$ is differentiable and $a \in \mathbb{R}$ is an extreme point of f (maximum or minimum), then

$$f'(a) = 0$$

Remarks:

- This is a necessary optimality criterion
- Points where the derivative is zero are called critical points
- A critical point does not have to have a global maximum or minimum. There can also be a local maximum or minimum or a saddle point
- To decide whether there is a local optimum or a saddle point, the second derivative can be considered (see convexity)

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Additional Notes

Additional Notes

Mathematical Foundation

Partial Derivative

- When training neural networks, functions $f : \mathbb{R}^n \to \mathbb{R}$ act as an error function, also known as loss function
- The input parameters are typically the weights of a neural network, the function value represents an error between the model prediction and the expected model output (ground truth) w.r.t. to the training data
- The aim is to determine the weights in such a way that the error is as small as possible

Partial derivative

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of several variables and let $\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ be given. If the limit value

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) := \lim_{h \to 0} \frac{f(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_n)}{h}$$

exists, it is called the partial derivative of f with respect to x_i at the point $\mathbf{x}$.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Gradient

- If you summarize all partial derivatives in a vector, you get the so-called gradient

Gradient

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real function whose partial derivatives exist at a point $\mathbf{x} \in \mathbb{R}^n$. Then

$$\nabla f(\mathbf{x}) := \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x})\right)^T$$

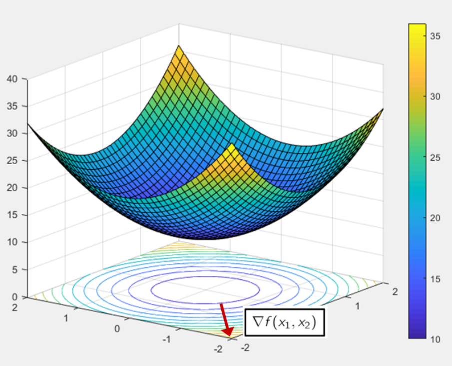
the **Gradient** of f at the point $\mathbf{x}$.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Interpretation Gradient

- The gradient at a point $\mathbf{x} = (x_1, x_2)$ always points in the direction of the steepest incline of the function f:



Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

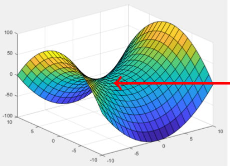
Optimality Condition

Criterion

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and $\mathbf{a} \in \mathbb{R}^n$ is an extreme point of f, then

$$\nabla f(\mathbf{a}) = \mathbf{0} = (0, \ldots, 0)^T$$

- Points at which the gradient disappears are called critical points
- A critical point does not have to be a minimum or maximum, e.g. for the function $f(x_1, x_2) = x_1^2 - x_2^2$ the point $\mathbf{x} = (0, 0)^T$ is a critical point, but it is neither a maximum nor a minimum $\to$ Saddle point



The zero-point here is a critical point, but neither a local maximum nor a local minimum. It is rather a saddle point

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Exercise:

$$f : \mathbb{R}^3 \to \mathbb{R}\ , \quad f(x_1, x_2, x_3) = 2x_1^2 - 3x_2^2 + x_1 x_3^2\ .$$

(a) Calculate the gradient of f

(b) Determine all critical points of f

(c) Investigate whether the critical points are local minima.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Additional Notes

OTH Lecture – Prof. Dr.-Ing. Christian Bergler

Additional Notes

OTH Lecture – Prof. Dr.-Ing. Christian Bergler

Additional Notes

Ostbayerische Technische Hochschule Amberg-Weiden

Semi-Positive-Definite (SPD) matrices

Let $A \in \mathbb{R}^{(n,n)}$ be a symmetric matrix. Then A is

positively definite,	if $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$
positive semidefinite,	if $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$
negative definite,	if $\mathbf{x}^T A\mathbf{x} < 0$ for all $\mathbf{x} \in \mathbb{R}^n$
negative semidefinite,	if $\mathbf{x}^T A\mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^n$

If A is neither positive nor negative semidefinite, it is called **indefinite**.

Definiteness and Eigenvalues

A symmetric matrix $A \in \mathbb{R}^{(n,n)}$ is

positive (semi-)definite	if all eigenvalues are positive (non-negative)
negative (semi-)definite	if all eigenvalues are negative (non-positive)
indefinite	if there are positive and negative eigenvalues

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Optimality criterion

Criterion

Let the function $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and let $\mathbf{a} \in \mathbb{R}^n$ be a critical point of f (i.e. $\nabla f(\mathbf{a}) = \mathbf{0}$) and let $H_f(\mathbf{a})$ be the Hessian matrix of f in $\mathbf{a}$. Is

$H_f(\mathbf{a})$	positive definite, then $\mathbf{a}$ is a strict local minimum of f ,
$H_f(\mathbf{a})$	negative definite, then $\mathbf{a}$ is a strict local maximum of f ,
$H_f(\mathbf{a})$	indefinite, then $\mathbf{a}$ is not a local extremum of f .

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

- Convex and concave functions are particularly interesting from an optimization perspective. The lack of convexity makes the optimization problems in connection with neural networks challenging to solve.

Convex Function

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if for all $\mathbf{x_1}, \mathbf{x_2} \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$ it holds that

$$f(\lambda \mathbf{x_1} + (1 - \lambda)\mathbf{x_2}) \leq \lambda f(\mathbf{x_1}) + (1 - \lambda) f(\mathbf{x_2})$$

It is called **strictly convex** if for all $\mathbf{x_1} \neq \mathbf{x_2}$ and $\lambda \in (0, 1)$ it holds that

$$f(\lambda \mathbf{x_1} + (1 - \lambda)\mathbf{x_2}) < \lambda f(\mathbf{x_1}) + (1 - \lambda) f(\mathbf{x_2})$$

Concave Function

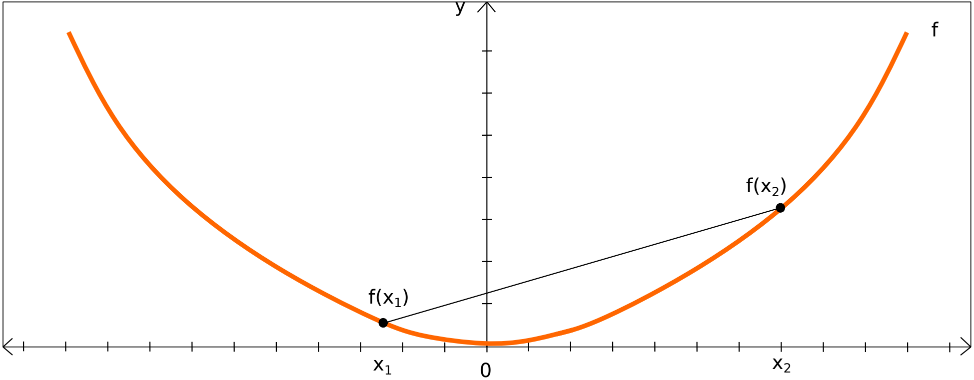
A function $f : \mathbb{R}^n \to \mathbb{R}$ is called (strictly) **concave** if $-f$ is (strictly) convex

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Visualization of Convexity der Konvexität

- In a one dimensional space, strictly convex functions have the property that the connecting line between any two points on the graph of the function always lies above the function:



Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Ostbayerische Technische Hochschule Amberg-Weiden

Local and global minimum of a real-valued function

Let $f : \mathbb{R}^n \to \mathbb{R}$. A point $\mathbf{x_1} \in \mathbb{R}^n$ is called

- **local minimum** of f if in a sufficiently small neighbourhood $\mathcal{U}$ of $\mathbf{x_1}$ it holds that

$$f(\mathbf{x_1}) \leq f(\mathbf{x_2}) \qquad \text{for all } \mathbf{x_2} \in \mathcal{U}$$

- **global minimum** of f if for all $\mathbf{x_2} \in \mathbb{R}^n$ it holds that

$$f(\mathbf{x_1}) \leq f(\mathbf{x_2})$$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Ostbayerische Technische Hochschule Amberg-Weiden

Local minima for convex functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then every local minimum is a global minimum

Local minima for strict convex functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be strictly convex. Then there is at most one local minimum $\bar{x} \in \mathbb{R}^n$ of f

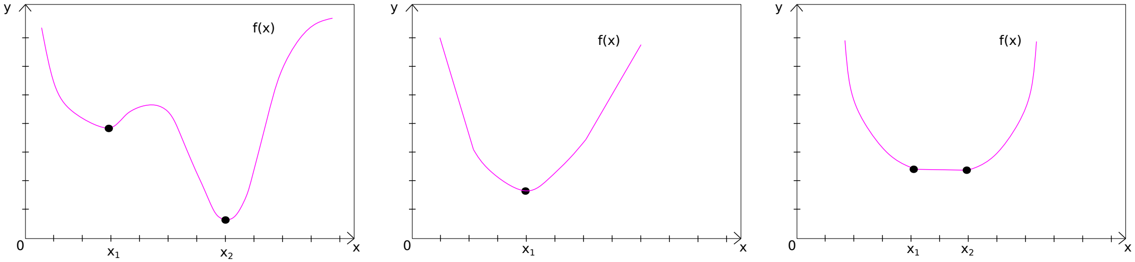
Proof: Suppose there are two different local minima $\tilde{x}$ and $\bar{x}$ of f. Because of the above theorem, both are also global minima and $f(\tilde{x}) = f(\bar{x})$ must hold. From the strict convexity of f it follows for any $\lambda \in (0, 1)$ that

$$f(\lambda\tilde{x} + (1-\lambda)\bar{x}) < \lambda f(\tilde{x}) + (1-\lambda)f(\bar{x}) = f(\tilde{x})$$

This is a contradiction to the actual assumption and thus every strict convex function possesses at most one local minimum!

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Examples



- Type of convexity and minimum?

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Ostbayerische Technische Hochschule Amberg-Weiden

Second order convexity criteria

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. For $x \in \mathbb{R}^n$ let $H(x)$ be the Hessian matrix at the point x. Then applies:

- If $H(x)$ is positive semidefinite for all $x \in \mathbb{R}^n$, then f is convex
- If $H(x)$ is even positive definite for all $x \in \mathbb{R}^n$, then f is strictly convex

Questions for Understanding:

- What does the above criterion mean for functions with a single variable?
- Does every strictly convex function have a global minimum? If not, give an example of a strictly convex function that does not have a global minimum.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

- The model function of neural networks can be written down compactly with the help of vectors and matrices.

Matrix

Let m and n be natural numbers. By an $m \times n$-matrix over the body $\mathbb{R}$ we mean a scheme of numbers of the form

$$A = (a_{ij})_{\substack{i=1,\dots,m \\ j=1,\dots,n}} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & & & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & \dots & \dots & a_{mn} \end{pmatrix}$$

The set of all $m \times n$ matrices over the body $\mathbb{R}$ is called $\mathbb{R}^{m \times n}$.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Addition and Scalar Multiplication for Matrices

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & & & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & \dots & \dots & a_{mn} \end{pmatrix} \text{ und } B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & & & b_{2n} \\ \vdots & & \ddots & \vdots \\ b_{m1} & \dots & \dots & b_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Addition

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & & & a_{2n} + b_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} + b_{m1} & \dots & \dots & a_{mn} + b_{mn} \end{pmatrix}$$

Scalar Multiplication

$$\lambda \cdot A = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & & & \lambda a_{2n} \\ \vdots & & \ddots & \vdots \\ \lambda a_{m1} & \dots & \dots & \lambda a_{mn} \end{pmatrix}, \quad \lambda \in \mathbb{R}$$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Transposed Matrix

Definition

The transpose A^T of an $m \times n$-matrix A is the $n \times m$ matrix whose columns are equal to the rows of A and whose rows are equal to the columns of A. The following therefore applies

$$(a_{ij})^T = (a_{ji}) \ .$$

Rule Set

For any $m \times n$-matrices A and B and $\lambda \in \mathbb{R}$ applies:

$$(A + B)^T = A^T + B^T \ ,$$
$$(\lambda A)^T = \lambda A^T \ ,$$
$$(A^T)^T = A \ .$$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Exercise:

$$A = \begin{pmatrix} 2 & -1 \\ 1 & 4 \\ 3 & 5 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

- Calculate $A + B$ and $(A + B)^T$
- Determine the matrices A^T and B^T
- Calculate $A^T + B^T$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

OTH Lecture – Prof. Dr.-Ing. Christian Bergler

Additional Notes

Ostbayerische Technische Hochschule Amberg-Weiden

Symmetric and adjoint matrix

- A square matrix $A \in \mathbb{R}^{n \times n}$ with real entries is called **symmetric** if

$$A = A^T$$

- A square matrix $A \in \mathbb{C}^{n \times n}$ with complex entries is called **adjoint** (refers to the conjugate transpose) if

$$A^* := \overline{A}^T = A$$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Multiplication of Matrices

- Besides the addition and scalar multiplication, matrices can also be multiplied under certain conditions

Matrix-matrix Multiplication

Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times k}$. The product $C := A \cdot B \in \mathbb{R}^{m \times k}$ is an $m \times k$-matrix with the entries

$$c_{ij} = \sum_{k=1}^{n} a_{ik} \cdot b_{kj} \enspace .$$

Remarks:

- The prerequisite for multiplying two matrices is that the number of columns in the left-hand matrix is the same as the number of rows in the right-hand matrix.
- For square matrices $A, B \in \mathbb{R}^{n \times m}$, both the product $A \cdot B$ and the product $B \cdot A$ are well-defined.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Example Matrix-Matrix Multiplication

- The following matrices are given:

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 1 \\ 2 & -1 \\ 1 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- Calculate all possible matrix products from these four matrices

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Additional Notes

- A special case of matrix-matrix multiplication is the multiplication of a matrix with a vector (= matrix with single column)

Matrix-Vector Multiplication

For $A \in \mathbb{R}^{m \times n}$ and $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ applies

$$Ax = \mathbf{a}^{(1)} \cdot x_1 + \mathbf{a}^{(2)} \cdot x_2 + \ldots + \mathbf{a}^{(n)} \cdot x_n$$

where $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)} \in \mathbb{R}^m$ denote the columns of the matrix A

- Note: When multiplying a matrix by a vector, linear combinations of the columns are formed

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Rules for Matrix-Matrix Multiplication

- If A, B, C are matrices with suitable dimensions and $\lambda \in \mathbb{R}$ is a scalar, then:

$$(\lambda A)B = \lambda(AB) = A(\lambda B)$$
$$A(BC) = (AB)C$$
$$(A + B)C = AC + BC$$
$$A(B + C) = AB + AC$$
$$(AB)^T = B^T A^T$$

Remarks:

- The matrix-matrix multiplication is generally not commutative.
- Example: calculate the products $A \cdot B$ and $B \cdot A$ for the following matrices:

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix} B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Ostbayerische Technische Hochschule Amberg-Weiden

Linear Mapping

A mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ is called **linear mapping** if the following applies to all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$:

(i) $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$

(ii) $f(\lambda \mathbf{x}) = \lambda f(\mathbf{x})$

Remarks:

- If $A \in \mathbb{R}^{m \times n}$ is a matrix, then the mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ given by $f(\mathbf{x}) = A\mathbf{x}$ is a linear mapping.
- If you add a contant vector to a linear mapping, the resulting mapping is called affine-linear. For example, the mapping $g : \mathbb{R}^n \to \mathbb{R}^m$ given by $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ is an affine-linear mapping.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Mathematical Foundation

Norms

- Norms are used to measure the lengths of vectors. These are functions that assign a non-negative number to a vector. The following properties must be fulfilled:

Norm on $\mathbb{R}^n$

A norm is a function $f : \mathbb{R}^n \to \mathbb{R}$ with the following properties:

- $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $f(\mathbf{x}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$ (positive definiteness)
- $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ (triangle inequality)
- $f(\lambda \mathbf{x}) = |\lambda| f(\mathbf{x})$ (positive homogeneity)

- L_2-Norm: $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T \cdot \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$
- L_1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
- L_p-norm: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$ for $p > 0$.

Source: OTH-AW, Electrical Engineering, Media and Computer Science, Fabian Brunner – Vorlesung Deep Learning, Mathematische Grundlagen

Further Questions?

https://www.oth-aw.de/hochschule/ueber-uns/personen/bergler-christian/

Source: https://emekaboris.medium.com/the-intuition-behind-100-days-of-data-science-code-c98402cdc92c