Neural Networks
Lecture Notes
Mathematical preliminaries

Marcel van Gerven

1 Learning Goals

After studying the lecture material you should:

1. understand the different branches of mathematics which neural network theory relies on.

2 Notes

Neural network theory relies on different branches of mathematics. Linear algebra is essential since most neural network algorithms can be written very efficiently in terms of matrix-vector multiplications. Probability theory is essential since many neural network algorithms and objective functions have a probabilistic interpretation. Calculus is essential since the most common ANN learning algorithms make use of gradient descent methods. This requires computing derivatives which allow small steps in weight space to improve an objective function. It also requires knowledge of numerical computing techniques to make these algorithms work on finite precision machines.

2.1 Notation

a	a scalar
$\mathbf{a}$	a vector
a_i	element i of a vector
$\mathbf{A}$	a matrix
$A_{i,j}$	element i, j of a matrix
$\mathbf{A}^{-1}$	matrix inverse
$\|\mathbf{A}\|$	matrix determinant
$\mathbf{I}$	identity matrix
$\|\|\mathbf{x}\|\|_p$	L^p norm
$\|\|\mathbf{x}\|\|$	Euclidean (L^2) norm
$\mathbf{A}^T$	matrix transpose
$f\colon \mathbb{A} \to \mathbb{B}$	a function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	composition of functions f and g
$\log x$	natural logarithm of x
$\sigma(x)$	sigmoid function
$\frac{dy}{dx}$	derivative of y with respect to x
$\frac{\partial y}{\partial x}$	partial derivative of y with respect to x
$\nabla_{\mathbf{x}}$ y	gradient of y with respect to $\mathbf{x}$
$p(x)$	probability distribution over a continuous variable
$p(x \mid y)$	conditional probability distribution
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathbb{E}_{\mathbf{x}\sim p}[f(\mathbf{x})]$	expectation of $f(\mathbf{x})$ with respect to $p(\mathbf{x})$

2.2 Linear algebra

Linear algebra is concerned with vector spaces and linear mappings between such spaces. Its objects of interest are scalars, vectors, matrices and tensors (i.e. multidimensional arrays).

The matrix transpose is defined as:

$$(\mathbf{A}^T)_{i,j} = \mathbf{A}_{j,i}\,.$$

The matrix inverse $\mathbf{A}^{-1}$ of $\mathbf{A}$ is defined as that matrix such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

with $\mathbf{I}$ the identity matrix.

The determinant $|\mathbf{A}|$ can be viewed as the scaling factor of the transformation described by $\mathbf{A}$. It can be computed explicitly but we will not make use of this in the course.

You need to be familiar with matrix operations such as addition and multiplication, as well as their properties. Recall that a matrix product $\mathbf{C} = \mathbf{AB}$ is defined by

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}$$

This is not the same as the product of individual elements. This latter operation is called the element-wise product or Hadamard product, and is written as $\mathbf{A} \odot \mathbf{B}$.

The L^p norm is defined as

$$||\mathbf{x}||_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

and is used to measure the size of a vector. The Euclidean norm $||\mathbf{x}||$ (i.e. $||\mathbf{x}||_2$) is the Euclidean distance from the origin to the point defined by $\mathbf{x}$. Note that the squared Euclidean norm is equal to $\mathbf{x}^T\mathbf{x}$.

2.3 Probability theory

Probability theory describes the world in terms of random variables that can be either continuous or discrete. Probabilities that random variables take on certain values are described by probability distributions. More formally, this is a probability mass function $P(x)$ for a discrete random variable and a probability density function $p(x)$ for a continuous random variable. A probability mass function $P(x)$ satisfies:

- the domain of P is the set of possible states of x.
- $0 \leq P(x) \leq 1$ for all x
- $\sum_x P(x) = 1$

A probability density function $p(x)$ satisfies:

- the domain of p is the set of possible states of x.
- $p(x) \geq 0$ for all x
- $\int p(x)dx = 1$

Note that the probability density does not give the probability of x but rather then probability of landing in an infinitesimal region with volume δx given by $p(x)\delta x$.

These definitions can be extended to the multivariate case. Moreover, probability theory defines several operations which you need to be familiar with, such as marginal and conditional probability and the chain rule.

We can also use probability distributions to compute expectations of functions:

$$\mathbb{E}_{\mathbf{x}\sim p}[f(\mathbf{x})] = \int p(\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

Several standard distributions have been defined, of which the (multivariate) Gaussian is an important example. It is given by

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

where $\mathbf{x}$ is an n-dimensional vector, $\boldsymbol{\mu}$ is a mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix, defined by $\mu_i = \mathbb{E}[x_i]$ and $\boldsymbol{\Sigma}_{i,j} = \text{Cov}(x_i, x_j) = \mathbb{E}[(x_i - \mathbb{E}(x_i))(x_j - \mathbb{E}(x_j))]$.

2.4 Calculus

Neural networks often require the computation of derivatives and partial derivatives. The derivative is written as

$$\frac{dy}{dx}$$

E.g.

$$\frac{df(x)}{dx} = f'(x) = 1$$

if $f(x) = x$.

The partial derivative is written as

$$\frac{\partial y}{\partial x}.$$

Recall that the partial derivative is the derivative of a function of two or more variables with respect to one variable, the other(s) being treated as constant. E.g.

$$\frac{\partial f(x,y)}{\partial x} = 1$$

if $f(x, y) = x + y$.

3 Reading material

For this lecture it is important to get up to speed in terms of mathematical background. In terms of mathematical techniques, please consult linear algebra and probability theory in case you don't understand particular concepts. Note that as these pages are quite advanced you don't need to study concepts that are not discussed in the course.