National University of Singapore
EE5904/ME5404 Neural Networks

Mathematical Review

1) Mathematical notation

$\exists$ - there exist
$\forall$ - for all
$\Leftrightarrow$ - if and only if

2) Euclidean Space, R^n

Euclidean space, R^n :
The set of all column vectors, $\mathbf{x}=(x_1,\ldots,x_n)^T$

Addition of two vectors, $\mathbf{x}+\mathbf{y}=(x_1+y_1,\ldots,x_n+y_n)^T$

Multiplication of a vector with a scalar, λ, are defined as : $\lambda\mathbf{x}=(\lambda x_1,\ldots, \lambda x_n)^T$

Inner product (or scalar product) of two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is denoted by $\mathbf{x}^T\mathbf{y}$:

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

A norm $\|\,.\,\|$ on R^n is a mapping that assigns a scalar $\|\mathbf{x}\|$ to every $\mathbf{x}\in R^n$ and that has the following properties:

(a) $\|\mathbf{x}\| \geq 0 \ \forall \ \mathbf{x} \in R^n$
(b) $\|c\mathbf{x}\| = |c| \, \|\mathbf{x}\|$ for every $c \in R$ and every $\mathbf{x} \in R^n$

(c) $||\mathbf{x}|| = 0 \Leftrightarrow$ if $\mathbf{x} = \mathbf{0}$

(d) $||\mathbf{x}+\mathbf{y}|| \le ||\mathbf{x}||+||\mathbf{y}|| \ \forall\ \mathbf{x},\mathbf{y} \in R^n$ (Triangular inequality)

Euclidean norm of $\mathbf{x}$, $||\mathbf{x}|| = \sqrt{(\mathbf{x}^T\mathbf{x})} = \left(\sum_{i=1}^{n} x_i^{\ 2}\right)^{1/2}$

Schwartz inequality (applies to the Euclidean norm):

$|\mathbf{x}^T\mathbf{y}| \le ||\mathbf{x}||.||\mathbf{y}||$

with equality holding if and only if $\mathbf{x} = \alpha\mathbf{y}$ for some scalar α

3) Matrices

m x n matrix => $\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & \\ a_{m1} & \cdots & & a_{mn} \end{bmatrix} = \mathbf{A} = [a_{ij}]$

Product of a matrix $\mathbf{A}$ and a scalar α, $\alpha\mathbf{A}$ or $\mathbf{A}\alpha = [\alpha a_{ij}]$

Product of two matrices $\mathbf{A}$ and $\mathbf{B}$:

Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ be a mxn, nxp and mxp matrices, respectively.

$\mathbf{AB} = \mathbf{C} = [c_{ij}]$, where $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$

Transpose of a mxn matrix $\mathbf{A}$ is the nxm matrix $\mathbf{A}^T = [a'_{ij}] = [a_{ji}]$

A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A}^T=\mathbf{A}$

Non-singular matrix

A square matrix (nxn) $\mathbf{A}$ is called *non-singular* or *invertible* if there is an nxn matrix called the inverse of $\mathbf{A}$ (denoted by $\mathbf{A}^{-1}$), such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$, where $\mathbf{I}$ is the nxn identity matrix.

Positive Definite and Semidefinite Matrices

A square nxn matrix $\mathbf{A}$ is said to be *positive semidefinite* if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0 \ \forall \mathbf{x} \in R^n$

A square nxn matrix $\mathbf{A}$ is said to be *positive definite* if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \ \forall \mathbf{x} \in R^n$

The matrix $\mathbf{A}$ is said to be *negative* semidefinite (definite) if $-\mathbf{A}$ is positive semidefinite (definite).

4) Derivatives

Let $f : R^n \to R$ be some function.

Partial Derivative:

For a fixed $\mathbf{x} \in R^n$, the first partial derivative of f at the point $\mathbf{x}$ in the ith coordinate is defined by

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \lim_{\alpha \to 0} \frac{f(\mathbf{x} + \alpha \mathbf{e}_i) - f(\mathbf{x})}{\alpha}$$

where $\mathbf{e}_i$ is the *i*th unit vector.

Gradient:

If the partial derivatives with respect to all coordinates exist, *f* is called differentiable at **x** and its *gradient* at **x** is defined as:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

Note: ∇f is orthogonal to the corresponding level surface $f(x_1,\ldots,x_n)=C$

Various Definitions

The function *f* is called *differentiable* if it is differentiable at every $\mathbf{x} \in R^n$.

The function *f* is said to be *continuously differentiable* if $\nabla f(\mathbf{x})$ exists for every **x** and is a continuous function of **x**.

The function *f* is said to be *twice differentiable* if the gradient $\nabla f(\mathbf{x})$ is itself a differentiable function.

Hessian Matrix

For this case, we can define Hessian matrix of f at $\mathbf{x}$,

$\nabla^2 f(\mathbf{x}) = \left[\dfrac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right]$, i.e. the elements are the second partial derivatives of f at $\mathbf{x}$.

The function f is *twice continuously differentiable* if $\nabla^2 f(\mathbf{x})$ exists and is continuous.

Vector-valued function

A vector-valued function $\mathbf{f} : R^n \to R^m$ is differentiable (continuously differentiable) if each component f_i of $\mathbf{f}$ is differentiable (continuously differentiable).

Gradient matrix of the vector-valued function

The gradient matrix of $\mathbf{f}$, $\nabla \mathbf{f}(\mathbf{x})$, is the nxm matrix whose ith column is $\nabla f_i(\mathbf{x})$,

$$\nabla \mathbf{f}(\mathbf{x}) = \left[\nabla f_1(\mathbf{x}) \quad \cdots \quad \nabla f_m(\mathbf{x}) \right]$$

Jacobian of the vector-valued function

Jacobian of $\mathbf{f} = (\nabla \mathbf{f}(\mathbf{x}))^T$

Chain rule

Let $\mathbf{f} : R^k \to R^m$ and $\mathbf{g} : R^m \to R^n$ be continuously differentiable functions, and let $\mathbf{h}(\mathbf{x}) = \mathbf{g}(\mathbf{f}(\mathbf{x}))$.

The chain rule for differentiation states that

$$\nabla \mathbf{h}(\mathbf{x}) = \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x})\, \nabla_{\mathbf{f}}\, \mathbf{g}(\mathbf{f}(\mathbf{x})),\ \forall\ \mathbf{x} \in R^k$$

Special examples

If $\mathbf{f} : R^n \rightarrow R^m$ is of the form $\mathbf{f}(\mathbf{x}) = \mathbf{Ax}$, where $\mathbf{A}$ is an m x n matrix,

$$\nabla \mathbf{f}(\mathbf{x}) = \mathbf{A}^T$$

If $f : R^n \rightarrow R$ is of the form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{Ax}/2$, where $\mathbf{A}$ is a symmetric n x n matrix,

$$\nabla f(\mathbf{x}) = \mathbf{Ax}$$
$$\nabla^2 f(\mathbf{x}) = \mathbf{A}$$

5) Unconstrained Optimization Techniques

Local and global minima

x^*: local min if $\exists\ \varepsilon>0$ s.t. $f(x^*)\leq f(x)$, $\forall x$ in $\|x-x^*\| \leq \varepsilon$

x^*: strictly local min if $\exists\ \varepsilon>0$ s.t. $f(x^*)<f(x)$, $\forall x$ in $\|x-x^*\| \leq \varepsilon$

x^*: global min if $\exists\ \varepsilon>0$ s.t. $f(x^*)\leq f(x)$, $\forall x\in R^n$