Applied Math Review for Deep Learning

UCSB CS165B W22 Section 1

Yijun Xiao

Table of Contents

1. Linear Algebra

2. Calculus

3. Probability

Vectors, Matrices and Tensors

Scalar	Vector	Matrix	Tensor
1	$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$	$\begin{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} & \begin{bmatrix} 3 & 2 \end{bmatrix} \\ \begin{bmatrix} 1 & 7 \end{bmatrix} & \begin{bmatrix} 5 & 4 \end{bmatrix} \end{bmatrix}$

Vectors, Matrices and Tensors

Notations

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}, \text{ or } \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \qquad \boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

We often denote the set of all possible real value vectors with d elements as $\mathbb{R}^d$. The shape of such vectors is $d \times 1$, i.e. they are column vectors.

Similarly, the set of real value matrices of shape $m \times n$ is denoted as $\mathbb{R}^{m \times n}$.

Matrix Transpose

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \rightarrow \boldsymbol{A}^\top = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}$$

Formally, the **transpose** of a matrix $\boldsymbol{A}$ is denoted as $\boldsymbol{A}^\top$. It is defined such that

$$(\boldsymbol{A}^\top)_{i,j} = \boldsymbol{A}_{j,i}$$

The transpose of a vector $\boldsymbol{x}$ therefore becomes a row vector.

Matrix Multiplication

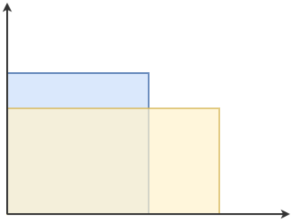
$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} & a_{21}b_{13} + a_{22}b_{23} \end{bmatrix}$$

For matrix $\boldsymbol{A}$ of shape $m \times n$ and matrix $\boldsymbol{B}$ of shape $n \times p$, the **matrix product** of the two is another matrix $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$ of shape $m \times p$ where

$$\boldsymbol{C}_{i,j} = \sum_k \boldsymbol{A}_{i,k}\boldsymbol{B}_{k,j}$$

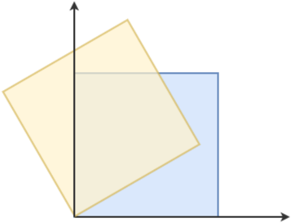
The dot product between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with the same dimensions can be written as $\boldsymbol{x}^\top \boldsymbol{y}$.

Matrix Multiplication as Linear Transformation



$$\begin{bmatrix} 1.5 & 0 \\ 0 & 0.75 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Scaling



$$\begin{bmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Rotation

Identity and Inverse Matrices

An n-dimensional **identity matrix** is denoted as $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$. All its diagonal elements are 1's and all other elements are 0's. For example,

$$\boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

It is called identity matrix because for any n-dimensional vector $\boldsymbol{x}$, $\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{x}$.

Identity and Inverse Matrices

An n-dimensional **identity matrix** is denoted as $\boldsymbol{I}_n \in \mathbb{R}^{n \times n}$. All its diagonal elements are 1's and all other elements are 0's. For example,

$$\boldsymbol{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

It is called identity matrix because for any n-dimensional vector $\boldsymbol{x}$, $\boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}$.

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

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

Finding the inverse of a matrix $\boldsymbol{A}$ helps us to solve linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. i.e. $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$.

Vector Norms

Norms are functions to measure the size of a vector. The L^p norm is given by

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

L^2 norm, or **Euclidean norm**, is frequently used in machine learning and simply represents the Euclidean distance from point $\boldsymbol{x}$ to the origin.

Table of Contents

1. Linear Algebra

2. Calculus

3. Probability

Derivatives and Gradients

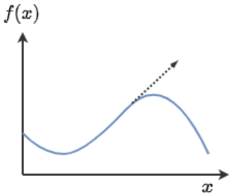
For a function $f : \mathbb{R} \rightarrow \mathbb{R}$, the **derivative** of f is defined as

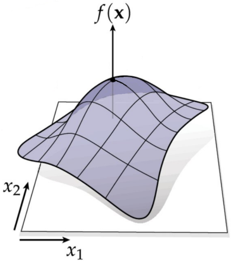
$$f'(x) = \frac{df}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

The derivative gives the slope of the function at x.

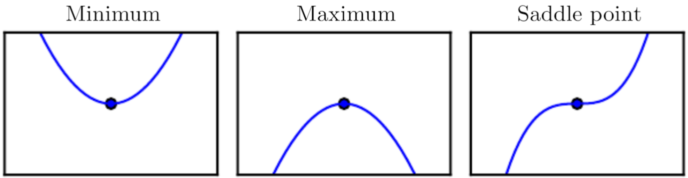
For a general function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the **gradient** of f with respect to the input $\boldsymbol{x}$ is defined as the vector of all partial derivatives

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right]^\top$$





Stationary Points



Points where $f'(x) = 0$ are called **stationary points**. Local minimum, local maximum, and saddle points are all stationary.

Derivative Calculation

Common Functions:
$$\frac{d}{dx}x^n = n \cdot x^{n-1},$$
$$\frac{d}{dx}e^x = e^x,$$
$$\frac{d}{dx}\log x = \frac{1}{x}$$

Product Rule:
$$\frac{d}{dx}f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$

Chain Rule:
$$\frac{d}{dx}f(g(x)) = f'(g(x)) \cdot g'(x)$$

Chapter 2 of the Matrix Cookbook[1] has all formula needed to compute derivatives with respect to vectors and matrices.

[1] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. "The matrix cookbook". In: *Technical University of Denmark* 7.15 (2008), p. 510.

Derivative Calculation

Exercise

Consider vector $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n$ and scalar b, find $\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})$ with the function f defined as

$$f(\boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w}^\top \boldsymbol{x} + b)}}$$

Derivative Calculation

Exercise

Consider vector $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n$ and scalar b, find $\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})$ with the function f defined as

$$f(\boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w}^\top \boldsymbol{x} + b)}}$$

f can be seen as a composite of $f_1(x) = \frac{1}{x}$, $f_2(x) = 1 + e^{-x}$, and $f_3(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$:

$$f(\boldsymbol{x}) = f_1(f_2(f_3(\boldsymbol{x})))$$

Now let's denote $y = f_3(\boldsymbol{x})$ and $z = f_2(y)$. Using chain rule:

$$\frac{\partial}{\partial \boldsymbol{x}} f_1(f_2(f_3(\boldsymbol{x}))) = \frac{\partial f_1(z)}{\partial z} \cdot \frac{\partial f_2(y)}{\partial y} \cdot \frac{\partial f_3(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

$$= -z^{-2} \cdot (-e^{-y}) \cdot \boldsymbol{w} = \frac{e^{-(\boldsymbol{w}^\top \boldsymbol{x} + b)}}{(1 + e^{-(\boldsymbol{w}^\top \boldsymbol{x} + b)})^2} \cdot \boldsymbol{w}$$

Table of Contents

1. Linear Algebra

2. Calculus

3. Probability

Key Concepts

Conditional Probability: $$p(y|x) = \frac{p(x, y)}{p(x)}$$

Marginal Probability: $$p(x) = \int p(x, y)\, dy$$

Independence: $$p(x, y) = p(x)p(y)$$

Expectation: $$\mathbb{E}_{x \sim p}\left[f(x)\right] = \int f(x)p(x)\, dx$$

Bayes' Rule

When we are interested in the value of $P(x|y)$, but only have access to $P(x)$ and $P(y|x)$, we can apply the Bayes' rule to compute it.

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)} = \frac{P(x)P(y|x)}{\sum_x P(x)P(y|x)}$$

$P(x)$ if often referred to as the **prior distribution**, and $P(x|y)$ is known as the **posterior distribution** of x.

Bayes' Rule Application

The distribution of Mark's body temperature is $\mathcal{N}(98, 0.5)$ under healthy conditions. When sick, the distribution is $\mathcal{N}(99, 0.7)$. We know Mark is sick 10% of the time, and his body temperature right now is 98.5. What is the probability that Mark is sick at the moment?

Bayes' Rule Application

The distribution of Mark's body temperature is $\mathcal{N}(98, 0.5)$ under healthy conditions. When sick, the distribution is $\mathcal{N}(99, 0.7)$. We know Mark is sick 10% of the time, and his body temperature right now is 98.5. What is the probability that Mark is sick at the moment?

We are essentially looking for the posterior distribution of "Mark is sick" given the prior distribution and the conditionals. Denote x as the event "Mark is sick" and y as Mark's body temperature, we have

Prior: $P(x = \mathsf{T}) = 0.1$

Conditionals: $y \mid (x = \mathsf{T}) \sim \mathcal{N}(99, 0.7), \quad y \mid (x = \mathsf{F}) \sim \mathcal{N}(98, 0.5)$

Posterior: $$P(x = \mathsf{T} \mid y = 98.5) = \frac{P(x = \mathsf{T})P(y = 98.5 \mid x = \mathsf{T})}{P(x = \mathsf{T})P(y = 98.5 \mid x = \mathsf{T}) + P(x = \mathsf{F})P(y = 98.5 \mid x = \mathsf{F})}$$