ML101: Introduction to AI & Reinforcement Learning **Lecture Date: January, 2021**

A Brief Review of Linear Algebra

Lecturer: GuangBing Yang *Scribe: Guangbing Yang*

1 General Concepts

- **Linear algebra** is a branch of mathematics providing a concise way to represent and operate on a set of linear equations via vectors and matrices.
- For example, consider the following two linear equations:

$$2x_1 + 3x_2 = 15 \tag{1}$$
$$-x_1 + 2x_2 = 6 \tag{2}$$

  There are two equations and two variables, so you are able to find a unique solution for x_1 and x_2 unless the equations are somehow degenerate, which means the Eq.(1) is the same as the Eq.(2). So, after simply calculation, we have $x_1 = 12/7$ and $x_2 = 27/7$ (you shall solve this in a minute).

- Here the point is not testing you to solve this question, but is to recap the concepts of matrix and vector in linear algebra. Remember that we can represent this question using matrix and matrix operation. In matrix notation, we can write this system compactly as:

$$Ax = y \tag{3}$$

$$A = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix}, y = \begin{bmatrix} 15 \\ 6 \end{bmatrix} \tag{4}$$

- Basic Notation
  1. $A \in \mathbb{R}^{m \times n}$, denotes a matrix, with m rows and n columns, where the entries of A are real numbers.
  2. $x \in \mathbb{R}^n$, denotes a vector with n entries. Normally, one can view a vector as a one column matrix. For example, an n-dimensional vector $x = x_1, x_2, ..., x_n$ can be treated as a matrix with n rows and 1 columns as:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix} \tag{5}$$

     known as a column vector.
  3. If we want to explicitly represent a row vector – a matrix with 1 row and n columns – we typically write x^T (here x^T, which is $x^T = [x_1, x_2, ..., x_n]$ denotes the transpose of x, which we will define shortly).

- For the example given above, the x is represented in the column vector form as:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{6}$$

1. Then, above Eq.(1, 2, 3) system equations can be represented as the matrix form:

$$Ax = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 15 \\ 6 \end{bmatrix} = y \tag{7}$$

- Use the notation a_{ij}, (or A_{ij} , $A_{i,j}$, etc) to denote the entry of A in the ith row and jth column:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \tag{8}$$

- denote the jth column of A by a_j or $A_{:,j}$:

$$A = \begin{bmatrix} | & | & & | & \\ a_1 & a_2 & \dots & a_n & \\ | & | & & | & \end{bmatrix} \tag{9}$$

- denote the ith row of A by a_i^T or $A_{i,:}$:

$$A = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & . & \\ & . & \\ - & a_m^T & - \end{bmatrix} \tag{10}$$

- Note the a_1 and a_1^T are not the same vector, one is all rows in the same one column, another is one row of all columns.

2 Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ is the matrix

$$C = AB \in \mathbb{R}^{m\times p} \tag{11}$$

where

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

Note that in order for the matrix product to exist, the number of columns in A must equal the number of rows in B.

2.1 Vector-Vector Multiplication

For two vectors a, $b \in \mathbb{R}^n$, the outcome of the product $a^T b$, also called the ***inner product*** or ***dot product*** of vectors, is a real number calculated by

$$a^T b \in \mathbb{R} = [a_1 a_2 ... a_n] \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix} = \sum_{i=1}^{n} a_i b_i = (a_1 b_1 + a_2 b_2 + ... + a_n b_n) \tag{12}$$

In this case, $a^T b = b^T a$ because the size of the vector a and b are the same, which is n. In contrast, for two vectors, $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$ with different size, the ***outer product***, $ab^T \in \mathbb{R}^{m \times n}$ is defined as

$$ab^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ a_m \end{bmatrix} [b_1 b_2 ... b_n] = \begin{bmatrix} a_1 b1 & a_1 b_2 & ... & a_1 b_n \\ a_2 b_1 & a_2 b_2 & ... & a_2 b_n \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ a_m b_1 & a_m b_2 & ... & a_m b_n \end{bmatrix} \tag{13}$$

2.2 Matrix-Vector Multiplication

The product of a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $x \in \mathbb{R}^n$ is given as a vector $y = Ax \in \mathbb{R}^m$

- A is given by rows, then

$$y = Ax = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & . & \\ & . & \\ - & a_m^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ . \\ . \\ a_m^T x \end{bmatrix} \tag{14}$$

  which means the ith entry of y is equal to the inner product of the ith row of A and x, $y_i = a_i^T x$.

- if A is given by columns, then

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & ... & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 \end{bmatrix} x_1 + \begin{bmatrix} a_2 \end{bmatrix} x_2 + ... + \begin{bmatrix} a_n \end{bmatrix} x_n \tag{15}$$

  The matrix over vector product y is a linear combination of the columns of A, where the coefficients of the linear combination are given by the entries of x.

- Above cases are matrix multiplying on the right by a column vector. It is also possible to multiply on the left by a row vector, which is given as $y^T = x^T A$ for $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$.

- As before, the y^T can be expressed in two obvious ways, depending on whether matrix A is expressed in terms on its rows or columns. In the first case A is expressed in terms of its columns, which gives

- if A is given by columns, then

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a_1 & a_2 & ... & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x^T a_1 & x^T a_2 & ... & x^T a_n \end{bmatrix} \tag{16}$$

which shows that the ith entry of y^T is equal to the inner product of x and the ith column of A.

- Let us expressing A in terms of rows, the final representation of the vector-matrix product is given as

$$y^T = x^T A = \begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix} \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & . & \\ & . & \\ - & a_m^T & - \end{bmatrix} = x_1 \begin{bmatrix} - & a_1^T & - \end{bmatrix} + x_2 \begin{bmatrix} - & a_2^T & - \end{bmatrix} + ... + x_n \begin{bmatrix} - & a_n^T & - \end{bmatrix} \tag{17}$$

Again, the y^T is a linear combination of the rows of A, where the coefficients for the linear combination are given by the entries of x.

3 Matrix-Matrix Products

- There are several different ways to viewing the matrix-matrix multiplication $C = AB$ as defined at the beginning of this section.
- Here, let us view the most popular way to express matrix-matrix multiplication as a set of vector-vector products. In this case, the (i, j)th entry of C is equal to the inner product of the ith row of A and the jth column of B.

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & . & \\ & . & \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & ... & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & ... & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & ... & a_2^T b_p \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ a_m^T b_1 & a_m^T b_2 & ... & a_m^T b_p \end{bmatrix} \tag{18}$$

Since $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, $a_i^T \in \mathbb{R}^n$ and $b_j \in \mathbb{R}^n$, so these inner products all make sense, where $i \in \{1, ..., m\}$ and $j \in \{1, ..., p\}$

- Other ways, like viewing matrix-matrix multiplication as a set of matrix-vector products will not discuss here, but one can refer any linear algebra text book for more details or go to the Wiki page of the linear algebra.

4 Operations and Properties

4.1 The Identity Matrix and Diagonal Matrix

- The **identity matrix** is a square matrix with ones on the diagonal and aeros everywhere else. It is denoted as $I \in \mathbb{R}^{n \times n}$,

$$I_{ij} = \begin{cases} 1 & \text{if i = j} \\ 0 & \text{if } i \neq j \end{cases} \tag{19}$$

For example, for an identity matrix $I \in \mathbb{R}^{4\times 4}$, the I is expressed as:

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{20}$$

It has a property for all $A \in \mathbb{R}^{m\times n}$,

$$AI = A = IA \tag{21}$$

Note that the two Is in above notation have different dimensions. When identity matrix I is in the right side of the matrix A, it is n-dimension square matrix. When it is in the left side of A, it is a m-dimension square matrix. Generally, the dimensions of I are inferred from context so as to make matrix multiplication possible.

- A ***diagonal matrix*** is a matrix where all non-diagonal elements are zeros, denoted as $D = \text{diag}(d_1, d_2, ..., d_n)$. It is not necessary a square matrix, with

$$D_{ij} = \begin{cases} d_i & \text{if i = j} \\ 0 & \text{if } i \neq j \end{cases} \tag{22}$$

One can see, the identity matrix $I = \text{diag}(1, 1, ..., 1)$

4.2 The Transpose

The transpose of a matrix results from flipping the rows and columns. Given a matrix $A \in \mathbb{R}^{m\times n}$, its transpose, written $A^T \in \mathbb{R}^{n\times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji} \tag{23}$$

For example,

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \tag{24}$$

It's transpose matrix is

$$A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \tag{25}$$

Properties of Matrix Transpose:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

4.3 Symmetric Matrices

Define a square matrix $A \in \mathbb{R}^{n\times n}$ as ***symmetric*** if $A = A^T$. If $A = -A^T$, is is called ***anti-symmetric***. For any square matrix $A \in \mathbb{R}^{n\times n}$, $A + A^T$ is a symmetric matrix, and $A - A^T$ is an anti-symmetric matrix.

Thus, any square matrix $A \in \mathbb{R}^{n\times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since: $A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$, the first part of the right side of the equation is the symmetric matrix and the second half is an anti-symmetric matrix.

4.4 The Trace and Determinant

The ***trace*** of a square matrix $A \in \mathbb{R}^{n\times n}$, denoted $tr(A)$ is the sum of diagonal elements in the matrix:

$$tr(A) = \sum_{i=1}^{n} A_{ii} \tag{26}$$

Properties of ***trace***:

- For $A \in \mathbb{R}^{n\times n}$, $trA = trA^T$
- For $A, B \in \mathbb{R}^{n\times n}$, $tr(A + B) = trA + trB$
- For $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}$, $tr(bA) = btrA$
- For A, B, if $AB \in \mathbb{R}^{n\times n}$ is square, $trAB = trBA$
- For A, B, C, if ABC is square, $trABC = trBAC = trCAB$.

The ***determinant***, denoted $|A|$, for a square matrix $A \in \mathbb{R}^{n\times n}$, is defined as

$$|A| = \sum(\pm 1)A_{1i_1}A_{2i_2}...A_{Ni_N} \tag{27}$$

in which the sum is taken over all products consisting of one element from each row and one element from each column, with a coefficient +1 or -1 according to wether the permutation $i_1, i_2, ..., i_N$ is even or odd, respectively. For example,

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{28}$$

it's determinant is

$$|A| = |1 \times 4 - 2 \times 3| = 2 \tag{29}$$

Properties of the determinant:

- For the determinant of a product of two matrices A and B, given as $|AB| = |A||B|$
- The determinant of an inverse matrix is given by $|A^{-1}| = \frac{1}{|A|}$
- For $A, B \in \mathbb{R}^{n\times m}$, then $|I_n + AB^T| = |I_m + A^T B|$
- As a special case, for a and b are n-dimansional vectors, $|I_n + ab^T| = 1 + a^T b$

4.5 Linear Independence and Rank

- A set of vectors $\{x_1, x_2, ...x_n\} \in \mathbb{R}^n$ is said to be (linearly) ***independent*** if no vector can be represented as a linear combination of the remaining vectors.
- The ***column rank*** of a matrix $A \in \mathbb{R}^{m\times n}$ is the size of the largest subset of columns of A that constitute a linearly independent set.
- For $A \in \mathbb{R}^{m\times n}$, $rank(A) \leq min(m, n)$. If rank(A) = min(m, n), then A is said to be ***full rank***.

4.6 The Inverse

The ***inverse*** of a square matrix $A \in \mathbb{R}^{n\times n}$ is denoted as A_{-1}. It is unique and

$$A^{-1}A = I = AA^{-1}$$

Note that not all matrices, including same square matrices, have inverses. By definition, non-square matrices have no inverses.

Particularly, if A^{-1} exists, we say A is ***invertible*** or ***non-singular***, otherwise, it is ***non-invertible*** or ***singular***. Properties of inverse:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$

Note that in order for a square matrix A to have an inverse A^{-1}, the A must be full rank, which means A's rank must be equal to minimum of row size and column size, which is $\mathrm{rank}(A) = \min(m, n)$.

For the example shown in the Eq.(3), we can solve the system equation in Eq.(1) and calculate values of x_1 and x_2 using the inverse matrix, like this:

$$Ax = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix} x = \begin{bmatrix} 15 \\ 6 \end{bmatrix} = y \tag{30}$$

with the inverse A^{-1},

$$A^{-1}Ax = A^{-1}y \tag{31}$$

$$x = A^{-1}y \tag{32}$$

Since

$$A^{-1} = \begin{bmatrix} 0.28571429 & -0.42857143 \\ 0.14285714 & 0.28571429 \end{bmatrix}, y = \begin{bmatrix} 15 \\ 6 \end{bmatrix} \tag{33}$$

So,

$$x = A^{-1}y = \begin{bmatrix} 0.28571429 & -0.42857143 \\ 0.14285714 & 0.28571429 \end{bmatrix} \begin{bmatrix} 15 \\ 6 \end{bmatrix} = \begin{bmatrix} 1.71428571 \\ 3.85714286 \end{bmatrix} = \begin{bmatrix} 12/7 \\ 27/7 \end{bmatrix} \tag{34}$$

5 Matrix Calculus

Matrix calculus is widely used in machine learning and reinforcement learning.

5.1 The Gradient

Given a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ that takes a matrix A as an input and outputs a real value. The the ***gradient*** of f is the matrix of partial derivatives, defined as

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \cdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}. \tag{35}$$

For example, an $m \times n$ matrix with partial derivatives over A_{ij} is given as:

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}} \tag{36}$$

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. If A is just a vetor $x \in \mathbb{R}^n$

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ . \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \tag{37}$$

It is very important to notice that the gradient of a function is only defined if the function is real-valued, that is, if it returns a scalar value. We can not, for example, take the gradient of Ax,$A \in \mathbb{R}^{n \times n}$ with respect to x, since this quantity is vector-valued.

5.2 The Hessian

For function $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the Hessian matrix with respect to x, written $nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives, given as

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial^2 x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial^2 x_2} & \cdots \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ . & . & . \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots \frac{\partial^2 f(x)}{\partial^2 x_n} \end{bmatrix} \tag{38}$$

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i} \tag{39}$$

Note that the gradient can be say as the first derivative for a function of vectors, but the Hessian is not simply a second derivatives of a function of vectors. It is actually say that $\nabla_x^2 f(x) = \nabla_x (\nabla_x f(x))^T$, that this really means taking the gradient of each **entry** of $(\nabla_x f(x))^T$ not the gradient of the **whole vector**. In other words, for the ith entry of the gradient $(\nabla_x f(x))_i = \partial f(x)/\partial x_i$, and take the gradient with respect to x we get

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_i \partial x_1} \\ \frac{\partial^2 f(x)}{\partial x_i \partial x_2} \\ . \\ \frac{\partial^2 f(x)}{\partial x_i \partial x_n} \end{bmatrix} \tag{40}$$