Math Review

Peifan Wu*

September 15, 2020

1 Basic Linear Algebra

1.1 Definitions of Vectors and Matrices

Vector $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$ is a column vector.

The transpose of x: $x^T = x' = [x_1, x_2, \dots, x_n]$ is a row vector.

Matrix $A = \begin{bmatrix} A_{11} & \dots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nm} \end{bmatrix}$ is a $n \times m$ matrix. Each element of the matrix is represented by $A_{ij}, i = 1, \dots, n, j = 1, \dots, m$.

The transpose of matrix A is denoted as A^T, A' where $A_{ij}^T = A_{ji}$. Notice that $\left(A^T\right)^T = A$.

1.2 Matrix Operations

Addition/Subtraction Addition/subtraction of matrices are applied element-wise. If A, B are both $n \times m$ matrices then

$$A + B = \begin{bmatrix} A_{11} + B_{11} & \dots & A_{1m} + B_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} + B_{n1} & \cdots & A_{nm} + B_{nm} \end{bmatrix}$$

$$A - B = \begin{bmatrix} A_{11} - B_{11} & \dots & A_{1m} - B_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} - B_{n1} & \cdots & A_{nm} - B_{nm} \end{bmatrix}$$

Properties:

- Commutativity: $A + B = B + A$

*I took most of them from Jesse Perla's notes

- Associativity: $(A+B)+C = A+(B+C)$

Multiplication For matrices A with size $n \times m$ and B with size $m \times p$, we have $C = A \times B$ where C is a $n \times p$ matrix,

$$C_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk}$$

Matrix multiplication is not commutative!
Properties:

- $(AB)^T = B^T A^T$
- Associativity: $(AB)\,C = A\,(BC)$
- Distributivity: $(A+B)\,C = AC + BC$, $C\,(A+B) = CA + CB$
- Communtativity only holds for scalar multiplication, but not for matrix multiplication:

$$\alpha A = A\alpha$$
$$AB \neq BA$$

- The product of a matrix and a vector can be dot product.

Inverse We first define Identity matrix:

$$I = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}$$

- $n \times n$ square matrix with "1"s on the diagonal, and "0"s everywhere
- $\forall A$ that is a $n \times n$ matrix, $A \times I = I \times A = A$

Then we define the inverse of a $n \times n$ square matrix A.

- A $n \times n$ matrix, denoted by A^{-1}, is called the inverse of A if $A^{-1}A = I$
- $A^{-1}A = I \Longrightarrow AA^{-1} = I$

Proof. Left multiply A to both sides of $A^{-1}A = I$ then we have $AA^{-1}A = A$. Move the matrix on the right-hand side to the left-hand side and we have $\left(AA^{-1} - I\right)A = 0$. Since A consists of n linearly independent vectors, $xA = 0$ only has a trivial solution $x = 0$. Therefore it has to be $AA^{-1} - I = 0$. □

- Such A^{-1} doesn't always exist. When it exists we call A invertible or non-singular. Otherwise, A is singular

1.3 Matrix and Linear Equations

A linear equation system can be expressed as $A \cdot x = b$ where A is a $m \times n$ matrix, x is a column vector of n elements, and b is a column vector of m elements.

- If $n = m$ (then A is a square matrix) and A is invertible (the rank of A equals n), then the solution to the linear equation system is $x = A^{-1}b$ by left multiplying A^{-1} to both sides of the equation
- If $n > m$, the number of unknowns x is more than the number of equations m. If the rank of A is greater than m as well, then x has multiple solutions
- If $n < m$, the number of unknowns x is less than the number of equations m. In general there's no solution. However, we can choose x subject to a loss function,

  $$\min_x ||Ax - b||$$

  gives $x = \left(A^T A\right)^{-1} A^T b$. A way to interpret the OLS regression.

2 Basic Optimization

2.1 Unconstrained Optimization

Assume x is a vector with n elements, and f is a function $\mathbb{R}^n \to \mathbb{R}$. We want x such that

$$\min_x f(x)$$

If f is differentiable then the first-order necessary condition is

$$\partial_i f(x) = f_{x_i}(x) = 0$$

i.e., the partial derivative of f with respect to any element in x is 0. This is only a necessary condition: you need to verify the point is local minimum instead of local maximum.

2.2 Constrained Optimization

Assume x is a vector with n elements. Almost all the constrained problems can be reduced to the same form,

$$\begin{aligned} &\min_x f(x) \\ s.t. &g(x) \leqslant 0 \\ &h(x) = 0 \end{aligned}$$

- $f : \mathbb{R}^n \to \mathbb{R}$, the objective function
- $g : \mathbb{R}^n \to \mathbb{R}^m$, constraints may or may not bind
- $h : \mathbb{R}^n \to \mathbb{R}^p$, constraints that are always binding

We use the Lagrangian to solve the optimization problem.

First, the definition of Lagragian,

$$\mathcal{L} = f(x) + \mu g(x) + \lambda h(x)$$

where μ and λ are called Lagrange multipliers. These are also the so called "shadow prices" under some economic setting. The values of μ and λ shows how tight the constraints are.

Then we write the first-order necessary condition, $\partial \mathcal{L}(x) = 0$. i.e.,

$$\begin{aligned} \partial f(x) + \mu \partial g(x) = 0 \\ s.t. \begin{cases} \mu > 0, g(x) = 0 & g \text{ constraints binding} \\ \mu = 0, g(x) < 0 & g \text{ constraints not binding} \end{cases} \end{aligned}$$

Check Kuhn-Tucker Theorem for more details.

Example A consumer optimizes her consumption bundle subject to her utility function

$$\max_{x,y} U(x, y)$$

subject to

$$p_x x + p_y y \leqslant I$$

We write the Lagrangian out,

$$\mathcal{L} = U(x, y) + \mu \left(p_x x + p_y y - I \right)$$

and the first-order necessary conditions are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x} = U_x + \mu p_x \\ \frac{\partial \mathcal{L}}{\partial y} = U_y + \mu p_y \end{aligned}$$

Therefore, if the constraint is not binding then $\mu = 0$; otherwise if constraint is binding then $p_x x + p_y y = I$ and

$$\frac{U_x}{U_y} = \frac{p_x}{p_y}$$

3 Probability and Statistics

3.1 Discrete Random Variable

- A **random variable** is a number whose value depends upon the outcome of a random experiment. Mathematically, a random variable X is a real-valued function on S, the space of outcomes (which can be a very abstract set)

$$X : S \to \mathbb{R}$$

- A **discrete random variable** X has finite or countably many values x_s for $s = 1, 2, \cdots$.
- The probabilities $\mathbb{P}(X = x_s)$ with $s = 1, 2, \cdots$ are called the **probability mass function** (PMF) of X which has the following properties:
  - For all s, $\mathbb{P}(X = x_s) \geqslant 0$
  - For any $B \subset S$, $\mathbb{P}(X \in B) = \sum_{x_s \in B} \mathbb{P}(X = x_s)$
  - $\sum_s \mathbb{P}(X = x_s) = 1$
- Assume that X is a discrete random variable with possible values x_s.Then the **expectation** of X is defined as

$$\mathbb{E}(X) = \sum_s x_s \mathbb{P}(X = x_s)$$

3.2 Expectations and Vectors

Assume that there are n states, i.e. $x_1, \ldots, x_n$. List of values for states of the world:

$$x \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Now, list out all of the probabilities in a vector, $\phi \in \mathbb{R}^n$

$$\phi \equiv \begin{bmatrix} \mathbb{P}(X = x_1) \\ \vdots \\ \mathbb{P}(X = x_n) \end{bmatrix}$$

then the expectation is a dot product of two vectors,

$$\mathbb{E}(X) = \sum \phi_s x_s = \phi \cdot x$$

More generally,

$$\mathbb{E}(f(X)) = \sum \phi_s f(x_s) = \phi \cdot f(x)$$

3.3 Joint Distributions

For discrete random variables, consider if there are multiple events yielding random variables X and Y. The joint probability distribution is the probability that both events occur,

$$\mathbb{P}\left(X = x_i \text{ and } Y = y_j\right)$$

such that $\sum_i \sum_j \mathbb{P}\left(X = x_i \text{ and } Y = y_j\right) = 1$.

- The **marginal probability** is the distribution of one random variable if we ignore the other one. For example, the probability that x_i occurs (regardless of the y_i outcome) just sums over the probabilities in the joint distribution with y_j.

$$\mathbb{P}\left(X = x_i\right) = \sum_j \mathbb{P}\left(X = x_i, Y = y_j\right)$$

- The **conditional probability** is the distribution of one random variable if we know the other has occurred. For example, if we know $Y = y_j$ then the probability that x_i occurs is written as

$$\mathbb{P}\left(X = x_i | Y = y_j\right) = \frac{\mathbb{P}\left(X = x_i, Y = y_j\right)}{\mathbb{P}\left(Y = y_j\right)} = \frac{\mathbb{P}\left(X = x_i, Y = y_j\right)}{\sum_k \mathbb{P}\left(X = x_k, Y = y_j\right)}$$

  and we have **Bayes Theorem** for this case,

$$\mathbb{P}\left(Y = y_j | X = x_i\right) \mathbb{P}\left(X = x_i\right) = \mathbb{P}\left(X = x_i, Y = y_j\right) = \mathbb{P}\left(X = x_i | Y = y_j\right) \mathbb{P}\left(Y = y_j\right)$$

$$\downarrow$$

$$\begin{aligned}
\mathbb{P}\left(Y = y_j | X = x_i\right) &= \frac{\mathbb{P}\left(X = x_i | Y = y_j\right) \mathbb{P}\left(Y = y_j\right)}{\mathbb{P}\left(X = x_i\right)} \\
&= \frac{\mathbb{P}\left(X = x_i | Y = y_j\right) \mathbb{P}\left(Y = y_j\right)}{\sum_j \mathbb{P}\left(X = x_k, Y = y_j\right)}
\end{aligned}$$

- A **conditional expectation** is when one of the multiple events is known (e.g. which Y occurred), and finds the expectation over the over event. It is denoted as $\mathbb{E}\left(X|Y\right)$
  - For example, in the above if we know that $Y = y_j$

$$\mathbb{E}\left(X | Y = y_j\right) = \sum_i x_i \mathbb{P}\left(X = x_i | Y = y_j\right)$$

  - This will be especially useful for agents making forecasts of the future given knowledge of events today

- Events X and Y has **statistical independence** if

$$\mathbb{P}\left(X = x_i, Y = y_j\right) = \mathbb{P}\left(X = x_i\right) \mathbb{P}\left(Y = y_j\right)$$

Independence implies

$$\mathbb{P}\left(X = x_i | Y = y_j\right) = \mathbb{P}\left(X = x_i\right)$$