Math Review Session

CS5785, 2019 Fall

Yichun Hu, Xiaojie Mao

Cornell Tech, Cornell Unviersity

Table of contents

Linear Algebra

Matrix and Vector

- Vector:

$$a = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

- Matrix:

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}.$$

- Vectorize a matrix:

$$\text{vec}(A) = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ 9 \end{bmatrix}.$$

Vector Operation

For two column vectors $v, u \in \mathbb{R}^n$,

- **Inner product**: $u \cdot v = u^\top v = \sum_{i=1}^n u_i v_i$.
- **Orthogonal vectors**: $u \cdot v = 0$.
- **Norm**: $\|u\| = \sqrt{u^\top u} = \sqrt{\sum_{i=1}^n u_i^2}$.
- **Euclidean distance**: $d(u, v) = \|u - v\| = \sqrt{\sum_{i=1}^n (u_i - v_i)^2}$.

Matrix Operation

- **Transpose**. For $A \in \mathbb{R}^{m \times n}$, $A^\top$ is a $n \times m$ matrix:

$$(A^\top)_{ij} = A_{ji}.$$

- **Matrix addition**. For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{m \times n}$, $C = A + B$ is a $m \times n$ matrix: for $1 \leq i \leq m$ and $1 \leq j \leq n$,

$$C_{ij} = A_{ij} + B_{ij}.$$

- **Matrix Multiplication**. For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $C = AB$ is a $m \times p$ matrix: for $1 \leq i \leq m$ and $1 \leq j \leq n$,

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj} = A_{i,:} \cdot B_{:,j}$$

Matrix Operation

- **Matrix inverse**. If A is square ($n \times n$), and invertible, then A^{-1} is the unique $n \times n$ matrix such that

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

- **Matrix trace**. If A is square ($n \times n$), then its trace is

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

- **Frobenius norm**: for a matrix $A \in \mathbb{R}^{m \times n}$,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} = \sqrt{\operatorname{tr}(A^\top A)} = \|\operatorname{vec}(A)\|.$$

Properties of Matrix Operation

- Transpose:
    - $(AB)^\top = B^\top A^\top$.
    - $(ABC)^\top = C^\top B^\top A^\top$.
    - $(A+B)^\top = A^\top + B^\top$.
- Multiplication:
    - Associative: $(AB)C = A(BC)$.
    - Distributive: $(A+B)C = AC + BC$.
    - Non-commutative: $AB \neq BA$ in general.

Properties of Matrix Operation

- Inverse:
  - $(AB)^{-1} = B^{-1}A^{-1}$.
  - $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$.
  - $(A^{-1})^{-1} = A$.
  - $(A^{-1})^\top = (A^\top)^{-1}$.
- Trace:
  - $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
  - $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$.

Special matrices

For $A \in \mathbb{R}^{n \times n}$,

- Diagonal matrix: $A_{ij} = 0$ for any $i \neq j$.
- Symmetric (Hermitian) matrix: $A = A^\top$ or $A_{ij} = A_{ji}$.
- Orthogonal matrix: $A^\top = A^{-1}$.
  - $AA^\top = A^\top A = I$.
  - Rows and Columns are orthogonal unit vectors, namely, for $i \neq j$,

$$A_{i,:} \cdot A_{j,:} = 0, \quad A_{:,i} \cdot A_{:,j} = 0,$$

    and for any i,

$$A_{i,:} \cdot A_{i,:} = 1, \quad A_{:,i} \cdot A_{:,i} = 1.$$

- Positive semidefinite matrix: for any $x \in \mathbb{R}^n$ with $x \neq 0$,

$$x^\top A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \geq 0.$$

Eigenvalues and Eigenvectors

For matrix $A \in \mathbb{R}^{n \times n}$, and nonzero vector $u \in \mathbb{R}^n$ $(u \neq 0)$ such that

$$Au = \lambda u,$$

u is an eigenvector of A, and λ is the corresponding eigenvalue.

Spectral decomposition theorem

If $A \in \mathbb{R}^{n\times n}$ is a real symmetric matrix, then

$$A = U\Lambda U^\top \Leftrightarrow A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \Leftrightarrow U^\top A U = \Lambda$$

where

- $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns are eigenvectors of A, i.e., $U_{:,i}$ and $U_{:,j}$ are orthogonal unit eigenvectors for $i \neq j$.
- Λ is a diagonal matrix whose entries are the corresponding eigenvalues.

Remark:

- $\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i$.
- Real symmetric A is positive semidefinite $\Leftrightarrow \lambda_i \geq 0$ for any $i = 1, \ldots, n$.

Singular Value Decomposition (SVD)

For $A \in \mathbb{R}^{m \times n}$, its singular value decomposition is

$$A = U\Sigma V^\top \Leftrightarrow A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top \Leftrightarrow U^\top AV = \Sigma$$

where

- $U \in \mathbb{R}^{m \times r}$ is an orthogonal matrix whose columns $\{u_i\}_{i=1}^r$ are the left singular vectors;
- $V \in \mathbb{R}^{r \times n}$ is an orthogonal matrix whose columns $\{v_i\}_{i=1}^r$ are the right singular vectors;
- $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix whose diagonal elements $\{\sigma_i\}_{i=1}^r$ are singular values.

Remark:

- r is the rank of matrix A;
- The maximum singular value $\sigma_{\max}(A)$ is called the spectral norm of A, which we denote as $\|A\|_2$.
- Connection between SVD and eigen-decomposion.

$$A = U\Sigma V^\top \Rightarrow AA^\top = U\Sigma^2 U^\top,$$
$$A = U\Sigma V^\top \Rightarrow A^\top A = V\Sigma^2 V^\top.$$

  Thus

  - The columns of U are eigenvectors of $AA^\top$, and the columns of V are eigenvectors of $A^\top A$;
  - $\sigma_i^2(A) = \lambda_i(AA^\top) = \lambda_i(A^\top A)$.

Calculus

Univariate Calculus

- Polynomial: $\frac{\partial}{\partial x} x^n = n x^{n-1}$.
- Exponential: $\frac{\partial}{\partial x} \exp(x) = \exp(x)$.
- Logarithm: $\frac{\partial}{\partial x} \log(x) = \frac{1}{x}$.
- Sum: $\frac{\partial}{\partial x}(f(x) + g(x)) = \frac{\partial}{\partial x} f(x) + \frac{\partial}{\partial x} g(x)$.
- Multiplication: $\frac{\partial}{\partial x}(f(x) \cdot g(x)) = f(x) \frac{\partial}{\partial x} g(x) + g(x) \frac{\partial}{\partial x} f(x)$.
- Chain Rule: $\frac{\partial}{\partial x}(f(g(x))) = f'(g(x)) \cdot g'(x)$.

Multivariate Calculus

Let f be a function of $x_1, x_2, \ldots, x_n$.

- **Partial derivative** $\frac{\partial}{\partial x_i} f(x_1, \ldots, x_n)$: treat other variables as constants and take derivative w.r.t. x_i.

$$\frac{\partial}{\partial x} f := \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \ldots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}, \frac{\partial}{\partial x^\top} f := \begin{pmatrix} \frac{\partial f}{\partial x_1} & \ldots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$

- **Gradient** of f with respect to x: $\nabla_x f := \frac{\partial}{\partial x} f$.
- **Hessian matrix** of f: $\mathcal{H}$ is a $n \times n$ matrix with $\mathcal{H}_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f$, or

$$\mathcal{H} = \frac{\partial^2}{\partial x \partial x^\top} f.$$

Let $f = (f_1, \ldots, f_m)$ be a multivariate vector function of $x_1, \ldots, x_n$.

- **Jacobian matrix** of f: $\mathcal{J}$ is a $m \times n$ matrix with $\mathcal{J}_{ij} = \frac{\partial f_i}{\partial x_j}$. The i-th row of $\mathcal{J}$ is $\frac{\partial}{\partial x^\top} f_i$.

Multivariate Calculus Rules

Here **a** and **A** are vector/matrix that do not depend on $\mathbf{x} = (x_1, \ldots, x_n)^\intercal$.

- $\frac{\partial}{\partial x}\mathbf{a} = \mathbf{0}$;
- $\frac{\partial}{\partial x}\mathbf{a}^\intercal\mathbf{x} = \frac{\partial}{\partial x}\mathbf{x}^\intercal\mathbf{a} = \mathbf{a}$;
- $\frac{\partial}{\partial x}(\mathbf{x}^\intercal\mathbf{a})^2 = 2\mathbf{a}\mathbf{a}^\intercal\mathbf{x}$;
- $\frac{\partial}{\partial x}\mathbf{A}\mathbf{x} = \mathbf{A}^\top$;
- $\frac{\partial}{\partial x}\mathbf{x}^\intercal\mathbf{A} = \mathbf{A}$;
- $\frac{\partial}{\partial x}\mathbf{x}^\top\mathbf{A}\mathbf{x} = (\mathbf{A}^\intercal + \mathbf{A})\mathbf{x}$.

For a detailed multivariate derivatives list, see
`https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`.

Example: Least Squares

Lets apply the equations to derive the least squares equations. Suppose we are given matrices $A \in \mathbb{R}^{m \times n}$ (for simplicity we assume A is full rank so that $(A^\top A)^{-1}$ exists) and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$. In this situation we will not be able to find a vector $x \in \mathbb{R}^n$ such that $Ax = b$, so instead we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $||Ax - b||_2^2$.

Using the fact that $||x||_2^2 = x^\intercal x$, we have

$$||Ax - b||_2^2 = (Ax - b)^\intercal (Ax - b) = x^\intercal A^\intercal Ax - 2b^\intercal Ax + b^\intercal b.$$

Taking the gradient with respect to x we have

$$\nabla_x (x^\intercal A^\intercal Ax - 2b^\intercal Ax + b^\intercal b) = \nabla_x x^\intercal A^\intercal Ax - \nabla_x 2b^\intercal Ax + \nabla_x b^\intercal b = 2A^\intercal Ax - 2A^\intercal b.$$

Setting this last expression equal to zero and solving for x gives the normal equations $x = (A^\intercal A)^{-1} A^\intercal b$.

Probability

Sample space

- **Sample space** Ω is the set of all possible outcomes of a random experiment;
- **Event** A is a subset of Ω, and the collection of all possible events is denoted as $\mathcal{F}$;
- **Probability measure** is a function $P : \mathcal{F} \to \mathbb{R}$ that maps an event into a real number which indicates the chance at which this event happens in the experiment.
- A and B are **independent events** if

$$P(A \cap B) = P(A)P(B).$$

Example: consider tossing a six-sided die,

- $\Omega = \{1, 2, 3, 4, 5, 6\}$;
- $A = \{1, 2, 3, 4\} \subset \Omega$ is an event;
- $P(A) = \frac{4}{6}$ for an even die.

Random Variable

- A **random variable** X is a function $X : \Omega \to \mathbb{R}$.
- Discrete random variable can only take countably many values, and

$$P(X = x) = P(\{w : X(w) = x\}).$$

- Continuous random variable can take uncountably many values, and

$$P(a \leq X \leq b) = P(\{w : a \leq X(w) \leq b\}).$$

Example: If the die gives value larger than 4, we set $X = 1$, and otherwise $X = 0$.

- $P(X = 1) = P(\{5, 6\}) = \frac{2}{6}$;
- $P(X = 0) = P(\{1, 2, 3, 4\}) = \frac{4}{6}$.

Distribution

- A **cumulative distribution function** (CDF) of a random variable X (either continuous or discrete) is a function $F_X : \mathbb{R} \to [0,1]$ such that

$$F_X(x) = P(X \leq x).$$

- A **probability mass function** (PMF) of a *discrete* random variable X is a function $p_X : \mathbb{R} \to [0,1]$ such that

$$p_X(x) = P(X = x).$$

- A **probability density function** (PDF) of a *continuous* random variable is a function $f_X : \mathbb{R} \to \mathbb{R}$ given by the derivative of CDF:

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}.$$

  As a result,

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx.$$

Expectation

- For a *discret* random variable X with PMF p_X and an aribitrary function $g : \mathbb{R} \to \mathbb{R}$, $g(X)$ is also a random variable whose expectation is given by

$$\mathbb{E}[g(X)] = \sum_x p_X(x)g(x).$$

- For a *continuous* random variable X with PDF f_X, $g(X)$ is also a random variable whose expectation is given by

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

- For two functions g_1 and g_2,

$$\mathbb{E}[g_1(X) + g_2(X)] = \mathbb{E}[g_1(X)] + \mathbb{E}[g_2(X)]$$

Variance

The variance of a random variable X is

$$\mathsf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2,$$

and the associated standard deviation is

$$\sigma(X) = \sqrt{\mathsf{Var}[X]}.$$

Exercise: uniform distribution

Consider $X \sim \text{uniform}(0, 1)$ whose PDF is

$$f_X(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

What's the expectation and variance of X?

Hint:

- $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$;
- $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Common distributions

- **Normal distribution**: $X \sim \mathcal{N}(\mu, \sigma^2)$ has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}.$$

  - $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.

- **Bernoulli distribution**: $X \sim \mathrm{Bernoulli}(p)$ with $0 \le p \le 1$ has PMF

$$P_X(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$$

  - $\mathbb{E}[X] = p$ and $\mathrm{Var}[X] = p(1-p)$.

Joint distributions

- For two random variables X and Y, their joint cumulative distribution function is

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

- For two *discrete* random variables X and Y, their joint probability mass function is

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

- For two *continuous* random variable X and Y, their joint probability density function is

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y},$$

so that for a set $A \in \mathbb{R}^2$ and a function $g : \mathbb{R}^2 \to \mathbb{R}$,

$$P((X, Y) \in A) = \iint_{(x,y) \in A} f_{X,Y}(x, y) dxdy,$$

$$\mathbb{E}[g(X, Y)] = \iint g(x, y) f_{X,Y}(x, y) dxdy.$$

Independence

- Random variables X, Y are independent if for any possible values x, y

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \text{for continuous} \quad X, Y,$$

$$\text{or} \quad p_{X,Y}(x,y) = p_X(x) p_Y(y), \quad \text{for discrete} \quad X, Y.$$

- For any set $A = \{(x, y) : x \in A_1, y \in A_2\} \subset \mathbb{R}^2$, independent random variables X, Y satisfy that

$$P((X, Y) \in A) = P(X \in A_1) P(Y \in A_2).$$

  or events $\{w : X(w) \in A_1\}$ and $\{w : Y(w) \in A_2\}$ are independent events for any A_1 and A_2.

Exercise: Independence

For example, consider toss two coins consecutively, and $X_1 = 1$ if the first coin heads up, otherwise $X_1 = 0$; $X_2 = 1$ if the second coin heads up, otherwise $X_2 = 0$.

- $\Omega = \{(T, T), (H, H), (T, H), (H, T)\}$.
- $P(X_1 = 1, X_2 = 1) = P(\{(H, H)\}) = \frac{1}{4}$;
- $P(X_1 = 1) = P(\{(H, T), (H, H)\}) = \frac{1}{2}$;
- $P(X_2 = 1) = P(\{(T, H), (H, H)\}) = \frac{1}{2}$.

Thus

$$P(X_1 = 1, X_2 = 1) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(X_1 = 1)P(X_2 = 1)..$$

Similarly, we can show that

$$\begin{aligned}
P(X_1 = 1, X_2 = 0) &= P(X_1 = 1)P(X_2 = 0) \\
P(X_1 = 0, X_2 = 1) &= P(X_1 = 0)P(X_2 = 1) \\
P(X_1 = 0, X_2 = 0) &= P(X_1 = 0)P(X_2 = 0).
\end{aligned}$$

Conditional Probability

Let A, B be two events.

- The conditional probability of A given B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- If A is independent of B, we have $P(A|B) = P(A)$, as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

- **Bayes Rule**:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

- **Chain Rule**:

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_n) \\ &= P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \ldots P(A_n|A_{n-1} \cap \cdots \cap A_1). \end{aligned}$$

Example: conditional probability

Consider toss a die once, and we define events

$$A = \{\text{The value is larger than } 4\}, B = \{\text{The value is larger than } 2\}.$$

Then

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\{5,6\} \cap \{3,4,5,6\})}{P(\{3,4,5,6\})} \\ &= \frac{\frac{2}{6}}{\frac{4}{6}} = \frac{1}{2}. \end{aligned}$$

Conditional Distribution

Conditional Density. The conditional probability density function of continuous random variable X given $Y = y$ is

$$f_X(x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Conditional Expectation. The conditional expectation of X given $Y = y$ is

$$\mathbb{E}(X|Y = y) = \int_{-\infty}^{\infty} x f_X(x|Y = y) dx \triangleq g(y)$$

Conditional Variance. The conditional variance of random variable X given $Y = y$ is

$$\mathrm{Var}[X|Y = y] = \mathbb{E}[(X - \mathbb{E}(X|Y = y))^2|Y = y] \triangleq h(y).$$

Both $\mathbb{E}[X \mid Y]$ and $\mathrm{Var}(X|Y)$ are random variables, and their distributions are determined by the distribution of Y.

Properties of Conditional Distributions

Iterated Expectation. Recall that $\mathbb{E}(X|Y)$ is a function of Y, i.e., a random variable. The law of iterative expectation states that

$$\mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}(X).$$

Law of Total Variance. Recall that $\mathbb{E}(X|Y)$ and $Var(X|Y)$ are both random variables that are functions of Y. We have

$$Var(Y) = \mathbb{E}[Var(X|Y)] + Var[\mathbb{E}(X|Y)].$$

Example: conditional distribution

Assume we throw two six-sided dice.

- What is the probability that the total of two dice will be greater than 8 given that the first die is a 6?
- What is the expectation of the total of two dice given that the first die is a 6?
- What is the variance of the total of two dice given that the first die is a 6?

Example: conditional distribution

We use X_1 to denote the value for the first die and X_2 the value for the second die.

- What is the probability that the total of two dice will be greater than 8 given that the first die is a 6?

$$
\begin{aligned}
P(X_1 + X_2 > 8 \mid X_1 = 6) &= \frac{P(X_1 + X_2 > 8, X_1 = 6)}{P(X_1 = 6)} \\
&= \frac{P(X_1 = 6, X_2 > 2)}{P(X_1 = 6)} \\
&= \frac{P(X_1 = 6)P(X_2 > 2)}{P(X_1 = 6)} \\
&= P(X_2 > 2) = \frac{4}{6}.
\end{aligned}
$$

Example: conditional distribution

- What is the expectation of the total of two dice given that the first die is a 6?
  Given that $X_1 = 6$, $X_1 + X_2$ can be $7, 8, 9, 10, 11, 12$, all with probability $\frac{1}{6}$. Thus

$$\mathbb{E}[X_1 + X_2 \mid X_1 = 6] = 7 * \frac{1}{6} + \cdots + 12 * \frac{1}{6} = \frac{57}{6}.$$

- What is the variance of the total of two dice given that the first die is a 6? Answer: $\frac{105}{36}$.

Law of large number

Consider i.i.d random variables $X_1, \cdots, X_n$, i.e., independent random variables with identical distributions, and an arbitrary function g. Suppose the common expectation $\mathbb{E}[g(X_1)] < \infty$ and common variance $\mathrm{Var}[g(X_1)] < \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \mathbb{E}[g(X_1)].$$

Actually

- $\mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} g(X_i)] = \mathbb{E}[g(X_1)]$.
- $\mathrm{Var}[\frac{1}{n} \sum_{i=1}^{n} g(X_i)] = \frac{1}{n} \mathrm{Var}[g(X)]$.

Resources

Resources

- Linear algebra:
  http://cs229.stanford.edu/summer2019/cs229-linalg.pdf
- Matrix calculus: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
- Probability:
  http://cs229.stanford.edu/summer2019/cs229-prob.pdf and All of Statistics by Larry Wasserman.