CAS751: IT Metods in Trustworthy ML September 3, 2025

Lecture 0: Basics of Probability

Instructor: Shahab Asoodeh

1 Elements of Probability

Probability theory is the study of uncertainty. Through this class, we will be relying on concepts from probability theory for deriving machine learning algorithms. These notes attempt to cover the basics of probability theory at a level appropriate for introductory course of differential privacy (such as CS 3DP3). The mathematical theory of probability is rather sophisticated, and delves into a branch of analysis known as measure theory. In these notes, we provide a basic treatment of probability that does not address these finer details.

1.1 Definition of probability space

In order to define a probability on a set we need a few basic elements:

- **Sample space** Ω: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.
- **Event space** $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called **events**) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment).[1]
- **Probability measure**: A function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties,
  - **Non-negativity**: $P(A) \geq 0$, for all $A \in \mathcal{F}$
  - **Completeness**: $P(\Omega) = 1$
  - **Countable Additivity**: If $A_1, A_2, \dots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

These three properties are called the **Axioms of Probability**.

Example 1. Consider the event of tossing a six-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. We can define different event spaces on this sample space. For example, the simplest event space is the trivial event space $\mathcal{F} = \{\emptyset, \Omega\}$. Another event space is the set of all subsets of Ω. For the first event space, the unique probability measure satisfying the requirements above is given by $P(\emptyset) = 0$, $P(\Omega) = 1$. For the second event space, one valid probability measure is to assign the probability of each set in the event space to be $\frac{i}{6}$ where i is the number of elements of that set; for example, $P(\{1, 2, 3, 4\}) = \frac{4}{6}$ and $P(\{1, 2, 3\}) = \frac{3}{6}$.

[1] The event space $\mathcal{F}$ is technically required to satisfy three properties: (1) $\emptyset \in \mathcal{F}$; (2) $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$; and (3) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_i A_i \in \mathcal{F}$.

1.2 Properties of probability

Proposition 2. *The following properties can be derived from the axioms of probability.*

- *If* $A \subseteq B$ *then* $P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$.
- $P(A^c) \triangleq P(\Omega \setminus A) = 1 - P(A)$.
- $P(A \cup B) \leq P(A) + P(B)$. *This property is known as the* ***union bound****.*
- *If* $A_1, \ldots, A_k$ *are a set of disjoint events such that* $\bigcup_{i=1}^k A_i = \Omega$*, then* $\sum_{i=1}^k P(A_k) = 1$*. This property is known as the* ***Law of Total Probability****.*

1.3 Conditional probability and independence

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as

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

In other words, $P(A|B)$ is the probability measure of the event A after observing the occurrence of event B. Two events are called **independent** if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A|B) = P(A)$). Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

In general, for multiple events, $A_1, \ldots, A_k$, we say that $A_1, \ldots, A_k$ are **mutually independent** if for any subset $S \subseteq \{1, 2, \ldots, k\}$, we have

$$P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P(A_i).$$

1.4 Law of total probability and Bayes' theorem

In practice, it is often helpful to compute the marginal probabilities from the conditional probabilities. The following **Law of total probability** expresses the total probability of an outcome which can be realized via several distinct events:

Theorem 3 (Law of total probability)**.** *Suppose* $A_1, \ldots, A_n$ *are disjoint events, and event* B *satisfies* $B \subseteq \bigcup_{i=1}^n A_i$*, then*

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i). \tag{1}$$

This theorem can be proved directly by applying the definition of the conditional probability. Note that Theorem 3 holds for any event B if $\bigcup_{i=1}^n A_i = \Omega$. As a common special case, for any event A, it is the case that

$$P(B) = P(A)P(B|A) + P(A^c)P(B|A^c). \tag{2}$$

An important corollary of the law of total probability is the following Bayes' theorem

Theorem 4 (Bayes' theorem)**.** *Suppose $A_1, \dots, A_n$ are disjoint events, and event B satisfies $B \subseteq \bigcup_{i=1}^n A_i$. Then if $P(B) > 0$, it is the case that*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^n P(A_j)P(B|A_j)}. \tag{3}$$

A special case of the Bayes' theorem is

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}. \tag{4}$$

Bayes' theorem is widely applied in various topics in statistics and machine learning.

2 Random Variables

2.1 Definition and examples

Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space Ω are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \{H, H, T, H, T, H, H, T, T, T\} \in \Omega$. However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as **random variables**.

More formally, a random variable X is a function $X : \Omega \to \mathbb{R}$.[2] Typically, we will denote random variables using upper case letters $X(\omega)$ or more simply X (where the dependence on the random outcome ω is implied). We will denote the value that a random variable may take on using lower case letters x.

In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω. Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a **discrete random variable**. Here, the probability of the set associated with a random variable X taking on some specific value k is

$$P(X = k) := P(\{\omega : X(\omega) = k\}).$$

As an additional example, suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay. In this case, $X(\omega)$ takes on a infinite number of possible values, so it is called a **continuous random variable**. We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as

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

[2] Technically speaking, not every function is acceptable as a random variable. From a measure-theoretic perspective, random variables must be Borel-measurable functions. Intuitively, this restriction ensures that given a random variable and its underlying outcome space, one can implicitly define the set of events on the original outcome space $\omega \in \Omega$ for which $X(\omega)$ satisfies some property (e.g., the event $\{\omega : X(\omega) \geq 3\}$).

2.2 Cumulative distribution functions

In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions (CDFs, PDFs, and PMFs) from which the probability measure governing an experiment immediately follows. In this section and the next two sections, we describe each of these types of functions in turn.

A **cumulative distribution function (CDF)** is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as,

$$F_X(x) \triangleq P(X \leq x). \tag{5}$$

By using this function one can calculate the probability of any event in $\mathcal{F}$.[3] Figure 1 shows a sample CDF function. A CDF function satisfies the following properties.

- $0 \leq F_X(x) \leq 1$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.
- $x \leq y \Rightarrow F_X(x) \leq F_X(y)$.

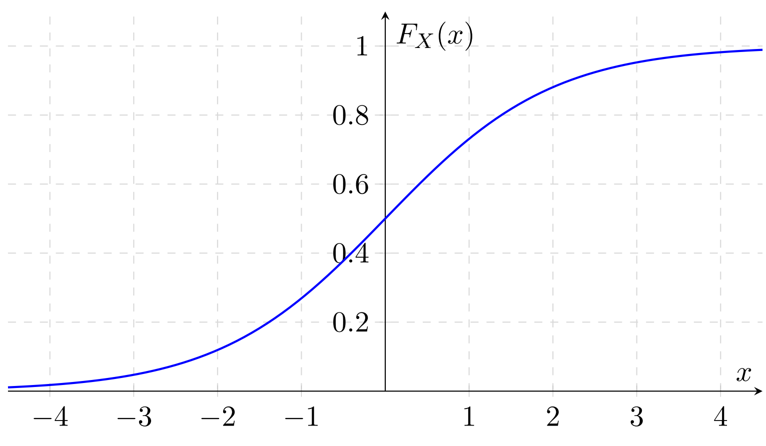

Figure 1: A cumulative distribution function (CDF).

2.3 Probability mass functions

When a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable), a simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume. In particular, a **probability mass function (PMF)** is a function $p_X : \Omega \to \mathbb{R}$ such that

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

[3]This is a remarkable fact and is actually a theorem that is proved in more advanced courses.

In the case of discrete random variable, we use the notation $\mathcal{X}$ for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $\mathcal{X} = \{0, 1, 2, \ldots, 10\}$.

A PMF function satisfies the following properties.

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in \mathcal{X}} p_X(x) = 1$.
- $\sum_{x \in A} p_X(x) = P(X \in A)$.

2.4 Probability density functions

For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. In these cases, we define the **Probability Density Function (PDF)** as the derivative of the CDF, i.e.,

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}. \tag{6}$$

Note here, that the PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere).

According to the properties of differentiation, for very small Δx,

$$P(x \leq X \leq x + \Delta x) \approx f_X(x)\Delta x. \tag{7}$$

Both CDFs and PDFs (when they exist!) can be used for calculating the probabilities of different events. But it should be emphasized that the value of PDF at any given point x is not the probability of that event, i.e. $f_X(x) \neq P(X = x)$. For example, $f_X(x)$ can take on values larger than one (but the integral of $f_X(x)$ over any subset of $\mathbb{R}$ will be at most one).

A PDF function satisfies the following properties.

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
- $\int_{x \in A} f_X(x)dx = P(X \in A)$.

2.5 Expectation

Suppose that X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the **expectation** or **expected value** of $g(X)$ as

$$\mathbb{E}[g(X)] \triangleq \sum_{x \in \mathcal{X}} g(x)p_X(x).$$

If X is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(X)$ is defined as

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

Intuitively, the expectation of $g(X)$ can be thought of as a "weighted average" of the values that $g(X)$ can take on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$. As a special case of the above, note that the expectation, $\mathbb{E}[X]$, of a random variable itself is found by letting $g(x) = x$; this is also known as the **mean** of the random variable X.

Expectation satisfies the following properties:

- $\mathbb{E}[a] = a$ for any constant $a \in \mathbb{R}$.
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$ for any constant $a \in \mathbb{R}$.
- $\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$. This property is known as the **linearity of expectation**.
- For a discrete random variable X, $\mathbb{E}[1_{\{X=k\}}] = P(X = k)$.

2.6 Variance

The **variance** of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as

$$\text{Var}[X] \triangleq \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Using the properties in the previous section, we can derive an alternate expression for the variance:

$$\begin{aligned}
\mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[\mathbb{E}[X]X] + \mathbb{E}[\mathbb{E}[X]^2] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2,
\end{aligned}$$

where the second equality follows from linearity of expectations and the fact that $\mathbb{E}[X]$ is actually a constant with respect to the outer expectation. We note the following properties of the variance.

- $\mathsf{Var}(a) = 0$ for any constant $a \in \mathbb{R}$.
- $\mathsf{Var}(af(X)) = a^2\mathsf{Var}(f(X))$ for any constant $a \in \mathbb{R}$.

Example 5. Calculate the mean and the variance of the uniform random variable X with PDF $f_X(x) = 1, \forall x \in [0, 1]$, and 0 elsewhere. The expectation of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x dx = \frac{1}{2}.$$

The variance of X can be computed by first computing the second moment of X:

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}.$$

Therefore

$$\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Example 6. Suppose that $g(x) = 1_{\{x \in \mathcal{A}\}}$ for some subset $\mathcal{A} \subseteq \Omega$. What is $\mathbb{E}[g(X)]$? Discrete case:

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}(\Omega)} 1_{\{x \in \mathcal{A}\}} P_X(x)dx = \sum_{x \in \mathcal{A}} P_X(x)dx = P(x \in \mathcal{A}).$$

Continuous case:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} 1_{\{x \in \mathcal{A}\}} f_X(x)dx = \int_{x \in \mathcal{A}} f_X(x)dx = P(x \in \mathcal{A}).$$

2.7 Some common distributions

In this subsection, we review several common discrete and continuous distributions that are commonly used throughout any ML courses.

Discrete random variables

- $X \sim \mathsf{Bernoulli}(p)$ (where $0 \le p \le 1$): one if a coin with heads probability p comes up heads, zero otherwise.

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

- $X \sim \mathsf{Binomial}(n, p)$ (where $0 \le p \le 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $X \sim \mathsf{Geometric}(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}$$

- $X \sim \mathsf{Poisson}(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Continuous random variables

- $X \sim \mathsf{Uniform}(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim \mathsf{Exponential}(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

- (**Very important for this course**) $X \sim \mathcal{L}(\mu, b)$: also known as the **Laplace distribution**

$$f(x) = \frac{1}{\sqrt{2b}} e^{-\frac{1}{b}|x-\mu|}$$

- (**Very important for this course**) $X \sim \mathcal{N}(\mu, \sigma^2)$: also known as the **Gaussian distribution**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The following table is the summary of some of the properties of these distributions.

Distribution	PDF or PMF	Mean	Variance
Bernoulli(p)	$\begin{cases} p, & \text{if } x = 1, \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
Binomial(n, p)	$\binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \le k \le n$	np	npq
Geometric(p)	$p(1-p)^{k-1}, \quad k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$e^{-\lambda} \frac{\lambda^x}{x!}, \quad k = 1, 2, \ldots$	λ	λ
Uniform(a, b)	$\frac{1}{b-a}, \quad x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$\mathcal{L}(\mu, b)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\lvert x-\mu \rvert}{2\sigma^2}}$	μ	$2b^2$
Exponential(λ)	$\lambda e^{-\lambda x}, \quad x \ge 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Table 1: PDF/PMF, mean, and variance of some common distributions.

3 Two Random Variables

Thus far, we have considered single random variables. In many situations, however, there may be more than one quantity that we are interested in knowing during a random experiment. For instance, in an experiment where we flip a coin ten times, we may care about both:

$$\begin{cases} X(\omega) = \text{the number of heads that come up,} \\ Y(\omega) = \text{the length of the longest run of consecutive heads.} \end{cases}$$

In this section, we consider the setting of two random variables.

3.1 Joint and marginal distributions

Suppose that we have two random variables X and Y. One way to work with these two random variables is to consider each of them separately. If we do that we will only need $F_X(x)$ and $F_Y(y)$. But if we want to know about the values that X and Y assume simultaneously during outcomes of a random experiment, we require a more complicated structure known as the **joint cumulative distribution function** of X and Y, defined by

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

It can be shown that by knowing the joint cumulative distribution function, the probability of any event involving X and Y can be calculated. The joint CDF $F_{XY}(x, y)$ and the joint distribution functions $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y) dy$$

$$F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y) dx.$$

Here, we call $F_X(x)$ and $F_Y(y)$ the **marginal cumulative distribution functions** of $F_{XY}(x, y)$. The joint CDF satisfies the following properties

- $0 \leq F_{XY}(x, y) \leq 1$.
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$.
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$.
- $F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$.

3.2 Joint and marginal probability mass functions

If X and Y are discrete random variables, then the **joint probability mass function** $p_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is defined by

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

Here, $0 \leq p_{XY}(x, y) \leq 1$ for all x, y, and $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x, y) = 1$. How does the joint PMF over two variables relate to the probability mass function for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x, y).$$

and similarly for $p_Y(y)$. In this case, we refer to $p_X(x)$ as the **marginal probability mass function** of X. In statistics, the process of forming the marginal distribution with respect to one variable by summing out the other variable is often known as "marginalization."

3.3 Joint and marginal probability density functions

Let X and Y be two continuous random variables with joint distribution function F_{XY}. In the case that $F_{XY}(x, y)$ is everywhere differentiable in both x and y, then we can define the **joint probability density function**,

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

Like in the single-dimensional case, $f_{XY}(x,y) \neq P(X=x, Y=y)$, but rather

$$\iint_A f_{XY}(x,y)dxdy = P((X,Y) \in \mathcal{A}).$$

Note that the values of the probability density function $f_{XY}(x,y)$ are always nonnegative, but they may be greater than 1. Nonetheless, it must be the case that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y) = 1$. Analogous to the discrete case, we define

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy,$$

as the **marginal probability density function** (or **marginal density**) of X, and similarly for $f_Y(y)$.

3.4 Conditional distributions

Conditional distributions seek to answer the question, what is the probability distribution over Y, when we know that X must take on a certain value x? In the discrete case, the conditional probability mass function of Y given X is simply

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)},$$

assuming that $p_X(x) \neq 0$. In the continuous case, the situation is technically a little more complicated because the probability that a continuous random variable X takes on a specific value x is equal to zero. Ignoring this technical point, we simply define, by analogy to the discrete case, the conditional probability density of Y given $X = x$ to be

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

An important relationship of conditional distribution and marginal distribution is the **law of total expectation**. This result can be viewed as an extension of the law of total probability discussed in Section 1.

Theorem 7 (Law of total expectation)**.** *Let X, Y be two random variables defied on the same probability space. Then*

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

3.5 Bayes' rule for random variables

We can derive the bayes' rule for random variables as follows. It arises when trying to derive expression for the conditional probability of one variable given another. In the case of discrete random variables X and Y,

$$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{y'} \in \mathcal{Y} P_{XY}(x,y')P_Y(y')}.$$

If the random variables X and Y are continuous,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{XY}(x|y')f_Y(y')dy'}.$$

3.6 Independence of random variables

Two random variables X and Y are **independent** if $F_{XY}(x,y) = F_X(x)F_Y(y)$ for all values of x and y. Equivalently,

- For discrete random variables, $p_{XY}(x,y) = p_X(x)p_Y(y)$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.
- For discrete random variables, $p_{Y|X}(y|x) = p_Y(y)$ whenever $p_X(x) \neq 0$ for all $y \in \mathcal{Y}$.
- For continuous random variables, $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.
- For continuous random variables, $f_{Y|X}(y|x) = f_Y(y)$ whenever $f_X(x) \neq 0$ for all $y \in \mathbb{R}$.

Informally, two random variables X and Y are independent if "knowing" the value of one variable will never have any effect on the conditional probability distribution of the other variable, that is, you know all the information about the pair (X,Y) by just knowing $f(x)$ and $f(y)$. The following lemma formalizes this observation:

Lemma 8. *If X and Y are independent then for any subsets $A, B \subseteq \mathbb{R}$, we have,*

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

By using the above lemma one can prove that if X is independent of Y then any function of X is independent of any function of Y.

3.7 Expectation and covariance

Suppose that we have two discrete random variables X, Y and $g : \mathbb{R}^2 \to \mathbb{R}$ is a function of these two random variables. Then the expected value of g is defined in the following way,

$$\mathbb{E}[g(X,Y)] \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x,y)p_{XY}(x,y).$$

For continuous random variables X, Y, the analogous expression is

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dx dy.$$

We can use the concept of expectation to study the relationship of two random variables with each other. In particular, the **covariance** of two random variables X and Y is defined as

$$\mathsf{Cov}(X,Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Using an argument similar to that for variance, we can rewrite this as,

$$\begin{aligned} \mathsf{Cov}(X,Y) &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned}$$

Here, the key step in showing the equality of the two forms of covariance is in the third equality, where we use the fact that $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are actually constants which can be pulled out of the expectation. When $\mathsf{Cov}(X,Y) = 0$, we say that X and Y are **uncorrelated**.

We note the following properties of expectation and covariance

- (Linearity of expectation) $\mathbb{E}[f(X,Y)+g(X,Y)] = \mathbb{E}[f(X,Y)] + \mathbb{E}[g(X,Y)]$.
- $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X,Y)$.
- If X and Y are independent, then $\mathsf{Cov}(X,Y) = 0$.
- If X and Y are independent, then $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$.

4 Multiple Random Variables

The notions and ideas introduced in the previous section can be generalized to more than two random variables. In this section, for simplicity of presentation, we focus only on the continuous case, but the generalization to discrete random variables works similarly.

4.1 Basic properties

Suppose that we have n continuous random variables, $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$. We can define the **joint distribution function** of $X_1, X_2, \ldots, X_n$, the **joint probability density function** of $X_1, X_2, \ldots, X_n$, the **marginal probability density function** of X_1, and the **conditional probability density function** of X_1 given $X_2, \ldots, X_n$, as

$$\begin{aligned}
F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) \\
f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \frac{\partial^n F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{\partial x_1 \ldots \partial x_n} \\
f_{X_1}(x_1) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) dx_2 \ldots dx_n \\
f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \frac{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{f_{X_2,\ldots,X_n}(x_2,\ldots,x_n)}.
\end{aligned}$$

To calculate the probability of an event $\mathcal{A} \subseteq \mathbb{R}^n$ we have,

$$P((x_1, x_2, \ldots, x_n) \in \mathcal{A}) = \int_{(x_1,\ldots,x_n)\in\mathcal{A}} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) dx_1 \ldots dx_n.$$

From the definition of conditional probabilities for multiple random variables, one can establish the following theorem of chain rule.

Theorem 9 (Chain rule)**.** *We have*

$$\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= f(x_n|x_1,\ldots,x_{n-1}) f(x_1,\ldots,x_{n-1}) \\
&= f(x_n|x_1,\ldots,x_{n-1}) f(x_{n-1}|x_1,\ldots,x_{n-2}) f(x_1,\ldots,x_{n-2}) \\
&= \cdots = f(x_1) \prod_{i=2}^{n} f(x_i|x_1,\ldots,x_{i-1}).
\end{aligned}$$

Particularly, we say that random variables $X_1, \ldots, X_n$ are **independent** if

$$f(x_1,\ldots,x_n) = f(x_1)f(x_2)\ldots f(x_n).$$

Here, the definition of mutual independence is simply the natural generalization of independence of two random variables to multiple random variables. Independent random variables arise often in machine learning algorithms where we assume that the training examples belonging to the training set represent independent samples from some unknown probability distribution. To make the significance of independence clear, consider a "bad" training set in which we first sample a single training example $(x^{(1)}, y^{(1)})$ from the some unknown distribution, and then add $m-1$ copies of the exact same training example to the training set. In this case, we have (with some abuse of notation)

$$P(x^{(1)}, y^{(1)}, \ldots, x^{(m)}, y^{(m)}) \neq \prod_{i=1}^{m} P(x^{(i)}, y^{(i)}).$$

Despite the fact that the training set has size m, the examples are not independent! While clearly the procedure described here is not a sensible method for building a training set for a machine learning algorithm, it turns out that in practice, non-independence of samples does come up often, and it has the effect of reducing the "effective size" of the training set.

4.2 Random vectors, expectation and covariance

Suppose that we have n random variables. When working with all these random variables together, we will often find it convenient to put them in a vector $X = [X_1, X_2, \ldots, X_n]^T$. We call the resulting vector a **random vector** (more formally, a random vector is a mapping from Ω to $\mathbb{R}^n$). It should be clear that random vectors are simply an alternative notation for dealing with n random variables, so the notions of joint PDF and CDF will apply to random vectors as well.

Expectation. Consider an arbitrary function from $g : \mathbb{R}^n \to \mathbb{R}$. The **expected value** of this function is defined as

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n) f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n,$$

where $\int_{\mathbb{R}^n}$ is n consecutive integrations from $-\infty$ to ∞. If g is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, then the expected value of g is the element-wise expected values of the output vector, i.e., if g is

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}.$$

Then,

$$\mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \mathbb{E}[g_2(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{bmatrix}.$$

Covariance. For a given random vector $X : \Omega \to \mathbb{R}^n$, its **covariance matrix** Σ is the $n \times n$ square matrix whose entries are given by $\Sigma_{ij} = \mathsf{Cov}(X_i, X_j)$. From the definition of covariance, we have

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \mathsf{Cov}(X_1, X_1) & \dots & \mathsf{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \dots & \mathsf{Cov}(X_n, X_n) \end{bmatrix} \\
&= \mathbb{E} \begin{bmatrix} \mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] & \dots & \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_n - \mathbb{E}[X_n])] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n - \mathbb{E}[X_n])(X_1 - \mathbb{E}[X_1])] & \dots & \mathbb{E}[(X_n - \mathbb{E}[X_n])^2] \end{bmatrix} \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]\mathbb{E}[X]^T - \mathbb{E}[X]\mathbb{E}[X]^T + \mathbb{E}[X]\mathbb{E}[X]^T \\
&= \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T],
\end{aligned}$$

where the matrix expectation is defined in the obvious way. As seen in the following proposition, the covariance matrix of *any* random vector must always be symmetric positive semidefinite.

Proposition 10. *Suppose that Σ is the covariance matrix corresponding to some random vector X. Then Σ is symmetric positive semidefinite.*

Proof The symmetry of Σ follows immediately from its definition. Next, for any vector $z \in \mathbb{R}^n$, observe that

$$\begin{aligned}
z^T \Sigma z &= \sum_{i=1}^n \sum_{j=1}^n \Sigma_{ij} z_i z_j \\
&= \sum_{i=1}^n \sum_{j=1}^n (\mathsf{Cov}(X_i, X_j)) z_i z_j \\
&= \sum_{i=1}^n \sum_{j=1}^n (\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]) z_i z_j \\
&= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j]) z_i z_j \right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) z_i \right) \left(\sum_{j=1}^n (X_j - \mathbb{E}[X_j]) z_j \right) \right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^n z_i (X_i - \mathbb{E}[X_i]) \right)^2 \right].
\end{aligned}$$

Here, the first equality follows from the formula for expanding a quadratic form (see section notes on linear algebra), and the last equality follows by linearity of expectations (see probability notes). To complete the proof, observe that $z^T \Sigma z \geq 0$ for any z, thus Σ is positive semidefinite. $\square$

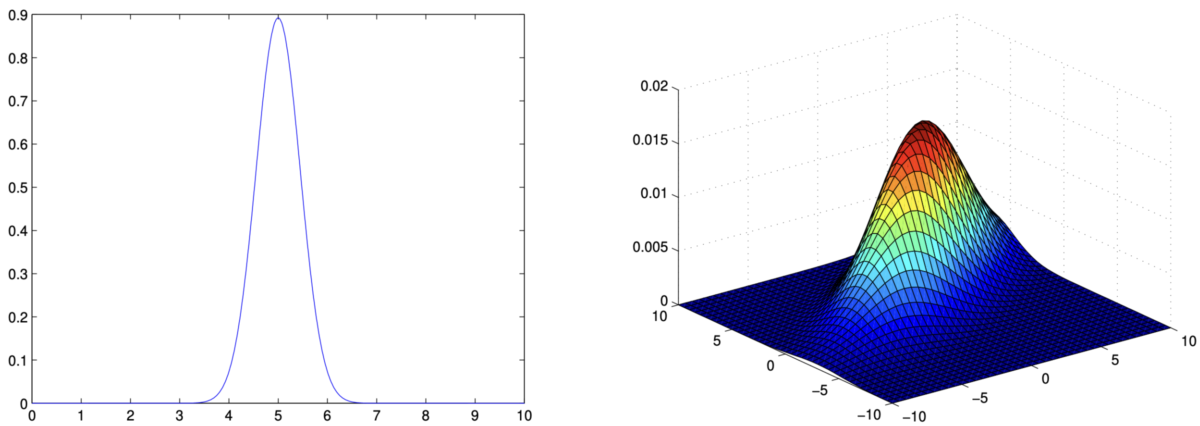

Figure 2: The figure on the left shows a univariate Gaussian density for a single variable X. The figure on the right shows a two-dimensional Gaussian density over two variables X_1 and X_2.

5 Multivariate Gaussian Distribution

One particularly important example of a probability distribution over random vectors X is called the **multivariate Gaussian** or **multivariate Normal distribution**. A random vector $X \in \mathbb{R}^d$ is said to have a multivariate normal (or Gaussian) distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{S}^d_{++}$ (where $\mathbb{S}^d_{++}$ refers to the space of symmetric positive definite $d \times d$ matrices) if

$$f_{X_1,X_2,\ldots,X_d}(x_1, x_2, \ldots, x_d; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

We write this as $X \sim \mathcal{N}(\mu, \Sigma)$. In this section, we describe multivariate Gaussians and some of their basic properties. Generally speaking, Gaussian random variables are extremely useful in machine learning and statistics for two main reasons. First, they are extremely common when modeling "noise" in statistical algorithms. Quite often, noise can be considered to be the accumulation of a large number of small independent random perturbations affecting the measurement process; by the Central Limit Theorem, summations of independent random variables will tend to "look Gaussian." Second, Gaussian random variables are convenient for many analytical manipulations, because many of the integrals involving Gaussian distributions that arise in practice have simple closed form solutions. We will encounter this later in the course.

Figure 2 illustrates the density of a two-dimensional Gaussian random variable.

5.1 Relationship to univariate Gaussian

Recall that the density function of a univariate normal (or Gaussian) distribution is given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

Here, the argument of the exponential function, $-\frac{1}{2\sigma^2}(x-\mu)^2$, is a quadratic function of the variable x. Furthermore, the parabola points downwards, as the coefficient of the quadratic term is negative.

The coefficient in front, $\frac{1}{\sqrt{2\pi}\sigma}$, is a constant that does not depend on x; hence, we can think of it as simply a "normalization factor" used to ensure that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = 1.$$

In the case of the multivariate Gaussian density, the argument of the exponential function, $-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$, is a **quadratic form** in the vector variable x. Since Σ is positive definite, and since the inverse of any positive definite matrix is also positive definite, then for any non-zero vector z, $z^T\Sigma^{-1}z > 0$. This implies that for any vector $x \neq \mu$,

$$-(x-\mu)^T\Sigma^{-1}(x-\mu) < 0$$
$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) < 0.$$

Like in the univariate case, you can think of the argument of the exponential function as being a downward opening quadratic bowl. The coefficient in front (i.e., $\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}$) has an even more complicated form than in the univariate case. However, it still does not depend on x, and hence it is again simply a normalization factor used to ensure that

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx_1 \ldots dx_d = 1.$$

5.2 Covariance matrix

The following proposition gives an alternative way to characterize the covariance matrix of a random vector X:

Proposition 11. *For any random vector X with mean μ and covariance matrix Σ, we have*

$$\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T] = \mathbb{E}[XX^T] - \mu\mu^T.$$

Proof We prove the first of the two equalities in the theorem; the proof of the other equality is similar. Recall that the covariance can be written as

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \mathsf{Cov}(X_1, X_1) & \ldots & \mathsf{Cov}(X_1, X_d) \\ \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_d, X_1) & \ldots & \mathsf{Cov}(X_d, X_d) \end{bmatrix} \\
&= \mathbb{E}\begin{bmatrix} (X_1-\mu_1)^2 & \ldots & (X_1-\mu_1)(X_d-\mu_d) \\ \vdots & \ddots & \vdots \\ (X_d-\mu_d)(X_1-\mu_1) & \ldots & (X_d-\mu_d)^2 \end{bmatrix} \\
&= \mathbb{E}[(X-\mu)(X-\mu)^T].
\end{aligned}$$

Here, the last equality follows from the fact that for any vector $z \in \mathbb{R}^d$,

$$zz^T = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_d \end{bmatrix} \begin{bmatrix} z_1 & z_2 & \ldots & z_d \end{bmatrix} = \begin{bmatrix} z_1z_1 & z_1z_2 & \ldots & z_1z_d \\ z_2z_1 & z_2z_2 & \ldots & z_2z_d \\ \vdots & \vdots & \ddots & \vdots \\ z_dz_1 & z_dz_2 & \ldots & z_dz_d \end{bmatrix}.$$

$\square$

In the definition of multivariate Gaussian, we require that the covariance matrix Σ be symmetric positive definite (i.e., $\Sigma \in \mathbf{S}^d_{++}$). Why does this restriction exist? First, Σ must be symmetric positive semidefinite in order for it to be a valid covariance matrix. However, in order for Σ^{-1} to exist (as required in the definition of the multivariate Gaussian density), then Σ must be invertible and hence full rank. Since any full rank symmetric positive semidefinite matrix is necessarily symmetric positive definite, it follows that Σ must be symmetric positive definite.

5.3 The diagonal covariance matrix case

To get an intuition for what a multivariate Gaussian is, consider the simple case where $d = 2$, and where the covariance matrix Σ is diagonal, i.e.,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

In this case, the multivariate Gaussian density has the form,

$$\begin{aligned} f(x;\mu,\Sigma) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix} \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1^2} \\ \frac{x_2-\mu_2}{\sigma_2^2} \end{bmatrix} \right), \end{aligned}$$

where we have relied on the explicit formula for the determinant of a 2×2 matrix[4], and the fact that the inverse of a diagonal matrix is simply found by taking the reciprocal of each diagonal entry. Continuing, we get

$$\begin{aligned} f(x;\mu,\Sigma) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right] \right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2} \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} \right) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2} \right). \end{aligned}$$

The last equation we recognize to simply be the product of two independent Gaussian densities; one with mean μ_1 and variance σ_1^2, and the other with mean μ_2 and variance σ_2^2. More generally, one can show that an d-dimensional Gaussian with mean $\mu \in \mathbb{R}^d$ and diagonal covariance matrix $\Sigma = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$ is the same as a collection of d independent Gaussian random variables with mean μ_i and variance σ_i^2, respectively.

[4]Namely, $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$.

5.4 Closure properties

A fancy feature of the multivariate Gaussian distribution is the following set of **closure properties**:

- The sum of independent Gaussian random variables is Gaussian.
- The marginal of a joint Gaussian distribution is Gaussian.
- The conditional of a joint Gaussian distribution is Gaussian.

The formal statement of these results are as follows.

Theorem 12. *Suppose that $Y \sim \mathcal{N}(\mu, \Sigma)$ and $Z \sim \mathcal{N}(\mu', \Sigma')$ are independent Gaussian distributed random variables, where $\mu, \mu' \in \mathbb{R}^d$ and $\Sigma, \Sigma' \in \mathbb{S}^d_{++}$. Then, their sum is also Gaussian:*

$$Y + Z \sim \mathcal{N}(\mu + \mu', \Sigma + \Sigma').$$

Theorem 13. *Suppose that*

$$\begin{bmatrix} X_A \\ X_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right)$$

where $X_A \in \mathbb{R}^{d_A}, X_B \in \mathbb{R}^{d_B}$, and the dimensions of the mean vectors and covariance matrix subblocks are chosen to match X_A and X_B. Then, the marginal densities,

$$p(x_A) = \int_{\mathbb{R}^{d_B}} p(x_A, x_B; \mu, \Sigma) dx_B$$

$$p(x_B) = \int_{\mathbb{R}^{d_A}} p(x_A, x_B; \mu, \Sigma) dx_A$$

are Gaussian;

$$X_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})$$

$$X_B \sim \mathcal{N}(\mu_B, \Sigma_{BB}).$$

Theorem 14. *Suppose that*

$$\begin{bmatrix} X_A \\ X_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right)$$

where X_A takes values in $\mathbb{R}^m$ and X_B takes values in $\mathbb{R}^n$, and the dimensions of the mean vectors and covariance matrix subblocks are chosen to match X_A and X_B. Then, the conditional densities

$$p_{X_A|X_B}(x_A|x_B) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_A \in \mathbb{R}^m} p(x_A, x_B; \mu, \Sigma) dx_A}$$

and

$$p_{X_B|X_A}(x_B|x_A) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B \in \mathbb{R}^n} p(x_A, x_B; \mu, \Sigma) dx_B}$$

are also Gaussian:

$$X_A|X_B \sim \mathcal{N}(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$$

$$X_B|X_A \sim \mathcal{N}(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB})$$

You don't need to memorize these theorems. Instead understand the closure properties of Gaussian distributions given at the top of the page.