Probability Theory Review and Reference

Arian Maleki (updated by Honglin Yuan and Tengyu Ma)

March 30, 2022

Contents

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 CS 229. The mathematical theory of probability is very 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 $P : \mathcal{F} \rightarrow \mathbf{R}$ that satisfies the following properties,
  - **Non-negativity**: $P(A) \geq 0$, for all $A \in \mathcal{F}$
  - **Completeness**: $P(\Omega) = 1$
  - **Countable Additivity**: If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

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

Example 1.1.: 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 $P(\{1, 2, 3\}) = \frac{3}{6}$.

1.2 Properties of probability

Proposition 1.2.: *The following properties can be derived from the axioms of probability.*

[1] $\mathcal{F}$ should satisfy three properties: (1) $\emptyset \in \mathcal{F}$; (2) $A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F}$; and (3) $A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_i A_i \in \mathcal{F}$.

- *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

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

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 1.3.: *[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*

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

Theorem 1.3 can be proved directly by applying the definition of the conditional probability. Note that Theorem 1.3 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 1.4.: *[Bayes' theorem] Suppose $A_1, \ldots, A_n$ are disjoint events, and event B satisfies $B \subset \bigcup_{i=1}^n A_i$. Then if $P(B) > 0$, it is the case that*

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

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. We will revisit this theorem many times throughout the course of CS 229.

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 $w_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$. 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 $X : \Omega \longrightarrow \mathbf{R}$.[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

[2] Technically speaking, not every function is not 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 each of the events of the event space as being sets of outcomes $\omega \in \Omega$ for which $X(\omega)$ satisfies some property (e.g., the event $\{\omega : X(\omega) \geq 3\}$).

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

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

2.2 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 $F_X : \mathbf{R} \to [0, 1]$ 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.

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

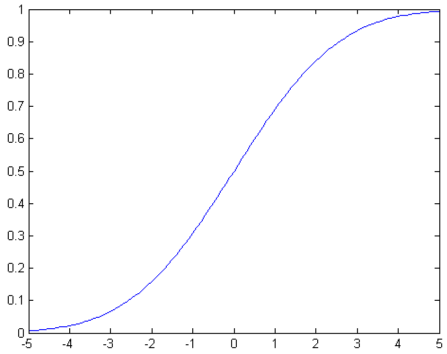

Figure 1: A cumulative distribution function (CDF).

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

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

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 $p_X : \Omega \to \mathbf{R}$ such that

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

In the case of discrete random variable, we use the notation $Val(X)$ 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 $Val(X) = \{0, 1, 2, \ldots, 10\}$.

A PMF function satisfies the following properties.

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

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 $\mathbf{R}$ will be at most one).

A PDF function satisfies the following properties.

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

2.5 Expectation

Suppose that X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbf{R} \longrightarrow \mathbf{R}$ 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

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

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

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

Intuitively, the expectation of $g(X)$ can be thought of as a "weighted average" of the values that $g(x)$ can taken 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, $E[X]$ 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:

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

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

$$Var[X] \triangleq E[(X - E(X))^2]$$

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

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

where the second equality follows from linearity of expectations and the fact that $E[X]$ is actually a constant with respect to the outer expectation.

We note the following properties of the variance.

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

Example 2.1.: 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

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

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

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

Therefore

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

Example 2.2.: Suppose that $g(x) = 1\{x \in A\}$ for some subset $A \subseteq \Omega$. What is $E[g(X)]$?

Discrete case:

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

Continuous case:

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

2.7 Some common distributions

In this subsection, we review several common discrete and continuous distributions that are commonly used throughout the CS 229 class.

Discrete random variables

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

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

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

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

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

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

- $X \sim Poisson(\lambda)$ (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

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

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

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

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

- $X \sim Normal(\mu, \sigma^2)$: also known as the Gaussian distribution

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

The shape of the PDFs and CDFs of some of these random variables are shown in Figure 2.

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

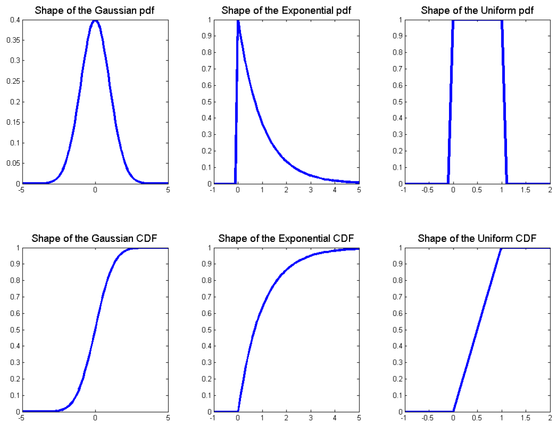

Figure 2: PDF and CDF of a couple of random variables.

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

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

$$\begin{aligned} F_X(x) &= \lim_{y \to \infty} F_{XY}(x, y) dy \\ F_Y(y) &= \lim_{x \to \infty} F_{XY}(x, y) dx. \end{aligned}$$

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** $p_{XY} : \mathbf{R} \times \mathbf{R} \to [0, 1]$ 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 $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} P_{XY}(x, y) = 1$.

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

$$\iint_{x \in A} f_{XY}(x, y) dx dy = P((X, Y) \in A).$$

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 $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) = 1$.

Analagous 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[4]. Ignoring this technical point, we simply define, by analogy to the discrete case, the

[4] To get around this, a more reasonable way to calculate the conditional CDF is,

$$F_{Y|X}(y, x) = \lim_{\Delta x \to 0} P(Y \leq y | x \leq X \leq x + \Delta x).$$

It can be easily seen that if $F(x, y)$ is differentiable in both x, y then,

$$F_{Y|X}(y, x) = \int_{-\infty}^{y} \frac{f_{X,Y}(x, \alpha)}{f_X(x)} d\alpha$$

conditional probability density of Y given $X = x$ to be

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

provided $f_X(x) \neq 0$.

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 disucssed in Section 1.

Theorem 3.1.: *Let X, Y be two random variables defiend on the same probability space, then*

$$E[X] = E[E[X|Y]]. \tag{8}$$

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,

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

If the random variables X and Y are continuous,

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

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

and therefore we define the conditional PDF of Y given $X = x$ in the following way,

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

- For continuous random variables, $f_{Y|X}(y|x) = f_Y(y)$ whenever $f_X(x) \neq 0$ for all $y \in \mathbf{R}$.

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 3.2.: *If X and Y are independent then for any subsets $A, B \subseteq \mathbf{R}$, 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 $g : \mathbf{R}^2 \longrightarrow \mathbf{R}$ is a function of these two random variables. Then the expected value of g is defined in the following way,

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

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

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

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

$$Cov[X,Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

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

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

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 $E[X]$ and $E[Y]$ are actually constants which can be pulled out of the expectation. When $Cov[X,Y] = 0$, we say that X and Y are **uncorrelated**[5].

[5] However, this is not the same thing as stating that X and Y are independent! For example, if $X \sim Uniform(-1,1)$ and $Y = X^2$, then one can show that X and Y are uncorrelated, even though they are not independent.

We note the following properties of expectation and covariance.

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

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

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

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

$$P((x_1,x_2,\ldots x_n) \in A) = \int_{(x_1,x_2,\ldots x_n)\in A} f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots x_n) dx_1 dx_2 \ldots dx_n \qquad (9)$$

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

Theorem 4.1.: *[Chain rule]*

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

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

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

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 $\mathbf{R}^n$). 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 $g : \mathbf{R}^n \to \mathbf{R}$. The **expected value** of this function is defined as

$$E[g(X)] = \int_{\mathbf{R}^n} g(x_1, x_2, \ldots, x_n) f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots x_n) dx_1 dx_2 \ldots dx_n, \tag{10}$$

where $\int_{\mathbf{R}^n}$ is n consecutive integrations from $-\infty$ to ∞. If g is a function from $\mathbf{R}^n$ to $\mathbf{R}^m$, 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,

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

Covariance. For a given random vector $X : \Omega \to \mathbf{R}^n$, its **covariance matrix** Σ is the $n \times n$ square matrix whose entries are given by $\Sigma_{ij} = Cov[X_i, X_j]$.

From the definition of covariance, we have

$$\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix} \\
&= E[XX^T] - E[X]E[X]^T = \ldots = E[(X - E[X])(X - E[X])^T].
\end{aligned}$$

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 4.2.: *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

$z \in \mathbf{R}^n$, observe that

$$
\begin{aligned}
z^T \Sigma z &= \sum_{i=1}^{n} \sum_{j=1}^{n} (\Sigma_{ij} z_i z_j) && (11)\\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (Cov[X_i, X_j] \cdot z_i z_j)\\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (E[(X_i - E[X_i])(X_j - E[X_j])] \cdot z_i z_j)\\
&= E\left[\sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - E[X_i])(X_j - E[X_j]) \cdot z_i z_j\right]. && (12)
\end{aligned}
$$

Here, (11) follows from the formula for expanding a quadratic form (see section notes on linear algebra), and (12) follows by linearity of expectations (see probability notes).

To complete the proof, observe that the quantity inside the brackets is of the form $\sum_i \sum_j x_i x_j z_i z_j = (x^T z)^2 \geq 0$. Therefore, the quantity inside the expectation is always nonnegative, and hence the expectation itself must be nonnegative. We conclude that $z^T \Sigma z \geq 0$. □

4.3 The law of large numbers and Central limit theorem

A common scenario to generate a series of random variables is to repeat the same experiment for a large number of times. An important probabilistic claim is that the average of the results obtained from a large number of trials should converge to its expected value (mean). This rule is called **The law of large numbers** (LLN) which we formally state as below

Theorem 4.3: *[Strong Law of Large Numbers] Let $X_1, X_2, \ldots,$ be a series of independent and identically distributed (usually abbreviated as i.i.d.) random variables for which $E[|X_1|] < \infty$. Then*

$$P\left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = E[X_1]\right) = 1. \tag{13}$$

The convergence stated in Theorem 4.3 is also known as the *almost surely* convergence in probabilistic literature, since the empirical average of sequence converges to the expected value with probability 1.

A natural follow-up question to ask is how *fast* the empirical average converges to its expected value. The **central limit theorem** (CLT) answers this question through a refinement of the law of the large numbers.

Theorem 4.4.: *[Central Limit Theorem] Let $X_1, X_2, \ldots$ be a series of iid random variables* *with mean μ and variance σ^2. Then the normalized partial sum*

$$\xi_n \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right) \tag{14}$$

satisfies

$$\lim_{n \to \infty} P(\xi_n \leq x) = \Phi(x) \tag{15}$$

for any x, where Φ is CDF of the standard normal distribution.

The convergence stated in (15) is also known as *convergence in distribution* in probabilistic literature. THeorem 4.4 shows that irrespective of what distribution X follows, its normalized partial sum (or average) is always a normal distribution! So when the number of samples are large, we can approximate any distribution using a normal distribtution.

Note that both LLN and CLT can be extended to multi-dimensional random vectors, and generalized to weaker assumptions. The proofs of both theorems are out of the scope of this class.

5 The 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 $X \in \mathbf{R}^d$ is said to have a multivariate normal (or Gaussian) distribution with mean $\mu \in \mathbf{R}^d$ and covariance matrix $\Sigma \in \mathbf{S}^d_{++}$ (where $\mathbf{S}^d_{++}$ refers to the space of symmetric positive definite $d \times d$ matrices)

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

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.

5.1 Relationship to univariate Gaussians

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

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

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

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

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 $x \neq \mu$,

$$\begin{aligned}(x-\mu)^T\Sigma^{-1}(x-\mu) &> 0 \\ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) &< 0.\end{aligned}$$

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

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx_1 dx_2 \cdots dx_d = 1.$$

5.2 The covariance matrix

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

Proposition 5.1.: *For any random vector X with mean μ and covariance matrix Σ,*

$$\Sigma = E[(X-\mu)(X-\mu)^T] = E[XX^T] - \mu\mu^T. \tag{16}$$

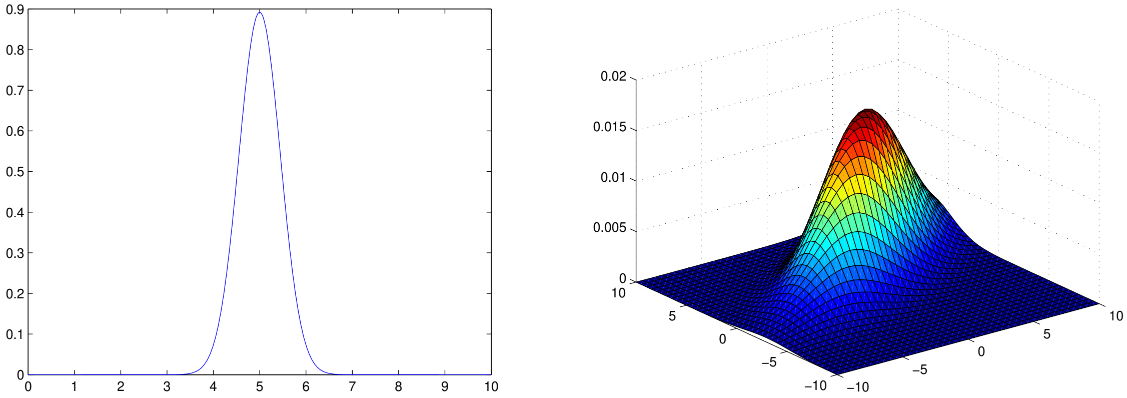

Figure 3: The figure on the left shows a univariate Gaussian density for a single variable X. The figure on the right shows a multivariate Gaussian density over two variables X_1 and X_2.

Proof. We prove the first of the two equalities in (16); the proof of the other equality is similar.

$$\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_d] \\ \vdots & \ddots & \vdots \\ Cov[X_d, X_1] & \cdots & Cov[X_d, X_d] \end{bmatrix} \\
&= \begin{bmatrix} E[(X_1 - \mu_1)^2] & \cdots & E[(X_1 - \mu_1)(X_d - \mu_d)] \\ \vdots & \ddots & \vdots \\ E[(X_d - \mu_d)(X_1 - \mu_1)] & \cdots & E[(X_d - \mu_d)^2] \end{bmatrix} \\
&= E \begin{bmatrix} (X_1 - \mu_1)^2 & \cdots & (X_1 - \mu_1)(X_d - \mu_d) \\ \vdots & \ddots & \vdots \\ (X_d - \mu_d)(X_1 - \mu_1) & \cdots & (X_d - \mu_d)^2 \end{bmatrix} && (17) \\
&= E \left[\begin{bmatrix} X_1 - \mu_1 \\ \vdots \\ X_d - \mu_d \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 & \cdots & X_d - \mu_d \end{bmatrix} \right] && (18) \\
&= E\left[(X - \mu)(X - \mu)^T \right].
\end{aligned}$$

Here, (17) follows from the fact that the expectation of a matrix is simply the matrix found by taking the componentwise expectation of each entry. Also, (18) follows from the fact that for any vector $z \in \mathbf{R}^d$,

$$zz^T = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_d \end{bmatrix} \begin{bmatrix} z_1 & z_2 & \cdots z_d \end{bmatrix} = \begin{bmatrix} z_1 z_1 & z_1 z_2 & \cdots & z_1 z_d \\ z_2 z_1 & z_2 z_2 & \cdots & z_2 z_d \\ \vdots & \vdots & \ddots & \vdots \\ z_d z_1 & z_d z_2 & \cdots & z_d z_d \end{bmatrix}.$$

□

In the definition of multivariate Gaussians, we required 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 $n = 2$, 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,

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

where we have relied on the explicit formula for the determinant of a 2 × 2 matrix[6], and the fact that the inverse of a diagonal matrix is simply found by taking the reciprocal of each diagonal entry. Continuing,

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

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.

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

More generally, one can show that an d-dimensional Gaussian with mean $\mu \in \mathbf{R}^d$ and diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$ is the same as a collection of d independent Gaussian random variables with mean μ_i and variance σ_i^2, respectively.

5.4 Isocontours

Another way to understand a multivariate Gaussian conceptually is to understand the shape of its **isocontours**. For a function $f : \mathbf{R}^2 \to \mathbf{R}$, an isocontour is a set of the form

$$\big\{x \in \mathbf{R}^2 : f(x) = c\big\}.$$

for some $c \in \mathbf{R}$.[7]

5.4.1 Shape of isocontours

What do the isocontours of a multivariate Gaussian look like? As before, let's consider the case where $d = 2$, and Σ is diagonal, i.e.,

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

As we showed in the last subsection,

$$p(x; \mu, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right). \tag{19}$$

Now, let's consider the level set consisting of all points where $f(x; \mu, \Sigma) = c$ for some constant $c \in \mathbf{R}$. In particular, consider the set of all $x_1, x_2 \in \mathbf{R}$ such that

$$\begin{aligned}
c &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \\
2\pi c\sigma_1\sigma_2 &= \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \\
\log(2\pi c\sigma_1\sigma_2) &= -\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \\
\log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right) &= \frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 + \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \\
1 &= \frac{(x_1 - \mu_1)^2}{2\sigma_1^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right)} + \frac{(x_2 - \mu_2)^2}{2\sigma_2^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right)}.
\end{aligned}$$

[7] Isocontours are often also known as **level curves**. More generally, a **level set** of a function $f : \mathbf{R}^d \to \mathbf{R}$, is a set of the form $\big\{x \in \mathbf{R}^2 : f(x) = c\big\}$ for some $c \in \mathbf{R}$.

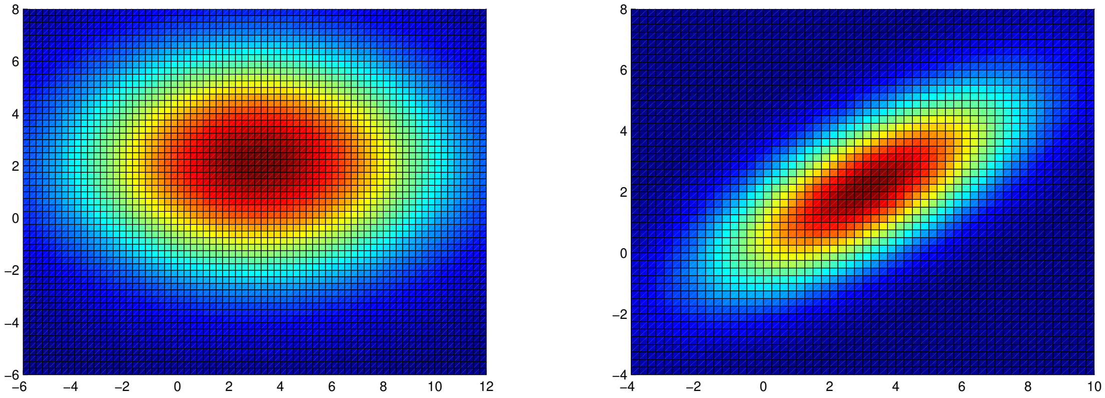

Figure 4:
The figure on the left shows a heatmap indicating values of the density function for an axis-aligned multivariate Gaussian with mean $\mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and diagonal covariance matrix $\Sigma = \begin{bmatrix} 25 & 0 \\ 0 & 9 \end{bmatrix}$. Notice that the Gaussian is centered at $(3, 2)$, and that the isocontours are all elliptically shaped with major/minor axis lengths in a 5:3 ratio. The figure on the right shows a heatmap indicating values of the density function for a non axis-aligned multivariate Gaussian with mean $\mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and covariance matrix $\Sigma = \begin{bmatrix} 10 & 5 \\ 5 & 5 \end{bmatrix}$. Here, the ellipses are again centered at $(3, 2)$, but now the major and minor axes have been rotated via a linear transformation.

Defining

$$r_1 = \sqrt{2\sigma_1^2 \log\left(\frac{1}{2\pi c \sigma_1 \sigma_2}\right)} \qquad\qquad r_2 = \sqrt{2\sigma_2^2 \log\left(\frac{1}{2\pi c \sigma_1 \sigma_2}\right)},$$

it follows that

$$1 = \left(\frac{x_1 - \mu_1}{r_1}\right)^2 + \left(\frac{x_2 - \mu_2}{r_2}\right)^2. \tag{20}$$

Equation (20) should be familiar to you from high school analytic geometry: it is the equation of an **axis-aligned ellipse**, with center (μ_1, μ_2), where the x_1 axis has length $2r_1$ and the x_2 axis has length $2r_2$!

5.4.2 Length of axes

To get a better understanding of how the shape of the level curves vary as a function of the variances of the multivariate Gaussian distribution, suppose that we are interested in the values of r_1 and r_2 at which c is equal to a fraction $1/e$ of the peak height of Gaussian density.

First, observe that maximum of Equation (19) occurs where $x_1 = \mu_1$ and $x_2 = \mu_2$. Substituting these values into Equation (19), we see that the peak height of the Gaussian density is $\frac{1}{2\pi\sigma_1\sigma_2}$.

Second, we substitute $c = \frac{1}{e}\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)$ into the equations for r_1 and r_2 to obtain

$$r_1 = \sqrt{2\sigma_1^2 \log\left(\frac{1}{2\pi\sigma_1\sigma_2 \cdot \frac{1}{e}\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)}\right)} = \sigma_1\sqrt{2}$$

$$r_2 = \sqrt{2\sigma_2^2 \log\left(\frac{1}{2\pi\sigma_1\sigma_2 \cdot \frac{1}{e}\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)}\right)} = \sigma_2\sqrt{2}.$$

From this, it follows that the axis length needed to reach a fraction $1/e$ of the peak height of the Gaussian density in the ith dimension grows in proportion to the standard deviation σ_i. Intuitively, this again makes sense: the smaller the variance of some random variable x_i, the more "tightly" peaked the Gaussian distribution in that dimension, and hence the smaller the radius r_i.

5.4.3 Non-diagonal case, higher dimensions

Clearly, the above derivations rely on the assumption that Σ is a diagonal matrix. However, in the non-diagonal case, it turns out that the picture is not all that different. Instead of being an axis-aligned ellipse, the isocontours turn out to be simply **rotated ellipses**. Furthermore, in the d-dimensional case, the level sets form geometrical structures known as **ellipsoids** in $\mathbf{R}^d$.

5.5 Linear transformation

In the last few subsections, we focused primarily on providing an intuition for how multivariate Gaussians with diagonal covariance matrices behaved. In particular, we found that an d-dimensional multivariate Gaussian with diagonal covariance matrix could be viewed simply as a collection of d independent Gaussian-distributed random variables with means and variances μ_i and σ_i^2, respectvely. In this section, we dig a little deeper and provide a quantitative interpretation of multivariate Gaussians when the covariance matrix is not diagonal.

The key result of this section is the following theorem (see proof in Appendix A).

Theorem 5.2.: *Let $X \sim \mathcal{N}(\mu, \Sigma)$ for some $\mu \in \mathbf{R}^d$ and $\Sigma \in \mathbf{S}^d_{++}$. Then, there exists a matrix $B \in \mathbf{R}^{d\times d}$ such that if we define $Z = B^{-1}(X - \mu)$, then $Z \sim \mathcal{N}(0, I)$.*

To understand the meaning of this theorem, note that if $Z \sim \mathcal{N}(0, I)$, then using the analysis from Section 5.4, Z can be thought of as a collection of d independent standard normal random variables (i.e., $Z_i \sim \mathcal{N}(0, 1)$). Furthermore, if $Z = B^{-1}(X - \mu)$ then $X = BZ + \mu$ follows from simple algebra.

Consequently, the theorem states that any random variable X with a multivariate Gaussian distribution can be interpreted as the result of applying a linear transformation ($X = BZ + \mu$) to some collection of d independent standard normal random variables (Z).

5.6 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.

In this subsection, we'll go through each of the closure properties, and we'll either prove the property or at least give some type of intuition as to why the property is true.

5.6.1 Sum of independent Gaussians is Gaussian

The formal statement of this rule is:

Theorem 5.3.: *Suppose that $y \sim \mathcal{N}(\mu, \Sigma)$ and $z \sim \mathcal{N}(\mu', \Sigma')$ are independent Gaussian distributed random variables, where $\mu, \mu' \in \mathbf{R}^d$ and $\Sigma, \Sigma' \in \mathbf{S}^d_{++}$. Then, their sum is also Gaussian:*

$$y + z \sim \mathcal{N}(\mu + \mu', \Sigma + \Sigma').$$

Before we prove anything, here are some observations:

1. The first thing to point out is that the importance of the independence assumption in the above rule. To see why this matters, suppose that $y \sim \mathcal{N}(\mu, \Sigma)$ for some mean vector μ and covariance matrix Σ, and suppose that $z = -y$. Clearly, z also has a Gaussian distribution (in fact, $z \sim \mathcal{N}(-\mu, \Sigma)$, but $y + z$ is identically zero!
2. The second thing to point out is a point of confusion for many students: if we add together two Gaussian densities ("bumps" in multidimensional space), wouldn't we get back some bimodal (i.e., "two-humped" density)? Here, the thing to realize is that the density of the random variable $y + z$ in this rule is NOT found by simply adding the densities of the individual random variables y and z. Rather, the density of $y + z$ will

actually turn out to be a *convolution* of the densities for y and z.[8] To show that the convolution of two Gaussian densities gives a Gaussian density, however, is beyond the scope of this class.

Instead, we will only show that the addition $y + z$ has mean $\mu + \mu'$ and covariance $\Sigma + \Sigma'$. For the mean, we have

$$E[y_i + z_i] = E[y_i] + E[z_i] = \mu_i + \mu_i'$$

from linearity of expectations. Therefore, the mean of $y + z$ is simply $\mu + \mu'$. Also, the (i, j)th entry of the covariance matrix is given by

$$\begin{aligned}
&E[(y_i + z_i)(y_j + z_j)] - E[y_i + z_i]E[y_j + z_j] \\
&\quad = E[y_iy_j + z_iy_j + y_iz_j + z_iz_j] - (E[y_i] + E[z_i])(E[y_j] + E[z_j]) \\
&\quad = E[y_iy_j] + E[z_iy_j] + E[y_iz_j] + E[z_iz_j] - E[y_i]E[y_j] - E[z_i]E[y_j] - E[y_i]E[z_j] - E[z_i][z_j] \\
&\quad = (E[y_iy_j] - E[y_i]E[y_j]) + (E[z_iz_j] - E[z_i]E[z_j]) \\
&\qquad + (E[z_iy_j] - E[z_i]E[y_j]) + (E[y_iz_j] - E[y_i]E[z_j]).
\end{aligned}$$

Using the fact that y and z are independent, we have $E[z_iy_j] = E[z_i]E[y_j]$ and $E[y_iz_j] = E[y_i]E[z_j]$. Therefore, the last two terms drop out, and we are left with,

$$\begin{aligned}
&E[(y_i + z_i)(y_j + z_j)] - E[y_i + z_i]E[y_j + z_j] \\
&\quad = (E[y_iy_j] - E[y_i]E[y_j]) + (E[z_iz_j] - E[z_i]E[z_j]) \\
&\quad = \Sigma_{ij} + \Sigma'_{ij}.
\end{aligned}$$

From this, we can conclude that the covariance matrix of $y + z$ is simply $\Sigma + \Sigma'$.

5.6.2 Marginal of a joint Gaussian is Gaussian

The formal statement of this rule is:

Theorem 5.4.: *Suppose that*

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

[8] For example, if y and z were univariate Gaussians (i.e., $y \sim \mathcal{N}(\mu, \sigma^2)$, $z \sim \mathcal{N}(\mu', \sigma'^2)$), then the convolution of their probability densities is given by

$$\begin{aligned}
p(y + z; \mu, \mu', \sigma^2, \sigma'^2) &= \int_{-\infty}^{\infty} p(w; \mu, \sigma^2)p(y + z - w; \mu', \sigma'^2)dw \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(w - \mu)^2\right) \cdot \frac{1}{\sqrt{2\pi}\sigma'} \exp\left(-\frac{1}{2\sigma'^2}(y + z - w - \mu')^2\right)dw
\end{aligned}$$

where $x_A \in \mathbf{R}^n$, $x_B \in \mathbf{R}^d$, and the dimensions of the mean vectors and covariance matrix subblocks are chosen to match x_A and x_B. Then, the marginal densities,

$$\begin{aligned}p(x_A) &= \int_{x_B \in \mathbf{R}^d} p(x_A, x_B; \mu, \Sigma) dx_B \\ p(x_B) &= \int_{x_A \in \mathbf{R}^n} p(x_A, x_B; \mu, \Sigma) dx_A\end{aligned}$$

are Gaussian:

$$\begin{aligned}x_A &\sim \mathcal{N}(\mu_A, \Sigma_{AA}) \\ x_B &\sim \mathcal{N}(\mu_B, \Sigma_{BB}).\end{aligned}$$

To justify this rule, let's just focus on the marginal distribution with respect to the variables x_A.[9]

First, note that computing the mean and covariance matrix for a marginal distribution is easy: simply take the corresponding subblocks from the mean and covariance matrix of the joint density. To make sure this is absolutely clear, let's look at the covariance between $x_{A,i}$ and $x_{A,j}$ (the ith component of x_A and the jth component of x_A). Note that $x_{A,i}$ and $x_{A,j}$ are also the ith and jth components of

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix}$$

(since x_A appears at the top of this vector). To find their covariance, we need to simply look at the (i, j)th element of the covariance matrix,

$$\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$$

The (i, j)th element is found in the Σ_{AA} subblock, and in fact, is precisely $\Sigma_{AA,ij}$. Using this argument for all $i, j \in \{1, \ldots, m\}$, we see that the covariance matrix for x_A is simply Σ_{AA}. A similar argument can be used to find that the mean of x_A is simply μ_A. Thus, the above argument tells us that if we knew that the marginal distribution over x_A is Gaussian, then we could immediately write down a density function for x_A in terms of the appropriate submatrices of the mean and covariance matrices for the joint density!

The above argument, though simple, however, is somewhat unsatisfying: how can we actually be sure that x_A has a multivariate Gaussian distribution? The argument for this is slightly long-winded, so rather than saving up the punchline, here's our plan of attack up front:

[9] In general, for a random vector x which has a Gaussian distribution, we can always permute entries of x so long as we permute the entries of the mean vector and the rows/columns of the covariance matrix in the corresponding way. As a result, it suffices to look only at x_A, and the result for x_B follows immediately.

1. Write the integral form of the marginal density explicitly.
2. Rewrite the integral by partitioning the inverse covariance matrix.
3. Use a "completion-of-squares" argument to evaluate the integral over x_B.
4. Argue that the resulting density is Gaussian.

Let's see each of these steps in action.

The marginal density in integral form Suppose that we wanted to compute the density function of x_A directly. Then, we would need to compute the integral,

$$\begin{aligned} p(x_A) &= \int_{x_B \in \mathbf{R}^d} p(x_A, x_B; \mu, \Sigma) dx_B \\ &= \frac{1}{(2\pi)^{\frac{n+n}{2}} \begin{vmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{vmatrix}^{1/2}} \int_{x_B \in \mathbf{R}^d} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) dx_B. \end{aligned}$$

Partitioning the inverse covariance matrix To make any sort of progress, we'll need to write the matrix product in the exponent in a slightly different form. In particular, let us define the matrix $V \in \mathbf{R}^{(m+n)\times(m+n)}$ as[10]

$$V = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \Sigma^{-1}.$$

It might be tempting to think that

$$V = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \text{``=''} \begin{bmatrix} \Sigma_{AA}^{-1} & \Sigma_{AB}^{-1} \\ \Sigma_{BA}^{-1} & \Sigma_{BB}^{-1} \end{bmatrix}$$

However, the rightmost equality does not hold! We'll return to this issue in a later step; for now, though, it suffices to define V as above without worrying what actual contents of each submatrix are.

Using this definition of V, the integral expands to

$$\begin{aligned} p(x_A) = \frac{1}{Z} \int_{x_B \in \mathbf{R}^d} \exp\Big(-\Big[&\frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) + \frac{1}{2}(x_A - \mu_A)^T V_{AB}(x_B - \mu_B) \\ &+ \frac{1}{2}(x_B - \mu_B)^T V_{BA}(x_A - \mu_A) + \frac{1}{2}(x_B - \mu_B)^T V_{BB}(x_B - \mu_B) \Big] \Big) dx_B, \end{aligned}$$

where Z is some constant not depending on either x_A or x_B that we'll choose to ignore for the moment. If you haven't worked with partitioned matrices before, then the expansion

[10] Sometimes, V is called the "precision" matrix.

above may seem a little magical to you. It is analogous to the idea that when defining a quadratic form based on some 2×2 matrix A, then

$$x^T A x = \sum_i \sum_j A_{ij} x_i x_j = x_1 A_{11} x_1 + x_1 A_{12} x_2 + x_2 A_{21} x_1 + x_2 A_{22} x_2.$$

Take some time to convince yourself that the matrix generalization above also holds.

Integrating out x_B To evaluate the integral, we'll somehow want to integrate out x_B. In general, however, Gaussian integrals are hard to compute by hand. Is there anything we can do to save time? There are, in fact, a number of Gaussian integrals for which the answer is already known. For example,

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \int_{\mathbf{R}^d} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) = 1. \tag{21}$$

The basic idea in this section, then, will be to transform the integral we had in the last section into a form where we can apply (21) in order to perform the required integration easily.

The key to this is a mathematical trick known as "completion of squares." Consider the quadratic function $z^T A z + b^T z + c$ where A is a symmetric, nonsingular matrix. Then, one can verify directly that

$$\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2}\big(z + A^{-1}b\big)^T A \big(z + A^{-1}b\big) + c - \frac{1}{2} b^T A^{-1} b.$$

This is the multivariate generalization of the "completion of squares" argument used in single variable algebra:

$$\frac{1}{2} a z^2 + bz + c = \frac{1}{2} a \left(z + \frac{b}{a}\right)^2 + c - \frac{b^2}{2a}$$

To apply the completion of squares in our situation above, let

$$\begin{aligned} z &= x_B - \mu_B \\ A &= V_{BB} \\ b &= V_{BA}(x_A - \mu_A) \\ c &= \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A). \end{aligned}$$

Then, it follows that the integral can be rewritten as

$$p(x_A) = \frac{1}{Z} \int_{x_B \in \mathbf{R}^d} \exp\Bigg(-\Bigg[\frac{1}{2}\big(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\big)^T V_{BB}\big(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\big)$$
$$+ \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) - \frac{1}{2}(x_A - \mu_A)^T V_{AB} V_{BB}^{-1} V_{BA}(x_A - \mu_A)\Bigg]\Bigg) dx_B$$

We can factor out the terms not including x_B to obtain,

$$p(x_A) = \exp\left(-\frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) + \frac{1}{2}(x_A - \mu_A)^T V_{AB} V_{BB}^{-1} V_{BA}(x_A - \mu_A)\right)$$
$$\cdot \frac{1}{Z} \int_{x_B \in \mathbf{R}^d} \exp\left(-\frac{1}{2}\left[\left(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\right)^T V_{BB}\left(x_B - \mu_B + V_{BB}^{-1} V_{BA}(x_A - \mu_A)\right)\right]\right) dx_B$$

At this point, we can now apply (21). We use this fact to get rid of the remaining integral in our expression for $p(x_A)$:

$$p(x_A) = \frac{1}{Z} \cdot (2\pi)^{d/2} |V_{BB}|^{-1/2} \cdot \exp\left(-\frac{1}{2}(x_A - \mu_A)^T (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})(x_A - \mu_A)\right).$$

Arguing that resulting density is Gaussian At this point, we are almost done! Ignoring the normalization constant in front, we see that the density of x_A is the exponential of a quadratic form in x_A. We can quickly recognize that our density is none other than a Gaussian with mean vector μ_A and covariance matrix $(V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1}$. Although the form of the covariance matrix may seem a bit complex, we have already achieved what we set out to show in the first place—namely, that x_A has a marginal Gaussian distribution. Using the logic before, we can conclude that this covariance matrix must somehow reduce to Σ_{AA}.

But, in case you are curious, it's also possible to show that our derivation is consistent with this earlier justification. To do this, we use the following result for partitioned matrices:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M^{-1} & -M^{-1}BD^{-1} \\ -D^{-1}CM^{-1} & D^{-1} + D^{-1}CM^{-1}BD^{-1} \end{bmatrix}.$$

where $M = A - BD^{-1}C$. This formula can be thought of as the multivariable generalization of the explicit inverse for a 2×2 matrix,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Using the formula, it follows that

$$\begin{aligned} \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} &= \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1} & -(V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1} V_{AB} V_{BB}^{-1} \\ -V_{BB}^{-1} V_{BA} (V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1} & (V_{BB} - V_{BA} V_{AA}^{-1} V_{AB})^{-1} \end{bmatrix} \end{aligned}$$

We immediately see that $(V_{AA} - V_{AB} V_{BB}^{-1} V_{BA})^{-1} = \Sigma_{AA}$, just as we expected!

5.6.3 Conditional of a joint Gaussian is Gaussian

The formal statement of this rule is:

Theorem 5.5.: *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 $x_A \in \mathbf{R}^n$, $x_B \in \mathbf{R}^d$, and the dimensions of the mean vectors and covariance matrix subblocks are chosen to match x_A and x_B. Then, the conditional densities

$$\begin{aligned} p(x_A \mid x_B) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_A \in \mathbf{R}^n} p(x_A, x_B; \mu, \Sigma) dx_A} \\ p(x_B \mid x_A) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B \in \mathbf{R}^d} p(x_A, x_B; \mu, \Sigma) dx_B} \end{aligned}$$

are also Gaussian:

$$\begin{aligned} x_A \mid x_B &\sim \mathcal{N}\big(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\big) \\ x_B \mid x_A &\sim \mathcal{N}\big(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}\big). \end{aligned}$$

The proof of Theorem 5.5 is similar to the marginal theorem above, which we defer to Appendix B.

6 Other Resources

A good textbook on probablity at the level needed for CS229 is the book, *A First Course on Probability* by Sheldon Ross.

A Proof of Theorem 5.2

The derivation of this theorem requires some advanced linear algebra and probability theory and can be skipped for the purposes of this class. Our argument will consist of two parts. First, we will show that the covariance matrix Σ can be factorized as $\Sigma = BB^T$ for some invertible matrix B. Second, we will perform a "change-of-variable" from X to a different vector valued random variable Z using the relation $Z = B^{-1}(X - \mu)$.

Step 1: Factorizing the covariance matrix. Recall the following two properties of symmetric matrices from the notes on linear algebra[11]:

[11] See section on "Eigenvalues and Eigenvectors of Symmetric Matrices."

1. Any real symmetric matrix $A \in \mathbf{R}^{d\times d}$ can always be represented as $A = U\Lambda U^T$, where U is a full rank orthogonal matrix containing of the eigenvectors of A as its columns, and Λ is a diagonal matrix containing A's eigenvalues.
2. If A is symmetric positive definite, all its eigenvalues are positive.

Since the covariance matrix Σ is positive definite, using the first fact, we can write $\Sigma = U\Lambda U^T$ for some appropriately defined matrices U and Λ. Using the second fact, we can define $\Lambda^{1/2} \in \mathbf{R}^{d\times d}$ to be the diagonal matrix whose entries are the square roots of the corresponding entries from Λ. Since $\Lambda = \Lambda^{1/2}(\Lambda^{1/2})^T$, we have

$$\Sigma = U\Lambda U^T = U\Lambda^{1/2}(\Lambda^{1/2})^T U^T = U\Lambda^{1/2}(U\Lambda^{1/2})^T = BB^T,$$

where $B = U\Lambda^{1/2}$.[12] In this case, then $\Sigma^{-1} = B^{-T}B^{-1}$, so we can rewrite the standard formula for the density of a multivariate Gaussian as

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|BB^T|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T B^{-T}B^{-1}(x-\mu)\right). \tag{22}$$

Step 2: Change of variables. Now, define the vector-valued random variable $Z = B^{-1}(X-\mu)$. A basic formula of probability theory, which we did not introduce in the section notes on probability theory, is the "change-of-variables" formula for relating vector-valued random variables:

> Suppose that $X = \begin{bmatrix} X_1 & \cdots & X_d\end{bmatrix}^T \in \mathbf{R}^d$ is a vector-valued random variable with joint density function $f_X : \mathbf{R}^d \to \mathbf{R}$. If $Z = H(X) \in \mathbf{R}^d$ where H is a bijective, differentiable function, then Z has joint density $f_Z : \mathbf{R}^d \to \mathbf{R}$, where
>
> $$f_Z(z) = f_X(x) \cdot \left|\det\left(\begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \cdots & \frac{\partial x_1}{\partial z_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial z_1} & \cdots & \frac{\partial x_d}{\partial z_d}\end{bmatrix}\right)\right|.$$

Using the change-of-variable formula, one can show (after some algebra, which we'll skip) that the vector variable Z has the following joint density:

$$p_Z(z) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}z^T z\right). \tag{23}$$

The claim follows immediately. □

[12] To show that B is invertible, it suffices to observe that U is an invertible matrix, and right-multiplying U by a diagonal matrix (with no zero diagonal entries) will rescale its columns but will not change its rank.

B Proof of Theorem 5.5

As before, we'll just examine the conditional distribution $x_B \mid x_A$, and the other result will hold by symmetry. Our plan of attack will be as follows:

1. Write the form of the conditional density explicitly.
2. Rewrite the expression by partitioning the inverse covariance matrix.
3. Use a "completion-of-squares" argument.
4. Argue that the resulting density is Gaussian.

Let's see each of these steps in action.

The conditional density written explicitly Suppose that we wanted to compute the density function of x_B given x_A directly. Then, we would need to compute

$$\begin{aligned} p(x_B \mid x_A) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B \in \mathbf{R}^n} p(x_A, x_B; \mu, \Sigma) dx_B} \\ &= \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \end{aligned}$$

where Z' is a normalization constant that we used to absorb factors not depending on x_B. Note that this time, we don't even need to compute any integrals – the value of the integral does not depend on x_B, and hence the integral can be folded into the normalization constant Z'.

Partitioning the inverse covariance matrix As before, we reparameterize our density using the matrix V, to obtain

$$\begin{aligned} p(x_B \mid x_A) &= \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \\ &= \frac{1}{Z'} \exp\Big(-\Big[\frac{1}{2}(x_A - \mu_A)^T V_{AA} (x_A - \mu_A) + \frac{1}{2}(x_A - \mu_A)^T V_{AB} (x_B - \mu_B) \\ &\qquad\qquad + \frac{1}{2}(x_B - \mu_B)^T V_{BA} (x_A - \mu_A) + \frac{1}{2}(x_B - \mu_B)^T V_{BB} (x_B - \mu_B) \Big] \Big). \end{aligned}$$

Use a "completion of squares" argument Recall that

$$\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2}\big(z + A^{-1} b\big)^T A \big(z + A^{-1} b\big) + c - \frac{1}{2} b^T A^{-1} b$$

provided A is a symmetric, nonsingular matrix. As before, to apply the completion of squares in our situation above, let

$$\begin{aligned}
z &= x_B - \mu_B \\
A &= V_{BB} \\
b &= V_{BA}(x_A - \mu_A) \\
c &= \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A).
\end{aligned}$$

Then, it follows that the expression for $p(x_B \mid x_A)$ can be rewritten as

$$\begin{aligned}
p(x_B \mid x_A) = \frac{1}{Z'} \exp\Bigg(- \bigg[&\frac{1}{2}\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big)^T V_{BB}\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big) \\
&+ \frac{1}{2}(x_A - \mu_A)^T V_{AA}(x_A - \mu_A) - \frac{1}{2}(x_A - \mu_A)^T V_{AB}V_{BB}^{-1}V_{BA}(x_A - \mu_A)\bigg]\Bigg)
\end{aligned}$$

Absorbing the portion of the exponent which does not depend on x_B into the normalization constant, we have

$$p(x_B \mid x_A) = \frac{1}{Z''} \exp\Bigg(-\frac{1}{2}\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big)^T V_{BB}\big(x_B - \mu_B + V_{BB}^{-1}V_{BA}(x_A - \mu_A)\big)\Bigg)$$

Arguing that resulting density is Gaussian Looking at the last form, $p(x_B \mid x_A)$ has the form of a Gaussian density with mean $\mu_B - V_{BB}^{-1}V_{BA}(x_A - \mu_A)$ and covariance matrix V_{BB}^{-1}. As before, recall our matrix identity,

$$\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} = \begin{bmatrix} (V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1} & -(V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1}V_{AB}V_{BB}^{-1} \\ -V_{BB}^{-1}V_{BA}(V_{AA} - V_{AB}V_{BB}^{-1}V_{BA})^{-1} & (V_{BB} - V_{BA}V_{AA}^{-1}V_{AB})^{-1} \end{bmatrix}.$$

From this, it follows that

$$\mu_{B|A} = \mu_B - V_{BB}^{-1}V_{BA}(x_A - \mu_A) = \mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A).$$

Conversely, we can also apply our matrix identity to obtain:

$$\begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} (\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})^{-1} & -(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})^{-1}\Sigma_{AB}\Sigma_{BB}^{-1} \\ -\Sigma_{BB}^{-1}\Sigma_{BA}(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})^{-1} & (\Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB})^{-1} \end{bmatrix},$$

from which it follows that

$$\Sigma_{B|A} = V_{BB}^{-1} = \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}.$$