Machine Learning Lecture Notes

Predrag Radivojac

January 13, 2015

1 Probability Theory

Probability theory can be seen as a branch of mathematics that deals with set functions. At the heart of probability theory is the concept of an *experiment*. An experiment can be the process of tossing a coin, rolling a die, checking the temperature tomorrow or figuring out the location of one's keys. When carried out, each experiment has an *outcome*, which is an element "drawn" from a set of predefined options, potentially infinite in size. The outcome of a roll of a die is a number between one and six; the temperature tomorrow is typically an integer or sometimes a real number; the outcome of the location of one's keys can be a discrete set of places such as a kitchen table, under a couch, in office etc. In many ways, the main goal of probabilistic modeling is to formulate a particular question or a hypothesis pertaining to the physical world as an experiment, collect the data, and then construct a model. Once a model is created, we can compute quantitative measures of sets of outcomes we are interested in and assess the confidence we should have in these measures.

1.1 Axioms of probability

We start by introducing the *axioms of probability*. Let the *sample space* (Ω) be a non-empty set of outcomes of the experiment and the *event space* ($\mathcal{F}$) be a non-empty set of subsets of Ω such that

1. $A \in \mathcal{F} \quad \Rightarrow \quad A^c \in \mathcal{F}$
2. $A_1, A_2, \ldots \in \mathcal{F} \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

where A and all A_i's are *events*, and A^c is the complement of A; i.e. $A^c = \Omega - A$. If both conditions hold, $\mathcal{F}$ is called a sigma field, or sigma algebra, and is a set of so-called measurable events. The tuple $(\Omega, \mathcal{F})$ is then called a *measurable space*.

It is important to emphasize that the definition of sigma field requires that $\mathcal{F}$ be closed under both finite and countably infinite number of basic set operations (union, intersection, and complementation; but not set difference). For example, through De Morgan's laws, i.e. $\cup A_i = (\cap A_i^c)^c$ and $\cap A_i = (\cup A_i^c)^c$, a sigma field

is also closed under intersection. Observe that all the above conditions imply that $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$.

Let now $(\Omega, \mathcal{F})$ be a measurable space. Any function $P : \mathcal{F} \to [0, 1]$ such that

1. $P(\Omega) = 1$
2. $\forall A, B \in \mathcal{F}$ and $A \cap B = \emptyset \quad \Rightarrow \quad P(A \cup B) = P(A) + P(B)$.

is called a *probability measure* or a *probability distribution*. The tuple $(\Omega, \mathcal{F}, P)$ is called the *probability space*.

The beauty of these axioms lies in their compactness and elegance. Many useful expressions can be derived from the axioms of probability. For example, it is nearly obvious that $P(\oslash) = 0$ or $P(A^c) = 1 - P(A)$. Another formula that is particularly important can be derived by considering a *partition* of the sample space; i.e. a set of k non-overlapping sets $\{B_i\}_{i=1}^k$ such that $\Omega = \cup_{i=1}^k B_i$. That is, if A is any set in Ω and if $\{B_i\}_{i=1}^k$ is a partition of Ω it follows that

$$\begin{aligned} P(A) &= P(A \cap \Omega) \\ &= P\left(A \cap \left(\cup_{i=1}^k B_i\right)\right) \\ &= P\left(\cup_{i=1}^k (A \cap B_i)\right) \\ &= \textstyle\sum_{i=1}^k P\left(A \cap B_i\right), \end{aligned} \tag{1}$$

where the last line followed from the axioms of probability. We will refer to this expression as the *sum rule*. Another important expression, shown here without a derivation, is that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

It is convenient to separately consider discrete (countable) and continuous (uncountable) sample spaces. A roll of a die draws numbers from a finite space $\Omega = \{1, 2, 3, 4, 5, 6\}$. For finite and other countable sample spaces, $\mathcal{F}$ is usually the power set $\mathcal{P}(\Omega)$. An example of the continuous sample space is the set of real numbers $\mathbb{R}$. As we shall see later, the situation with events is somewhat complicated here since there exist sets over which one cannot integrate. For that reason, $\mathcal{F}$ must be a proper subset of $\mathcal{P}(\Omega)$, i.e. $\mathcal{F} \subset \mathcal{P}(\Omega)$. Technically, sample spaces can also be mixed; e.g. $\Omega = [0, 1] \cup \{2\}$, but we will not consider such spaces here. Discrete and continuous sample spaces give rise to discrete and continuous probability distributions, respectively.

Owing to many constraints in defining the distribution function P, it is clear that it cannot be chosen arbitrarily. For example, if $\Omega = [0, 1]$ and $P([0, \frac{1}{2})) = \frac{1}{2}$, we cannot arbitrarily assign $P([\frac{1}{2}, 1]) = \frac{1}{3}$ because probabilities of complement sets must sum to one. It turns out, in practice it is easier to define P indirectly, by selecting a probability mass function or a probability density function. These functions are defined directly on the sample space where we have fewer restrictions to be concerned with compared to the event space. We address these two ways of defining probability distributions next.

1.2 Probability mass functions

Let Ω be a discrete (finite or countably infinite) sample space and $\mathcal{F} = \mathcal{P}(\Omega)$. A function $p : \Omega \to [0, 1]$ is called a *probability mass function* (pmf) if

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

The probability of any event $A \in \mathcal{F}$ is defined as

$$P(A) = \sum_{\omega \in A} p(\omega).$$

It is straightforward to verify that P satisfies the axioms of probability and, thus, is a probability distribution.

Example 1. Consider a roll of a fair six-sided die, i.e. $\Omega = \{1, 2, 3, 4, 5, 6\}$, and the event space $\mathcal{F} = \mathcal{P}(\Omega)$. What is the probability that the outcome is a number greater than 4?

First, because the die is fair, we know that $p(\omega) = \frac{1}{6}$ for $\forall \omega \in \Omega$. Now, let A be an event in $\mathcal{F}$ that the outcome is greater than 4, i.e. $A = \{5, 6\}$. Thus,

$$\begin{aligned} P(A) &= \sum_{\omega \in A} p(\omega) \\ &= \frac{1}{3}. \end{aligned}$$

It is important to note that P is defined on the elements of $\mathcal{F}$, whereas p is defined on the elements of Ω. That is, $P(\{1\}) = p(1)$, $P(\{2\}) = p(2)$, $P(\{1, 2\}) = p(1) + p(2)$, etc.

□

In discrete cases, $P(\{\omega\}) = p(\omega)$ for every $\omega \in \Omega$, and the probability of a set is always equal to the sum of probabilities of individual elements. We can define a discrete probability space by providing the tuple $(\Omega, \mathcal{F}, P)$, but it is often much simpler to define P indirectly by assuming that $\mathcal{F} = \mathcal{P}(\Omega)$ and providing a probability mass function p. In this case we say that the probability measure P is induced by a pmf p. In fact, we rarely define $(\Omega, \mathcal{F}, P)$ directly.

1.2.1 A few useful pmfs

Let us now look at some families of functions that are often used as discrete probability distributions. This is by no means a comprehensive review of the subject; we shall simply focus on a few basic concepts and will later introduce other distributions as needed. For simplicity, we will often refer to both pmfs and probability distributions they induce as distribution functions.

The *Bernoulli distribution* derives from the concept of a Bernoulli trial, an experiment that has two possible outcomes: success and failure. In a Bernoulli

trial, a success occurs with probability α and, thus, failure occurs with probability $1-\alpha$. A toss of a coin (heads/tails), a basketball game (win/loss), or a roll of a die (even/odd) can all be seen as Bernoulli trials. We model this distribution by setting a sample space to two elements and defining the probability of one of them as α. More specifically, $\Omega = \{S, F\}$ and

$$p(\omega) = \begin{cases} \alpha & \omega = S \\ 1 - \alpha & \omega = F \end{cases}$$

where $\alpha \in (0, 1)$ is a parameter. If $\Omega = \{0, 1\}$, we can compactly write the Bernoulli distribution as $p(k) = \alpha^k \cdot (1 - \alpha)^{1-k}$ for $k \in \Omega$. Observe that we replaced ω with k, which is a more common notation when the sample space is comprised of integers. To be precise, the Bernoulli distribution as presented above is actually a family of discrete probability distributions, one for each α. We shall refer to each such distribution as Bernoulli(α). However, we do not need to concern ourselves with semantics because the correct interpretation of a family vs. individual distributions will usually be clear from the context. The Bernoulli distribution is sometimes referred to as binary distribution.

The *Binomial distribution* is used to describe a sequence of n independent and identically distributed (i.i.d.) Bernoulli trials. At each value k in the sample space the distribution gives the probability that the success happened exactly k times out of n trials, where of course $0 \leq k \leq n$. More formally, given $\Omega = \{0, 1, \ldots, n\}$, for $\forall k \in \Omega$ the binomial pmf is defined as

$$p(k) = \binom{n}{k} \alpha^k (1 - \alpha)^{n-k},$$

where $\alpha \in (0, 1)$, as before, is the parameter indicating the probability of success in a single trial. Here, the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

enumerates all ways in which one can pick k elements from a list of n elements (e.g. there are 3 different ways in which one can pick 2 elements from a group of 3 elements). We will refer to a binomial distribution with parameters n and α as Binomial(n, α). The experiment leading to a binomial distribution can be generalized to a situation with more than two possible outcomes. This experiment results in a multidimensional probability mass function (one dimension per possible outcome) called the multinomial distribution.

The *Poisson distribution* can be derived as a limit of the binomial distribution as $n \to \infty$ with a fixed expected number of successes (λ). Here, $\Omega = \{0, 1, \ldots\}$ and for $\forall k \in \Omega$

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

where $\lambda \in (0, \infty)$ is a parameter (the relationship with the binomial distribution can be obtained by taking $\alpha = {}^{\lambda}/_{n}$). The Poisson distribution is often used to

model counts of events occurring sequentially and independently but with a fixed average (λ) in a particular time interval. Unlike the previous two distributions, Poisson(λ) is defined over an infinite sample space.

The geometric distribution is also used to model a sequence of independent Bernoulli trials with the probability of success α. At each point $k \in \Omega$, it gives the probability that the first success occurs exactly in the k-th trial. Here, $\Omega = \{1, 2, \ldots\}$ and for $\forall k \in \Omega$

$$p(k) = (1 - \alpha)^{k-1}\alpha,$$

where $\alpha \in (0, 1)$ is a parameter. The geometric distribution, Geometric(α), is defined over an infinite sample space, i.e. $\Omega = \mathbb{N}$.

Hypergeometric distribution. Consider a finite population of N elements of two types (e.g. success and failure), K of which are of one type (e.g. success). The experiment consists of drawing n elements, without replacement, from this population such that the elements remaining in the population are equiprobable in terms of being selected in the next draw. The probability of drawing k successes out of n trials can be described as

$$p(k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}},$$

where $0 \leq n \leq N$ and $k \leq n$. The hypergeometric distribution is intimately related to the binomial distribution where the elements are drawn with replacement ($\alpha = {}^{K}/_{N}$). There, the probability of drawing a success does not change in subsequent trials. We will refer to the hypergeometric distribution as Hypergeometric(n, N, K).

Uniform distribution: the uniform distribution for discrete sample spaces is defined over a finite set of outcomes each of which is equally likely to occur. Here we can set $\Omega = \{1, \ldots, n\}$; then for $\forall k \in \Omega$

$$p(k) = \frac{1}{n}.$$

The uniform distribution does not contain parameters; it is defined by the size of the sample space. We refer to this distribution as Uniform(n). We will see later that the uniform distribution can also be defined over finite intervals in continuous spaces.

All of the functions above satisfy the definition of a probability mass function, which we can verify by summing over all possible outcomes in the sample space. Four examples are shown in Figure 1.

1.3 Probability density functions

We shall see soon that the treatment of continuous probability spaces is analogous to that of discrete spaces, with probability density functions replacing

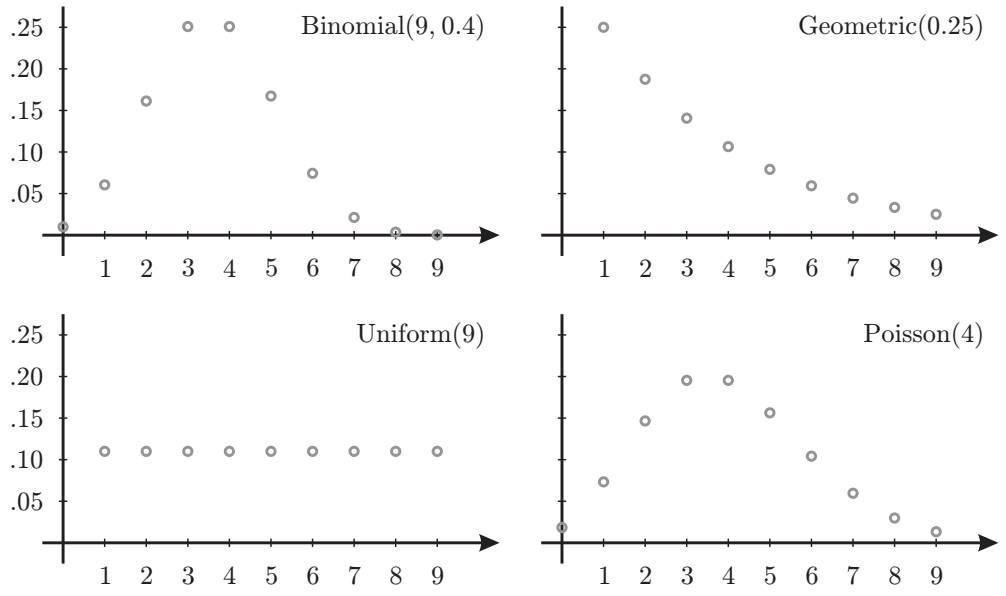

Figure 1: Four discrete probability mass functions.

probability mass functions and integrals replacing sums. Mathematically, however, there are fundamental differences between the two situations which we should keep in mind whenever working with continuous spaces. The main obstacle in generalizing the theory to uncountable sample spaces lies in addressing mathematical nuances involving infinitesimal calculus, the countable nature of a sigma field, and limitations of the definition of integrals. For example, there exist sets over which we cannot integrate and thus the set of events $\mathcal{F}$ cannot be the power set of any uncountable set (e.g. $\mathbb{R}$). It is therefore necessary to define an adequate event space which would be applicable to a vast majority of practically important situations.

To operate with continuous sample spaces we proceed by taking $\Omega = \mathbb{R}$ and defining the Borel field. The Borel field on $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is a set that contains all points in $\mathbb{R}$, all open, semi-open and closed intervals in $\mathbb{R}$, as well as sets that can be obtained by a countable number of basic set operations on them. By definition, $\mathcal{B}(\mathbb{R})$ is a sigma field; $\mathcal{B}(\mathbb{R})$ is an uncountably infinite set, but still smaller than $\mathcal{P}(\mathbb{R})$. The construction of subsets of $\mathbb{R}$ that are not in $\mathcal{B}(\mathbb{R})$ is difficult and only of theoretical importance (e.g., google Vitali sets), but nevertheless, the use of $\mathcal{P}(\mathbb{R})$ as the event space would lead to a flawed theory. Therefore, when discussing probability distributions over continuous sample spaces, we will usually take $\Omega = \mathbb{R}$ to be the sample space and implicitly consider $\mathcal{F} = \mathcal{B}(\mathbb{R})$ to be the event space.

Let now Ω be a continuous sample space and $\mathcal{F} = \mathcal{B}(\Omega)$. A function $p : \Omega \to [0, \infty)$ is called a *probability density function* (pdf) if

$$\int_{\Omega} p(\omega) d\omega = 1.$$

The probability of an event $A \in \mathcal{B}(\Omega)$ is defined as

$$P(A) = \int_A p(\omega) d\omega.$$

There are a few mathematical nuances associated with this definition. First, interestingly, the standard Riemann integration does not work for some sets in the Borel field (e.g. how would you integrate over the set of rational or irrational numbers within $[0, 1]$ for any pdf?). For that reason, probability density functions are formally defined using Lebesgue integration which allows us to integrate over all sets in $\mathcal{B}(\Omega)$. Luckily, Riemann integration, when possible, provides identical results as Lebesgue's integrals; thus, it will suffice to use Riemann integration in all situations of our interest.

Second, we mentioned before for pmfs that the probability of a singleton event $\{\omega\}$ is the value of the pmf at the sample point ω, i.e. $P(\{\omega\}) = p(\omega)$. In contrast, the value of a pdf at point ω is not a probability; it can be greater than 1. The probability at any single point, but also over any finite or countably infinite set is 0. One way to think about the probabilities in continuous spaces is to look at small intervals $A = [x, x + \Delta x]$ as

$$\begin{aligned} P(A) &= \int_x^{x+\Delta x} p(\omega) d\omega \\ &\approx p(x)\Delta x. \end{aligned}$$

Here, a potentially large value of the density function is compensated by the small interval Δx to result in a number between 0 and 1.

1.3.1 A few useful pdfs

Some important probability density functions are reviewed below. As before, the sample space will be defined for each distribution and the Borel field will be implicitly assumed as the event space.

Uniform distribution: the uniform distribution is defined by an equal value of a probability density function over a finite interval in $\mathbb{R}$. Thus, for $\Omega = [a, b]$ the uniform probability density function for $\forall \omega \in [a, b]$ is defined as

$$p(\omega) = \frac{1}{b - a}.$$

Note that one can also define Uniform(a, b) by taking $\Omega = \mathbb{R}$ and setting $p(\omega) = 0$ whenever ω is outside of $[a, b]$. This form is convenient because $\Omega = \mathbb{R}$ can then be used consistently for all one-dimensional probability distributions.

Exponential distribution: the exponential distribution is defined over a set of non-negative numbers, i.e. $\Omega = [0, \infty)$. Its probability density function is

$$p(\omega) = \lambda e^{-\lambda \omega},$$

where $\lambda > 0$ is a parameter. As before, the sample space can be extended to all real numbers, in which case we would set $p(\omega) = 0$ for $\omega < 0$.

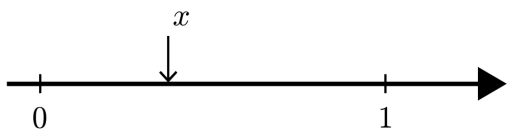

Figure 2: Selection of a random number (x) from the unit interval $[0, 1]$.

Gaussian distribution: the Gaussian or normal distribution is one of the most frequently used probability distributions. It is defined as

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

for any point $\omega \in \mathbb{R}$. The Gaussian distribution has two parameters, $\mu \in \mathbb{R}$ and $\sigma > 0$. We will refer to this distribution as Gaussian(μ, σ^2) or $N(\mu, \sigma^2)$. Both Gaussian and exponential distribution are members of a broader family of distributions called the exponential family. We will see a general definition of this family later.

Example 2. Consider selecting a number (x) between 0 and 1 uniformly randomly (Figure 2). What is the probability that the number is greater than $\frac{3}{4}$ or lower than $\frac{1}{4}$?

We know that $\Omega = [0, 1]$. We define an event of interest as $A = \left[0, \frac{1}{4}\right) \cup (\frac{3}{4}, 1]$ and calculate its probability as

$$\begin{aligned} P(A) &= \int_0^{1/4} d\omega + \int_{3/4}^{1} d\omega \\ &= \frac{1}{2}. \end{aligned}$$

Because the probability of any individual event in a continuous case is 0, there is no difference in integration if we consider open or closed intervals.

□

1.4 Multidimensional distributions

It is often convenient to think of the sample space as a multidimensional space. In the discrete case, one can think of the sample space Ω as a k-dimensional matrix. That is, $\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_k$, where Ω_i can be seen as the sample space along dimension i. Then, any function $p : \Omega_1 \times \Omega_2 \times \ldots \times \Omega_k \to [0, 1]$ is called a multidimensional probability mass function if

$$\sum_{\omega_1 \in \Omega_1} \cdots \sum_{\omega_k \in \Omega_k} p(\omega_1, \omega_2, \ldots, \omega_k) = 1.$$

One example of the multidimensional pmf is the multinomial distribution, which generalizes the binomial distribution to the case when the number of outcomes in any trial is a positive integer $k \geq 2$.

The multinomial distribution is used to model a sequence of n independent and identically distributed (i.i.d.) trials with k outcomes. At each point $(n_1, n_2, \ldots, n_k)$ in the sample space, the multinomial pmf provides the probability that the outcome 1 occurred n_1 times, outcome 2 occurred n_2 times, etc. Of course, $0 \leq n_i \leq n$ for $\forall i$ and $\sum_{i=1}^{k} n_i = n$. More formally, given the sample space $\Omega = \{0, 1, \ldots, n\}^k$, the multinomial pmf is defined as

$$p(n_1, n_2, \ldots, n_k) = \begin{cases} \binom{n}{n_1, n_2, \ldots, n_k} \alpha_1^{n_1} \alpha_2^{n_2} \ldots \alpha_k^{n_k} & n_1 + n_2 + \cdots + n_k = n \\ \\ 0 & \text{otherwise} \end{cases}$$

where α_i's are positive coefficients such that $\sum_{i=1}^{k} \alpha_i = 1$. That is, each coefficient α_i gives the probability of outcome i in any trial. The multinomial coefficient

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

generalizes the binomial coefficient by enumerating all ways in which one can distribute n balls into k boxes such that the first box contains n_1 balls, the second box n_2 balls, etc. An experiment consisting of n tosses of a fair six-sided die and counting the number of occurrences of each number can be described by a multinomial distribution. Clearly, in this case $\alpha_i = 1/6$, for each $i \in \{1, 2, 3, 4, 5, 6\}$.

In the continuous case, we can think of the sample space as the k-dimensional Euclidean space; i.e. $\Omega = \mathbb{R}^k$ and an event space as $\mathcal{F} = \mathcal{B}(\mathbb{R})^k$. Then, the k-dimensional probability density function can be defined as any function $p : \mathbb{R}^k \to [0, \infty)$ such that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\omega_1, \omega_2, \ldots, \omega_k) d\omega_1 \cdots d\omega_k = 1.$$

The multivariate Gaussian distribution is a generalization of the Gaussian or normal distribution to the k-dimensional case. It is defined as

$$p(\boldsymbol{\omega}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\boldsymbol{\omega} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu}) \right),$$

for any point $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_k) \in \mathbb{R}^k$. The multidimensional Gaussian distribution has two sets of parameters: $\boldsymbol{\mu} \in \mathbb{R}^k$ and a positive definite k-by-k matrix $\boldsymbol{\Sigma}$ (note that $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$). We will refer to this distribution as Gaussian$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1.5 Elementary conditional probabilities

Let $(\Omega, \mathcal{F}, P)$ be a probability space and B an event that already occurred. We are interested in the probability that event A also occurred, i.e. $P(A|B)$. The elementary conditional probability is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{2}$$

where $P(B) > 0$. From this expression, which is sometimes referred to as *product rule*, we can now derive two important formulas. The first one is *Bayes' rule*

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

The second formula, referred to as *chain rule*, applies to a collection of k events $\{A_i\}_{i=1}^{k}$ and can be derived by recursively applying the product rule. Then,

$$P(A_1 \cap A_2 \ldots \cap A_k) = P(A_1)P(A_2|A_1) \ldots P(A_k|A_1 \cap A_2 \ldots \cap A_{k-1}).$$

In some situations we refer to the probability $P(A)$ as *prior probability* because it quantifies the likelihood of occurrence of event A in absence of any other information or evidence. The probability $P(A|B)$ is referred to as *posterior probability* because it quantifies the likelihood about A in the presence of additional information (event B). The product rule from Eq. (2) has long history; it was derived by Abraham de Moivre in 1718.

One way to think about conditional probabilities is to consider that the experiment has already been conducted, but that we do not know the outcome yet. For example, a fair die has been rolled and we are interested in an event that the outcome was 4; i.e. $A = \{4\}$. The prior probability of event A is $P(A) = \frac{1}{6}$. But imagine that someone had observed the experiment and told us that the number was even ($B = \{2, 4, 6\}$). The probability after hearing this news becomes $P(A|B) = \frac{1}{3}$. Proper estimation of posterior probabilities from data is central to statistical inference.

1.6 Independence of events

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two events A and B from $\mathcal{F}$ are defined as *independent* if

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

or, alternatively, if $P(A|B) = P(A)$. More broadly, two or more events are (collectively) independent, if the probability of intersection of any group of events (of size two, three, etc.) can be expressed as the product of probabilities of individual events. For k events, there are $2^k - k - 1$ independence tests, one for each subset excluding the empty set and singletons.

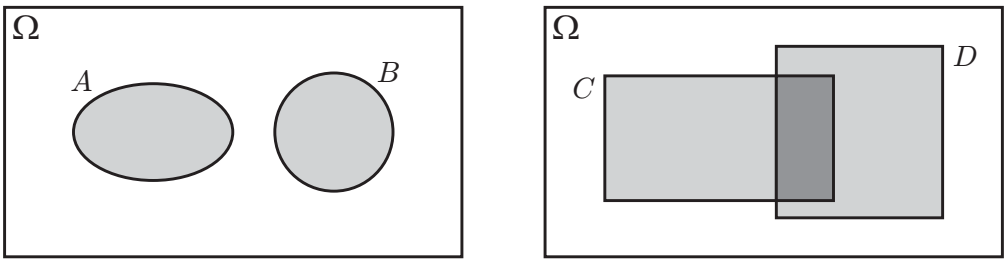

Figure 3: Visualization of dependent and independent events. Events A and B on the left are dependent because the occurrence of one excludes the occurrence of the other one. Events C and D on the right are independent. Each event occupies $1/4$ of the sample space Ω, while their intersection occupies $1/16$ of the sample space.

It is important to distinguish between mutually exclusive events and independent events. Mutually exclusive events are in fact never independent because the knowledge that the outcome of the experiment belongs to event A excludes the possibility that it is in B (Figure 3). It is often difficult, and quite non-intuitive, to simply look at events and conclude whether they are independent or not. One should (almost) always calculate $P(A \cap B)$ and $P(A) \cdot P(B)$ and numerically verify independence. Sometimes there may exist deep physical reasons why particular events are independent or assumed to be independent. In other occasions it may just be a numerical coincidence.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and A, B, and C some events from $\mathcal{F}$. Events A and B are defined as *conditionally independent* given C if

$$P(A \cap B|C) = P(A|C) \cdot P(B|C)$$

or, alternatively, if $P(A|B \cap C) = P(A|C)$. Independence between events does not imply conditional independence and, likewise, conditional independence between events does not imply their independence. We shall see an example later.

1.7 Interpretation of probability

There are two different interpretations of probability: a *frequentist* (*objectivist*) and a *subjectivist* one. In the frequentist view, to determine probabilities one needs repetitive experiments, e.g. Bernoulli trials. For example, in rolling a die 100 times, one could obtain a good approximation of $p(6)$ and even a better one if they rolled it $100,000$ times. As the number of experiments increases, the fraction of favorable outcomes converges to the probability of an event. Thus, probability is defined through a limit in an infinite series of trials.

However, consider the following question: What is the probability of war between England and France next year? They certainly have been in many wars before (e.g. the Hundred Years' War) so perhaps it is not zero. Such an event cannot be tested multiple times, as there is only one next year. Yet we may have

a pretty good sense of probability of war and can come up with a reasonable number (a subjectivist view). This number is our *degree of belief* or *conviction* derived from particular parameters of the situation. In the frequentist view, this probability would be undefined, or we would have to resort to *ad hoc* methods; for example, the probability of war is ϵ, where ϵ is the fraction of war days since formation of England and France (this is not a pure frequentist view). Even further, we could take probabilities of war between any two neighboring countries during their existence and apply this knowledge to England and France.

Interestingly, it can be argued that both frequentist and subjectivist interpretations of probability originated around the same time, perhaps in the same work. In *The Art of Conjecturing* Bernoulli proved the weak law of large numbers (undoubtably a frequentist's concept) but also introduced a subjective notion of probability as something that depends on a person's knowledge.

The good news is, the philosophical debate has almost no bearing on the use of probability theory in practice. No matter how probabilities are assigned, the mechanics of probabilistic manipulations are the same and valid, as long as the assignments adhere to the axioms of probability.