# Note Set 1: Review of Basic Concepts in Probability

Padhraic Smyth,
Department of Computer Science
University of California, Irvine
January 2024

This set of notes is intended as a brief refresher on probability. As a student reading these notes you will likely have seen (in other classes) most or all of the ideas discussed below. Nonetheless, even though these ideas are relatively straightforward, they are the key building blocks for more complex ideas in probabilistic learning. Thus, it is important to be familiar with these ideas in order to understand the material we will be discussing later in the course.

## 1 Discrete Random Variables and Distributions

Consider a variable $A$ that can take a finite set of values $\{a_1, \ldots, a_K\}$. We can define a **probability distribution** (also referred to as a probability mass function) for $A$ by specifying a set of numbers $\{P(A = a_1), \ldots, P(A = a_K)\}$, where $0 \leq P(A = a_k) \leq 1$ and where $\sum_{k=1}^{m} P(A = a_k) = 1$. We can think of $A = a_k$ as an **event** that either is or is not true, and $P(A = a_k)$ is the probability of this particular event being true. We can also think of $A = a_k$ as a **proposition** or **logical statement** about the world that is either true or false. Probability expresses our uncertainty about whether the proposition $A = a_k$ is true or is false.

Notational comment: for convenience, we follow a typical convention in probability notation and will often shorten $P(A = a_k)$ to just $P(a_k)$, and will also often use $a$ (and $P(a)$) to indicate a generic value for $A$, i.e., one of the possible $a_k$ values.

We call $A$ a *random variable* if the set of values $\{a_1, \ldots, a_K\}$ are **mutually exclusive and exhaustive**. *Mutually exclusive* means that the random variable $A$ can only take one value at a time—another way to state this is that the $P(A = a_j \text{ AND } A = a_k) = 0, \forall\, j, k, j \neq k$, i.e., the probability of the event occurring where $A$ takes two different values is 0 (i.e., it is deemed impossible)[1]. *Exhaustive* means that the variable $A$ always takes one of the values in the set $\{a_1, \ldots, a_K\}$, i.e., there are no other values it can take.

[1] the symbol $\forall$ means "for all"

**Example 1:** In medical diagnosis we often want to be able to predict whether a patient has a particular disease or not, given a set of measurements—this is a particular type of prediction problem known as **classification** that we will discuss later in the course in more detail. Let $Y$ be a variable representing what disease a particular patient has. Consider for example $Y$ taking values in $\{y_1, y_2\}$, where $y_1 =$ *has the flu* and $y_2 =$ *does not have the flu*. In this case $Y$ is a random variable since $y_1$ and $y_2$ are clearly mutually exclusive and exhaustive by definition. The randomness in "random variable" arises from the fact that in general we may be uncertain about the actual true value of $Y$ for any particular patient.

Now consider another set of events: $d_1 =$ *has the flu* and $d_2 =$ *has malaria* and $d_3$ = *healthy*. Could $\{d_1, d_2, d_3\}$ be used as the set of events to define a random variable? The answer is no since the set of events are not mutually exclusive: a person could have both the flu and malaria (this may have a very small probability of being true for any random patient, but this probability is not zero, i.e., it is possible). Nor are the events exhaustive: a person could be in none of these states since they could in general have some condition other than being healthy or having flu or malaria. One way to model this type of situation would be to model *has the flu* and *has malaria* as two different binary variables.

**Example 2:** Consider modeling the probability distribution of English words in a particular set of text documents. Let $W$ be a random variable representing a word and let $w_i$ represent a particular word. For example, we might want to know the probability that a particular word $w_i$ will come next in a sentence. Such probabilities are used in practice in a variety of applications such as language models, speech recognition, machine translation, etc. To treat $W$ as a random variable we need to define $W$ appropriately to ensure that the words being spoken or written are *mutually exclusive* and *exhaustive*. For example, for mutual exclusion, we could define $W$ to be the next word in a sequence of words, which by definition means there will always be a single unique "next word"—and we would need to include in our vocabulary a special symbol for the end of a sequence.
The requirement for exhaustivity is more difficult to satisfy. We could define the set of possible words $\{w_1, \ldots, w_K\}$ as the set of all words in an English dictionary—but which dictionary should we use? and what about words not in the dictionary such as regional variations of words, slang words, newly-emerging words, proper nouns such as "California," and so on? This is a problem commonly faced with text modeling, since it is impossible to know all future words that might occur in practice[a]. One way to get around this is to limit the set of words being modeled to (say) the $m = 20,000$ most frequent words in a training corpus and then an additional event $w_{K+1}$ is used to represent "all other words," ensuring that $\sum_k P(w_k) = 1$. An interesting problem in this context, that we will not discuss at this point, is how much probability mass we should assign to the event "all other words," or equivalently, how likely are we to see unseen words in the future?

[a] Sidenote: in the past few years, researchers working on large language models have come up with clever ways to deal with this: they use tokens as values for $W$, where tokens can include not only words but also single characters or common combinations of characters, allowing for any new unseen word to be treated as a combination of known tokens).

Before we finish with discrete random variables we note that not all discrete random variables necessarily take values from a finite set, but instead could take values from a countably infinite set. For example, we can define a random variable $A$ taking values from the set of positive integers $\{1, 2, 3, \ldots\}$. The sum $\sum_{i=1}^{\infty} P(A = i)$ must converge to 1 of course.

For random variables taking values in a countably infinite set it is impossible to explicitly represent a probability distribution directly by a table of numbers—and even in the finite case it may be often inconvenient or inefficient. In such cases we can use a *parametric model* for the distribution, i.e., a function that describes how $P(A = k)$ varies as a function of $k$ where the function includes one or more **parameters** of the model.

**Example 3:** An example of a parametric probability distribution defined on the positive integers is the geometric distribution,

$$P(A = k) \;=\; (1-\alpha)^{k-1}\,\alpha, \quad k = 1, 2, \ldots$$

where $\alpha$ is a parameter of the model and $0 < \alpha < 1$. This can be used for example to describe the distribution of the number of consecutive tail events that we will see before we see the first "heads event" in a coin-tossing experiment, where $\alpha$ is the probability of a head occurring on each coin toss.

## 2 Continuous Random Variables and Density Functions

The variables we discussed above such as $A$ and $Y$ take values that can be put in one-to-one correspondence with sets of integers, and are often referred to as **discrete random variables**. It is also useful to be able to build and use probability models for **real-valued or continuous random variables**, e.g., a random variable $X$ that can take values $x$ anywhere on the real line. (Sidenote: as is often done elsewhere in the literature we may occasionally use the term "distribution" rather than "density" when referring to a probability density function—the terminology can sometimes be somewhat loose, but the key distinction to keep in mind is whether a random variable is discrete or continuous.)

A probability density function $p(x)$ for a random variable $X$ taking values $x$ must satisfy two properties:

1. $p(x) \geq 0, \;\; \forall x$, i.e., $p(x)$ can never be negative;
2. $\int_{-\infty}^{\infty} p(x)dx = 1$, i.e., the function $p(x)$ must integrate to 1. (Generally speaking, whenever we have sums for discrete random variables we usually will have integrals for continuous random variables).

Any function $p(x)$ that satisfies these two properties is a valid probability density function. We can calculate the probability that a random variable $X$ has a value between two values $a$ and $b$ by integrating $p(x)$ between $a$ and $b$, i.e., $P(a \leq X \leq b) = \int_a^b p(x)dx$.

**Example 4:** The **uniform density** is defined as $p(x) = \frac{1}{b-a}$ for $a \leq x \leq b$ and 0 otherwise. We sometimes use the notation $X \sim U(a, b)$ to denote that the random variable $X$ has a particular type of density distribution (in this case the uniform density $U$). Here $a$ and $b$ are the **parameters** of the uniform density function. Note that while we had the option of describing a discrete probability distribution (for a finite set) with a list or table of probabilities, we can't define a continuous density function this way—instead we must parametrize the function and describe it via its functional form and its parameters.

Note also that although $p(x)$ integrates to 1, this does not imply that $p(x)$ itself needs to be less than 1 (unlike probability distributions). For example, for the uniform density if $a = 0$ and $b = 0.1$, then

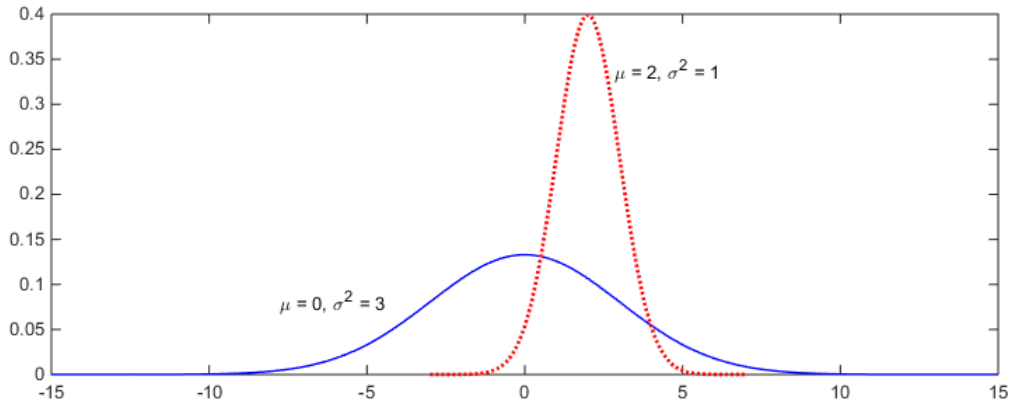


Figure 1: An example of two Gaussian density functions with different parameter values.

$p(x) = \frac{1}{0.1-0.0} = 10$ for $a \leq x \leq b$. The key point is that the **area** under the density function, $p(x)$, is constrained to be 1, but the height of the function can be any non-negative quantity.

**Example 5:** The well-known **Gaussian or Normal density** function is defined as:

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

Here the parameters of the model are the mean $\mu$ and the variance $\sigma^2$. We can also say $X \sim N(\mu, \sigma^2)$ for shorthand. Figure 1 shows an example of two Gaussians, each with different parameters—the dotted one is much narrower and concentrated than the solid one. We will examine the Gaussian density function in more detail later in this class.

# 3 Multiple Random Variables

For simplicity we will initially focus our discussion of multiple random variables on the case of discrete random variables and generalize to continuous random variables later on.

## 3.1 Conditional Probabilities

Modeling single random variables is a useful starting point but the real power in probabilistic modeling comes from modeling sets of random variables. Assume for the moment that we have two discrete variables $A$ and $B$. We define the **conditional probability** $P(A = a|B = b) = P(a|b)$ as the probability that $A$ takes value $a$ given that $B$ takes value $b$. The vertical bar "|" means that the probability of the proposition to the left of the bar (in this case $A = a$) is conditioned on knowing or assuming that the proposition to the right of the bar (in this case $B = b$) is true.

There are two obvious interpretations of what a conditional probability means. In the first case, $P(a|b)$ could be interpreted as "having measured (or observed that) $B = b$, then the conditional probability is the updated probability of $A = a$ now that we know $B = b$". The second case allows for hypothetical reasoning, i.e., $P(a|b)$ can be interpreted as "the probability of $A = a$ **if** hypothetically $B = b$." The value of the conditional probability $P(a|b)$ and the associated mathematics will be the same for both cases, but the semantic interpretation is a little different.

We can define a *conditional probability distribution* over all values for $A$, i.e., the list of probabilities $P(a_1|b), P(a_2|b), \ldots, P(a_K|b)$, for some value $b$ of the variable $B$. A key point is that a conditional probability distribution is just another probability distribution (for the variable that is the argument of the distribution function, here $A$), but where now the distribution of $A$ is conditioned on some event (here $B = b$). In particular, we have $\sum_{k=1}^{m} P(a_k|b) = 1$.

In many respects conditional probabilities are the key central concept in probabilistic information processing. A conditional probability distribution $P(A|b)$ gives us a quantitative way to represent how much information the event $B = b$ provides about $A$, i.e., how much does knowing that $B = b$ change our distribution $P(A)$ (which represents the uncertainty about the value of $A$ before we know that $B$ takes value $b$). For example, when a variable $B$ provides no information about another variable $A$, i.e., $p(A|B = b) = p(A),\ \forall b$, this statement is equivalent to saying that the two variables $A$ and $B$ are independent. Information and dependence are closely intertwined concepts in probability modeling and we will explore this further in later notes.

## 3.2 Joint Probability Models

We can define the joint probability $P(a, b)$, which is short for $P(A = a \text{ AND } B = b)$, to represent the probability that random variable $A$ takes value $a$ and random variable $B$ takes value $b$. Again note that the terms inside the parentheses, i.e, $a, b$ in $P(a, b)$, represent a logical statement about the world, namely that $A$ takes value $a$ and variable $B$ takes value $b$. We can think of the "joint variable" $AB$ as taking values from the Cartesian product of the sets of values of each of $A$ and $B$—it is sometimes conceptually useful to think of $AB$ as a "super-variable" defined as the combination of individual variables $A$ and $B$. By definition, since $A$ and $B$ are random variables, then values of the joint variable $AB$ must also be mutually exclusive and exhaustive: so, if $A$ can take one of $K$ values and $B$ can take one of $M$ values, then we must have that

$$\sum_{k=1}^{K} \sum_{m=1}^{M} P(a_k, b_m) = 1.$$

Note that there are $K \times M$ different possible combinations of the $a$ and $b$ values, and the joint probability table will contain $KM$ numbers, which could be very large for large values of $K$ and $M$. This hints at a combinatorial issue that arises when we build probability models involving multiple variables, namely that the number of entries in the joint distribution will grow exponentially fast with the number of variables in the model.

### 3.3 Relating Conditional and Joint Probabilities

From the basic axioms of probability one can show straightforwardly that the joint probability $P(a, b)$ is related to the conditional probability $P(a|b)$ and the probability $P(b)$ (often referred to as the **marginal probability** of $b$) in the following manner:

$$P(a, b) = P(a|b)P(b)$$

This is always true and is one of the basic rules of probability. There is a simple intuitive interpretation in words, namely, the probability that $A$ takes value $a$ and $B$ takes value $b$ can be decomposed into (1) the probability that $A$ takes value $a$ given that $B$ takes value $b$, (2) times the probability that $B = b$ is true. This argument makes intuitive sense (although its not a formal proof).

### 3.4 More than Two Variables

We can directly extend our definitions above beyond just 2 variables, to 3, 4, and indeed thousands of variables. As an example, say we have 4 random variables $A, B, C, D$.

- We can define the conditional probability $P(a|b, c, d)$, which is interpreted as the conditional probability that $A$ takes value $a$, given that we know (or assume) that $B = b$ and $C = c$ and $D = d$. (Note again that the comma "," notation is used as shorthand for the conjunction AND).
- We can define a joint distribution on all 4 variables, with joint probabilities $P(a, b, c, d)$, where the "joint variable" takes values from the Cartesian product of the individual value sets of each variable.
- Similarly we could define $P(a, b|c, d)$. This is the conditional probability of $A = a$ and $B = b$ given that $C = c$ and $D = d$. If $A$ has $M$ possible values and $B$ has $N$ possible values, then $P(A, B|c, d)$ is a conditional joint distribution for any fixed values of $c$ and $d$, defined as a table of $M \times N$ probabilities $P(A = a_m, B = b_n|c, d)$, with

$$\sum_{a,b} P(a, b|c, d) = \sum_{m} \sum_{n} P(A = a_m, B = b_n|c, d) = 1.$$

Probability models with multiple variables, such as $P(A, B, C, D)$ or $P(A, B|c, d)$ above, can be referred to as **joint probability models** or **multivariate probability models** (as opposed to **univariate probability models** for a single variable). As mentioned earlier, one issue with multivariate models is that the size of the table required to specify the joint distribution (for discrete random variables) grows exponentially. For example, if we have $K$ random variables each taking $M$ values, then the $K$-dimensional table for the joint distribution will contain $M^K$ entries (each of which is a joint probability).

## 4 Computing with Probabilities

We are often interested in computing the conditional probability of some proposition of interest (e.g., $A = a$), using a joint probability table (e.g., $P(a, b, c, d)$), and given some observed evidence (e.g., $B = b$).

### 4.1 The Law of Total Probability: Computing Marginals from Joint Probabilities

We can relate the probability distribution of any random variable $A$ to that of any other random variable $B$ as follows:

$$P(a) = \sum_b P(a, b) \; = \; \sum_b P(a|b)P(b)$$

This is known as the **law of total probability** and it tells us how to compute a marginal probability $P(a)$ from a joint distribution such as $P(a, b)$ or from a conditional and another marginal[2]. To motivate this idea, consider a case where we would like to know the marginal probability distribution $P(a)$, but don't have it directly: instead we only know the joint probabilities $P(a, b)$. So, in order to be able to get from the joint to the marginal, we make use of this joint distribution by introducing the variable $B$ into the problem and express $P(a)$ in terms of what we know about $A$ and $B$ together. This "summing out of the other variable," also known as **marginalization**, is very useful and we will use it in various contexts in this course.

We can of course generalize this idea. For example, if we have a table specifying the joint probabilities $P(a, b, c, d)$, then to get $P(a, b)$ we can use the law of total probability in the following general manner:

$$P(a, b) \; = \; \sum_{c,d} P(a, b, c, d) \; = \; \sum_{c,d} P(a, b|c, d)P(c, d).$$

(Make sure you can convince yourself that the equation above makes sense: it may for example help to think of $AB$ as a "super variable" that is conditioned on some value of another "super variable" $CD$.)

If we want to compute a conditional probability $P(a|b)$, from the joint distribution $P(a, b, c, d)$, we can do this as follows: compute the joint distribution $P(a, b)$ by summing out over $c, d$ as above (for every possible value of $a$, with $b$ fixed); then compute $P(b) = \sum_a P(a, b)$; and finally compute $P(a|b) = P(a, b)/P(b)$ for any value $a$ (here we used Bayes' rule, discussed further below).

Note that the types of marginalization operations described above (summing out of values of variables) are applicable to any set of random variables, i.e., it is not an assumption, but instead a basic property of random variables.

### 4.2 Bayes' Rule

We have seen already that we can express $P(a, b)$ as $P(a|b)P(b)$. Since there is nothing special about which argument $a$ or $b$ comes first in $P(a, b)$ then it is clear that we can also write $P(a, b) = P(b, a) = P(b|a)P(a)$. In fact if we equate these two different expressions for $P(a, b)$ we get

$$P(a|b)P(b) = P(b|a)P(a)$$

and if we divide each side by $P(b)$ we can derive **Bayes' rule**, i.e.,

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

[2]Note that the term "marginal" is generally used to refer to any unconditional distribution such as $P(a)$ or $P(b)$.

Furthermore, from the law of total probability we know that we can re-express the denominator to yield

$$P(a|b) = \frac{P(b|a)P(a)}{\sum_j P(b|a_j)P(a_j)}.$$

Bayes' rule expresses how we can "reason in reverse", i.e., given a forward model connecting $b$ to $a$, namely $P(b|a)$, and given a marginal distribution on $a$ (i.e., $P(a)$), we can make inferences about $a$ given $b$.

**Example 6:** Consider a medical diagnosis problem where a random variable $Y$ can take two values, 0 meaning a patient *does not have the disease* and 1 meaning a patient *has the disease*. The random variable $T$ represents the outcome of a test for the disease, where $T$ can take values $t = 0$ (negative) and $t = 1$ (positive). Assume we know (based on past medical data or prior experience) the following:

$$P(T = 1|Y = 0) = 0.01, \quad P(T = 1|Y = 1) = 0.9,$$

from which we can deduce that $P(T = 0|Y = 0) = 0.99$ and $P(T = 0|Y = 1) = 0.1$. In words this tells us that for healthy people ($Y = 0$) the test is negative 99% of the time—or equivalently there is a 1% chance of a false positive. And for people with the disease ($Y = 1$), the test is negative only 10% of the time. To use Bayes rule to calculate probabilities like $P(Y = 1|T = 1)$ we will also need to know what the value of the marginal $P(Y = 1)$ is. Suppose we know that $P(Y = 1) = 0.001$ (i.e., only 1 in a thousand people on average have the disease). Given this information we can now use Bayes rule to compute the conditional probability that a person has the disease given (a) that the test outcome is negative (has value 0), and (b) given that the test outcome is positive (has value 1). (Calculation of the actual conditional probabilities for this problem is left as an exercise).

### 4.3 Factorization and the Chain Rule

We have seen earlier that we can write $P(a, b)$ as $P(a|b)P(b)$. In fact one can do a similar decomposition of a joint distribution defined on any set of $K$ random variables, decomposing the joint into a product of conditional probabilities and a marginal, e.g.,

$$\begin{aligned} P(a, b, c, d) &= P(a|b, c, d)P(b, c, d) \\ &\qquad \text{treating } b, c, d \text{ as a conjunctive value from } B \times C \times D \\ &= P(a|b, c, d)P(b|c, d)P(c, d) \\ &\qquad \text{factorizing again, by repeating the same trick with } P(b, c, d) \\ &= P(a|b, c, d)P(b|c, d)P(c|d)P(d). \\ &\qquad \text{and now factorizing } P(c, d) \end{aligned}$$

Note that this works for any ordering of random variables, i.e., there is nothing special about the ordering $a, b, c, d$—we could have just as easily have decomposed as $P(a, b, c, d) = P(d|a, b, c)P(b|a, c)P(c|a)P(a)$

or using any other ordering. Also note that this works for any number of variables. There is no assumption being made here, this factoring always holds (as long as we are working with random variables). These types of factored representations can often be useful when we are working with a joint distribution and we wish to break it down into simpler factors that may be easier to manipulate and interpret.

# 5 Real-valued Variables

All of the properties and equations above extend in a natural manner to real-valued variables and density functions[3]: in general we can just replace our distributions $P(.)$ with density functions $p(.)$ to get equivalent concepts of conditional densities, joint densities, Bayes rule, marginalization, factorization, and so on. For example,

- **Conditional density functions:** Given two real-valued random variables $X$ and $Y$, we can define the conditional density $p(x|y)$, which can be interpreted as the density of the continuous random variable $X$ conditioned on a world where $Y = y$ is true. As with a conditional probability distribution, a conditional density function is itself a density function and obeys the laws of density functions, i.e., $p(x|y) \geq 0$ and $\int p(x|y)dx = 1$. We can also define conditional density functions of the form $p(x|a)$ where the conditioning is now on a discrete random variable taking a particular value, $A = a$.

- **Joint density functions:** Naturally, we can define joint probability density functions over multiple real-valued variables. For example $p(x, y)$ is a joint density function for two real-valued random variables $X$ and $Y$. Again, we can think of this as a "super-variable" taking values (a two-component vector) in the Cartesian product of the values of $X$ and the values of $Y$. If $X$ and $Y$ can each take values anywhere on the real line (i.e., $x \in \mathcal{R}, y \in \mathcal{R}$), then the Cartesian product of the two defines a two-dimensional plane (i.e., $(x, y) \in \mathcal{R}^2$). The density function $p(x, y)$ can then be thought of as a scalar-valued function (or surface, pointing into the 3rd dimension) that is defined over the 2d-plane, where $p(x, y) \geq 0$ and $\int_x \int_y p(x, y)dxdy = 1$. We could plot $p(x, y)$ as a contour plot or heatmap in two dimensions. We can extend this to any number of variables: with $K$ real-valued variables we have a joint density defined as a scalar function of $K$ real-valued variables (and, of course, although we can define such density functions for higher-dimensional sets of variables, we won't be able to plot or visualize them).

- Other properties such as the law of total probability, Bayes rule, and factorization, are extended to real-valued variables in the manner one might expect. For example, extending the law of total probability to probability densities we have $p(x) = \int p(x, y)dy$ with integration replacing summation. For factorization we can write $p(x, y, z) = p(x|y, z)p(y|z)p(z)$ for three real-valued random variables $X, Y, Z$, and so on.

[3] If this were a mathematics or statistics course we might be interested in rigorosly proving some of these statements.

# 6 Mixing Discrete and Continuous Random Variables

Up to this point we have been discussing sets of variables that are either all discrete-valued or real-valued. In practice of course we will frequently encounter mixed sets of random variables where some are discrete-valued and some are real-valued, e.g., in machine learning the inputs to a classification model might all be real-valued while the output (the class) is discrete. The general theory and principles that we discussed above can be extended in a natural way to mixed discrete-continuous situations, as long we are careful to interpret which parts of our expressions are distributions and which parts are densities.

To illustrate the ideas consider an example where $Y$ is a discrete random variable and $X$ is a real-valued random variable. For example, $Y$ might be binary, taking values either 0 or 1. Lets say that $Y = 0$ represents the event that a patient does not have the flu, and $Y = 1$ represents the event that a patient does have the flu. And lets say that $X$ represents this patient's real-valued body temperature (taken at some point in time). We can define various conditional and joint distributions and densities in this context:

- $P(y = 0|x)$ is a conditional probability, ranging between values 0 and 1 (since its a probability), where this conditional probability is a function of $x$, i.e., it is a function defined over the real-line (for some range of possible human body temperatures) where the function takes values between 0 and 1. And $P(y = 0|x) = 1 - P(y = 1|x)$ by definition. For example, in a simple model, for high-values of $x$ (high temperature) we might have that $P(y = 1|x)$ (the probability of flu) is high (close to 1), and the probability of flu could decrease monotonically as the temperature $x$ decreases. Much of machine learning is concerned with learning models for conditional probability functions of this type, where $P(y|x)$ is parametrized in some manner, ranging from simple models such as logistic models to more complex models such as neural networks.

- $p(x|y = 1)$ and $p(x|y = 2)$ are two **conditional density functions** (and these two densities are not necessarily related or coupled in any specific way). For example, $p(x|y = 1)$ might be a Gaussian model with a mean temperature of 102 degrees, conditioned on the patient having the flu ($y = 1$), and $p(x|y = 0)$ could be another Gaussian model with a mean temperature of 98 degrees, conditioned on the patient not having the flu ($y = 0$).

- We can define the marginal distribution of $p(x)$ as $p(x) = p(x|y = 0)P(y = 0) + p(x|y = 1)P(y = 1) = \alpha p(x|y = 0) + (1 - \alpha)p(x|y = 1)$, where $\alpha = P(y = 0)$. This is an example of a **finite mixture model**, i.e., the marginal distribution of the temperature $x$ can be expressed as a weighted combination of two conditional distributions (corresponding to the two subpopulations, people with the flu and people without the flu). We can also see that this way of writing $p(x)$ above (as a mixture) as just another application of the law of total probability, where one variable $Y$ is discrete and the other $X$ is real-valued.

- Bayes rule also works in this context, e.g.,

$$\begin{aligned} P(y=1|x) &= \frac{p(x|y=1)P(y=1)}{p(x|y=0)P(y=0)+p(x|y=1)P(y=1)} \\ &= \frac{p(x|y=1)P(y=1)}{p(x)} \end{aligned}$$

**Example 7:** In the example above with $Y$ and $x$, say we make an assumption that $p(x|y=1) = N(\mu_1, \sigma^2)$ and $p(x|y=0) = N(\mu_0, \sigma^2)$, i.e., that the conditional densities on patients' temperatures are both Gaussian, with different means and a common variance. Assume that the value of $P(y=1)$ is also known. With a little algebraic manipulation and the use of Bayes rule, one can show (homework problem) that

$$P(y=1|x) = \frac{1}{1+e^{-(\alpha_0+\alpha x)}}$$

where $\alpha_0$ and $\alpha$ are scalars that are functions of $P(y=1)$ and of the parameters of the two Gaussian models. This functional form for $P(y=1|x)$ is known as the **logistic model** and we will encounter it again in the future when we discuss classification models for machine learning.

# 7 Expectation

The **expected value** of a discrete-valued random variable is defined as

$$E[A] = \sum_{k=1}^{K} P(a_k)\, a_k.$$

This only makes sense of course if the values $a_k$ are numerical, e.g., the integers $1, 2, 3, \ldots$. On the other hand, if $A$ represents a variable such as *job category*, with possible values such as *engineer*, *teacher*, *cook*, etc., then taking the expected value of these categories is not meaningful. Such types of discrete random variables, with values that cannot meaningfully be ordered or mapped to numerical values, are referred to as **categorical variables**. To treat such variables in a numerical manner, in machine learning we often represent such variables with "one-hot" encodings, i.e., we can represent the value of a categorical variable that takes $K$ possible values by using a vector of binary indicator functions of length $K$, with a value of 1 for observed categorical value and a value of 0 for all the others.

For discrete random variables we can also define the expected value in a conditional manner, i.e., the conditional expectation:

$$E_{A|b}[A|b] = \sum_{k=1}^{K} P(a_k|b)\, a_k.$$

Note that the averaging here occurs with respect to the conditional distribution $P(a|b)$ rather than $P(a)$. Its often clear from the context what distribution the expectation is being taken with respect to, so its common to write expectations as just $E[A]$ (for example) rather than $E_A[A]$, etc, when its clear from the context what distribution is being used.

We can similarly define expectation and conditional expectation for real-valued random variables, e.g.,

$$E_X[X] \;=\; \int p(x)\; x\; dx \quad \text{and} \quad E_{X|y}[X|y] \;=\; \int p(x|y)\; x\; dx.$$

The expected value in an intuitive sense represents the "center of mass" of the density function, i.e., the point on the real line where it would be balanced if it consisted of actual mass. Note above that $Y$ could be a real-valued or a discrete-valued random variable: since we are conditioning on some particular value $Y = y$ it doesn't matter which (for the purposes of defining conditional probabilities, conditional expectations, etc).

A useful property to know about expectations is that they have the property of linearity, i.e.,

$$E[aX + b] \;=\; aE[X] + b$$

where $a$ and $b$ are arbitrary constants and where the random variable $X$ can be either real-valued or discrete. (Proof left as an exercise).

Finally, we can also take **expectations of functions of random variables**, $g(X)$, i.e., $E_X[g(X)] = \int_x g(x)p(x)dx$. Note that in general $E[g(X)] \neq g(E[X])$, i.e., the expectation of a function is not the same as the function of the expectation.

A well known example of a function $g(x)$ is the function defining the squared difference between $X$ and the expected value of $X$, i.e., $g(x) = (x - E[X])^2$. The expected value of this function $g(x)$ turns out to be the variance of $X$, i.e.,

$$Var(X) = E_X[g(x)] = E_X[(x - E_X[X])^2] = \int_x p(x)(x - E_X[X])^2 dx.$$

The variance measures how "spread out" a density function is (also known as a "scale parameter" in statistics). A useful identity is the following: $Var(X) = E_X[X^2] - (E_X[X])^2$.

# 8 The Semantic Interpretation of Probability

We conclude this section with a brief discussion on the semantic interpretation of probability. For simplicity consider a discrete random variable with some probability distribution such that event $a$ has probability $P(a)$. As we discussed earlier, we can think of $a$ as a logical statement about the world that is either true or false, and $P(a)$ expresses our uncertainty about $a$ being true (e.g., $a$ could represent the statement "it will rain in Irvine tomorrow.") But what exactly do we mean when we say for example that $P(a) = 0.3$? Of interest here is our interpretation of the number 0.3.

The term "probability" is taken to mean that the proposition $a$ is assigned a number between 0 and 1 representing our uncertainty about whether or not $a$ is true. We can agree that it makes sense to interpret $P(a) = 0$ as meaning that $a$ is logically false or impossible, and $P(a) = 1$ as stating that $a$ is logically true.

Numbers between 0 and 1 are a little more open to interpretation. The classical and traditional viewpoint, that most of us learned as undergraduates, is that $P(a)$ represents the relative frequency with which $a$ is true, in the infinite limit over a series of experiments or trials. This interpretation works well for problems like tossing a coin or throwing a pair of dice—we can imagine repeating such experiments until the observed frequency can be trusted as being a good estimate of the true probability of occurrence of a particular event, e.g., estimating the probability of "heads" for a particular coin by tossing it repeatedly and observing the outcome. This is known as the **frequentist interpretation of probability**.

However, there are other propositions $a$ for which it does not make sense conceptually to imagine an infinite series of experiments. For example, the event $a$ could be defined as the following proposition: *the US soccer team will win the World Cup within the next 20 years*. This is a proposition for which we can't easily imagine conducting repeated trials. Similarly imagine propositions such as $a =$ *life exists on other planets* or $a =$ *Alexander the Great played the harp*. We could come up with subjective estimates (our best guesses) of probabilities for any of these propositions, even though there is clearly no notion of repeated trials.

This way of thinking leads to the **subjective or Bayesian interpretation of probability**, which can perhaps best be summarized as thinking of $P(a)$ as the **degree of belief** that an agent (a human or an AI agent)) attaches to the likelihood that the proposition $a$ is true. Note that there is no notion of repeated trials required here: the probability is interpreted as a number that reflects an agent's degree of belief. More formally, degree of belief can be stated as a conditional probability $P(a|I)$ where $I$ is the background information available to the agent (although $I$ is often implicitly assumed rather than explicitly included in notation). So, for example, $I$ could be a person's model or assumptions or past data for a problem involving repeated coin tossing. Thus, our interpretation of probability from a Bayesian perspective could be thought to include the frequentist approach as a special case, in the sense that $I$ represents whatever assumptions or background knowledge we use to construct a frequentist model for $P$.

Other ways to think about Bayesian interpretation of probability statements is to attach them to an economic argument: an agent's degree of belief can be thought of as being related to the odds that the agent (e.g., you) would be willing to accept a bet that the proposition is true. Depending on the agent's internal model of the world and what information is available to the agent, the odds that the agent is willing to accept (and the corresponding degree of belief) might be quite different between different agents, depending on their world models and what information they have available to them.

While the Bayesian view of probability as a degree of belief might seem somewhat informal and imprecise, one can make the concept of degree of belief quite precise. In this course, as in much of machine learning research, we will tend to use the Bayesian degree of belief interpretation of probability since it covers a broader range of situations than the frequentist approach. This Bayesian interpretation of probability maps well to problems in artificial intelligence and machine learning: the probabilities computed by our AI agents (computer programs or robots) correspond to "subjective estimates" by the agent of how likely

it is that particular events are true, conditioned on the information available to the agent. It is natural that different AI agents may have access to different information or make different assumptions about the world, and that they corresponding degrees of belief will be different. For example, think of two different autonomous vehicles designed by two different auto companies, each inferring the probability of a pedestrian being present when presented with the same real-world scene.

One very important point, however, is that whether we use a frequentist or Bayesian interpretation, the rules and properties of probability are the same for both! i.e., we use the same equations and definitions. Only the semantic interpretation of the probabilities changes, not the mathematics of how we combine them and work with them.

# 9 Summary of Key Points

- The ideas above can be summarized in terms of a few relatively general principles: definition and properties of a random variable, distributions and densities, conditional and joint probabilities, Bayes rule, the law of total probability, factorization, expectation. These basic ideas are extremely powerful and well worth knowing: we will use them extensively in probabilistic learning.
- Note that all of the above principles hold *in general* for all distributions and density functions, irrespective of the functional form or numerical specification of the distributions and densities. Thus, we can reason with and manipulate distributions and densities at an abstract level, independently of the detailed specifications. Note also that we have made no assumptions above in stating the general principles and properties (apart from some specific assumptions in our illustrative examples)—so the properties stated above are quite general.
- The Bayesian interpretation of probability gives us more latitude in modeling of machine learning problems, compared to more traditional frequentist approaches.
- Two significant open problems remain that we have not yet discussed:
    1. Where do the numbers come from? the probabilities in the tables for discrete distributions and the parameters in the density functions? This is where probabilistic learning comes in—we can frequently learn these numbers (or parameters) from observed data.
    2. As we include more variables in our multivariate models, how will we avoid the problems of computation in high-dimensional spaces? e.g., for marginalization, how can we avoid computationally expensive high-dimensional integration for real-valued densities or the exponential increase in the number of terms in the sums for discrete distributions? We will see that there ways to impose *independence structure* on our probability models that allow us to simplify models in a systematic way.