Chapter 4

Uncertainty

Robots operate in the real world, and the real world is messy and noisy. It is critical that real-world robotics applications incorporate an understanding of **uncertainty**, whether to model the likely outcomes of an action, better estimate the state of the world, or make predictions about the future.

4.1 Nondeterminism and Stochasticity

We will begin by defining two key pieces of terminology: **nondeterminism** and **stochasticity**. These terms have slightly different meanings in a robotics context than you may be accustomed to from other fields.

Definition 4.1 A system is **nondeterministic** if we cannot predict what it will do in the future.

Definition 4.2 A system is **stochastic** or **probabilistic** if we have some notion of how it is likely to behave.

These definitions, which are drawn from section 9.2.2 of LaValle [4], reflect the significance of randomness to robotics applications. To inform a robot's decision-making capabilities, we typically would like to model the environment that the robot is in (including other agents in the environment like people), and thus we want to characterize these systems in terms of our ability to predict them. As an example, consider a robot attempting to navigate a crowded hallway without colliding with moving pedestrians. Without gathering statistics about the trajectories of the pedestrians, the behavior of a person in the hallway is *nondeterministic*; that is, we cannot predict where the person might go from one timestep to another.

In the absence of information about the behavior of the environment and the agents within it, one option is to make a decision assuming that the world is acting in the robot's worst interest. More specifically, we can apply a **minimax** solution that attempts to *mini*mize the cost of the robot's next action by assuming that the world is trying to drive that cost to its *max*imum. Although it is rather pessimistic, choosing actions based on this worst-case outlook should guarantee that the robot's interaction with the environment cannot have a higher cost than anticipated.

However, we know some things about the hallway environment. We know that humans walk in continuous paths, and we may even have some statistical knowledge of trajectories, destinations, or other relevant information to the problem at hand, which makes the hallway scenario *probabilistic* (or *stochastic*). We can use this information to build a more informed model than the worst-case assumption. For example, we can create a time-varying distribution of paths and choose our own trajectory based on a minimization of the projected likelihood of intersecting any of the paths of the pedestrians. If our information is statistically valid, then we can expect our robot to make better decisions given the current environment.

For more about nondeterministic and probabilistic models in the context of robot motion planning, please see section 9.2.2 of LaValle [4].

4.2 Probability Basics

A **probabilistic model** is a mathematical description of an uncertain situation, i.e. a situation in which several possible outcomes may occur. The process underlying the probabilistic model is called the **experiment** and will produce exactly one of the possible outcomes. Although the term "experiment" seems to suggest a controlled laboratory setting, it is used broadly to mean anything from flipping a coin to taking sensor readings. The set of all possible outcomes of the experiment is called the **sample space** and is denoted by $\mathcal{S}$. The smallest sample space is one that contains two possible outcomes; for example, the sample space for flipping a coin one time is $\mathcal{S} = \{H, T\}$, where H represents heads and T represents tails. If we flipped the coin two times, then the sample space would be all sequences of H's and T's of length two, i.e. $\mathcal{S} = \{HH, HT, TH, TT\}$. Note that in our formulation of a probabilistic model, there is only a single experiment, so flipping a coin twice constitutes one experiment rather than two separate experiments.

In our study of probability, we are interested in not only single outcomes but also collections of outcomes. We refer to the subset of a sample space as an **event**, which can either be **simple** if it contains exactly one outcome or **compound** if it contains more than one outcome. In the experiment where we flip a coin twice, an example of a simple event is "the event that heads comes up exactly twice," which we could write as $A = \{HH\}$, and an example of a compound event is "the event that tails comes up at least once," which we could write as $A = \{HT, TH, TT\}$.

Example 4.1 A laser rangefinder is a sensor that bounces a frequency-modulated laser off of objects in order to determine their distance from the robot. A typical laser rangefinder might return an integer distance in centimeters in the range $[10, 1000]$, whereas a return value of 0 would indicate that no return signal was detected. An experiment occurs each time the device emits laser light and tries to detect a return signal. Note that the physics of light travel mean that after emitting a laser signal, there is a small range of times during which a return signal could possibly be detected, so that the experiment has a definite start and end time. The value r of the return signal is an outcome. The sample space would be $\{0, 10, 11, 12, \ldots, 1000\}$. An important compound event occurs when $r \in \{10, 11, 12, \ldots, 1000\}$. In words, we would describe this event as "the robot detected an obstacle." This event carries significance because it may require that the robot perform obstacle avoidance to prevent a collision. Note that the opposite event – that the robot did not detect an obstacle – is a simple event since the single reading 0 is used to indicate no return signal. □

With these definitions in place, we can now state a probability law and three important axioms. Given an experiment with sample space $\mathcal{S}$, the probability law assigns to each event A a number $\mathrm{P}(A)$ called the probability of A. The probability of A quantifies the likelihood that A will occur and satisfies the following axioms:

1. **(Nonnegativity)** $\mathrm{P}(A) \geq 0$ for any event A.

2. **(Normalization)** $\mathrm{P}(\mathcal{S}) = 1$.

3. **(Additivity)** If A and B are two disjoint events (i.e. they have no outcomes in common), then the probability of their union satisfies

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).$$

In general, if $\mathcal{S}$ contains an infinite collection of disjoint events $A_1, A_2, \ldots,$ then the probability of their union satisfies

$$\mathrm{P}(A_1 \cup A_2 \cup \cdots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \cdots.$$

The nonnegativity axiom formalizes the intuitive notion that an event cannot have negative probability. The normalization axiom states that the probability of $\mathcal{S}$ must be equal to 1, the maximum probability that can be assigned to an event, since by definition $\mathcal{S}$ contains all possible outcomes and therefore must always occur when the experiment is performed. Finally, the additivity axiom states that if we have a collection of events which cannot occur simultaneously, then the probability of one of them occurring is the sum of the probabilities of the individual events.

What about the case where we have events that are *not* disjoint? Suppose for example that we are rolling a fair six-sided die, and let event A be the event that we roll a two. Since the die is fair, we have $\text{P}(A) = 1/6$. Now let event B be the event that we roll an even number. There are three outcomes contained within this compound event (i.e. we roll a two, a four, or a six), so the probability of B is $\text{P}(B) = 1/6 + 1/6 + 1/6 = 1/2$. Now let's ask a slightly different question about these two events. Suppose that we know that B occurred, i.e. we rolled an even number, and we want to ask the question "What is the probability that I rolled a two *given that* I rolled an even number?" To answer this question, we need a new probability law that accounts for available knowledge; that is, we need a law that specifies the **conditional probability of A given that B has occurred**, which we denote as $\text{P}(A \mid B)$ and define as

$$\text{P}(A \mid B) = \frac{\text{P}(A \cap B)}{\text{P}(B)}, \tag{4.1}$$

where we assume that the conditioning event B has probability greater than zero. Intuitively, this definition follows from the fact that $\text{P}(A \mid B)$ must be proportional to $\text{P}(A \cap B)$, the probability that both A and B occur. Since the set of possible outcomes now consists only of the outcomes in B rather than the entire sample space $\mathcal{S}$, we use the proportionality constant $1/\text{P}(B)$ to ensure that $\text{P}(B \mid B) = 1$ and to scale $\text{P}(A \mid B)$ accordingly. Thus, the conditional probability that we rolled a two given that we rolled an even number is

$$\text{P}(A \mid B) = \frac{1/6}{1/2} = \frac{1}{3}. \tag{4.2}$$

Note that it is important to know which event is the conditioning event. If we made A the conditioning event, then we would be computing $\text{P}(B \mid A)$, the conditional probability that we rolled an even number given that we rolled a two, and the result would be

$$\text{P}(B \mid A) = \frac{1/6}{1/6} = 1. \tag{4.3}$$

Thus, in general, $\text{P}(A \mid B) \neq \text{P}(B \mid A)$. Later in this chapter, we will introduce a crucial theorem called **Bayes' rule** that gives the relationship between $\text{P}(A \mid B)$ and $\text{P}(B \mid A)$, allowing us to compute one conditional probability given the other.

One final important note about conditional probabilities is that if events A and B are **independent**, then the occurrence of B has no effect on the probability of A. In other words, A and B are independent events if and only if $\text{P}(A \mid B) = \text{P}(A)$. It follows from (4.1) that if A and B are independent, then the probability that they both occur is

$$\text{P}(A \cap B) = \text{P}(A)\text{P}(B). \tag{4.4}$$

We adopt the relation in (4.4) as our definition of independence since it implies that independence is a symmetric property and can also be used when $\mathrm{P}(B) = 0$.

Example 4.2 Returning to the laser rangefinder from Example 4.1, let us consider the probability of every simple event. In the absence of any knowledge about the situation in which the laser rangefinder experiment was performed (robot location, presence of obstacles in the vicinity), we might assign an equal probability to every range reading as well as the "no obstacle detected" reading. Thus, if we let A be the event that $r_1 = 500$, then we would say $\mathrm{P}(A) = \frac{1}{992}$ because there are 992 different possible return values.

Having taken a reading and received the return value $r_1 = 500$ (event A), let B be the event that a second reading gives a return value $r_2 = 500$. There is a good chance that these readings occurred because the laser struck an obstacle at a distance of 500 cm. Since the two events are caused by a common mechanism, they are not independent. Whereas $\mathrm{P}(A) = \mathrm{P}(B) = \frac{1}{992}$, the conditional probability $\mathrm{P}(B|A)$ is much higher (nearly one) because laser rangefinder readings are very repeatable. □

4.3 Random Variables and Probability Distributions

Often times, we would like to define a rule that associates each possible outcome of an experiment with a numeric value. Such a rule of association is called a **random variable** and in mathematical terms is a **real-valued function of the experimental outcome**. We denote a random variable with an uppercase letter like X and use the notation $X(s) = x$ to mean that x is the numeric value associated with the outcome s by the random variable X. For example, for a coin flip, we might define

$$X = \begin{cases} 1 & \text{if the outcome is heads,} \\ 0 & \text{if the outcome is tails.} \end{cases} \tag{4.5}$$

Note that this kind of random variable — one where the only two possible values are 0 and 1 — is called a **Bernoulli random variable**. In practice, the Bernoulli random variable is often used to model probabilistic situations with two outcomes.

An important characteristic of a random variable is whether it is **discrete** or **continuous**. If a random variable is discrete, its possible values are either *finite* or *countably infinite*. The Bernoulli random variable, with its finite number of values, is an example of a discrete random variable. Continuous random variables, on the other hand, can take on an *uncountably infinite* number of values, such as all of the numbers in the interval $[0, 1]$. The distinction between discrete and continuous random variables is important when defining its **probability distribution**, which describes how the total probability of the sample space (recall that this must be 1 by the normalization axiom) is distributed among the possible values of the random variable. Probability distributions are the subject of the rest of this section.

4.3.1 Probability Mass Functions

Given a *discrete* random variable X, we define its probability distribution with a **probability mass function**, or a **pmf** for short. For each possible value x of X, the pmf assigns a **probability mass**, which we notate as follows:

$$p(x) \;=\; \text{the probability that } X \text{ will take on value } x \;=\; \mathrm{P}(X = x).$$

For example, suppose we flip a fair coin twice and define X to be the number of times the outcome is heads. Then the pmf of X is

$$p(x) = \begin{cases} 1/4 & \text{if } x \in \{0, 2\}, \\ 1/2 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

Notice that

$$\sum_x p(x) = 1, \tag{4.7}$$

where x ranges over all possible values of X. For any pmf, this property must hold by the normalization and additivity axioms of probability.

In a similar manner, we can use the pmf of X to determine the probability of compound events. Suppose for example that we would like to know the probability of the outcome being heads at least once. We compute this as

$$\mathrm{P}(X > 0) = \sum_{x=1}^{2} p(x) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}. \tag{4.8}$$

Thus, the probability of heads coming up either once or twice is $3/4$.

4.3.2 Probability Density Functions

Given a *continuous* random variable X, we define its probability distribution with a **probability density function**, or a **pdf** for short. The pdf of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$, we have

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\, dx, \tag{4.9}$$

which can be interpreted as the area under the graph of the pdf. Note that this means **no single value of a continuous random variable has positive probability**, i.e.

$$\mathrm{P}(X = a) = \int_a^a f(x)\,dx = 0. \tag{4.10}$$

Thus, when a random variable is continuous, we always talk about the probability of *intervals* of values rather than the probability of single values.

To be a legitimate pdf, the function $f(x)$ must be nonnegative, i.e. $f(x) \geq 0$ for every x, and must have the normalization property

$$\int_{-\infty}^{\infty} f(x)\,dx = 1, \tag{4.11}$$

i.e. the area under the entire graph of the pdf must be equal to 1.

4.3.3 Cumulative Distribution Functions

To describe both discrete random variables and continuous random variables with a single mathematical concept, we can use the **cumulative distribution function**, or **cdf** for short. The cdf is defined as follows:

$$F(x) = \mathrm{P}(X \leq x) = \begin{cases} \displaystyle\sum_{y \leq x} p(y) & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{x} f(y)\,dy & \text{if } X \text{ is continuous.} \end{cases} \tag{4.12}$$

An intuitive way to understand the cdf is that it is the accumulation of probability "up to" the value x. Any random variable, whether discrete or continuous, has a cdf since $\{X \leq x\}$ is always an event and consequently has a well-defined probability.

Notice that we can use the cdf as a means of obtaining the pmf or pdf of X. Suppose that X is discrete. In this case, we can obtain the probability mass for any value y by differencing, as

$$p(y) = \mathrm{P}(X \leq y) - \mathrm{P}(X \leq y - 1) = F(y) - F(y - 1). \tag{4.13}$$

In the case where X is continuous, we can obtain the pdf by differentiating, as

$$f(x) = \frac{dF}{dx}(x), \tag{4.14}$$

which is valid at every x at which the derivative of the cdf exists. Also note that we can compute the probability of an interval of values with the cdf, as

$$\mathrm{P}(a \leq X \leq b) = F(b) - F(a), \tag{4.15}$$

by the Fundamental Theorem of Calculus.

4.4 Statistical Moments

Although we can understand how a random variable behaves by studying the many individual numbers that constitute its probability distribution, it is frequently useful to summarize this information into a few representative numbers. These summary statistics are called **moments**, and we define several of them — expected value, variance, skewness, and kurtosis — in this section.

4.4.1 Expected Value

To motivate the definition of the expected value of a random variable, let us suppose that we are at a fair and are deciding whether to play a game that costs one ticket. We are told that we will double our tickets back with probability 0.2 and triple our tickets back with probability 0.1 (and we infer from this that we earn back no tickets with probability 0.7). Let X be the net gain or loss from playing one round of this game. The possible values of X are therefore 1, 2, and -1, and the pmf is

$$p(x) = \begin{cases} 0.7 & \text{if } x = -1, \\ 0.2 & \text{if } x = 1, \\ 0.1 & \text{if } x = 2, \\ 0.0 & \text{otherwise.} \end{cases} \tag{4.16}$$

In order to decide whether to play, we might consider what to “expect per round” if we played the game many times. Suppose that we play the game n times and that k_x is the number of times that the outcome is x. The net gain or loss in tickets averaged over the n rounds is then given by

$$\frac{-k_{-1} + k_1 + 2k_2}{n}. \tag{4.17}$$

It is reasonable for us to anticipate that if n is quite large, then the fraction of rounds where the outcome is x is approximately the probability of x, i.e.

$$\frac{k_x}{n} \approx p(x), \tag{4.18}$$

and thus the net gain or loss we “expect per round” is approximately

$$-p(-1) + p(1) + 2p(2) = -0.7 + 0.2 + 2(0.1) = -0.3. \tag{4.19}$$

We would therefore expect to lose about 0.3 tickets per round if we played this game a large number of times (which we clearly should not do). The intuitive principle behind our method for reaching this result is formalized by the **Law of Large**

Numbers, which states that as the number of times an experiment is performed approaches infinity, the average of the outcomes approaches the expected value.

Motivated by this example, we define the **expected value** (also known as the **expectation** or **mean**) of a random variable X to be

$$\mathrm{E}(X) = \begin{cases} \displaystyle\sum_x x \cdot p(x) & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{\infty} x \cdot f(x)\,dx & \text{if } X \text{ is continuous.} \end{cases} \tag{4.20}$$

Hence, $\mathrm{E}(X)$ is a probability-weighted average over all possible values of X, and by the Law of Large Numbers is the number that we will approach if we repeat the experiment associated with X many times and average the results. Another useful way to view $\mathrm{E}(X)$ is that it is the **center of mass** of the probability distribution.

Sometimes we would like to determine the expected value not of X itself but of a *function* of X. For instance, suppose that in our motivating example we instead wanted to determine how many tickets we could expect to win each round rather than the expected net gain or loss in tickets. Keeping the same definition of X, we could define a function $h(X) = X + 1$ and then compute the expected value of that function. It turns out that if $h(X)$ is any function of X (it does not necessarily need to be linear), then the expected value of the random variable $h(X)$ is given by

$$\mathrm{E}(h(X)) = \begin{cases} \displaystyle\sum_x h(x) \cdot p(x) & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{\infty} h(x) \cdot f(x)\,dx & \text{if } X \text{ is continuous;} \end{cases} \tag{4.21}$$

that is, we compute the expected value in the same manner as in (4.20) except that we replace x with $h(x)$. We will make use of this rule shortly.

4.4.2 Variance

Aside from the expected value, the most important moment of a random variable X is its **variance**, denoted as $\mathrm{Var}(X)$ or σ_X^2. Whereas the expected value describes where the probability distribution is centered, the variance measures how much the probability distribution is dispersed about the center. We can informally understand the variance $\mathrm{Var}(X)$ as a measure of how much we expect X to deviate from $\mathrm{E}(X)$. More formally, we define the variance of X as

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right]. \tag{4.22}$$

In other words, $\mathrm{Var}(X)$ is the expected squared deviation of X from the center of mass. A related measure of variability is the **standard deviation**, denoted as σ_X, which is simply the square root of the variance,

$$\sigma_X = \sqrt{\mathrm{Var}(X)}. \tag{4.23}$$

It is often easier to interpret the standard deviation of X as a measure of variability since it uses the same units as X rather than the square of the units. For example, if X is measured in meters, then the standard deviation σ_X is measured in meters, whereas the variance σ_X^2 is measured in meters squared. The standard deviation is thus often preferred to describe a distribution in a real-world setting, whereas the variance is typically more useful and conducive to work with mathematically.

From the definition of variance given in (4.22), computing the variance seems somewhat expensive since we need to compute the distribution of $(X - \mathrm{E}(X))^2$. Fortunately, we can reduce the number of arithmetic operations required by using the rule for the expected value of a function of a random variable given in (4.21). In particular, suppose that X is a random variable and that Y is a linear function of X; that is, $Y = aX + b$, where a and b are scalars. Then

$$\mathrm{E}(Y) = a\,\mathrm{E}(X) + b \qquad \text{and} \qquad \mathrm{Var}(Y) = a^2\,\mathrm{Var}(X). \tag{4.24}$$

We can use the first of these results and the definition of variance in (4.22) to derive

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] \\
&= \mathrm{E}\left[X^2 - 2\,\mathrm{E}(X)X + (\mathrm{E}(X))^2\right] \\
&= \mathrm{E}(X^2) - \mathrm{E}(2\,\mathrm{E}(X)X) + (\mathrm{E}(X))^2 \\
&= \mathrm{E}(X^2) - 2\,\mathrm{E}(X)\mathrm{E}(X) + (\mathrm{E}(X))^2 \\
&= \mathrm{E}(X^2) - 2(\mathrm{E}(X))^2 + (\mathrm{E}(X))^2 \\
&= \mathrm{E}(X^2) - (\mathrm{E}(X))^2,
\end{aligned} \tag{4.25}$$

a formula for computing the variance that is often less costly and more convenient.

4.4.3 Skewness

Another statistical moment that describes an aspect of the probability distribution about its center of mass is **skewness**. Intuitively, the skewness of a random variable X measures the degree to which its distribution departs from horizontal symmetry, as well as the direction of that departure. We define the skewness of X as

$$\mathrm{Skew}(X) = \frac{\mathrm{E}\left[(X - \mathrm{E}(X))^3\right]}{\left(\mathrm{E}\left[(X - \mathrm{E}(X))^2\right]\right)^{3/2}}. \tag{4.26}$$

Unlike expected value and variance, skewness does not have a unit, so it can be hard to interpret. Even so, we can glean an important property — the direction of the skew — just from looking at the sign of the skewness score:

- If the skewness is **positive**, then the distribution is positively (or right) skewed, meaning that **the right tail is longer than the left**.
- If the skewness is **negative**, then the distribution is negatively (or left) skewed, meaning that **the left tail is longer than the right**.

Note that the positivity or negativity of the skewness indicates which tail is longer, *not* which direction the distribution visually appears to be leaning. Also note that a skewness score of zero means that the distribution is perfectly symmetric about its center of mass. In practice, random variables very rarely have perfectly symmetric distributions, so it is useful to have heuristics for interpreting the skewness score. Bulmer [2] suggests this rule of thumb:

- If $|\text{Skew}(X)| > 1$, then the distribution is *highly skewed*.
- If $1/2 < |\text{Skew}(X)| \leq 1$, then the distribution is *moderately skewed*.
- If $0 < |\text{Skew}(X)| \leq 1/2$, then the distribution is *fairly symmetrical*.

Note that a normal (or Gaussian) distribution has a skewness of zero. Thus, a high skewness score indicates that a distribution is non-normal.

4.4.4 Kurtosis

The final statistical moment that we will introduce in this chapter is **kurtosis**, which is similar to skewness in that it describes the shape of the distribution, has no unit, and is somewhat hard to interpret. The mathematical definition is also the same as that of skewness except that the powers of three are replaced by powers of four:

$$\text{Kurt}(X) = \frac{\text{E}\left[(X - \text{E}(X))^4\right]}{(\text{E}\left[(X - \text{E}(X))^2\right])^2}. \tag{4.27}$$

What the kurtosis of a random variable X means in terms of distribution shape has been disputed, but the interpretation that we adopt in these notes is that the kurtosis is related to tail extremity. As a result, the kurtosis of X gives us information about the kind of outliers we can expect X to produce. In particular,

- a **higher** kurtosis indicates more **extreme outliers**, whereas
- a **lower** kurtosis indicates more **modestly-sized outliers**.

Note that the kurtosis of a normal distribution is 3. It is common for the kurtosis of a given distribution to be compared to this value, thus giving a sense of how the outliers of the given distribution compare to a normal distribution.

4.5 Multiple Random Variables

Up until now, our probabilistic models have only involved a single random variable, but we often would like to develop a model for the simultaneous behavior of several random variables. For example, suppose that we have two random variables X and Y that represent a robot's distance to each of two people walking down a hallway together. The trajectory of one pedestrian influences the trajectory of the other, and thus to make informed judgments about how X and Y will behave over time, we need to understand how X and Y relate to each other. Motivated by examples such as these, in this section we discuss probabilistic models involving multiple random variables simultaneously. We start by generalizing previous concepts from this chapter to cases where we have two random variables, and then at the end of the section we further generalize to more than two random variables.

4.5.1 Joint Probability Mass Functions

The pmf of a single discrete random variable X tells us how much probability mass is placed on each possible value x. Similarly, the **joint pmf** of two discrete random variables X and Y that are associated with the same experiment tells us how much probability mass is placed upon each possible pair of values (x, y). We define the joint pmf of X and Y as

$$p(x, y) = \mathrm{P}(X = x \text{ and } Y = y). \tag{4.28}$$

Hence, if we have a subset of the sample space $A \subseteq \mathcal{S}$ that is the set of all pairs (x, y) with a certain property, then the joint pmf satisfies

$$\mathrm{P}((X, Y) \in A) = \sum_{(x,y)\in A} p(x, y). \tag{4.29}$$

In addition, the joint pmf must satisfy

$$p(x, y) \geq 0 \qquad \text{and} \qquad \sum_{(x,y)\in\mathcal{S}} p(x, y) = 1. \tag{4.30}$$

To gain some intuition about joint distributions, let us consider an example. Suppose that the local bakery Wakeful Cookies is open from midnight to 3:00am and sells cookies in packs of six, twelve, eighteen, and twenty-four. For simplicity, assume that every customer purchases one pack of cookies. Let X be the number of cookies that the customer purchased, and let Y be the hour that the customer bought the cookies, where 0 means "between midnight and 12:59am," 100 means "between 1:00am and 1:59am," and 200 means "between 2:00am and 2:59am." Suppose that the joint pmf is given by the following table:

	$X = 6$	$X = 12$	$X = 18$	$X = 24$
$Y = 0$	0.07	0.10	0.09	0.14
$Y = 100$	0.04	0.14	0.07	0.10
$Y = 200$	0.04	0.11	0.04	0.06

Using this table, we can read off the probability of events involving both X and Y, e.g. the probability that a given customer of Wakeful Cookies bought 24 cookies between midnight and 12:59am is $p(24, 0) = 0.14$. We can also use this table to calculate the pmf of each variable individually. In the context of jointly-distributed random variables, the pmf of an individual variable is called the **marginal pmf** to distinguish it from the joint pmf. The marginal pmf of X at value x, denoted $p_X(x)$, is found by summing all the values in the corresponding column:

$p_X(6)$	$p_X(12)$	$p_X(18)$	$p_X(24)$
0.07	0.10	0.09	0.14
$+0.04$	$+0.14$	$+0.07$	$+0.10$
$+0.04$	$+0.11$	$+0.04$	$+0.06$
$= \mathbf{0.15}$	$= \mathbf{0.35}$	$= \mathbf{0.20}$	$= \mathbf{0.30}$

Likewise, the marginal pmf of Y at value y, denoted $p_Y(y)$, is found by summing all the values in the corresponding row:

$p_Y(0)$	$0.07 + 0.10 + 0.09 + 0.14 = \mathbf{0.40}$
$p_Y(100)$	$0.04 + 0.14 + 0.07 + 0.10 = \mathbf{0.35}$
$p_Y(200)$	$0.04 + 0.11 + 0.04 + 0.06 = \mathbf{0.25}$

Thus, in general, the marginal pmfs of two random variables X and Y associated with the same experiment are given by

$$p_X(x) = \sum_y p(x, y) \qquad \text{and} \qquad p_Y(y) = \sum_x p(x, y). \tag{4.31}$$

These formulas can be derived from (4.29), but it is helpful to remember that they correspond to column totals and row totals when the joint pmf is given as a table. In fact, the name "marginal pmf" comes from this idea, as one often writes column totals and row totals in the margins of a table.

4.5.2 Joint Probability Density Functions

When we have two continuous random variables X and Y that are associated with the same experiment, we cannot write a table for all of the possible pairs of values, but the interpretation is analogous to the discrete case. A **joint pdf** of two continuous random variables X and Y is a function $f(x, y)$ that satisfies

$$\mathrm{P}((X, Y) \in A) = \iint\limits_{(x,y)\in A} f(x, y)\, dx\, dy \tag{4.32}$$

for every subset A of the two-dimensional plane. In particular, if A is the two-dimensional rectangle $\{(x, y) \mid a \leq x \leq b,\ c \leq y \leq d\}$, then we have

$$\mathrm{P}(a \leq X \leq b,\ c \leq Y \leq d) = \int_c^d \int_a^b f(x, y)\, dx\, dy. \tag{4.33}$$

To be a valid joint pdf, $f(x, y)$ must also satisfy

$$f(x, y) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1. \tag{4.34}$$

Just as in the discrete case, we can use the joint pdf to compute the **marginal pdf** of each random variable. The marginal pdf of X, denoted $f_X(x)$, is given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \tag{4.35}$$

and the marginal pdf of Y, denoted $f_Y(y)$, is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx. \tag{4.36}$$

In other words, integrating the joint pdf with respect to y yields the marginal pdf of X, and integrating the joint pdf with respect to x yields the marginal pdf of Y.

4.5.3 Joint Cumulative Distribution Functions

If X and Y are two random variables associated with the same experiment, whether discrete or continuous, we define their **joint cdf** as

$$\begin{aligned} F(x,y) &= \mathrm{P}(X \le x, Y \le y) \\ &= \begin{cases} \displaystyle\sum_{\{(u,v) \mid u \le x, v \le y\}} p(u,v) & \text{if } X \text{ and } Y \text{ are discrete,} \\ \displaystyle\int_{-\infty}^{y} \int_{-\infty}^{x} f(u,v)\,du\,dv & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases} \end{aligned} \tag{4.37}$$

Although we do not consider it here, note that $F(x,y) = \mathrm{P}(X \le x, Y \le y)$ is also well-defined in the case where one of the random variables is discrete and the other is continuous, and thus any pair of random variables has a joint cdf.

As you might expect, we can obtain the **marginal cdf** of each variable from the joint cdf. To find the marginal cdf of X, we evaluate the joint cdf at $y = \infty$, as

$$F_X(x) = F(x, \infty), \tag{4.38}$$

and to find the marginal cdf of Y, we evaluate the joint cdf at $x = \infty$, as

$$F_Y(y) = F(\infty, y). \tag{4.39}$$

Just as in the case of a single random variable, the benefit of working with the cdf is that it is a single mathematical concept that applies equally well to both discrete and continuous random variables.

4.5.4 Conditional Distributions

Let X and Y be two random variables associated with the same experiment, and suppose we know that the value of Y is some particular y with $p_Y(y) > 0$. Having this knowledge about Y may influence what we know about X, knowledge which is captured by a *conditional distribution*. In the case where X and Y are discrete, the **conditional pmf of** X **given** Y is

$$p_{X|Y}(x \mid y) = \frac{\mathrm{P}(X = x \text{ and } Y = y)}{\mathrm{P}(Y = y)} = \frac{p(x,y)}{p_Y(y)}. \tag{4.40}$$

Note how this is analogous to (4.1), the conditional probability of an event given that another event occurred, as is the justification for this definition. To understand how $p_{X|Y}(x \mid y)$ relates to $p(x,y)$, suppose we fix some y with $p_Y(y) > 0$ and consider $p_{X|Y}(x \mid y)$ to be a function of x. Then this function of x is the slice $Y = y$ of the joint pmf scaled by $1/p_Y(y)$ to abide by the normalization axiom.

In the case where X and Y are continuous, the **conditional pdf of X given Y** is defined in a similar manner, as

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}. \tag{4.41}$$

Along the same lines as before, the conditional pdf of X given Y can be interpreted by visualizing the joint pdf along the slice $Y = y$ for a fixed y, where the slice is normalized with the factor $1/f_Y(y)$ so that it integrates to 1.

To better understand conditional distributions, let us return to Wakeful Cookies, the example from the beginning of this section. Suppose that a customer purchased a pack of cookies between 1:00am and 1:59am. Given this information, we would like to determine the probability that the customer bought a dozen cookies. In other words, we seek to calculate

$$p_{X|Y}(12 \mid 100) = \frac{p(12,100)}{p_Y(100)}. \tag{4.42}$$

If we take the slice $Y = 100$ of the table representing the joint pmf, we get

	$X = 6$	$X = 12$	$X = 18$	$X = 24$
$Y = 100$	0.04	0.14	0.07	0.10

Note that this slice of $p(x,y)$ is *not* a valid pmf on its own because the sum of all of the probability mass is less than 1. Thus, we need to divide the probability mass assigned to each value of X by $p_Y(100)$ in order to renormalize. We obtain $p_Y(100)$ by summing all of the mass in the slice, as

$$p_Y(100) = 0.04 + 0.14 + 0.07 + 0.10 = 0.35. \tag{4.43}$$

Now we can compute $p_{X|Y}(x \mid 100)$ at any value of X using (4.40) directly, as

$$p_{X|Y}(6 \mid 100) = \frac{0.04}{0.35} = \frac{4}{35},$$

$$p_{X|Y}(12 \mid 100) = \frac{0.14}{0.35} = \frac{2}{5},$$

$$p_{X|Y}(18 \mid 100) = \frac{0.07}{0.35} = \frac{1}{5},$$

$$p_{X|Y}(24 \mid 100) = \frac{0.10}{0.35} = \frac{2}{7}.$$

Thus, the probability that the customer purchased a dozen cookies given that she bought the cookies between 1:00am and 1:59am is $p_{X|Y}(12 \mid 100) = 2/5 = 0.4$.

4.5.5 Independence of Random Variables

Intuitively, if any two random variables X and Y are independent of one another, then knowing the value of one of these variables provides no knowledge about the other variable. More formally, if X and Y are independent random variables, then the events $\{X = x\}$ and $\{Y = y\}$ are independent for every x and y. One way to define independence is therefore

$$p_{X|Y}(x \mid y) = p_X(x) \quad \text{for all } y \text{ with } p_Y(y) > 0 \text{ and all } x. \tag{4.44}$$

However, just as we derived a more symmetric definition for the independence of two events in (4.4), we can do exactly the same here using (4.40), as

$$p(x, y) = p_X(x)\, p_Y(y) \quad \text{for all } x \text{ and } y. \tag{4.45}$$

Likewise, we say that two continuous random variables X and Y are independent if and only if

$$f(x, y) = f_X(x)\, f_Y(y) \quad \text{for all } x \text{ and } y. \tag{4.46}$$

To obtain a general definition of independence, we can use the joint cdf, i.e. we say that two random variables X and Y are independent if and only if

$$F(x, y) = F_X(x)\, F_Y(y) \quad \text{for all } x \text{ and } y. \tag{4.47}$$

Note that this holds even in the case where we have one discrete random variable and one continuous random variable.

4.5.6 Joint Expectation and Covariance

Suppose we have a function $h(X, Y)$ of two random variables X and Y associated with the same experiment. Then $h(X, Y)$ is itself a random variable, and therefore we can compute its expected value, as

$$\mathrm{E}(h(X,Y)) = \begin{cases} \displaystyle\sum_{(x,y)\in\mathcal{S}} h(x,y)\cdot p(x,y) & \text{if } X \text{ and } Y \text{ are discrete,} \\[2ex] \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\cdot f(x,y)\,dx\,dy & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases} \tag{4.48}$$

We can use this formula to determine another quantity that is frequently of interest: the strength and the direction of the relationship between X and Y. We call this measure the **covariance** of X and Y and define it as

$$\text{Cov}(X, Y) = \text{E}\left[(X - \text{E}(X))(Y - \text{E}(Y))\right]. \tag{4.49}$$

Thus, the covariance of X and Y is the expected product of the deviations of X and Y from their respective expected values. Note that

$$\text{Cov}(X, X) = E\left[(X - E(X))^2\right] = \text{Var}(X), \tag{4.50}$$

i.e. the covariance of a random variable with itself is equal to the variance of that random variable. To gain further insight into the definition of covariance and how to interpret it, consider the following three scenarios:

- Suppose X and Y have a strong **positive relationship**, meaning that

  - **small** x-values tend to occur with **small** y-values, and
  - **large** x-values tend to occur with **large** y-values.

  Then the variables tend to exhibit **similar behavior**, and as a result, the signs of $X - \text{E}(X)$ and $Y - \text{E}(Y)$ tend to be either both positive or both negative, resulting in a **positive covariance**.

- Suppose X and Y have a strong **negative relationship**, meaning that

  - **small** x-values tend to occur with **large** y-values, and
  - **large** x-values tend to occur with **small** y-values.

  Then the variables tend to exhibit **opposite behavior**, and as a result, the signs of $X - \text{E}(X)$ and $Y - \text{E}(Y)$ tend to be opposite of each other, resulting in a **negative covariance**.

- Suppose X and Y are not strongly correlated. Then positive and negative products tend to cancel each other out, resulting in a covariance near zero. In the case where $\text{Cov}(X, Y) = 0$, we say that X and Y are **uncorrelated**. Independent random variables are always uncorrelated, but the converse is not true, i.e. uncorrelated random variables are not necessarily independent.

Finally, to reduce the number of computations that we need to make to calculate the covariance, we can derive an alternate formula by expanding the product in the definition of covariance and taking the expected value of each individual term:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}\left[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right]\\ &= \mathrm{E}\left[XY - X\mathrm{E}(Y) - \mathrm{E}(X)Y + \mathrm{E}(X)\mathrm{E}(Y)\right]\\ &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) - \mathrm{E}(X)\mathrm{E}(Y) + \mathrm{E}(X)\mathrm{E}(Y)\\ &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).\end{aligned} \tag{4.51}$$

By using this formula, we replace all of the intermediate subtractions with a single subtraction at the end of the computation.

4.5.7 Random Vectors and Covariance Matrices

All of the concepts introduced in this section can be extended to account for cases with more than two random variables. For example, if we have three continuous random variables X, Y, Z that are associated with the same experiment, we define their joint pmf to be a function $f(x,y,z)$ satisfying

$$\mathrm{P}((X,Y,Z)\in A) = \iiint\limits_{(x,y,z)\in A} f(x,y)\,dx\,dy\,dz \tag{4.52}$$

as well as the usual nonnegativity and normalization constraints. It is often helpful to collect all of the random variables associated with the same experiment into a vector, which we call a **random vector**. Specifically, if we have n random variables $X_1, X_2, \ldots, X_n$, we define their corresponding random vector to be

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}. \tag{4.53}$$

Whether $\mathbf{X}$ consists of discrete random variables, continuous random variables, or a combination thereof, it has a well-defined joint cdf denoted as

$$F(\mathbf{x}) = \mathrm{P}(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n). \tag{4.54}$$

We can make other generalizations as well. For example, the expected value of $\mathbf{X}$ is a vector containing the expected values of the elements of $\mathbf{X}$; that is,

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_n) \end{bmatrix}. \tag{4.55}$$

The covariance of $\mathbf{X}$ is slightly more complicated because it must encode the relationship between each element of $\mathbf{X}$ with every other element of $\mathbf{X}$. It is defined analogously to the covariance of two random variables, as

$$\mathrm{Cov}(\mathbf{X}) = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\right]. \tag{4.56}$$

To better understand this definition, we can expand the vector notation to get

$$\mathrm{Cov}(\mathbf{X}) = \mathrm{E}\left[\begin{pmatrix} X_1 - \mathrm{E}(X_1) \\ \vdots \\ X_n - \mathrm{E}(X_n) \end{pmatrix} \begin{pmatrix} X_1 - \mathrm{E}(X_1) & \cdots & X_n - \mathrm{E}(X_n) \end{pmatrix}\right], \tag{4.57}$$

and then, letting $C_{i,j} = (X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))$, carry out the multiplication:

$$\mathrm{Cov}(\mathbf{X}) = \mathrm{E}\begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,n} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n,1} & C_{n,2} & \cdots & C_{n,n} \end{bmatrix}. \tag{4.58}$$

Akin to how we defined the expected value of a vector in (4.55), the expected value of a matrix M is a matrix that contains the expected values of the elements of M. Accordingly, we note that $\mathrm{E}(C_{i,j}) = \mathrm{Cov}(X_i, X_j)$ and rewrite (4.58) as

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}. \tag{4.59}$$

We call this matrix the **covariance matrix**, also known as the variance-covariance matrix since $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$ and thus the entries along the main diagonal are the variances of each element of $\mathbf{X}$. The covariance matrix is used extensively in robotics and a variety of other fields.

4.6 Bayes' Rule

In the preceding sections, we have seen how the notion of conditional probability allows us to take existing knowledge into account when determining the likelihood of an event. What we have not yet seen is how to relate conditional probabilities to each other; that is, given two events A and B associated with the same experiment, what is the relationship between $P(A \mid B)$ and $P(B \mid A)$? To motivate finding an answer to this question, suppose that a personal assistant robot is trying to *localize* itself (i.e. determine its own location) in a house. Using its vision sensors, the robot detects a spatula. A reasonable question to ask based on this discovery is

> "What is the probability that I'm in room r given that I saw a spatula?"

Having a probability of this form readily available for any arbitrary household item is infeasible, so it is unlikely that the robot has direct access to this probability. But if the robot has some knowledge about spatulas, then it might have direct access to a probability of the form

> "What is the probability of seeing a spatula given that I'm in room r?"

in which the conditioning order of the events is reversed. We can derive a relationship between these two conditional probabilities as follows. Let R be the event that the robot is in room r, and let S be the event that the robot sees a spatula. Recall that the conditional probability of R given that S has occurred is

$$\mathrm{P}(R \mid S) = \frac{\mathrm{P}(R \cap S)}{\mathrm{P}(S)}, \tag{4.60}$$

and therefore the probability that both R and S occur is

$$\mathrm{P}(R \cap S) = \mathrm{P}(R \mid S)\,\mathrm{P}(S). \tag{4.61}$$

Similarly,

$$\mathrm{P}(S \cap R) = \mathrm{P}(S \mid R)\,\mathrm{P}(R). \tag{4.62}$$

Since the intersection operation is commutative, i.e. $R \cap S = S \cap R$, Equations (4.61) and (4.62) are equal to each other, and we can rewrite (4.60) as

$$\mathrm{P}(R \mid S) = \frac{\mathrm{P}(S \mid R)\,\mathrm{P}(R)}{\mathrm{P}(S)}. \tag{4.63}$$

Equation (4.63) is the simplest form of a crucial theorem called **Bayes' rule**. Each component of Bayes' rule is given a name and meaning in the context of **Bayesian inference**, an approach to statistical inference in which we update our beliefs about the world based on our observations. These names and meanings are as follows:

- R is the **hypothesis** being considered,
- S is a new piece of **evidence** that may affect the probability of the hypothesis,
- $\mathrm{P}(R)$ is the **prior probability** (or just the **prior**), which is the probability of the hypothesis *before* observing the new evidence,
- $\mathrm{P}(R \mid S)$ is the **posterior probability** (or just the **posterior**), which is the probability of the hypothesis *after* observing the evidence,

- P$(S \mid R)$ is the **likelihood model**, which tells us how likely we are to see the evidence if the hypothesis is true, and
- P(S) is the **model evidence**, which tells us how likely we are to see the evidence in general.

To interpret the meaning of Bayes' rule, it is important to understand the effect that the prior probability and the model evidence have in scaling the likelihood model. We consider both below.

- To say anything about the likelihood of R occurring given that S occurred, we need to take into account the prior probability P(R). Without this factor, we could easily assign too high a value to the posterior probability. For example, suppose that room r is the kitchen, in which case the likelihood model would probably be quite high (i.e. the robot is likely to see a spatula given that it is in the kitchen). But now suppose that the kitchen is under renovation and thus is completely blocked off to the robot. Then P(R), the probability that the robot is in the kitchen, would be very low. If we did not scale down the likelihood model by multiplying it by the low-probability P(R), then the posterior probability would end up being too high.
- The normalization constant $1/$P(S) also plays an important role in scaling the likelihood model. Suppose that the robot's vision sensor is faulty and tends to detect spatulas in all rooms of the house. In this case, the model evidence P(S) would be very high, and knowing that S occurred would give us little to no information about the robot's location. Thus, multiplying the likelihood model by $1/$P(S) would scale the probability down accordingly. On the other hand, if seeing a spatula is a rare event, then the multiplication would scale the probability up accordingly.

Example 4.3 Returning once again to the laser rangefinder from Example 4.2, we ask whether there really is a wall located at a distance of $d = 500$ that the laser is bouncing off of. We have two readings $r_1 = 500$ (event A) and $r_2 = 500$ (event B) that each constitute evidence that there is actually a wall at $d = 500$ (event W). The event W cannot be directly observed, but it can be estimated from the occurrence of other events like A and B. So we might run tests and collect data on all possible robot positions and wall positions and determine that on average P$(W) = 0.0011$ — that is, even without evidence there will still sometimes be a wall at $d = 500$ just by chance. However, we want to know P$(W \mid A)$ — having gotten a return value consistent with event W, how does that change the chance that W is true? This is a hard question to answer directly, but using Bayes' rule we can transform it into concepts that we can more readily answer. We know P(W) and P(A), and we can get the number P$(A \mid W) = 0.9$ from the manufacturer — if there is a wall, how likely will the laser rangefinder return a reading consistent with that fact? From these numbers, we can estimate

$$\mathrm{P}(W \mid A) = \frac{\mathrm{P}(A \mid W)\,\mathrm{P}(W)}{\mathrm{P}(A)} = \frac{0.9 \cdot 0.0011}{\frac{1}{992}} = 0.982.$$

This technique can be used repeatedly to further refine the estimate of an event that cannot be directly observed. □

Although the Bayesian approach to statistical inference is quite powerful, it is not the only method of inference that we might want to use. The most prevalent alternate approach is called **frequentist inference** and involves taking repeated samples in order to estimate the probabilistic model that underlies an experiment. Rather than treating an unknown model as a random variable with a known prior probability distribution, the frequentist approach views the model as a deterministic quantity that happens to be unknown. Consequently, frequentist inference does not assign beliefs to events that cannot be measured since the event is either true or false with certainty (but it is unknown which). For example, the event W from Example 4.3 above could not have a probability of 0.982 using the frequentist approach since it cannot be directly observed. Rather, frequentist inference would assign P$(W) = 1$ since the two samples taken both indicated that W is true.

When might we choose to use frequentist inference over Bayesian inference? The most obvious instance is when we cannot obtain or do not account for any prior probability, as Bayesian inference requires the selection of a prior. Another reason that we may choose to use frequentist inference is when we require certain performance guarantees and Bayesian inference is too expensive. In many robotics applications though, Bayesian inference is the preferred approach since it allows us to use our beliefs to make estimations about unobservable random variables.

4.7 Bayes Filters

In the previous section, we discussed the concept of having a *belief* about a random variable that we cannot observe directly. We may for example have a belief about whether or not there is a wall in front of us based on our sensor readings, but we cannot directly observe whether or not there is a wall. Such a belief is represented by a conditional probability distribution that assigns a probability to each possible hypothesis about the true state of the random variable. We denote a belief as

$$bel(x_t) = p(x_t | z_{1:t}, u_{1:t}), \tag{4.64}$$

where x_t is the state of the random variable X at time t, $z_{1:t}$ is all measurements of the state up until time t, and $u_{1:t}$ is all control inputs up until time t.

In some cases, we may wish to formulate a belief about the state right after the control u_t is executed, before taking the measurement z_t. We denote this distribution as

$$\overline{bel}(x_t) = p(x_t | z_{1:t-1}, u_{1:t}) \tag{4.65}$$

and refer to the process of calculating it as **prediction** or as performing a **control update**. When we use $\overline{bel}(x_t)$ to compute $bel(x_t)$, the procedure is called **filtering**, a name that has its origin in signal processing in the sense of filtering out noise in order to estimate the underlying properties of a signal. The filtering computation is also referred to as making a **correction** or a **measurement update**.

To efficiently compute beliefs, we can use a **recursive estimation** algorithm in which we use the previous belief $bel(x_{t-1})$ to calculate the current belief $bel(x_t)$. The most general of these recursive estimation algorithms is called a **Bayes filter**; a single iteration of this algorithm is given in Algorithm 1.

Algorithm 1 bayes_filter($bel(x_{t-1}), u_t, z_t$)

```
1: for all $x_t$ do
2:     $\overline{bel}(x_t) = \int p(x_t \mid u_t, x_{t-1})\, bel(x_{t-1})\, dx_{t-1}$
3:     $bel(x_t) = \eta\, p(z_t \mid x_t)\, \overline{bel}(x_t)$
4: return $bel(x_t)$
```

The first step of Bayes filtering, called the *prediction step*, is given on line 2. In this step, $\overline{bel}(x_t)$ is calculated by integrating the product of two distributions: the probability that the control u_t will cause a transition from state x_{t-1} to state x_t, and the prior probability of state x_t. Once $\overline{bel}(x_t)$ has been obtained, the next step to take is the *correction step*, given on line 3. The correction step multiplies the probability of getting measurement z_t given state x_t, multiplied by $\overline{bel}(x_t)$. Since this product may not integrate to 1, it is normalized by the normalization constant η, which then leads to the desired belief $bel(x_t)$ that is returned in line 4.

Note that since the belief is computed recursively, there always needs to be an initial belief $bel(x_0)$ at time $t = 0$. If the value of x_0 is known, then $bel(x_0)$ should be initialized with a probability of 1 on that value and 0 elsewhere. If the value of x_0 is entirely unknown, then $bel(x_0)$ may be initialized with equal probability on all values (i.e. a uniform distribution). Partial knowledge about x_0 may also be incorporated when selecting the initial belief distribution.

Further Reading

For more on Bayes filters, please see Section 2.4 of Thrun et al. [5].

4.8 The Markov Assumption

All Bayesian filters make an important assumption when updating beliefs. Recall that we defined a belief as

$$bel(x_t) = p(x_t | z_{1:t}, u_{1:t}) \tag{4.66}$$

and note that this definition incorporates all measurements and all control inputs up until time t. However, you may have noticed that Bayesian filters do not compute beliefs using all of this information. Rather, a belief is computed as

$$bel(x_t) = p(x_t | z_t, u_t, bel(x_{t-1})). \tag{4.67}$$

In other words, when we use a Bayesian filter, we assume that the current state x_t depends on no variables prior to those at time t unless that dependence is mediated through the previous state x_{t-1}. Intuitively, this means we assume that the present state contains enough information to make the next state conditionally independent of all information from the past given the present state. We refer to this assumption as the **Markov assumption** and call a stochastic temporal process (that is, a set of random variables over time) satisfying this assumption a **Markov chain**.

Why would we want to make such an assumption? The key insight is that the time and space requirements for updating a belief must be held constant if a robot is to keep track of its current belief distribution. Without the Markov assumption, the cost of a belief update would increase with every timestep and eventually become *intractable*, meaning that too many resources would be required to make the update feasible. Making the Markov assumption allows us to avoid this expense since the number of variables that a belief depends on remains fixed over time.

The next important question to ask is whether making the Markov assumption is valid. In practice, there are many factors that give rise to violations of the Markov assumption. One such factor is having too few state variables to account for all the dynamics of the environment. For example, consider a robot trying to localize itself as it moves about the world, and suppose that the state of the robot is modeled by its position and velocity. Are these variables enough to maintain the Markov assumption? The answer is likely no: if the robot is battery-powered, for example, then battery depletion impacts the robot's change in velocity. In turn, battery level is impacted by all of the robot's previous control inputs, and as a result the Markov assumption is violated. To rectify this, we can include battery level in our set of state variables, but adding more variables makes updating beliefs more complex. Thus it is important to consider carefully the trade-offs when deciding which and how many state variables to model. As a general rule, it is best to try to define the model such that unmodeled state variables have random or near-random effects, in contrast to a state variable like battery level that has a systematic effect.

Let us consider a simple example process that satisfies the Markov assumption. Suppose that a robot DJ has a catalog of pop, rock, and folk music and continuously plays songs back-to-back. Since the robot has a limited amount of memory, it does not retain a history of the songs it has played and selects the next song based only on the song that is currently playing. If a pop song is currently playing, then the robot will select either a rock song or a folk song next with equal probability. If a rock song is currently playing, the robot will pick another rock song half of the time and the other half of the time pick a pop song or a folk song with equal probability. Finally, if a folk song is currently playing, then the robot will choose among the three genres with equal probability when selecting the next song.

It is easiest to express the transition probabilities from genre to genre described above in a matrix. Let g_i be the genre of the song that the robot is currently playing. Then the transition probabilities to genre g_{i+1} are given as follows:

		g_{i+1}		
		Pop	Rock	Folk
g_i	Pop	-	$\frac{1}{2}$	$\frac{1}{2}$
	Rock	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	Folk	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Since the robot only considers the current song when picking the next song, the Markov assumption holds, and the state of the genre over time is a Markov chain. We often visualize Markov chains as graphs where the nodes are the possible states and the weights on the edges are the transition probabilities. Such a visualization for the Markov chain describing the state of the genre is given in Fig. 4.1.

Now suppose that the robot considers both the current song and the song played previously when choosing the genre of the next song. Is it still possible to describe this process as a Markov chain? The answer is yes, but we must decide how much information to model. If we consider only the current genre as a state variable, then we cannot model the effect of the previous song since doing so would violate the Markov assumption. Thus, to both consider this effect and maintain the Markov assumption, we would need to include the genre of the previous song in our state as well as the genre of the current song. Doing this would in turn make our model more complicated: rather than having nine transition probabilities to contend with,

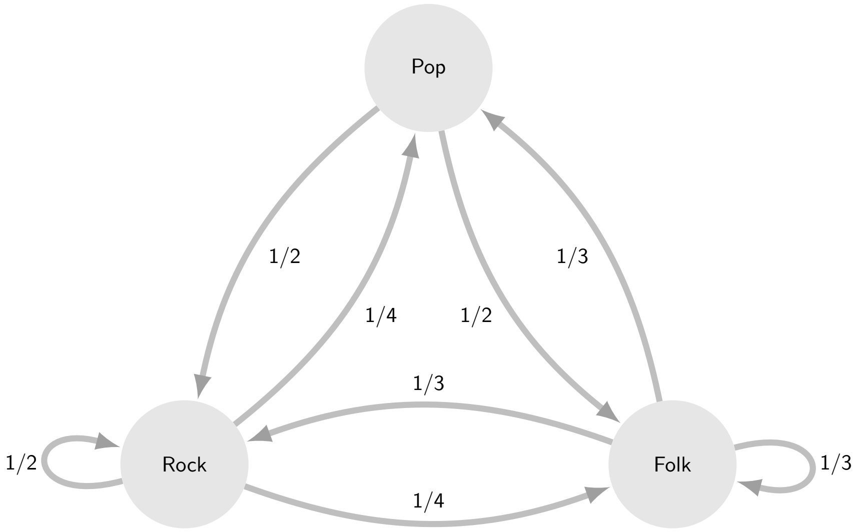

Figure 4.1: The Markov chain describing the song genre transition process.

we would have 27. We would therefore need to decide whether we wanted to deal with the added complexity of another state variable or whether we preferred to have a simpler but incomplete view of the genre transition process.

4.9 Entropy

In an uncertain world, it is useful to know how much information a robot can expect to receive upon making a particular observation. To quantify this amount, we define the **entropy** of a distribution $p(x)$ to be the expected amount of information that x carries. Intuitively, the more certain we are about x, the less information it carries. For example, if we let the random variable X be the outcome of flipping a biased coin that will always be heads with 100% certainty, then x carries no information: we already know for sure that x is heads without making an observation. If instead the coin is biased but only comes up heads with, say, 98% certainty, then x does not carry very much information since we are fairly certain that it is heads, but the entropy will be nonzero since there is a small amount of uncertainty. Finally, if the coin is fair, our entropy will be as high as possible for a random variable with two outcomes since both outcomes are equally likely. In this case, we are maximally uncertain about x, and thus it carries a higher quantity of information.

Given this analysis, we can see that outcomes with high probability correspond to low information content whereas outcomes with low probability correspond to high information content, motivating the mathematical definition of entropy

$$H_p(x) = \mathrm{E}[-\log_2 p(x)], \tag{4.68}$$

which in the discrete case resolves to

$$H_p(x) = -\sum_x p(x) \log_2 p(x). \tag{4.69}$$

and in the continuous case is the same except that the sum is replaced by an integral. The choice of base 2 for the logarithm means that entropy and information content is measured in **bits** (zeros and ones). To understand how the number of bits relates to the amount of information received, consider how many bits are required to transmit a message with eight possible outcomes. If all of the outcomes were equally likely, then we would need $\log_2(8) = 3$ bits to encode the eight outcomes: 000, 001, 010, 011, 100, 101, 110, 111. Our formula for entropy confirms this:

$$H_p(x) = -8 \left(\frac{\log_2 \frac{1}{8}}{8} \right) = 3. \tag{4.70}$$

But what if some outcomes are more likely than others? For example, suppose that the eight outcomes had probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}$ respectively. In this scenario, the entropy would be

$$H_p(x) = -\frac{\log_2 \frac{1}{2}}{2} - \frac{\log_2 \frac{1}{4}}{4} - \frac{\log_2 \frac{1}{8}}{8} - \frac{\log_2 \frac{1}{16}}{16} - \frac{4 \log_2 \frac{1}{64}}{64} = 2, \tag{4.71}$$

which means that on average, we should need only two bits to transmit a message. How can we encode eight outcomes with just two bits? We cannot do this exactly, but we can be clever in our encoding and choose to use fewer bits to represent likely outcomes at the expense of using more bits to represent unlikely outcomes, with the goal being to use fewer bits on average. For instance, we may choose the bit strings 0, 10, 110, 1110, 111100, 111101, 111110, 111111 to represent the eight outcomes with the respective probabilities given above. These messages have lengths 1, 2, 3, 4, 6, 6, 6, 6 respectively, and thus the average message length is

$$\frac{1}{2} + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{16}\right) + 6\left(\frac{4}{64}\right) = 2, \tag{4.72}$$

which is the same number of bits as the entropy. Note that we cannot make the length of the bit strings shorter since a sequence of messages, e.g. 11011100, must decode uniquely to the sequence of outcomes 110, 1110, 0.

Now let us return to our example of flipping a coin. In the case where the coin is completely biased to come up heads with 100% certainty, the entropy is

$$H_p(x) = -\log_2(1) = 0. \tag{4.73}$$

If the coin is biased to come up heads with probability 0.98, then the entropy is

$$H_p(x) = -0.98 \log_2 0.98 - 0.02 \log_2 0.02 \approx 0.14. \tag{4.74}$$

Lastly, if the coin is fair, then the entropy is

$$H_p(x) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1. \tag{4.75}$$

To compare the entropies of related distributions, we define **information gain** to be the expected reduction in entropy when we make an observation that changes our belief distribution. For instance, we may start out believing that a coin is fair, but then we may observe something that leads us to believe the coin is biased to come up heads with probability 0.98. We would then say that the information gain resulting from this observation is approximately $1 - 0.14 = 0.86$ bits. In robotics, we may take a minimum-entropy approach to making decisions, i.e. always select the action that we believe will result in the maximum information gain.

4.10 Further Reading

For more on applications of probability theory to robotics, see Thrun et al. [5]. For more discussion of probability and statistics in general, see Bertsekas and Tsitsiklis [1] and Devore [3].

Bibliography

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2nd edition, 2008.

[2] M. G. Bulmer. *Principles of Statistics*. Dover, 1979.

[3] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 8th edition, 2011.

[4] S. M. LaValle. *Planning Algorithms*. Cambridge University Press, 2006.

[5] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.