2

Preliminaries

The foundations of probabilistic graphical modeling lie centrally in *probability theory* and *graph theory*.[1] This chapter provides a non-exhaustive review of both. Sections 2.1 and 2.2 cover the basics of probability and random variables, including probability spaces, notions of variance and expectation, and joint, marginal, and conditional probability distributions. Section 2.3 provides an overview of basic graph theoretic concepts, such as directed and undirected graphs, their properties and common substructures, and special categories of graphs that are useful for inference and learning (e.g., tree-structured graphs).

A Note on Notation Moving forward, we will adhere to the following notation for clarity. A random variable will be denoted by a capital letter (e.g., X), while the specific values it takes on will be in lowercase (e.g., $X = x$). Sets and multivariate random variables (i.e., random vectors or sets of variables) will be denoted by boldface capitals (e.g., $\mathbf{X}$), with vector values in bold lowercase (e.g., $\mathbf{X} = \mathbf{x}$). Graphs will be denoted by calligraphic letters (e.g., $\mathcal{G}$). As the nodes in graph objects represent

[1] As for most machine learning topics, the reader may also benefit from a review of statistics, linear algebra, calculus, and information theory. See Cover (2006), Murphy (2022), and Strang (2023) for useful treatments of these topics.

random variables, these will generally be uppercase (e.g., $X \leftarrow Z \rightarrow Y$). Probability density functions associated with distributions will often be expressed with lowercase letters (e.g., $p(\mathbf{y} \mid \mathbf{x})$). Point estimates, such as maximum likelihood or maximum a posteriori (MAP) estimates, will be denoted using a hat (e.g., $\hat{\mathbf{x}}$). Model parameters will be denoted by Greek letters (e.g., θ), whether scalar or vector-valued.

2.1 Elements of Probability

The concept of probability has been explored through various lenses. In one sense, probability can be interpreted as a *frequency of occurrence*. For example, we might think of probability as the percentage of "successes" in a series of repeated trials that can succeed or fail. This is often referred to as the *frequentist* interpretation. In another sense, probability can be conceptualized as a *measure of uncertainty*: a degree of *subjective belief* or *reasonable expectation* informed by prior knowledge (Bertsekas and Tsitsiklis, 2008; Murphy, 2022). This is often referred to as the *Bayesian*[2] interpretation.

Each of these viewpoints can be useful, though some cases do not accommodate every interpetation. For example, if we wish to model the uncertainty of an event that does not have a long-run frequency (e.g., an event that can strictly occur once at most), the Bayesian interpretation of probability will be more natural than the frequentist perspective. Throughout this tutorial, we will draw from both the Bayesian and frequentist traditions. We will occasionally take a formal Bayesian stance, as in Section 6.5 on learning Bayesian models.

2.1.1 Set Theory

We begin with the basic elements of probability to establish the definition of probabilities on *sets*. In one view of probability theory, we can treat probability as a *measure of the size of a set* (Chan, 2021). We will briefly review the basic language of set theory, as this will crop up throughout this tutorial.

[2] In reference to the 18^{th} century mathematician Thomas Bayes and Bayes' rule (Definition 2.29), which we will discuss in this chapter.

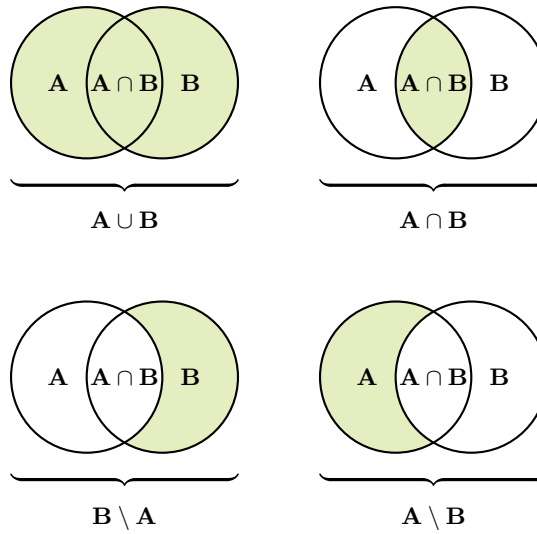


Figure 2.1: Set operations on two sets.

Definition 2.1 (Set)**.** A set is a collection of *objects* or *elements*, e.g.,

$$\mathbf{A} \coloneqq \{A_i\}_{i=1}^{n} = \{A_1, \ldots, A_n\}.$$

For each element A_i belonging to $\mathbf{A}$, we say that $A_i \in \mathbf{A}$. Note that elements themselves can be sets (yielding a set of sets). We can define operations on sets, such as intersection ($\cap$), union ($\cup$), and difference ($\setminus$) (Figure 2.1). For a more extensive review of set theory, see Kunen (1980) and Jech (2002).

2.1.2 Probability Spaces

The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a fundamental framework for expressing a random process. The sample space Ω is the set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment. The event space $\mathcal{F}$ is a subset of all possible sets of outcomes. It represents the collection of subsets of possible interest to us, where we denote elements of $\mathcal{F}$ as *events*. The mapping $\mathbb{P}$ assigns probabilities to each event $A \in \mathcal{F}$. Although each $A \in \mathcal{F}$ is itself a set,

we omit boldface here to align with standard probability notation and avoid clutter. For a probability space, $\mathcal{F}$ is furthermore a *σ-algebra*, and $\mathbb{P}$ is a *probability measure*, which we define as follows.

Definition 2.2 (σ-algebra)**.** Let 2^Ω denote the power set of Ω. We call $\mathcal{F} \subseteq 2^\Omega$ a σ-algebra if the following holds.

- $\Omega \in \mathcal{F}$.
- *Closed under complement.* If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$ (where $A^C \coloneqq \Omega \backslash A$ is the complement of A).
- *Closed under countable unions.* If $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \ldots, n$, then $\bigcup_{i=1}^n A_i \in \mathcal{F}$.

A pair $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is a σ-algebra is called a *measurable space*.

Definition 2.3 (Probability measure)**.** Given a measurable space $(\Omega, \mathcal{F})$, a measure μ is any set function $\mu : \mathcal{F} \to [0, \infty]$ that satisfies the following properties.

- $\mu(A) \geq \mu(\emptyset) \coloneqq 0$ for all $A \in \mathcal{F}$.
- $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$ for any countable collection of disjoint sets $A_i \in \mathcal{F}$.

When the above properties are satisfied and $\mu(\Omega) = 1$, we call μ a *probability measure* and denote it as $\mathbb{P}$.

A probability measure $\mathbb{P}$ boasts many useful properties. We enumerate some of these properties below.

Definition 2.4 (Axioms of probability, Kolmogorov 1933)**.** We define the following three axioms on probability measure $\mathbb{P}$.

1. *Nonnegativity.* For any event A, $0 \leq \mathbb{P}(A) \leq 1$.
2. *Normalization.* $\mathbb{P}(\Omega) = 1$.
3. *Countable additivity.* For pairwise disjoint sets $A_n \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Note that these axioms are consistent but not complete (Kolmogorov and Bharucha-Reid, 2018). Additionally, we remark the following useful properties.

Properties

- *Monotonicity.* $A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$.
- *Intersection.* $\mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$.
- *Union Bound.* $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.
- *Complement.* $\mathbb{P}(\Omega - A) = 1 - \mathbb{P}(A)$.
- *Law of Total Probability.* If $A_1, \ldots, A_k$ are a set of disjoint events such that $\bigcup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} \mathbb{P}(A_i) = 1$.

Example 2.1. Consider tossing a six-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. We can define different event spaces and probability measures on this sample space. For example, the simplest event space is the trivial event space $\mathcal{F} = \{\emptyset, \Omega\}$. Note that this $\mathcal{F}$ is a σ-algebra, as $\emptyset$ and Ω are complements of each other. The unique probability measure for this $\mathcal{F}$ satisfying the requirements above is given by $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$. Another event space is the set of all subsets of Ω. We can construct a valid probability measure for this $\mathcal{F}$ by assigning the probability of each set in the event space to be $\frac{i}{6}$ where i is the number of elements of that set; for example, $\mathbb{P}(\{1, 2, 3, 4\}) = \frac{4}{6}$ and $\mathbb{P}(\{1, 2, 3\}) = \frac{3}{6}$. Intuitively, this probability measure could correspond to the probability that a random fair die roll belongs to a given subset of Ω.

2.1.3 Independence of Events

Now that we understand the notion of an *event*, we can make statements on how events relate to each other. In particular, the concepts of *dependence* and *independence* are fundamental in probability and statistics.

Definition 2.5 (Independence of events)**.** Let A and B be events. A and B are independent (denoted $A \perp\!\!\!\perp B$) if and only if the following statements are true.

$$A \perp\!\!\!\perp B \iff \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$
$$A \perp\!\!\!\perp B \iff \mathbb{P}(A \mid B) = \mathbb{P}(A).$$

In words, A and B are independent if and only if their joint probability is equal to the product of the probabilities of each individual event. Alternatively, we can say that the conditional probability of A given that we known B is equal to the probability of A alone. Intuitively, if A and B are independent, then observing B does not have any effect on the probability of A.

Independence generalizes beyond the two-event case. A finite set of events is *pairwise independent* if every pair in the set is independent. A finite set of events is *mutually independent* if every event is independent of the intersection of any subset of events. Additionally, we have the fundamental notion of *conditional independence*: events A and B are conditionally independent if they are independent given another finite set of events (i.e., once we have already observed the other set of events, observing B does not have any effect on the probability of A). Later, we will see how the notion of conditional independence of random variables plays a pivotal role in graphical modeling and structure learning.

2.1.4 Conditional Probability

Conditional probability is a measure of the probability of an event given that another event has already occurred. Conditional probabilities show up frequently throughout probability theory and probabilistic graphical modeling, in such laws as Baye's rule (Definition 2.29) and the chain rule (Definitions 2.7, 2.28).

Definition 2.6 (Conditional probability of events)**.** Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

In words, $\mathbb{P}(A \mid B)$ is the probability measure of the event A after observing the occurrence of event B.

Definition 2.7 (Chain rule of probability)**.** Let $A_1, \ldots, A_k$ be events, $\mathbb{P}(A_i) > 0$. Then the chain rule states that

$$\begin{aligned}&\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_k) \\ &= \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_2 \cap A_1) \cdots \mathbb{P}(A_k|A_1 \cap A_2 \cap \cdots \cap A_{k-1}).\end{aligned}$$

Note that for $k = 2$ events, this is just the definition of conditional probability:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1).$$

In general, the chain rule is derived by applying the definition of conditional probability multiple times, as in the following example.

Example 2.2 (Chain rule applied to four events)**.** Given events A_1, A_2, A_3, A_4:

$$\begin{aligned}&\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) \\ &= \mathbb{P}(A_1 \cap A_2 \cap A_3)\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) \\ &= \mathbb{P}(A_1 \cap A_2)\mathbb{P}(A_3 \mid A_1 \cap A_2)\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) \\ &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2)\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3).\end{aligned}$$

2.2 Random Variables

Oftentimes, we do not care to know the probability of a particular event. Instead, we want to know probabilities over some function of these events. For example, consider an experiment in which we flip 10 coins. Here, the elements of the sample space Ω are length-10 sequences of heads and tails, and the event space $\mathcal{F}$ is all subsets of Ω. We observe sequences of coin flips; for instance, $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$. However, in practice we may not care to directly know the particular probability $\mathbb{P}(\omega_0)$ of a sequence or even the probability over a set of sequences in $\mathcal{F}$. Instead, we might want to know the number of coins that come up heads or the length of the longest run of tails. These quantities are functions of $\omega \in \Omega$, which we refer to as *random variables*.

More formally, define a mapping $X : \Omega \to E$ between two measurable spaces $(\Omega, \mathcal{F})$ and $(E, \mathcal{E})$, where $\mathcal{E}$ is a σ-algebra on E. Then, X is a

random variable if $X^{-1}(B) := \{\omega : X(\omega) \in B\} \in \mathcal{F}$ for all $B \in \mathcal{E}$. Intuitively, this means that every set B is associated with a set of outcomes that belongs to $\mathcal{F}$ and has a well-defined probability. Typically, we denote random variables using upper case letters $X(\omega)$ or more simply X (where the dependence on the random outcome ω is implied). We denote the value that a random variable may take on using lower case letters x. Thus, $X = x$ denotes the event that the random variable X takes on the value $x \in E$.

Example 2.3. In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω. Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values (0 through 10), so it is known as a discrete random variable. Here, the probability of the set associated with a random variable X taking on some specific value k is $\mathbb{P}(X = k) := \mathbb{P}(\{\omega : X(\omega) = k\}) = \mathbb{P}(\omega \in$ all sequences with k heads). Note that the set of all sequences with k heads is an element of $\mathcal{F}$, given that $\mathcal{F}$ consists of all subsets of Ω.

Example 2.4. Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay (ω for this example could be some underlying characterization of the particle that changes as it decays). In this case, $X(\omega)$ takes on an infinite number of possible values, so it is called a continuous random variable. We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as $\mathbb{P}(a \leq X \leq b) := \mathbb{P}(\{\omega : a \leq X(\omega) \leq b\})$.

When describing the event that a random variable takes on a certain value, we often use the *indicator function* $\mathbf{1}\{A\}$ which takes value 1 when event A happens and 0 otherwise. For example, for a random variable X,

$$\mathbf{1}\{X > 3\} = \begin{cases} 1, & \text{if } X > 3 \\ 0, & \text{otherwise} \end{cases}$$

In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions from which the probability measure governing an experiment immediately follows. In the following three sections, we describe these

functions: the cumulative distribution function (CDF), the probability mass function (PMF) for discrete random variables, and the probability density function (PDF) for continuous random variables. For the rest of this section, we suppose that X takes on real values, i.e., $E = \mathbb{R}$.

2.2.1 Cumulative Distribution Functions

Definition 2.8 (Cumulative distribution function (CDF))**.** A CDF is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as

$$F_X(x) \coloneqq \mathbb{P}(X \leq x).$$

By using this function, one can calculate the probability that X takes on a value between any two real constants a and b (where $a < b$).

Properties

- $0 \leq F_X(x) \leq 1$. This follows from the definition of the probability measure.
- $\lim_{x \to -\infty} F_X(x) = 0$. As x approaches $-\infty$, the corresponding set of ω where $X(\Omega) \leq x$ approaches $\emptyset$, for which $\mathbb{P}(\emptyset) = 0$.
- $\lim_{x \to +\infty} F_X(x) = 1$. As x approaches ∞, the corresponding set of ω where $X(\omega) \leq x$ approaches Ω, for which $\mathbb{P}(\Omega) = 1$.
- $x \leq y \implies F_X(x) \leq F_X(y)$. This follows from the fact that the event that $X \leq x$ is a subset of $X \leq y$ for $x \leq y$.

2.2.2 Probability Mass Functions

Suppose a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable). A simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume. Let $\mathrm{Val}(X)$ refer to the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $\mathrm{Val}(X) =$

$\{0, 1, 2, \ldots, 10\}$. We can then define a probability mass function with respect to X.

Definition 2.9 (Probability mass function (PMF))**.** A PMF is a function $p_X : \text{Val}(X) \to [0, 1]$ such that $p_X(x) = \mathbb{P}(X = x)$.

Properties

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in A} p_X(x) = \mathbb{P}(X \in A)$, as probability measures apply over countable unions of disjoint sets.
- $\sum_{x \in \text{Val}(X)} p_X(x) = 1$. Applying the previous property, we have that $\sum_{x \in \text{Val}(X)} p_X(x) = \mathbb{P}(X \in \text{Val}(X)) = \mathbb{P}(\Omega) = 1$.

2.2.3 Probability Density Functions

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

Definition 2.10 (Probability density function (PDF))**.**

$$f_X(x) \coloneqq \frac{dF_X(x)}{dx}.$$

Note that the PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere). According to the properties of differentiation, for very small δx,

$$\mathbb{P}(x \leq X \leq x + \delta x) \approx f_X(x)\delta x.$$

Both CDFs and PDFs (when they exist) can be used for calculating the probabilities of different events. But it should be emphasized that the value of PDF at any given point x is not the probability of that event, i.e., $f_X(x) \neq \mathbb{P}(X = x)$. Because X can take on infinitely many values, it holds that $\mathbb{P}(X = x) = 0$. On the other hand, $f_X(x)$ can take on values larger than one (but the integral of $f_X(x)$ over any subset of $\mathbb{R}$ will be at most one).

Properties

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

2.2.4 Expectation

The *expectation* or *expected value* of a random variable tells us something about what outcome to expect in the average case. In the discrete setting, the expectation is a generalization of the weighted average: it is the arithmetic mean of the possible outcomes of a random variable weighted by the probabilities of observing these outcomes. Given this arithmetic operation, the expectation can take on a value that itself is never observed in reality. In the continuous case, summation is replaced by integration. We can take expected values of functions of random variables, or of random variables themselves.

Definition 2.11 (Expected value)**.** Let X denote a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ denote an arbitrary function. In this case, $g(X)$ can be considered a random variable, with an associated *expectation* or *expected value*. We define the expected value of $g(X)$ as

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

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

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

Intuitively, the expectation of $g(X)$ can be thought of as a weighted average of the values that $g(x)$ can take on, where the weights are given by $p_X(x)$ or $f_X(x)$ which add up to 1 over all x. As a special case of the above, note that the expectation, $\mathbb{E}[X]$ of a random variable itself is found by letting $g(x) = x$; this is also known as the mean of the random variable X.

Properties

- $\mathbb{E}[a] = a$ for any constant $a \in \mathbb{R}$.
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$ for any constant $a \in \mathbb{R}$.
- *Linearity of Expectation.*

$$\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)].$$

- For a discrete random variable X,

$$\mathbb{E}[\mathbf{1}\{X = k\}] = \mathbb{P}(X = k).$$

2.2.5 Variance

Definition 2.12 (Variance)**.** The variance of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined

$$\mathrm{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

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

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

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

Properties

- $\mathrm{Var}[a] = 0$ for any constant $a \in \mathbb{R}$.

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

Example 2.5. Calculate the mean and the variance of the uniform random variable X with PDF $f_X(x) = 1, \forall x \in [0, 1], 0$ elsewhere.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x dx = \frac{1}{2}$$
$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}$$
$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Example 2.6. Suppose that $g(x) = \mathbf{1}\{x \in A\}$ for some subset $A \subseteq \Omega$. What is $\mathbb{E}[g(X)]$?

Discrete case.

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

Continuous case.

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

2.2.6 Common Random Variables

Here, we define several useful random variables that are commonly encountered in probabilistic modeling. See Ross (2010) and Murphy (2012) for additional examples and elaboration on their properties.

Discrete Random Variables

Definition 2.13 (Bernoulli random variables)**.** The Bernoulli distribution models the outcome of an experiment (or *Bernoulli trial*) with potential values 0 and 1, which can be construed as $1 =$ *success* and $0 =$ *failure*. A random variable X is said to be Bernoulli if its PMF $p_X(x)$ is given

	Distribution		Mean	Variance
Discrete	Bernoulli	$\text{Ber}(p)$	p	$p(1-p)$
	Binomial	$\text{Bin}(n,p)$	np	$np(1-p)$
	Geometric	$\text{Geo}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
	Poisson	$\text{Poi}(\lambda)$	λ	λ
Continuous	Uniform	$\text{U}(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	Gaussian	$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
	Exponential	$\text{Expon}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Table 2.1: Means and variances for common probability distributions.

by the following form.

$$X \sim \text{Ber}(p), \quad 0 \leq p \leq 1$$
$$p_X(x) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$$

For example, a Bernoulli random variable can represent the outcome of a coin flip ($H = 1, T = 0$) that comes up heads with probability p.

Definition 2.14 (Binomial random variables)**.** The binomial distribution models the number of successes in n successive independent Bernoulli trials. A random variable X is said to be binomial if its PMF $p_X(x)$ takes the following form.

$$X \sim \text{Bin}(n,p), \quad 0 \leq p \leq 1$$
$$p_X(x) = \binom{n}{x} \cdot p^x (1-p)^{n-x}.$$

For example, a binomial random variable can represent the number of heads in n independent flips of a coin with heads probability p. The Bernoulli distribution can also be viewed as a binomial distribution with $n = 1$.

Definition 2.15 (Geometric random variables)**.** Imagine that successive independent Bernoulli trials are performed until we reach the first success, where each trial has a probability p of success. A geometric random variable X represents the number of trials required until the first success, with PMF $p_X(x)$ expressed as follows.

$$X \sim \text{Geo}(p), \quad 0 \leq p \leq 1$$
$$p_X(x) = p(1-p)^{x-1}.$$

Continuing with the coin flip example, a geometric random variable can represent the number of flips of a coin until the first heads, for a coin that comes up heads with probability p.

Definition 2.16 (Poisson random variables)**.** The Poisson distribution is a probability distribution over the nonnegative integers $\{0, 1, 2, \dots\}$ that models the frequency of rare events. A Poisson random variable with PMF $p_X(x)$ is given by

$$X \sim \text{Poi}(\lambda), \quad \lambda > 0$$
$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

where x is the number of times the event of interest occurs, e is Euler's number, and parameter λ is the expected number of events. The term $e^{-\lambda}$ is a normalization constant that guarantees that the distribution sums to 1.

Continuous Random Variables

Definition 2.17 (Uniform random variables)**.** A uniformly distributed random variable with PDF $f_X(x)$ assigns equal probability density to every value between a and b on the real line.

$$X \sim \text{U}(a, b), \quad a < b$$
$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

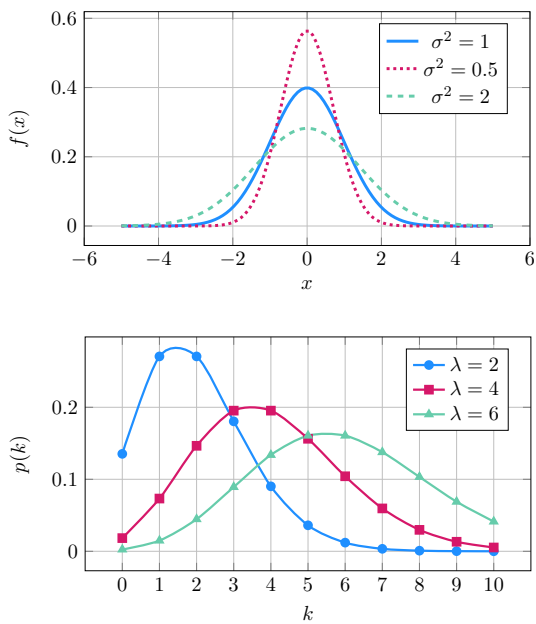


Figure 2.2: Gaussian PDF (top) and Poisson PMF (bottom) with varying parameters. Note that the Poisson PMF is defined over integer values, and smoothed curves are visual guides only.

Definition 2.18 (Gaussian random variables)**.** The Gaussian or *normal* distribution is parameterized by mean μ and variance σ^2. Its PDF $f_X(x)$ takes the form of a bell-shaped curve that is symmetric around μ.

$$X \sim \mathcal{N}(\mu, \sigma^2)$$
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The properties of univariate and multivariate Gaussian distributions offer many mathematical conveniences. Historically, Gaussian distributions have been commonly assumed to model continuous random variables of unknown form in the natural and social sciences.

Definition 2.19 (Exponential random variables)**.** The exponential distribution models the distance (or time) between events in a Poisson process with a constant average rate of change λ, wherein events occur

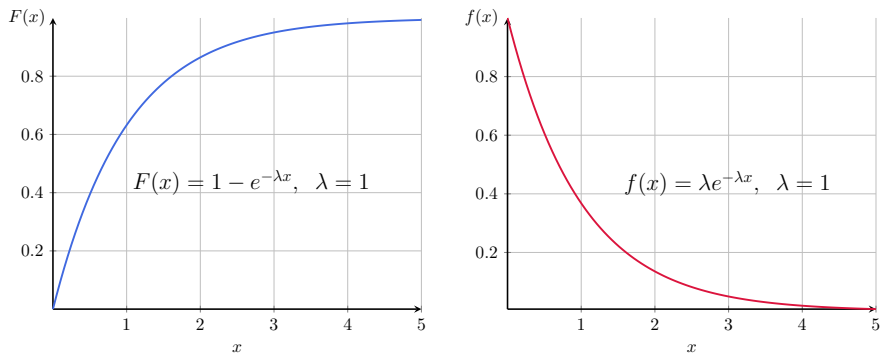


Figure 2.3: The CDF (left) and PDF (right) of an exponential random variable with $\lambda = 1$.

continuously and independently. It is a special case of the gamma distribution and the continuous analog to the geometric distribution. Its PDF $f_X(x)$ is a decaying probability density over the nonnegative reals.

$$X \sim \text{Expon}(\lambda), \quad \lambda > 0$$

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

2.2.7 Joint and Marginal Cumulative Distribution Functions

Thus far, we have considered single random variables. In many situations, however, there may be more than one quantity that we are interested in knowing during a random experiment. For instance, in an experiment where we flip a coin ten times, we may care about both the number of heads that come up (denoted $X(\omega)$) as well as the length of the longest run of consecutive heads (denoted $Y(\omega)$).

For ease of exposition, the remainder of this chapter will consider the setting of two random variables. However, the general multivariate case can consider arbitrarily many variables. We first discuss joint and marginal CDFs, then joint and marginal PMFs and PDFs.

Suppose that we have two random variables, X and Y. One way to work with these two random variables is to consider each of them separately. If we do that we will need the CDFs $F_X(x)$ and $F_Y(y)$. But if

we want to know about the values that X and Y assume simultaneously during outcomes of a random experiment, we require a more complicated structure known as the *joint CDF* of X and Y.

Definition 2.20 (Joint cumulative distribution function of two random variables)**.** The joint CDF is given by

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

Given the joint CDF, the probability of any event involving X and Y can be calculated. The joint CDF $F_{XY}(x, y)$ and the CDFs $F_X(x)$ and $F_Y(y)$ of each variable separately are related as follows.

Definition 2.21 (Marginal cumulative distribution function)**.** The marginal CDFs $F_X(x)$ and $F_Y(y)$ of $F_{XY}(x, y)$ are given by

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

Properties

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

2.2.8 Joint and Marginal Probability Mass Functions

Like the CDF, the notion of joint and marginal distributions can be applied to the probability mass functions of discrete random variables.

Definition 2.22 (Joint probability mass function of two random variables)**.** If X and Y are discrete random variables, then the joint PMF $p_{XY} : \text{Val}(X) \times \text{Val}(Y) \to [0, 1]$ is defined by

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

where $0 \leq p_{XY}(x, y) \leq 1$ for all x, y, and

$$\sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} p_{XY}(x, y) = 1.$$

The joint PMF over two variables relates to the PMF for each variable separately as follows.

Definition 2.23 (Marginal probability mass function)**.** The marginal PMF $p_X(x)$ of X is given by

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

and analogously for $p_Y(y)$ of Y.

In statistics, the process of forming the marginal distribution with respect to one variable by summing out the other variable is often known as *marginalization*.

2.2.9 Joint and Marginal Probability Density Functions

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

Definition 2.24 (Joint probability density function of two random variables)**.** The joint PDF for random variables X and Y takes the form

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

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

$$\int \int_{(x,y)\in A} f_{XY}(x, y)dxdy = \mathbb{P}((X, Y) \in A).$$

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

Analogous to the discrete case, we can define the *marginal PDFs* or *marginal densities* of X and Y.

Definition 2.25 (Marginal probability density function)**.** The marginal PDF for X is given as

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

and analogously for $f_Y(y)$ of Y.

2.2.10 Conditional Distributions

Conditional distributions seek to answer the question, *what is the probability distribution over Y, when we know that X must take on a certain value x?*

Definition 2.26 (Conditional probability mass function)**.** In the discrete case, the conditional PMF of Y given X is simply

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

assuming that $p_X(x) \neq 0$.

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

Definition 2.27 (Conditional probability density function)**.** For continuous random variables X and Y, the conditional PDF of Y given X is simply

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

provided $f_X(x) \neq 0$.

2.2.11 Chain Rule

The chain rule that we previously derived for events (Definition 2.7) can also be applied to random variables.

Definition 2.28 (Chain rule for random variables)**.** Let $\mathbf{X} = \{X_i\}_{i=1}^n$ be set of random variables. By the chain rule of probability, the joint distribution over $\mathbf{X}$ factorizes as

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid x_1, \cdots, x_{i-1}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{x}_{<i}).$$

Thus, for $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$, we have the factorization

$$p(\mathbf{x}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3).$$

2.2.12 Bayes' Rule

A useful formula that often arises when deriving expressions for conditional probability is *Bayes' rule*, also known as *Bayes' theorem*. This formula arises from the chain rule.

Definition 2.29 (Bayes' rule)**.** For discrete random variables X and Y, Bayes' rule states

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

When X and Y are continuous, we have

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

2.2.13 Independence of Random Variables

The notion of independence among random variables is central to factorizing joint distributions, learning graphical structures, and many other fundamental tasks in probabilistic graphical modeling. Variables can be *marginally independent* or *conditionally independent* given a set of additional variables. We denote independence with the symbol $\perp\!\!\!\perp$, and dependence with $\not\perp\!\!\!\perp$.

Definition 2.30 (Independence of random variables)**.** Two random variables X and Y are independent if the following holds for all values of x and y.

$$F_{XY}(x, y) = F_X(x)F_Y(y) \Longrightarrow X \perp\!\!\!\perp Y$$

Equivalently,

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

– $f_{Y|X}(y \mid x) = f_Y(y)$ whenever $f_X(x) \neq 0$ for all $y \in \mathbb{R}$.

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

Lemma 2.1. If X and Y are independent, then for any subsets $A, B \subseteq \mathbb{R}$, we have

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

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

2.2.14 Expectation and Covariance

Suppose that we have two random variables, X and Y. Let $g : \mathbb{R}^2 \to \mathbb{R}$ denote a function of these two random variables.

Definition 2.31 (Expected value of a function over two random variables)**.** When X and Y are discrete, the expected value of g is defined as

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

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

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

We can use the concept of expectation to study the relationship of two random variables with each other. In particular, we can examine their *covariance*, a measure of their joint variability.

Definition 2.32 (Covariance of two random variables)**.** The covariance of two random variables X and Y is defined as

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Intuitively, the covariance of X and Y measures how often and by how much X and Y are both greater than or less than their respective means. If larger values of X correspond with larger values of Y and vice versa, then covariance is positive. If larger values of X correspond with smaller values of Y and vice versa, then covariance is negative. Using an argument similar to that for variance, we can rewrite this as

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

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

Properties

- *Linearity of expectation.*

$$\mathbb{E}[f(X,Y) + g(X,Y)] = \mathbb{E}[f(X,Y)] + \mathbb{E}[g(X,Y)].$$

- $\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}[X,Y]$.
- If X and Y are independent, then $\operatorname{Cov}[X,Y] = 0$. However, if $\operatorname{Cov}[X,Y] = 0$, it is not necessarily true that X and Y are independent. For example, let $X \sim \text{Uniform}(-1,1)$ and let $Y = X^2$. Then, $\operatorname{Cov}[X,Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = \mathbb{E}[X^3] - 0 \cdot \mathbb{E}[X^2] = 0$ even though X and Y are not independent.
- If X and Y are independent, then $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$.

Further Reading

- C. M. Grinstead and J. L. Snell. (1997). *Introduction to Probability*. American Mathematical Soc.
- D. Bertsekas and J. N. Tsitsiklis. (2008). *Introduction to Probability*. Athena Scientific.
- K. P. Murphy. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

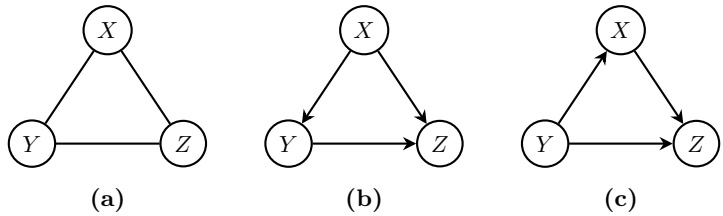


Figure 2.4: Undirected **(a)** and directed graphs **(b,c)** with the same adjacencies.

2.3 Basic Graph Theory

We will now review some basic graph theoretic concepts. Graph theory is the branch of mathematics that is broadly concerned with (1) the properties of graphical objects, which model the pairwise relationships among entities in a system; (2) the operations and transformations that can be performed on these objects; and (3) the design and analysis of algorithms that act on graph objects.

Fundamentally, a *graph* is an abstract mathematical object defined by a set of *nodes* and a set of *edges* that link those nodes. Nodes denote entities in the system, while edges model their relationships.

Definition 2.33 (Graph)**.** Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a graph, where $\mathbf{V} = \{V_i\}_{i=1}^{n}$ is the node set (also called vertices) and $\mathbf{E} = \{E_i\}_{i=1}^{m}$ is the edge set.

Undirected, Directed, and Mixed Graphs Many kinds of graphs can arise. A basic dichotomy in graph theory is that of undirected graphs versus directed graphs. If the edges of $\mathcal{G}$ have no orientation or directionality (as signified by a lack of arrowheads), we say that $\mathcal{G}$ is an *undirected graph* (Figure 2.4a). If the edges of $\mathcal{G}$ do display a fixed orientation (e.g., $X \to Y$), we say that $\mathcal{G}$ is a *directed graph* or *digraph* (Figure 2.4b). This is a false dichotomy, as there exist several noteworthy *mixed graph* classes that contain both directed and undirected edge types. Mixed graphs can even contain *bidirected edges* and edges with alternative endpoints denoting uncertain directionality. Such mixed graphs are famously useful for expressing uncertainty and unmeasured variables in *causal graphical modeling*. For the remainder of this chapter,

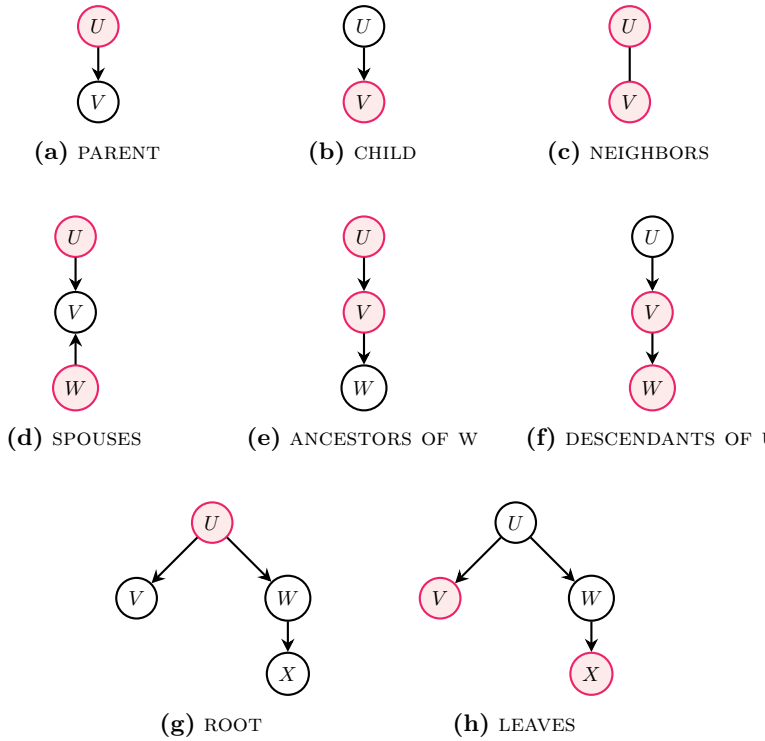


Figure 2.5: Roles that a given (highlighted) node can play with respect to others.

we will restrict our attention to directed and undirected graphs, but we will revisit mixed graphs in Section 6.2.2.

Properties and Substructures An enormous range of graph, node, and edge properties can be enumerated, but here we will focus on basic vocabulary. Nodes that are connected by an edge are said to be *adjacent* or *neighbors*. The *degree* of a node is the number of neighbors that it has. In a directed graph, the *in-degree* of a node denotes the total number of neighbors with edges pointing into it (e.g., $\cdots \rightarrow Z$). Analogously, the *out-degree* is equal to the number of edges pointing out of a node (e.g., $Z \rightarrow \cdots$). In Figure 2.4b, the in-degree of Z is 2 and its out-degree is 0, while the in-degree of X is 0 and its out-degree is 2.

Sets of adjacent nodes can form various noteworthy substructures.

A sequence of linked nodes can form a *path*: $V_i - V_{i+1} - \cdots - V_{i+n}$ in the undirected case, or $V_i \to V_{i+1} \to \cdots \to V_{i+n}$ in the directed case. A path that starts and ends at the same node is called a *cycle*.

Definition 2.34 (Cycle)**.** A cycle is a path of vertices such that only the first and last vertices are the same: $V_i - V_{i+1} - \cdots - V_{i+n} - V_i$ in the undirected case, or $V_i \to V_{i+1} \to \cdots \to V_{i+n} \to V_i$ in the directed case.

Thus, when the cycle is traversed, it loops back to its beginning. The graph in Figure 2.4a is an undirected cycle, while the graph in Figure 2.4b is *not* a directed cycle. Directed and undirected cycles (or, often, their *absence*) play important roles throughout graph theory and probabilistic graphical modeling. We will see throughout this tutorial that *acyclic* graphs are the focus of many inference and learning algorithms.

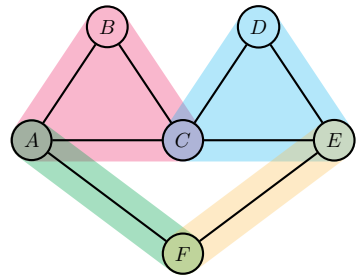
Many other relationships exist among groups of nodes (Figure 2.5). Let U, V be two neighbors. When U has a directed edge into V, we say that U is the *parent* and V is the *child*. If a third variable W is also a parent of V, we say that U and W are *spouses*. If W is instead a child of V, we say that U is the *ancestor* of *descendant* W. Ancestors and descendants generalize the parent-child relationship to paths of length greater than 1. When a node has no *parents*, we say that it is a *root*. When a node has no children, we say that it is a *leaf*. In an undirected graph, a leaf is a node with degree 1. In general, we will use the notation $\mathbf{pa}(\cdot)$, $\mathbf{ch}(\cdot)$, $\mathbf{an}(\cdot)$, $\mathbf{de}(\cdot)$, $\mathbf{sp}(\cdot)$, $\mathbf{ne}(\cdot)$ to refer to the parent, child, ancestor, descendant, spouse, and neighbor sets of a given node, respectively.

In undirected graphs, it is often useful to perform operations over *fully connected subgraphs* called *cliques*.

Definition 2.35 (Clique)**.** Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be an undirected graph. A *clique* is a subset $\mathbf{C} \subseteq \mathbf{V}$ where every pair of nodes $V_i, V_j \in \mathbf{C}$ are neighbors.

Definition 2.36 (Maximal clique)**.** A *maximal clique* is a clique to which no additional nodes in $\mathbf{V}$ can be added without violating the definition of a clique.

In Figure 2.6, $\{A, B\}$, $\{A, C\}$, and $\{B, C\}$ form pairwise cliques, while $\{A, B, C\}$ forms a maximal clique. We can also define *unary*



UNDIRECTED GRAPH $\mathcal{G}$

Figure 2.6: Maximal cliques in $\mathcal{G}$ are $\{A, C, B\}$, $\{C, D, E\}$, $\{A, F\}$, and $\{E, F\}$.

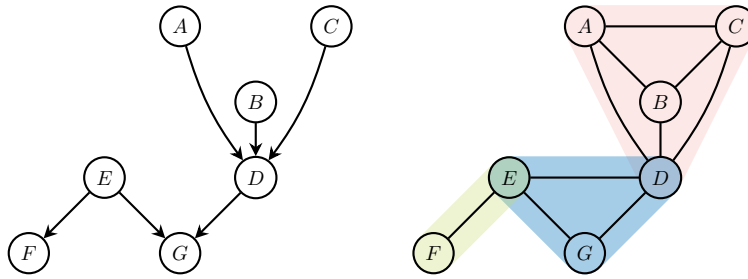
cliques, consisting of a single node each. Later, cliques will be integral to our discussion of undirected graphs (Section 3.2.1) and inference (Chapter 4).

We will introduce a final property of directed graphs: the *topological ordering* or *topological sort*.

Definition 2.37 (Topological ordering)**.** Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a directed graph. A topological ordering of the nodes in $\mathcal{G}$ is a linear ordering such that for every edge $V_i \to V_j \in \mathbf{E}$, V_i precedes V_j in the ordering.

Topological orderings can convey significant structural information. In general, however, this information cannot uniquely specify the graph. It is possible for multiple directed graphs to satisfy the same topological ordering, and multiple orderings can describe the same graph. For example, (U, V, W, X) and (U, W, X, V) are both valid topological orderings for the graph in Figure 2.5g. We will return to this ambiguity in Section 6.2.2. Nevertheless, we will see that topological orderings can be useful in both inference and learning.

Converting Directed Graphs to Undirected Graphs In this tutorial, we will reference several methods and concepts that require the transformation of a directed graph to an undirected graph. These primarily rely on two different approaches, which serve different purposes. Most simply, we can remove directionality by deleting the arrowheads from each edge. This yields the *undirected skeleton* of the directed graph. For



(a) DIGRAPH **(b)** MORAL GRAPH

Figure 2.7: Directed graph **(a)** and its corresponding moral graph **(b)**, with shading denoting maximal cliques.

example, Figure 2.4a is the skeleton of Figure 2.4b. However, some use cases require an additional manipulation: *moralization* (Figure 2.7).

When a node has two or more parents that are not adjacent to each other, we historically refer to this as an *immorality* (an anachronistic and value-laden term). This brings us to our notion of a *moral graph*.

Definition 2.38 (Moral graph)**.** The moral graph corresponding to a directed acyclic graph $\mathcal{G}$ is an undirected graph with an edge for every adjacency in $\mathcal{G}$ and for every pair of non-adjacent spouses (i.e., nodes in $\mathcal{G}$ that share a child).

Moralization ensures that each child node and all of its parents form a single clique in the resulting undirected graph (Bishop, 2006). To *moralize* a directed graph $\mathcal{G}$, we proceed in two steps:

1. For every pair of spouses in $\mathcal{G}$ that are not already adjacent, add an undirected edge. This removes all immoralities.

2. Remove the directionality for all edges in $\mathcal{G}$.

In addition to moralization, we may wish to *chordalize* the undirected skeleton of our directed graph $\mathcal{G}$ (Figure 2.8). We can transform the undirected skeleton of $\mathcal{G}$ to a *chordal graph* (or *triangulated graph*) by adding edges such that the following definition is met.

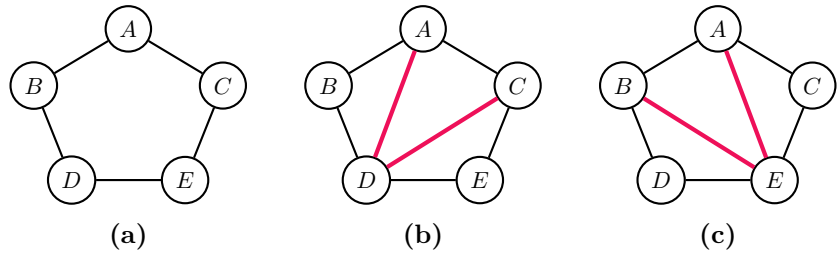


Figure 2.8: An undirected graph (**a**) and two possible chordalized graphs corresponding to it (**b**, **c**; chords in red).

Definition 2.39 (Chordal graph, Vandenberghe and Andersen 2015)**.** An undirected graph is chordal if any cycle of length greater than three has a chord (i.e., an edge connecting any two nonconsecutive nodes). Thus, the longest minimal loop is a triangle.

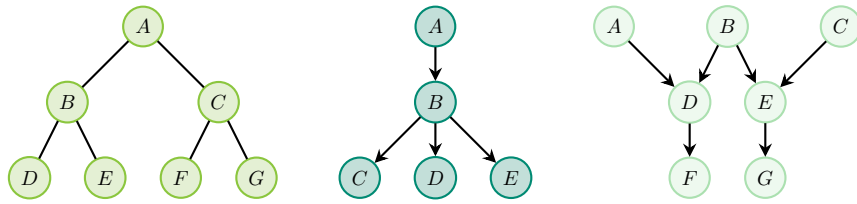
Both moralization and chordalization are important steps in the junction tree algorithm, which we will explore in Chapter 4. Chordal graphs have played pivotal roles in combinatorial optimization, semidefinite optimization, nonlinear optimization, linear algebra, statistics, and signal processing. For further discussion of the theory and applications of chordal graphs, see Vandenberghe and Andersen (2015).

2.3.1 Special Categories of Graphs

There are certain categories of graphs that we will encounter frequently in graphical modeling. One such kind of graph is the *directed acyclic graph* (DAG). This is simply a digraph with no cycles (e.g., Figure 2.4b). DAGs can be used for both probabilistic and causal modeling, and have played a central role in modeling in the fundamental sciences. Another important kind of graph is the *tree* (Figure 2.9). Many of the probabilistic inference algorithms that we will discuss apply to tree-structured graphs.

Definition 2.40 (Undirected tree)**.** An *undirected tree* is a connected, undirected graph with no cycles.

The following properties apply to undirected trees. See Chapter 3 of Deo (1974) for proof of these theorems.



(a) **(b)** **(c)**

Figure 2.9: (a) An undirected tree, **(b)** a directed tree, and **(c)** a polytree.

Theorem 2.2. Any two nodes in a tree are connected by exactly one path. If any two nodes in a graph are connected by exactly one path, then that graph is a tree.

Theorem 2.3. A tree with n nodes has $n - 1$ edges. If a connected graph with n nodes has $n - 1$ edges, then it is a tree.

Theorem 2.4. A graph is a tree if and only if it is minimally connected (i.e., it is connected, and the removal of any edge disconnects the graph).

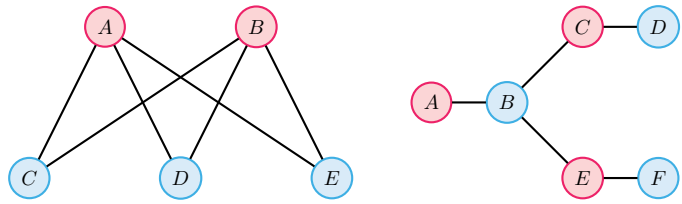
We can also define various tree-structured graphs with directed edges: *directed trees* and *polytrees*.

Definition 2.41 (Directed tree)**.** A directed graph is a *directed tree* if it is connected and if each node has at most one parent.

Definition 2.42 (Polytree)**.** A *polytree* is a DAG with no cycles in its undirected skeleton. A polytree is distinct from a tree, as nodes can have more than one parent.

We will also make use of the concept of *treewidth*, which is a number that characterizes how far a graph is from being a tree. The smaller the treewidth, the more "tree-like" the graph is, and true trees have a treewidth of 1.

Definition 2.43 (Treewidth)**.** A graph has *treewidth* k if it can be broken into overlapping groups of at most $k+1$ vertices (called bags), arranged in a tree-like structure, such that: (a) every graph vertex is in at least one bag; (b) for every edge, both endpoints appear together in some



(a) **(b)**

Figure 2.10: (a) A complete bipartite graph and **(b)** an acyclic bipartite graph with partite sets in red and blue.

bag; and (c) for each vertex, the bags containing it form a connected subtree of the tree-like structure over bags.

Finally, we introduce the *bipartite graph* (Figure 2.10). These will make an appearance in Chapter 3 when we discuss factor graphs.

Definition 2.44 (Bipartite graph)**.** A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is bipartite if $\mathbf{V}$ can be partitioned into two disjoint sets $\mathbf{V}_1$, $\mathbf{V}_2$ (i.e., *partite sets*), such that every edge in $\mathbf{E}$ has one endpoint in $\mathbf{V}_1$ and one endpoint in $\mathbf{V}_2$.

Additionally, bipartite graphs have the following property. See Harris *et al.* (2008) for full proof of this theorem.

Theorem 2.5. A graph containing at least two nodes is bipartite if and only if it contains no odd cycles.

Following from Theorem 2.5, we can see that all trees are bipartite (as they contain no cycles).

Further Reading

- N. Deo. (1974). *Graph Theory with Applications to Engineering and Computer Science*. Prentice-Hall, Inc.
- J. Harris *et al.* (2008). *Combinatorics and Graph Theory*. Springer Science & Business Media.