Chapter 3

Multivariate Random Variables

Probabilistic models usually include multiple uncertain numerical quantities. In this chapter we describe how to specify random variables to represent such quantities and their interactions. In some occasions, it will make sense to group these random variables as **random vectors**, which we write using uppercase letters with an arrow on top: $\vec{X}$. Realizations of these random vectors are denoted with lowercase letters: $\vec{x}$.

3.1 Discrete random variables

Recall that discrete random variables are numerical quantities that take either finite or countably infinite values. In this section we explain how to manipulate multiple discrete random variables that share a common probability space.

3.1.1 Joint probability mass function

If several discrete random variables are defined on the same probability space, we specify their probabilistic behavior through their **joint probability mass function**, which is the probability that each variable takes a particular value.

Definition 3.1.1 (Joint probability mass function)**.** *Let $X : \Omega \to R_X$ and $Y : \Omega \to R_Y$ be discrete random variables (R_X and R_Y are discrete sets) on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The joint pmf of X and Y is defined as*

$$p_{X,Y}(x, y) := \mathrm{P}(X = x, Y = y) \quad . \tag{3.1}$$

In words, $p_{X,Y}(x, y)$ is the probability of X and Y being equal to x and y respectively.

Similarly, the joint pmf of a discrete random vector of dimension n

$$\vec{X} := \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix} \tag{3.2}$$

with entries $X_i : \Omega \to R_{X_i}$ ($R_1, \ldots, R_n$ are all discrete sets) belonging to the same probability space is defined as

$$p_{\vec{X}}(\vec{x}) := \mathrm{P}(X_1 = \vec{x}_1, X_2 = \vec{x}_2, \ldots, X_n = \vec{x}_n). \tag{3.3}$$

As in the case of the pmf of a single random variable, the joint pmf is a valid probability measure if we consider a probability space where the sample space is $R_X \times R_Y$[1] (or $R_{X_1} \times R_{X_2} \cdots \times R_{X_n}$ in the case of a random vector) and the σ-algebra is just the power set of the sample space. This implies that the joint pmf *completely* characterizes the random variables or the random vector, we don't need to worry about the underlying probability space.

By the definition of probability measure, the joint pmf must be nonnegative and its sum over all its possible arguments must equal one,

$$p_{X,Y}(x,y) \geq 0 \quad \text{for any } x \in R_X, y \in R_Y, \tag{3.4}$$

$$\sum_{x \in R_X} \sum_{y \in R_Y} p_{X,Y}(x,y) = 1. \tag{3.5}$$

By the Law of Total Probability, the joint pmf allows us to obtain the probability of X and Y belonging to any set $\mathcal{S} \subseteq R_X \times R_Y$,

$$\begin{aligned} \mathrm{P}((X,Y) \in \mathcal{S}) &= \mathrm{P}\left(\cup_{(x,y)\in\mathcal{S}} \{X = x, Y = y\}\right) \quad \text{(union of disjoint events)} && (3.6)\\ &= \sum_{(x,y)\in\mathcal{S}} \mathrm{P}(X = x, Y = y) && (3.7)\\ &= \sum_{(x,y)\in\mathcal{S}} p_{X,Y}(x,y). && (3.8) \end{aligned}$$

These properties also hold for random vectors (and groups of more than two random variables). For any random vector $\vec{X}$,

$$p_{\vec{X}}(\vec{x}) \geq 0, \tag{3.9}$$

$$\sum_{\vec{x}_1 \in R_1} \sum_{\vec{x}_2 \in R_2} \cdots \sum_{\vec{x}_n \in R_n} p_{\vec{X}}(\vec{x}) = 1. \tag{3.10}$$

The probability that $\vec{X}$ belongs to a discrete set $\mathcal{S} \subseteq \mathbb{R}^n$ is given by

$$\mathrm{P}\left(\vec{X} \in \mathcal{S}\right) = \sum_{\vec{x}\in\mathcal{S}} p_{\vec{X}}(\vec{x}). \tag{3.11}$$

3.1.2 Marginalization

Assume we have access to the joint pmf of several random variables in a certain probability space, but we are only interested in the behavior of one of them. To compute the value of its pmf for a particular value, we fix that value and sum over the remaining random variables. Indeed, by the Law of Total Probability

$$\begin{aligned} p_X(x) &= \mathrm{P}(X = x) && (3.12)\\ &= \mathrm{P}\left(\cup_{y\in R_Y} \{X = x, Y = y\}\right) \quad \text{(union of disjoint events)} && (3.13)\\ &= \sum_{y\in R_Y} \mathrm{P}(X = x, Y = y) && (3.14)\\ &= \sum_{y\in R_Y} p_{X,Y}(x,y). && (3.15) \end{aligned}$$

[1] This is the Cartesian product of the two sets, defined in Section A.2, which contains all possible pairs (x, y) where $x \in R_X$ and $y \in R_y$.

When the joint pmf involves more than two random variables the argument is exactly the same. This is called **marginalizing** over the other random variables. In this context, the pmf of a single random variable is called its **marginal pmf**. Table 3.1 shows an example of a joint pmf and the corresponding marginal pmfs.

If we are interested in computing the joint pmf of several entries in a random vector, instead of just one, the marginalization process is essentially the same. The pmf is again obtained by summing over the rest of the entries. Let $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ be a subset of $m < n$ entries of an n-dimensional random vector $\vec{X}$ and $\vec{X}_{\mathcal{I}}$ the corresponding random subvector. To compute the joint pmf of $\vec{X}_{\mathcal{I}}$ we sum over all the entries that are not in $\mathcal{I}$, which we denote by $\{j_1, j_2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$

$$p_{\vec{X}_{\mathcal{I}}}(\vec{x}_{\mathcal{I}}) = \sum_{\vec{x}_{j_1} \in R_{j_1}} \sum_{\vec{x}_{j_2} \in R_{j_2}} \cdots \sum_{\vec{x}_{j_{n-m}} \in R_{j_{n-m}}} p_{\vec{X}}(\vec{x}). \tag{3.16}$$

3.1.3 Conditional distributions

Conditional probabilities allow us to update our uncertainty about the quantities in a probabilistic model when new information is revealed. The conditional distribution of a random variable specifies the behavior of the random variable when we assume that other random variables in the probability space take a fixed value.

Definition 3.1.2 (Conditional probability mass function)**.** *The conditional probability mass function of Y given X, where X and Y are discrete random variables defined on the same probability space, is given by*

$$p_{Y|X}(y|x) = \mathrm{P}(Y = y | X = x) \tag{3.17}$$

$$= \frac{p_{X,Y}(x, y)}{p_X(x)} \qquad \text{if } p_X(x) > 0 \tag{3.18}$$

and is undefined otherwise.

The conditional pmf $p_{X|Y}(\cdot|y)$ characterizes our uncertainty about X conditioned on the event $\{Y = y\}$. This object is a valid pmf of X, so that if R_X is the range of X

$$\sum_{x \in R_X} p_{X|Y}(x|y) = 1 \tag{3.19}$$

for any y for which it is well defined. However, it is *not* a pmf for Y. In particular, there is no reason for $\sum_{y \in R_Y} p_{X|Y}(x|y)$ to add up to one!

We now define the joint conditional pmf of several random variables (equivalently of a subvector of a random vector) given other random variables (or entries of the random vector).

Definition 3.1.3 (Conditional pmf)**.** *The conditional pmf of a discrete random subvector $\vec{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$, given another subvector $\vec{X}_{\mathcal{J}}$ is*

$$p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}) := \frac{p_{\vec{X}}(\vec{x})}{p_{\vec{X}_{\mathcal{J}}}(\vec{x}_{\mathcal{J}})}, \tag{3.20}$$

where $\{j_1, j_2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$.

R

L \ $p_{L,R}$	0	1
0	$\frac{14}{20}$	$\frac{1}{20}$
1	$\frac{2}{20}$	$\frac{3}{20}$

p_L
$\frac{15}{20}$
$\frac{5}{20}$

$p_{L\|R}(\cdot\|0)$
$\frac{7}{8}$
$\frac{1}{8}$

$p_{L\|R}(\cdot\|1)$
$\frac{1}{4}$
$\frac{3}{4}$

p_R	$\frac{16}{20}$	$\frac{4}{20}$

$p_{R\|L}(\cdot\|0)$	$\frac{14}{15}$	$\frac{1}{15}$

$p_{R\|L}(\cdot\|1)$	$\frac{2}{5}$	$\frac{3}{5}$

Table 3.1: Joint, marginal and conditional pmfs of the random variables L and R defined in Example 3.1.5.

The conditional pmfs $p_{Y|X}(\cdot|x)$ and $p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}(\cdot|\vec{x}_{\mathcal{J}})$ are valid pmfs in the probability space where $X = x$ or $\vec{X}_{\mathcal{J}} = \vec{x}_{\mathcal{J}}$ respectively. For instance, they must be nonnegative and add up to one.

From the definition of conditional pmfs we derive a chain rule for discrete random variables and vectors.

Lemma 3.1.4 (Chain rule for discrete random variables and vectors)**.**

$$p_{X,Y}(x,y) = p_X(x)\, p_{Y|X}(y|x), \tag{3.21}$$

$$p_{\vec{X}}(\vec{x}) = p_{X_1}(\vec{x}_1)\, p_{X_2|X_1}(\vec{x}_2|\vec{x}_1) \dots p_{X_n|X_1,\dots,X_{n-1}}(\vec{x}_n|\vec{x}_1,\dots,\vec{x}_{n-1}) \tag{3.22}$$

$$= \prod_{i=1}^{n} p_{X_i|\vec{X}_{\{1,\dots,i-1\}}}\left(\vec{x}_i|\vec{x}_{\{1,\dots,i-1\}}\right), \tag{3.23}$$

where the order of indices in the random vector is arbitrary (any order works).

The following example illustrates the definitions of marginal and conditional pmfs.

Example 3.1.5 (Flights and rains (continued))**.** Within the probability space described in Example 1.2.1 we define a random variable

$$L = \begin{cases} 1 & \text{if plane is late,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.24}$$

to represent whether the plane is late or not. Similarly,

$$R = \begin{cases} 1 & \text{it rains,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.25}$$

represents whether it rains or not. Equivalently, these random variables are just the indicators $R = 1_{\text{rain}}$ and $L = 1_{\text{late}}$. Table 3.1 shows the joint, marginal and conditional pmfs of L and R.

△

3.2 Continuous random variables

Continuous random variables allow us to model continuous quantities without having to worry about discretization. In exchange, the mathematical tools to manipulate them are somewhat more complicated than in the discrete case.

3.2.1 Joint cdf and joint pdf

As in the case of univariate continuous random variables, we characterize the behavior of several continuous random variables defined on the same probability space through the probability that they belong to Borel sets (or equivalently unions of intervals). In this case we are considering multidimensional Borel sets, which are Cartesian products of one-dimensional Borel sets. Multidimensional Borel sets can be represented as unions of multidimensional intervals or hyperrectangles (defined as Cartesian products of one-dimensional intervals). The **joint cdf** compiles the probability that the random variables belong to the Cartesian product of intervals of the form $(-\infty, r]$ for every $r \in \mathbb{R}$.

Definition 3.2.1 (Joint cumulative distribution function)**.** *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $X, Y : \Omega \to \mathbb{R}$ random variables. The **joint cdf** of X and Y is defined as*

$$F_{X,Y}(x, y) := \mathrm{P}(X \leq x, Y \leq y). \tag{3.26}$$

In words, $F_{X,Y}(x, y)$ is the probability of X and Y being smaller than x and y respectively.

Let $\vec{X} : \Omega \to \mathbb{R}^n$ be a random vector of dimension n on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The joint cdf of $\vec{X}$ is defined as

$$F_{\vec{X}}(\vec{x}) := \mathrm{P}\left(\vec{X}_1 \leq \vec{x}_1, \vec{X}_2 \leq \vec{x}_2, \ldots, \vec{X}_n \leq \vec{x}_n\right). \tag{3.27}$$

In words, $F_{\vec{X}}(\vec{x})$ is the probability that $\vec{X}_i \leq \vec{x}_i$ for all $i = 1, 2, \ldots, n$.

We now record some properties of the joint cdf.

Lemma 3.2.2 (Properties of the joint cdf)**.**

$$\lim_{x \to -\infty} F_{X,Y}(x, y) = 0, \tag{3.28}$$

$$\lim_{y \to -\infty} F_{X,Y}(x, y) = 0, \tag{3.29}$$

$$\lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1, \tag{3.30}$$

$$F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2) \quad \text{if } x_2 \geq x_1, y_2 \geq y_1, \quad \text{i.e. } F_{X,Y} \text{ is nondecreasing.} \tag{3.31}$$

Proof. The proof follows along the same lines as that of Lemma 2.3.2. □

The joint cdf completely specifies the behavior of the corresponding random variables. Indeed, we can decompose any Borel set into a union of disjoint n-dimensional intervals and compute their probability by evaluating the joint cdf. Let us illustrate this for the bivariate case:

$$\begin{aligned} \mathrm{P}\left(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\right) &= \mathrm{P}\left(\{X \leq x_2, Y \leq y_2\} \cap \{X > x_1\} \cap \{Y > y_1\}\right) && (3.32)\\ &= \mathrm{P}\left(X \leq x_2, Y \leq y_2\right) - \mathrm{P}\left(X \leq x_1, Y \leq y_2\right) && (3.33)\\ &\quad - \mathrm{P}\left(X \leq x_2, Y \leq y_1\right) + \mathrm{P}\left(X \leq x_1, Y \leq y_1\right) && (3.34)\\ &= F_{X,Y}\left(x_2, y_2\right) - F_{X,Y}\left(x_1, y_2\right) - F_{X,Y}\left(x_2, y_1\right) + F_{X,Y}\left(x_1, y_1\right). \end{aligned}$$

This means that, as in the univariate case, to define a random vector or a group of random variables all we need to do is define their joint cdf. We don't have to worry about the underlying probability space.

If the joint cdf is differentiable, we can differentiate it to obtain the **joint probability density function** of X and Y. As in the case of univariate random variables, this is often a more convenient way of specifying the joint distribution.

Definition 3.2.3 (Joint probability density function)**.** *If the joint cdf of two random variables X, Y is differentiable, then their joint pdf is defined as*

$$f_{X,Y}\left(x, y\right) := \frac{\partial^2 F_{X,Y}\left(x, y\right)}{\partial x \partial y}. \tag{3.35}$$

If the joint cdf of a random vector $\vec{X}$ is differentiable, then its joint pdf is defined as

$$f_{\vec{X}}\left(\vec{x}\right) := \frac{\partial^n F_{\vec{X}}\left(\vec{x}\right)}{\partial \vec{x}_1 \, \partial \vec{x}_2 \cdots \partial \vec{x}_n}. \tag{3.36}$$

The joint pdf should be understood as an n-dimensional density, *not* as a probability (for instance, it can be larger than one). In the two-dimensional case,

$$\lim_{\Delta_x \to 0, \Delta_y \to 0} \mathrm{P}\left(x \leq X \leq x + \Delta_x, y \leq Y \leq y + \Delta_y\right) = f_{X,Y}\left(x, y\right) \Delta_x \Delta_y. \tag{3.37}$$

Due to the monotonicity of joint cdfs in every variable, joint pmfs are always nonnegative.

The joint pdf of X and Y allows us to compute the probability of any Borel set $\mathcal{S} \subseteq \mathbb{R}^2$ by integrating over $\mathcal{S}$

$$\mathrm{P}\left((X, Y) \in \mathcal{S}\right) = \int_{(x,y) \in \mathcal{S}} f_{X,Y}\left(x, y\right) \, \mathrm{d}x \, \mathrm{d}y. \tag{3.38}$$

Similarly, the joint pdf of an n-dimensional random vector $\vec{X}$ allows to compute the probability that $\vec{X}$ belongs to a set Borel set $\mathcal{S} \subseteq \mathbb{R}^n$,

$$\mathrm{P}\left(\vec{X} \in \mathcal{S}\right) = \int_{\vec{x} \in \mathcal{S}} f_{\vec{X}}\left(\vec{x}\right) \, \mathrm{d}\vec{x}. \tag{3.39}$$

In particular, if we integrate a joint pdf over the whole space $\mathbb{R}^n$, then it must integrate to one by the Law of Total Probability.

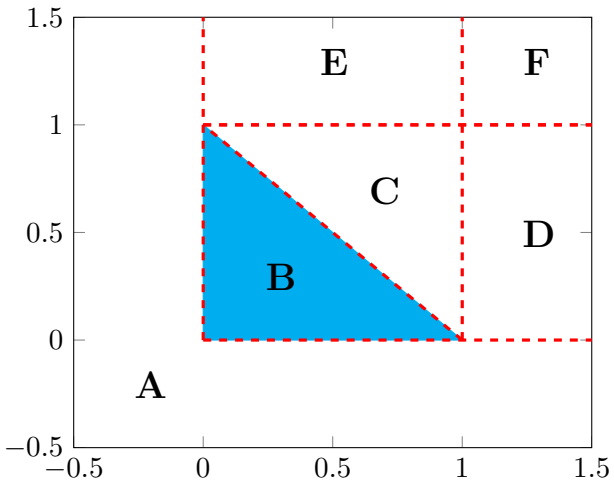


Figure 3.1: Triangle lake in Example 3.2.12.

Example 3.2.4 (Triangle lake)**.** A biologist is tracking an otter that lives in a lake. She decides to model the location of the otter probabilistically. The lake happens to be triangular as shown in Figure 3.1, so that we can represent it by the set

$$\text{Lake} := \{\vec{x} \mid \vec{x}_1 \geq 0, \vec{x}_2 \geq 0, \vec{x}_1 + \vec{x}_2 \leq 1\}. \tag{3.40}$$

The biologist has no idea where the otter is, so she models the position as a random vector $\vec{X}$ which is uniformly distributed over the lake. In other words, the joint pdf of $\vec{X}$ is constant,

$$f_{\vec{X}}(\vec{x}) = \begin{cases} c & \text{if } \vec{x} \in \text{Lake}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.41}$$

To find the normalizing constant c we use the fact that to be a valid joint pdf $f_{\vec{X}}$ should integrate to 1.

$$\int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} c \, \mathrm{d}x_1 \, \mathrm{d}x_2 = \int_{x_2=0}^{1} \int_{x_1=0}^{1-x_2} c \, \mathrm{d}x_1 \, \mathrm{d}x_2 \tag{3.42}$$

$$= c \int_{x_2=0}^{1} (1 - x_2) \, \mathrm{d}x_2 \tag{3.43}$$

$$= \frac{c}{2} = 1, \tag{3.44}$$

so $c = 2$.

We now compute the cdf of $\vec{X}$. $F_{\vec{X}}(\vec{x})$ represents the probability that the otter is southwest of the point $\vec{x}$. Computing the joint cdf requires dividing the range into the sets shown in Figure 3.1 and integrating the joint pdf. If $\vec{x} \in A$ then $F_{\vec{X}}(\vec{x}) = 0$ because $\mathrm{P}\left(\vec{X} \leq \vec{x}\right) = 0$. If $(\vec{x}) \in B$,

$$F_{\vec{X}}(\vec{x}) = \int_{u=0}^{\vec{x}_2} \int_{v=0}^{\vec{x}_1} 2 \, \mathrm{d}v \, \mathrm{d}u = 2\vec{x}_1\vec{x}_2. \tag{3.45}$$

If $\vec{x} \in C$,

$$F_{\vec{X}}(\vec{x}) = \int_{u=0}^{1-\vec{x}_1} \int_{v=0}^{\vec{x}_1} 2\,\mathrm{d}v\,\mathrm{d}u + \int_{u=1-\vec{x}_1}^{\vec{x}_2} \int_{v=0}^{1-u} 2\,\mathrm{d}v\,\mathrm{d}u = 2\vec{x}_1 + 2\vec{x}_2 - \vec{x}_2^2 - \vec{x}_1^2 - 1. \tag{3.46}$$

If $\vec{x} \in D$,

$$F_{\vec{X}}(\vec{x}) = \mathrm{P}\left(\vec{X}_1 \leq \vec{x}_1, \vec{X}_2 \leq \vec{x}_2\right) = \mathrm{P}\left(\vec{X}_1 \leq 1, \vec{X}_2 \leq \vec{x}_2\right) = F_{\vec{X}}(1, \vec{x}_2) = 2\vec{x}_2 - \vec{x}_2^2, \tag{3.47}$$

where the last step follows from (3.46). Exchanging $\vec{x}_1$ and $\vec{x}_2$, we obtain $F_{\vec{X}}(\vec{x}) = 2\vec{x}_1 - \vec{x}_1^2$ for $\vec{x} \in E$ by the same reasoning. Finally, for $\vec{x} \in F$ $F_{\vec{X}}(\vec{x}) = 1$ because $\mathrm{P}\left(\vec{X}_1 \leq x_1, \vec{X}_2 \leq x_2\right) = 1$. Putting everything together,

$$F_{\vec{X}}(\vec{x}) = \begin{cases} 0 & \text{if } \vec{x}_1 < 0 \text{ or } \vec{x}_2 < 0, \\ 2\vec{x}_1\vec{x}_2, & \text{if } \vec{x}_1 \geq 0, \vec{x}_2 \geq 0, \vec{x}_1 + \vec{x}_2 \leq 1, \\ 2\vec{x}_1 + 2\vec{x}_2 - \vec{x}_2^2 - \vec{x}_1^2 - 1, & \text{if } \vec{x}_1 \leq 1, \vec{x}_2 \leq 1, \vec{x}_1 + \vec{x}_2 \geq 1, \\ 2\vec{x}_2 - \vec{x}_2^2, & \text{if } \vec{x}_1 \geq 1, 0 \leq \vec{x}_2 \leq 1, \\ 2\vec{x}_1 - \vec{x}_1^2, & \text{if } 0 \leq \vec{x}_1 \leq 1, \vec{x}_2 \geq 1, \\ 1, & \text{if } \vec{x}_1 \geq 1, \vec{x}_2 \geq 1. \end{cases} \tag{3.48}$$

△

3.2.2 Marginalization

We now discuss how to characterize the marginal distributions of individual random variables from a joint cdf or a joint pdf. Consider the joint cdf $F_{X,Y}(x, y)$. When $x \to \infty$ the limit of $F_{X,Y}(x, y)$ is by definition the probability of Y being smaller than y, which is precisely the marginal cdf of Y. More formally,

$$\lim_{x\to\infty} F_{X,Y}(x, y) = \lim_{n\to\infty} \mathrm{P}\left(\cup_{i=1}^{n} \{X \leq i, Y \leq y\}\right) \tag{3.49}$$

$$= \mathrm{P}\left(\lim_{n\to\infty} \{X \leq n, Y \leq y\}\right) \tag{3.50}$$

$$= \mathrm{P}(Y \leq y) \tag{3.51}$$

$$= F_Y(y). \tag{3.52}$$

If the random variables have a joint pdf, we can also compute the marginal cdf by integrating over x

$$F_Y(y) = \mathrm{P}(Y \leq y) \tag{3.53}$$

$$= \int_{u=-\infty}^{y} \int_{x=-\infty}^{\infty} f_{X,Y}(x, u)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.54}$$

Differentiating the latter equation with respect to y, we obtain the marginal pdf of Y

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x. \tag{3.55}$$

Similarly, the marginal pdf of a subvector $\vec{X}_{\mathcal{I}}$ of a random vector $\vec{X}$ indexed by $\mathcal{I} := \{i_1, i_2, \ldots, i_m\}$ is obtained by integrating over the rest of the components $\{j_1, j_2, \ldots, j_{n-m}\} := \{1, 2, \ldots, n\} / \mathcal{I}$,

$$f_{\vec{X}_{\mathcal{I}}}(\vec{x}_{\mathcal{I}}) = \int_{\vec{x}_{j_1}} \int_{\vec{x}_{j_2}} \cdots \int_{\vec{x}_{j_{n-m}}} f_{\vec{X}}(\vec{x}) \, \mathrm{d}\vec{x}_{j_1} \, \mathrm{d}\vec{x}_{j_2} \cdots \mathrm{d}\vec{x}_{j_{n-m}}. \tag{3.56}$$

Example 3.2.5 (Triangle lake (continued))**.** The biologist is interested in the probability that the otter is south of x_1. This information is encoded in the cdf of the random vector, we just need to take the limit when $x_2 \to \infty$ to marginalize over x_2.

$$F_{X_1}(x_1) = \begin{cases} 0 & \text{if } x_1 < 0, \\ 2x_1 - x_1^2 & \text{if } 0 \leq x_1 \leq 1, \\ 1 & \text{if } x_1 \geq 1. \end{cases} \tag{3.57}$$

To obtain the marginal pdf of X_1, which represents the latitude of the otter's position, we differentiate the marginal cdf

$$f_{X_1}(x_1) = \frac{\mathrm{d}F_{X_1}(x_1)}{\mathrm{d}x_1} = \begin{cases} 2(1 - x_1) & \text{if } 0 \leq x_1 \leq 1, \\ 0, & \text{otherwise.} \end{cases}. \tag{3.58}$$

Alternatively, we could have integrated the joint uniform pdf over x_2 (we encourage you to check that the result is the same).

△

3.2.3 Conditional distributions

In this section we discuss how to obtain the conditional distribution of a random variable given information about other random variables in the probability space. To begin with, we consider the case of two random variables. As in the case of univariate distributions, we can define the joint cdf and pdf of two random variables given events of the form $\{(X, Y) \in \mathcal{S}\}$ for any Borel set in $\mathbb{R}^2$ by applying the definition of conditional probability.

Definition 3.2.6 (Joint conditional cdf and pdf given an event)**.** *Let X, Y be random variables with joint pdf $f_{X,Y}$ and let $\mathcal{S} \subseteq \mathbb{R}^2$ be any Borel set with nonzero probability, the conditional cdf and pdf of X and Y given the event $(X, Y) \in \mathcal{S}$ is defined as*

$$F_{X,Y|(X,Y)\in\mathcal{S}}(x, y) := \mathrm{P}(X \leq x, Y \leq y | (X, Y) \in \mathcal{S}) \tag{3.59}$$

$$= \frac{\mathrm{P}(X \leq x, Y \leq y, (X, Y) \in \mathcal{S})}{\mathrm{P}((X, Y) \in \mathcal{S})} \tag{3.60}$$

$$= \frac{\int_{u \leq x, v \leq y, (u,v) \in \mathcal{S}} f_{X,Y}(u, v) \; du \, dv}{\int_{(u,v)\in\mathcal{S}} f_{X,Y}(u, v) \; du \, dv}, \tag{3.61}$$

$$f_{X,Y|(X,Y)\in\mathcal{S}}(x, y) := \frac{\partial^2 F_{X,Y|(X,Y)\in\mathcal{S}}(x, y)}{\partial x \partial y}. \tag{3.62}$$

This definition only holds for events with nonzero probability. However, events of the form $\{X = x\}$ have probability equal to zero because the random variable is continuous. Indeed, the

range of X is uncountable, so the probability of almost every event $\{X = x\}$ must be zero, as otherwise the probability their union would be unbounded.

How can we characterize our uncertainty about Y given $X = x$ then? We define a **conditional pdf** that captures what we are trying to do in the limit and then integrate it to obtain a conditional cdf.

Definition 3.2.7 (Conditional pdf and cdf)**.** *If $F_{X,Y}$ is differentiable, then the conditional pdf of Y given X is defined as*

$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{if } f_X(x) > 0 \tag{3.63}$$

and is undefined otherwise.

The conditional cdf of Y given X is defined as

$$F_{Y|X}(y|x) := \int_{u=-\infty}^{y} f_{Y|X}(u|x)\, du \quad \text{if } f_X(x) > 0 \tag{3.64}$$

and is undefined otherwise.

We now justify this definition, beyond the analogy with (3.18). Assume that $f_X(x) > 0$. Let us write the definition of the conditional pdf in terms of limits. We have

$$f_X(x) = \lim_{\Delta_x \to 0} \frac{\mathrm{P}(x \le X \le x + \Delta_x)}{\Delta_x}, \tag{3.65}$$

$$f_{X,Y}(x,y) = \lim_{\Delta_x \to 0} \frac{1}{\Delta_x} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\partial y}. \tag{3.66}$$

This implies

$$\frac{f_{X,Y}(x,y)}{f_X(x)} = \lim_{\Delta_x \to 0, \Delta_y \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\partial y}. \tag{3.67}$$

We can now write the conditional cdf as

$$F_{Y|X}(y|x) = \int_{u=-\infty}^{y} \lim_{\Delta_x \to 0, \Delta_y \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le u)}{\partial y}\, \mathrm{d}u \tag{3.68}$$

$$= \lim_{\Delta_x \to 0} \frac{1}{\mathrm{P}(x \le X \le x + \Delta_x)} \int_{u=-\infty}^{y} \frac{\partial \mathrm{P}(x \le X \le x + \Delta_x, Y \le u)}{\partial y}\, \mathrm{d}u \tag{3.69}$$

$$= \lim_{\Delta_x \to 0} \frac{\mathrm{P}(x \le X \le x + \Delta_x, Y \le y)}{\mathrm{P}(x \le X \le x + \Delta_x)} \tag{3.70}$$

$$= \lim_{\Delta_x \to 0} \mathrm{P}(Y \le y | x \le X \le x + \Delta_x). \tag{3.71}$$

We can therefore interpret the conditional cdf as the limit of the cdf of Y at y conditioned on X belonging to an interval around x when the width of the interval tends to zero.

Remark 3.2.8. *Interchanging limits and integrals as in (3.69) is not necessarily justified in general. In this case it is, as long as the integral converges and the quantities involved are bounded.*

An immediate consequence of Definition 3.2.7 is the chain rule for continuous random variables.

Lemma 3.2.9 (Chain rule for continuous random variables)**.**

$$f_{X,Y}(x,y) = f_X(x)\, f_{Y|X}(y|x). \tag{3.72}$$

Applying the same ideas as in the bivariate case, we define the conditional distribution of a subvector given the rest of the random vector.

Definition 3.2.10 (Conditional pdf)**.** *The conditional pdf of a random subvector $\vec{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1,2,\ldots,n\}$, given the subvector $\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}$ is*

$$f_{\vec{X}_{\mathcal{I}}|\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\{1,\ldots,n\}/\mathcal{I}}\right) := \frac{f_{\vec{X}}(\vec{x})}{f_{\vec{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\vec{x}_{\{1,\ldots,n\}/\mathcal{I}}\right)}. \tag{3.73}$$

It is often useful to represent the joint pdf of a random vector by factoring it into conditional pdfs using the chain rule for random vectors.

Lemma 3.2.11 (Chain rule for random vectors)**.** *The joint pdf of a random vector $\vec{X}$ can be decomposed into*

$$f_{\vec{X}}(\vec{x}) = f_{\vec{X}_1}(\vec{x}_1)\, f_{\vec{X}_2|\vec{X}_1}(\vec{x}_2|\vec{x}_1) \ldots f_{\vec{X}_n|\vec{X}_1,\ldots,\vec{X}_{n-1}}(\vec{x}_n|\vec{x}_1,\ldots,\vec{x}_{n-1}) \tag{3.74}$$

$$= \prod_{i=1}^{n} f_{\vec{X}_i|\vec{X}_{\{1,\ldots,i-1\}}}\left(\vec{x}_i|\vec{x}_{\{1,\ldots,i-1\}}\right). \tag{3.75}$$

Note that the order is arbitrary, you can reorder the components of the vector in any way you like.

Proof. The result follows from applying the definition of conditional pdf recursively. □

Example 3.2.12 (Triangle lake (continued))**.** The biologist spots the otter from the shore of the lake. She is standing on the west side of the lake at a latitude of $x_1 = 0.75$ looking east and the otter is right in front of her. The otter is consequently also at a latitude of $x_1 = 0.75$, but she cannot tell at what distance. The distribution of the location of the otter given its latitude X_1 is characterized by the conditional pdf of the longitude X_2 given X_1,

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \tag{3.76}$$

$$= \frac{1}{1-x_1}, \quad 0 \leq x_2 \leq 1 - x_1. \tag{3.77}$$

The biologist is interested in the probability that the otter is closer than x_2 to her. This probability is given by the conditional cdf

$$F_{X_2|X_1}(x_2|x_1) = \int_{-\infty}^{x_2} f_{X_2|X_1}(u|x_1)\, \mathrm{d}u \tag{3.78}$$

$$= \frac{x_2}{1-x_1}. \tag{3.79}$$

The probability that the otter is less than x_2 away is $4x_2$ for $0 \leq x_2 \leq 1/4$.

△

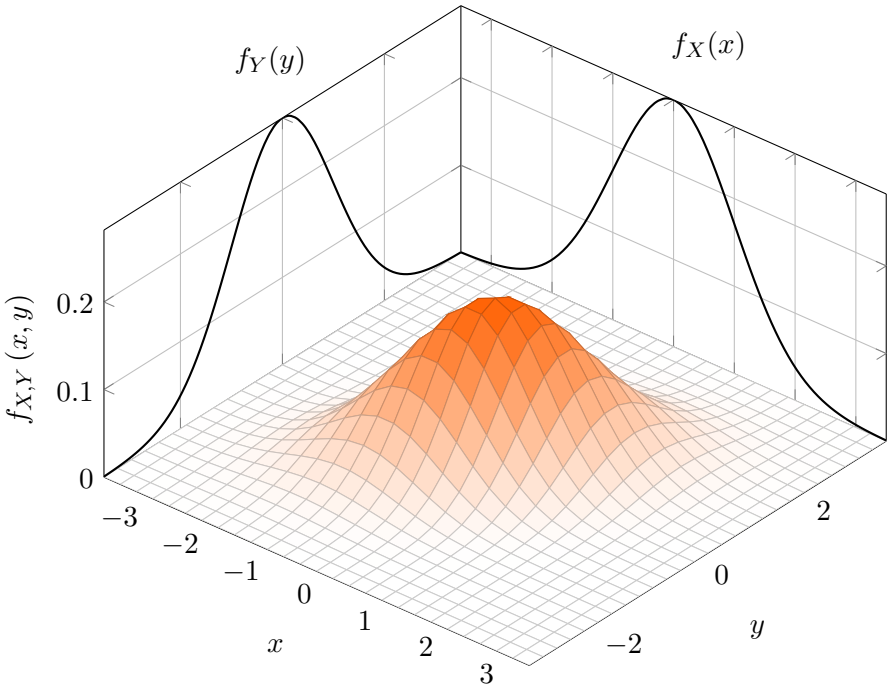


Figure 3.2: Joint pdf of a bivariate Gaussian random variable (X, Y) together with the marginal pdfs of X and Y.

3.2.4 Gaussian random vectors

Gaussian random vectors are a multidimensional generalization of Gaussian random variables. They are parametrized by a vector and a matrix that correspond to their mean and covariance matrix (we define these quantities for general multivariate random variables in Chapter 4).

Definition 3.2.13 (Gaussian random vector)**.** *A Gaussian random vector $\vec{X}$ is a random vector with joint pdf*

$$f_{\vec{X}}\left(\vec{x}\right) = \frac{1}{\sqrt{\left(2\pi\right)^n \left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)\right) \tag{3.80}$$

where the mean vector $\vec{\mu} \in \mathbb{R}^n$ and the covariance matrix Σ, which is symmetric and positive definite, parametrize the distribution. A Gaussian distribution with mean $\vec{\mu}$ and covariance matrix Σ is usually denoted by $\mathcal{N}\left(\vec{\mu}, \Sigma\right)$.

A fundamental property of Gaussian random vectors is that performing linear transformations on them always yields vectors with joint distributions that are also Gaussian. We will not prove this result formally, but the proof is similar to Lemma 2.5.1 (in fact this is a multidimensional generalization of that result).

Theorem 3.2.14 (Linear transformations of Gaussian random vectors are Gaussian)**.** *Let $\vec{X}$ be a Gaussian random vector of dimension n with mean $\vec{\mu}$ and covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$ $\vec{Y} = A\vec{X} + \vec{b}$ is a Gaussian random vector with mean $A\vec{\mu} + \vec{b}$ and covariance matrix $A\Sigma A^T$.*

A corollary of this result is that the joint pdf of a subvector of a Gaussian random vector is also a Gaussian vector.

Corollary 3.2.15 (Marginals of Gaussian random vectors are Gaussian)**.** *The joint pdf of any subvector of a Gaussian random vector is Gaussian. Without loss of generality, assume that the subvector $\vec{X}$ consists of the first m entries of the Gaussian random vector,*

$$\vec{Z} := \begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix}, \qquad \text{with mean} \quad \vec{\mu} := \begin{bmatrix} \mu_{\vec{X}} \\ \mu_{\vec{Y}} \end{bmatrix} \tag{3.81}$$

and covariance matrix

$$\Sigma_{\vec{Z}} = \begin{bmatrix} \Sigma_{\vec{X}} & \Sigma_{\vec{X}\vec{Y}} \\ \Sigma_{\vec{X}\vec{Y}}^T & \Sigma_{\vec{Y}} \end{bmatrix}. \tag{3.82}$$

Then $\vec{X}$ is a Gaussian random vector with mean $\mu_{\vec{X}}$ and covariance matrix $\Sigma_{\vec{X}}$.

Proof. Note that

$$\vec{X} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \begin{bmatrix} \vec{X} \\ \vec{Y} \end{bmatrix} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \vec{Z}, \tag{3.83}$$

where $I \in \mathbb{R}^{m\times m}$ is an identity matrix and $0_{c\times d}$ represents a matrix of zeros of dimensions $c \times d$. The result then follows from Theorem 3.2.14. □

Figure 3.2 shows the joint pdf of a bivariate Gaussian random variable along with its marginal pdfs.

3.3 Joint distributions of discrete and continuous variables

Probabilistic models often include both discrete and continuous random variables. However, the joint pmf or pdf of a discrete and a continuous random variable is not well defined. In order to specify the joint distribution in such cases we use their marginal and conditional pmfs and pdfs.

Assume that we have a continuous random variable C and a discrete random variable D with range R_D. We define the conditional cdf and pdf of C given D as follows.

Definition 3.3.1 (Conditional cdf and pdf of a continuous random variable given a discrete random variable)**.** *Let C and D be a continuous and a discrete random variable defined on the same probability space. Then, the conditional cdf and pdf of C given D are of the form*

$$F_{C|D}(c|d) := \mathrm{P}(C \leq c|d), \tag{3.84}$$

$$f_{C|D}(c|d) := \frac{\mathrm{d}F_{C|D}(c|d)}{\mathrm{d}c}. \tag{3.85}$$

We obtain the marginal cdf and pdf of C from the conditional cdfs and pdfs by computing a weighted sum.

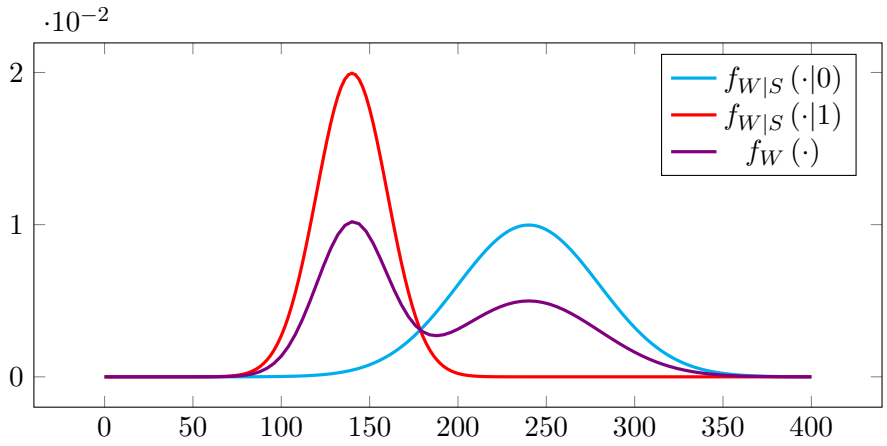


Figure 3.3: Conditional and marginal distributions of the weight of the bears W in Example 3.3.3.

Lemma 3.3.2. *Let $F_{C|D}$ and $f_{C|D}$ be the conditional cdf and pdf of a continuous random variable C given a discrete random variable D. Then,*

$$F_C(c) = \sum_{d \in R_D} p_D(d) F_{C|D}(c|d), \tag{3.86}$$

$$f_C(c) = \sum_{d \in R_D} p_D(d) f_{C|D}(c|d). \tag{3.87}$$

Proof. The events $\{D = d\}$ are a partition of the whole probability space (one of them must happen and they are all disjoint), so

$$F_C(c) = \mathrm{P}(C \leq c) \tag{3.88}$$

$$= \sum_{d \in R_D} \mathrm{P}(D = d)\, \mathrm{P}(C \leq c|d) \quad \text{by the Law of Total Probability} \tag{3.89}$$

$$= \sum_{d \in R_D} p_D(d) F_{C|D}(c|d). \tag{3.90}$$

Now, (3.87) follows by differentiating. □

Combining a discrete marginal pmf with a continuous conditional distribution allows us to define **mixture models** where the data is drawn from a continuous distribution whose parameters are chosen from a discrete set. If a Gaussian is used as the continuous distribution, this yields a Gaussian mixture model. Fitting Gaussian mixture models is a popular technique for clustering data.

Example 3.3.3 (Grizzlies in Yellowstone)**.** A scientist is gathering data on the bears in Yellowstone. It turns out that the weight of the males is well modeled by a Gaussian random variable with mean 240 kg and standard variation 40 kg, whereas the weight of the females is well modeled by a Gaussian with mean 140 kg and standard deviation 20 kg. There are about the same number of females and males.

The distribution of the weights of all the grizzlies can consequently be modeled by a Gaussian mixture that includes a continuous random variable W to represent the weight and a discrete random variable S to represent the sex of the bears. S is Bernoulli with parameter $1/2$, W given $S = 0$ (male) is $\mathcal{N}(240, 1600)$ and W given $S = 1$ (female) is $\mathcal{N}(140, 400)$. By (3.87) the pdf of W is consequently of the form

$$f_W(w) = \sum_{s=0}^{1} p_S(s) f_{W|S}(w|s) \tag{3.91}$$

$$= \frac{1}{2\sqrt{2\pi}} \left(\frac{e^{-\frac{(w-240)^2}{3200}}}{40} + \frac{e^{-\frac{(w-140)^2}{800}}}{20} \right). \tag{3.92}$$

Figure 3.3 shows the conditional and marginal distributions of W.

△

Defining the conditional pmf of a discrete random variable D given a continuous random variable C is challenging because the probability of the event $\{C = c\}$ is zero. We follow the same approach as in Definition 3.2.7 and define the conditional pmf as a limit.

Definition 3.3.4 (Conditional pmf of a discrete random variable given a continuous random variable)**.** *Let C and D be a continuous and a discrete random variable defined on the same probability space. Then, the conditional pmf of D given C is defined as*

$$p_{D|C}(d|c) := \lim_{\Delta \to 0} \frac{\mathrm{P}(D = d, c \leq C \leq c + \Delta)}{\mathrm{P}(c \leq C \leq c + \Delta)}. \tag{3.93}$$

Analogously to Lemma 3.3.2, we obtain the marginal pmf of D from the conditional pmfs by computing a weighted sum.

Lemma 3.3.5. *Let $p_{D|C}$ be the conditional pmf of a discrete random variable D given a continuous random variable C. Then,*

$$p_D(d) = \int_{c=-\infty}^{\infty} f_C(c) p_{D|C}(d|c) \, dc. \tag{3.94}$$

Proof. We will not give a formal proof but rather an intuitive argument that can be made rigorous. If we take a grid of values for c which are on a grid $\ldots, c_{-1}, c_0, c_1, \ldots$ of width Δ, then

$$p_D(d) = \sum_{i=-\infty}^{\infty} \mathrm{P}(D = d, c_i \leq C \leq c_i + \Delta) \tag{3.95}$$

by the Law of Total probability. Taking the limit as $\Delta \to 0$ the sum becomes an integral and we have

$$p_D(d) = \int_{c=-\infty}^{\infty} \lim_{\Delta \to 0} \frac{\mathrm{P}(D = d, c \leq C \leq c + \Delta)}{\Delta} \mathrm{d}c \tag{3.96}$$

$$= \int_{c=-\infty}^{\infty} \lim_{\Delta \to 0} \frac{\mathrm{P}(c \leq C \leq c + \Delta)}{\Delta} \cdot \frac{\mathrm{P}(D = d, c \leq C \leq c + \Delta)}{\mathrm{P}(c \leq C \leq c + \Delta)} \mathrm{d}c \tag{3.97}$$

$$= \int_{c=-\infty}^{\infty} f_C(c) p_{D|C}(d|c) \, \mathrm{d}c. \tag{3.98}$$

since $f_C(c) = \lim_{\Delta \to 0} \frac{\mathrm{P}(c \leq C \leq c + \Delta)}{\Delta}$. □

Combining continuous marginal distributions with discrete conditional distributions is particularly useful in Bayesian statistical models, as illustrated in the following example (see Chapter 10 for more information). The continuous distribution is used to quantify our uncertainty about the parameter of a discrete distribution.

Example 3.3.6 (Bayesian coin flip)**.** Your uncle bets you ten dollars that a coin flip will turn out heads. You suspect that the coin is biased, but you are not sure to what extent. To model this uncertainty you represent the bias as a continuous random variable B with the following pdf:

$$f_B(b) = 2b \quad \text{for } b \in [0,1] \ . \tag{3.99}$$

You can now compute the probability that the coin lands on heads denoted by X using Lemma 3.3.5. Conditioned on the bias B, the result of the coin flip is Bernoulli with parameter B.

$$\begin{aligned} p_X(1) &= \int_{b=-\infty}^{\infty} f_B(b)\, p_{X|B}(1|b)\ \mathrm{d}b & (3.100)\\ &= \int_{b=0}^{1} 2b^2\, \mathrm{d}b & (3.101)\\ &= \frac{2}{3}. & (3.102) \end{aligned}$$

According to your model the probability that the coin lands heads is $2/3$. △

The following lemma provides an analogue to the chain rule for jointly distributed continuous and discrete random variables.

Lemma 3.3.7 (Chain rule for jointly distributed continuous and discrete random variables)**.** *Let C be a continuous random variable with conditional pdf $f_{C|D}$ and D a discrete random variable with conditional pmf $p_{D|C}$. Then,*

$$p_D(d)\, f_{C|D}(c|d) = f_C(c)\, p_{D|C}(d|c)\,. \tag{3.103}$$

Proof. Applying the definitions,

$$\begin{aligned} p_D(d)\, f_{C|D}(c|d) &= \lim_{\Delta\to 0} \mathrm{P}(D=d)\, \frac{\mathrm{P}(c\le C\le c+\Delta|D=d)}{\Delta} & (3.104)\\ &= \lim_{\Delta\to 0} \frac{\mathrm{P}(D=d, c\le C\le c+\Delta)}{\Delta} & (3.105)\\ &= \lim_{\Delta\to 0} \frac{\mathrm{P}(c\le C\le c+\Delta)}{\Delta} \cdot \frac{\mathrm{P}(D=d, c\le C\le c+\Delta)}{\mathrm{P}(c\le C\le c+\Delta)} & (3.106)\\ &= f_C(c)\, p_{D|C}(d|c)\,. & (3.107) \end{aligned}$$

□

Example 3.3.8 (Grizzlies in Yellowstone (continued))**.** The scientist observes a bear with her binoculars. From their size she estimates that its weight is 180 kg. What is the probability that the bear is male?

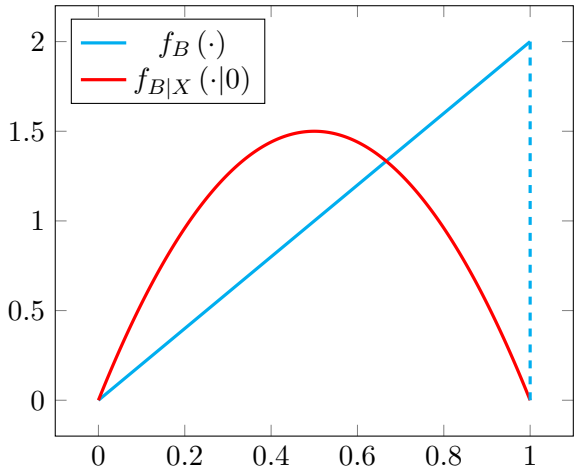


Figure 3.4: Conditional and marginal distributions of the bias of the coin flip in Example 3.3.9.

We apply Lemma 3.3.7 to compute

$$p_{S|W}(0|180) = \frac{p_S(0)\, f_{W|S}(180|0)}{f_W(180)} \tag{3.108}$$

$$= \frac{\frac{1}{40}\exp\left(-\frac{60^2}{3200}\right)}{\frac{1}{40}\exp\left(-\frac{60^2}{3200}\right) + \frac{1}{20}\exp\left(-\frac{40^2}{800}\right)} \tag{3.109}$$

$$= 0.545. \tag{3.110}$$

According to the probabilistic model, the probability that it's a male is 0.545.

△

Example 3.3.9 (Bayesian coin flip (continued))**.** The coin lands on tails. You decide to recompute the distribution of the bias conditioned on this information. By Lemma 3.3.7

$$f_{B|X}(b|0) = \frac{f_B(b)\, p_{X|B}(0|b)}{p_X(0)} \tag{3.111}$$

$$= \frac{2b(1-b)}{1/3} \tag{3.112}$$

$$= 6b(1-b). \tag{3.113}$$

Conditioned on the outcome, the pdf of the bias is now centered instead of concentrated near one as before, as shown in Figure 3.4.

△

3.4 Independence

In this section we define independence and conditional independence for random variables and vectors.

3.4.1 Definition

When knowledge about a random variable X does not affect our uncertainty about another random variable Y, we say that X and Y are **independent**. Formally, this is reflected by the marginal and conditional cdf and the conditional pmf or pdf which must be equal, i.e.

$$F_Y(y) = F_{Y|X}(y|x) \tag{3.114}$$

and

$$p_Y(y) = p_{Y|X}(y|x) \quad \text{or} \quad f_Y(y) = f_{Y|X}(y|x), \tag{3.115}$$

depending on whether the variable is discrete or continuous, for any x and any y for which the conditional distributions are well defined. Equivalently, the joint cdf and the conditional pmf or pdf factors into the marginals.

Definition 3.4.1 (Independent random variables)**.** *Two random variables X and Y are independent if and only if*

$$F_{X,Y}(x,y) = F_X(x) F_Y(y), \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{3.116}$$

If the variables are discrete, the following condition is equivalent

$$p_{X,Y}(x,y) = p_X(x) p_Y(y), \quad \text{for all } x \in R_X, y \in R_Y. \tag{3.117}$$

If the variables are continuous have joint and marginal pdfs, the following condition is equivalent

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{3.118}$$

We now extend the definition to account for several random variables (or equivalently several entries in a random vector) that do not provide information about each other.

Definition 3.4.2 (Independent random variables)**.** *The n entries X_1, X_2, ..., X_n in a random vector $\vec{X}$ are independent if and only if*

$$F_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} F_{X_i}(\vec{x}_i), \tag{3.119}$$

which is equivalent to

$$p_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} p_{X_i}(\vec{x}_i) \tag{3.120}$$

for discrete vectors and

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} f_{X_i}(\vec{x}_i) \tag{3.121}$$

for continuous vectors, if the joint pdf exists.

The following example shows that pairwise independence does *not imply* independence.

Example 3.4.3 (Pairwise independence does not imply joint independence)**.** Let X_1 and X_2 be the outcomes of independent unbiased coin flips. Let X_3 be the indicator of the event $\{X_1 \text{ and } X_2 \text{ have the same outcome}\}$,

$$X_3 = \begin{cases} 1 & \text{if } X_1 = X_2, \\ 0 & \text{if } X_1 \neq X_2. \end{cases} \tag{3.122}$$

The pmf of X_3 is

$$p_{X_3}(1) = p_{X_1,X_2}(1,1) + p_{X_1,X_2}(0,0) = \frac{1}{2}, \tag{3.123}$$

$$p_{X_3}(0) = p_{X_1,X_2}(0,1) + p_{X_1,X_2}(1,0) = \frac{1}{2}. \tag{3.124}$$

X_1 and X_2 are independent by assumption. X_1 and X_3 are independent because

$$p_{X_1,X_3}(0,0) = p_{X_1,X_2}(0,1) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(0), \tag{3.125}$$

$$p_{X_1,X_3}(1,0) = p_{X_1,X_2}(1,0) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(0), \tag{3.126}$$

$$p_{X_1,X_3}(0,1) = p_{X_1,X_2}(0,0) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(1), \tag{3.127}$$

$$p_{X_1,X_3}(1,1) = p_{X_1,X_2}(1,1) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(1). \tag{3.128}$$

X_2 and X_3 are independent too (the reasoning is the same).

However, are X_1, X_2 and X_3 all independent?

$$p_{X_1,X_2,X_3}(1,1,1) = \mathrm{P}(X_1 = 1, X_2 = 1) = \frac{1}{4} \neq p_{X_1}(1)\, p_{X_2}(1)\, p_{X_3}(1) = \frac{1}{8}. \tag{3.129}$$

They are not, which makes sense since X_3 is a function of X_1 and X_2. △

Conditional independence indicates that two random variables do not depend on each other, as long as an additional random variable is known.

Definition 3.4.4 (Conditionally independent random variables)**.** *Two random variables X and Y are independent with respect to another random variable Z if and only if*

$$F_{X,Y\,|\,Z}(x,y\,|\,z) = F_{X\,|\,Z}(x\,|\,z)\, F_{Y\,|\,Z}(y\,|\,z), \quad \textit{for all } (x,y) \in \mathbb{R}^2, \tag{3.130}$$

and any z for which the conditional cdfs are well defined. If the variables are discrete, the following condition is equivalent

$$p_{X,Y\,|\,Z}(x,y\,|\,z) = p_{X\,|\,Z}(x\,|\,z)\, p_{Y\,|\,Z}(y\,|\,z), \quad \textit{for all } x \in R_X, y \in R_Y, \tag{3.131}$$

and any z for which the conditional pmfs are well defined. If the variables are continuous have joint and marginal pdfs, the following condition is equivalent

$$f_{X,Y\,|\,Z}(x,y\,|\,z) = f_{X\,|\,Z}(x\,|\,z)\, f_{Y\,|\,Z}(y\,|\,z), \quad \textit{for all } (x,y) \in \mathbb{R}^2, \tag{3.132}$$

and any z for which the conditional pmfs are well defined.

The definition can be extended to condition on several random variables.

Definition 3.4.5 (Conditionally independent random variables)**.** *The components of a subvector $\vec{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ are conditionally independent given another subvector $\vec{X}_{\mathcal{J}}$, $\mathcal{J} \subseteq \{1, 2, \ldots, n\}$, if and only if*

$$F_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} F_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right), \tag{3.133}$$

which is equivalent to

$$p_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} p_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right) \tag{3.134}$$

for discrete vectors and

$$f_{\vec{X}_{\mathcal{I}}|\vec{X}_{\mathcal{J}}}\left(\vec{x}_{\mathcal{I}}|\vec{x}_{\mathcal{J}}\right) = \prod_{i\in\mathcal{I}} f_{X_i|\vec{X}_{\mathcal{J}}}\left(\vec{x}_i|\vec{x}_{\mathcal{J}}\right) \tag{3.135}$$

for continuous vectors if the conditional joint pdf exists.

As established in Examples 1.3.5 and 1.3.6, independence does **not** imply conditional independence or vice versa.

3.4.2 Variable dependence in probabilistic modeling

A fundamental consideration when designing a probabilistic model is the dependence between the different variables, i.e. what variables are independent or conditional independent from each other. Although it may sound surprising, if the number of variables is large, introducing some independence assumptions may be necessary to make the model tractable, even if we know that all the variables are dependent. To illustrate this, consider a model for the US presidential election where there are 50 random variables, each representing a state. If the variables only take two possible values (representing what candidate wins that state), the joint pmf of their distribution has $2^{50} - 1 \geq 10^{15}$ degrees of freedom. We wouldn't be able to store the pmf with all the computer memory in the world! In contrast, if we assume that all the variables are independent, then the distribution only has 50 free parameters. Of course, this is not necessarily a good idea because failing to represent dependencies may severely affect the prediction accuracy of a model, as illustrated in Example 3.5.1 below. Striking a balance between tractability and accuracy is a crucial challenge in probabilistic modeling.

We now illustrate how the dependence structure of the random variables in a probabilistic model can be exploited to reduce the number of parameters describing the distribution through an appropriate factorization of their joint pmf or pdf. Consider three Bernoulli random variables A, B and C. In general, we need $7 = 2^3 - 1$ parameters to describe the pmf. However, if B and C are conditionally independent given A we can perform the following factorization

$$p_{A,B,C} = p_A \, p_{B\,|\,A} \, p_{C\,|\,A} \tag{3.136}$$

which only depends on five parameters (one for p_A and two each for $p_{B\,|\,A}$ and $p_{C\,|\,B}$). It is important to note that there are many other possible factorizations that do not exploit the dependence assumptions, such as for example

$$p_{A,B,C} = p_B \, p_{A\,|\,B} \, p_{C\,|\,A,B}. \tag{3.137}$$

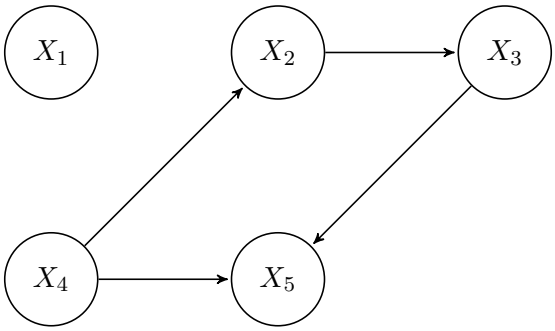


Figure 3.5: Example of a directed acyclic graphic representing a probabilistic model.

For large probabilistic models it is crucial to find factorizations that reduce the number of parameters as much as possible.

3.4.3 Graphical models

Graphical models are a tool for characterizing the dependence structure of probabilistic models. In this section we give a brief description of **directed** graphical models, which are also called Bayesian networks. Undirected graphical models, known as Markov random fields, are out of the scope of these notes. We refer the interested reader to more advanced texts in probabilistic modeling and machine learning for a more in-depth treatment of graphical models.

Directed acyclic graphs, known as DAGs, can be interpreted as diagrams representing a factorization of the joint pmf or pdf of a probabilistic model. In order to specify a valid factorization, the graphs are constrained to not have any cycles (hence the term acyclic). Each node in the DAG represents a random variable. The edges between the nodes indicate the dependence between the variables. The factorization corresponding to a DAG contains:

- The marginal pmf or pdf of the variables corresponding to all nodes with no incoming edges.
- The conditional pmf or pdf of the remaining random variables given their *parents*. A is a parent of B if there is a directed edge from (the node assigned to) A to (the node assigned to) B.

To be concrete, consider the DAG in Figure 3.5. For simplicity we denote each node using the corresponding random variable and assume that they are all discrete. Nodes X_1 and X_4 have no parents, so the factorization of the joint pmf includes their marginal pmfs. Node X_2 only descends from X_4 so we include $p_{X_2 \,|\, X_4}$. Node X_3 descends from X_2 so we include $p_{X_3 \,|\, X_2}$. Finally, node X_5 descends from X_3 and X_4 so we include $p_{X_5 \,|\, X_3, X_4}$. The factorization is of the form

$$p_{X_1,X_2,X_3,X_4,X_5} = p_{X_1} \, p_{X_4} \, p_{X_2 \,|\, X_4} \, p_{X_3 \,|\, X_2} \, p_{X_5 \,|\, X_3,X_4}. \tag{3.138}$$

This factorization reveals some dependence assumptions. By the chain rule another valid factorization of the joint pmf is

$$p_{X_1,X_2,X_3,X_4,X_5} = p_{X_1} \, p_{X_4 \,|\, X_1} \, p_{X_2 \,|\, X_1,X_4} \, p_{X_3 \,|\, X_1,X_2,X_4} \, p_{X_5 \,|\, X_1,X_2,X_3,X_4}. \tag{3.139}$$

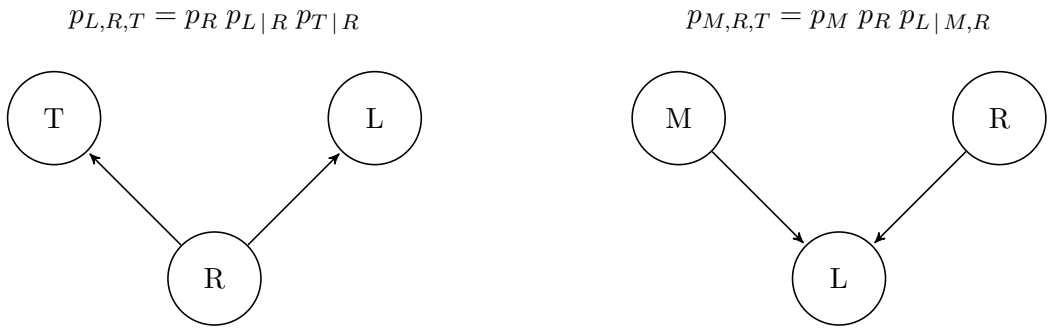


Figure 3.6: Directed graphical models corresponding to the variables in Examples 1.3.5 and 1.3.6.

Comparing both expressions, we see that X_1 and all the other variables are independent, since $p_{X_4\,|\,X_1} = p_{X_4}$, $p_{X_2\,|\,X_1,X_4} = p_{X_2\,|\,X_4}$ and so on. In addition, X_3 is conditionally independent of X_4 given X_2 since $p_{X_3\,|\,X_2,X_4} = p_{X_3\,|\,X_2}$. These dependence assumptions can be read directly from the graph, using the following property.

Theorem 3.4.6 (Local Markov property)**.** *The factorization of the joint pmf or pdf represented by a DAG satisfies the local Markov property: each variable is conditionally independent of its non-descendants given all its parent variables. In particular, if it has no parents, it is independent of its non-descendants. To be clear, B is a non-descendant of A if there is no directed path from A to B.*

Proof. Let X_i be an arbitrary variable. We denote by X_N the set of non-descendants of X_i, by X_P the set of parents and by X_D the set of descendants. The factorization represented by the graphical model is of the form

$$p_{X_1,\ldots,X_n} = p_{X_N}\, p_{X_P|X_N}\, p_{X_i|X_P}\, p_{X_D|X_i}. \tag{3.140}$$

By the chain rule another valid factorization is

$$p_{X_1,\ldots,X_n} = p_{X_N}\, p_{X_P|X_N}\, p_{X_i|X_P,X_N}\, p_{X_D|X_i,X_P,X_N}. \tag{3.141}$$

Comparing both expressions we conclude that $p_{X_i|X_P,X_N} = p_{X_i|X_P}$ so X_i is conditionally independent of X_N given X_P. □

We illustrate these ideas by showing the DAGs for Examples 1.3.5 and 1.3.6.

Example 3.4.7 (Graphical model for Example 1.3.5)**.** We model the different events in Example 1.3.5 using indicator random variables. T represents whether a taxi is available ($T = 1$) or not ($T = 0$), L whether the plane is late ($L = 1$) or not ($L = 0$), and R whether it rains ($R = 1$) or not ($R = 0$). In the example, T and L are conditionally independent given R. We can represent the corresponding factorization using the graph on the left of Figure 3.6.

△

Example 3.4.8 (Graphical model for Example 1.3.6)**.** We model the different events in Example 1.3.6 using indicator random variables. M represents whether a mechanical problem occurs

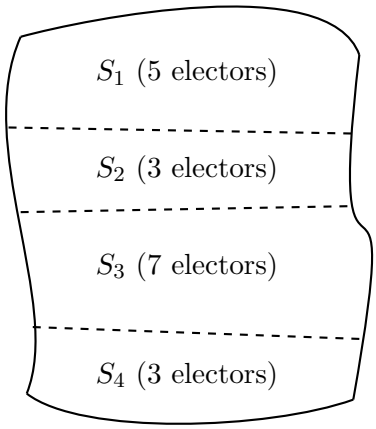


Figure 3.7: Fictitious country considered in Example 3.4.9.

($M = 1$) or not ($M = 0$) and L and R are the same as in Example 3.4.7. In the example, M and R are independent, but L depends on both of them. We can represent the corresponding factorization using the graph on the right of Figure 3.6.

△

The following example that introduces an important class of graphical models called Markov chains, which we will discuss at length in Chapter 7.

Example 3.4.9 (Election)**.** In the country shown in Figure 3.7 the presidential election follows the same system as in the United States. Citizens cast ballots for *electors* in the Electoral College. Each state is entitled to a number of electors (in the US this is usually the same as the members of Congress). In every state, the electors are pledged to the candidate that wins the state. Our goal is to model the election probabilistically. We assume that there are only two candidates A and B. Each state is represented by a random variable S_i, $1 \leq i \leq 4$,

$$S_i = \begin{cases} 1 & \text{if candidate A wins state } i, \\ -1 & \text{if candidate B wins state } i. \end{cases} \tag{3.142}$$

An important decision to make is what independence assumptions to assume about the model. Figure 3.8 shows three different options. If we model each state as independent, then we only need to estimate a single parameter for each state. However, the model may not be accurate, as the outcome in states with similar demographics is bound to be related. Another option is to estimate the full joint pmf. The problem is that it may be quite challenging to compute the parameters. We can estimate the marginal pmfs of the individual states using poll data, but conditional probabilities are more difficult to estimate. In addition, for larger models it is not tractable to consider fully dependent models (for instance in the case of the US election, as mentioned previously). A reasonable compromise could be to model the states that are not adjacent as conditionally independent given the states between them. For example, we assume that the outcome of states 1 and 3 are only related through state 2. The corresponding graphical model, depicted on the right of Figure 3.8, is called a Markov chain. It corresponds

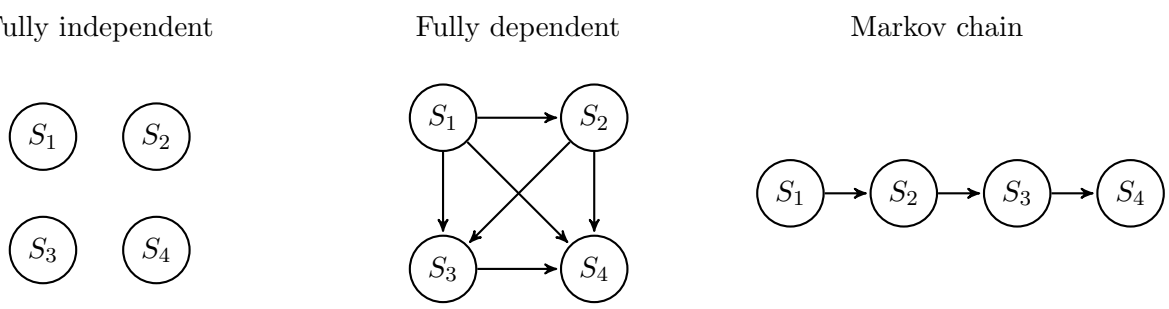


Figure 3.8: Graphical models capturing different assumptions about the distribution of the random variables considered in Example 3.4.9.

to a factorization of the form

$$p_{S_1,S_2,S_3,S_4} = p_{S_1}\, p_{S_2\,|\,S_1}\, p_{S_3\,|\,S_2}\, p_{S_4\,|\,S_3}. \tag{3.143}$$

Under this model we only need to worry about estimating pairwise conditional probabilities, as opposed to the full joint pmf. We discuss Markov chains at length in Chapter 7.

△

We conclude the section with an example involving continuous variables.

Example 3.4.10 (Desert)**.** Dani and Felix are traveling through the desert in Arizona. They become concerned that their car might break down and decide to build a probabilistic model to evaluate the risk. They model the time until the car breaks down as an exponential random variable T with a parameter that depends on the state of the motor M and the state of the road R. These three quantities are represented by random variables in the same probability space.

Unfortunately they have no idea what the state of the motor is so they assume that it is uniform between 0 (no problem with the motor) and 1 (the motor is almost dead). Similarly, they have no information about the road, so they also assume that its state is a uniform random variable between 0 (no problem with the road) and 1 (the road is terrible). In addition, they assume that the states of the road and the car are independent and that the parameter of the exponential random variable that represents the time in hours until there is a breakdown is equal to $M + R$. The corresponding graphical model is shown in Figure 3.9

To find the joint distribution of the random variables, we apply the chain rule to obtain,

$$\begin{aligned} f_{M,R,T}\,(m,r,t) &= f_M\,(m)\, f_{R|M}\,(r|m)\, f_{T|M,R}\,(t|m,r) && (3.144)\\ &= f_M\,(m)\, f_R\,(r)\, f_{T|M,R}\,(t|m,r) \quad \text{(by independence of } M \text{ and } R) && (3.145)\\ &= \begin{cases} (m+r)\, e^{-(m+r)t} & \text{for } t \geq 0,\ 0 \leq m \leq 1,\ 0 \leq r \leq 1,\\ 0 & \text{otherwise.} \end{cases} && (3.146) \end{aligned}$$

Note that we start with M and R because we know their marginal distribution, whereas we only know the conditional distribution of T given M and R.

After 15 minutes, the car breaks down. The road seems OK, about a 0.2 in the scale they defined for the value of R, so they naturally wonder about the state of the motor. Given their

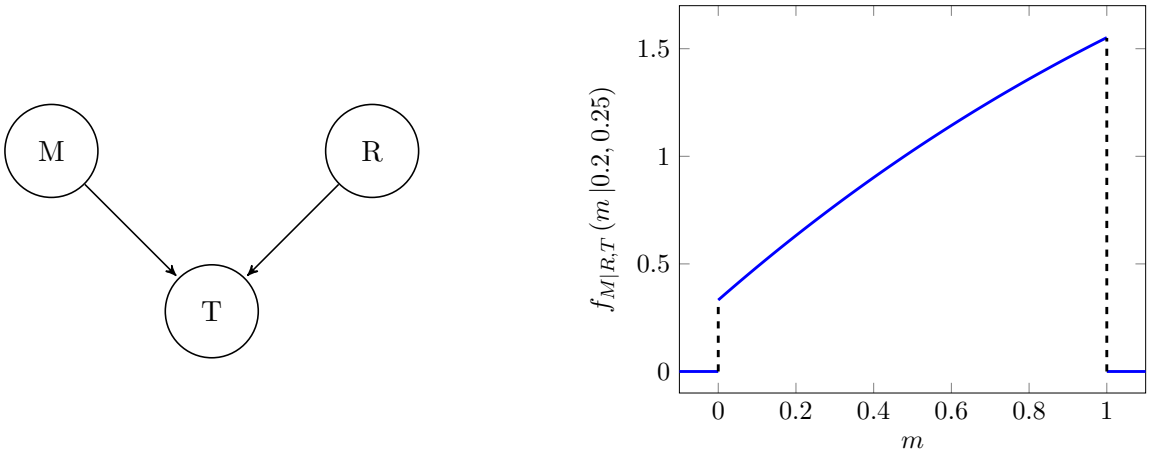


Figure 3.9: The left image is a graphical model representing the random variables in Example 3.4.10. The right plot shows the conditional pdf of M given $T = 0.25$ and $R = 0.2$.

probabilistic model, their uncertainty about the motor given all of this information is captured by the conditional distribution of M given T and R.

To compute the conditional pdf, we first need to compute the joint marginal distribution of T and R by marginalizing over M. In order to simplify the computations, we use the following simple lemma.

Lemma 3.4.11. *For any constant $c > 0$,*

$$\int_0^1 e^{-cx}\,dx = \frac{1 - e^{-c}}{c}, \tag{3.147}$$

$$\int_0^1 xe^{-cx}\,dx = \frac{1 - (1 + c)\,e^{-c}}{c^2}. \tag{3.148}$$

Proof. Equation (3.147) is obtained using the antiderivative of the exponential function (itself), whereas integrating by parts yields (3.148). △

We have

$$f_{R,T}(r, t) = \int_{m=0}^{1} f_{M,R,T}(m, r, t)\,\mathrm{d}m \tag{3.149}$$

$$= e^{-tr}\left(\int_{m=0}^{1} me^{-tm}\,\mathrm{d}m + r\int_{m=0}^{1} e^{-tm}\,\mathrm{d}m\right) \tag{3.150}$$

$$= e^{-tr}\left(\frac{1 - (1 + t)\,e^{-t}}{t^2} + \frac{r\left(1 - e^{-t}\right)}{t}\right) \quad \text{by (3.147) and (3.148)} \tag{3.151}$$

$$= \frac{e^{-tr}}{t^2}\left(1 + tr - e^{-t}\left(1 + t + tr\right)\right), \tag{3.152}$$

for $t \geq 0,\ 0 \leq r \leq 1$.

The conditional pdf of M given T and R is

$$f_{M|R,T}(m|r,t) = \frac{f_{M,R,T}(m,r,t)}{f_{R,T}(r,t)} \tag{3.153}$$

$$= \frac{(m+r)\,e^{-(m+r)t}}{\frac{e^{-tr}}{t^2}\left(1+tr-e^{-t}\left(1+t+tr\right)\right)} \tag{3.154}$$

$$= \frac{(m+r)\,t^2 e^{-tm}}{1+tr-e^{-t}\left(1+t+tr\right)}, \tag{3.155}$$

for $t \geq 0,\ 0 \leq m \leq 1,\ 0 \leq r \leq 1$. Plugging in the observed values, the conditional pdf is equal to

$$f_{M|R,T}(m|0.2,0.25) = \frac{(m+0.2)\,0.25^2 e^{-0.25m}}{1+0.25\cdot 0.2-e^{-0.25}\left(1+0.25+0.25\cdot 0.2\right)} \tag{3.156}$$

$$= 1.66\,(m+0.2)\,e^{-0.25m}. \tag{3.157}$$

for $0 \leq m \leq 1$ and to zero otherwise. The pdf is plotted in Figure 3.9. According to the model, it seems quite likely that the state of the motor was not good. △

3.5 Functions of several random variables

The pmf of a random variable $Y := g(X_1, \ldots, X_n)$ defined as a function $g : \mathbb{R}^n \to \mathbb{R}$ of several discrete random variables $X_1, \ldots, X_n$ is given by

$$p_Y(y) = \sum_{y=g(x_1,\ldots,x_n)} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n). \tag{3.158}$$

This follows directly from (3.11). In words, the probability that $g(X_1, \ldots, X_n) = y$ is the sum of the joint pmf over all possible values such that $y = g(x_1, \ldots, x_n)$.

Example 3.5.1 (Election)**.** In Example 3.4.9 we discussed several possible models for a presidential election for a country with four states. Imagine that you are trying to predict the result of the election using poll data from individual states. The goal is to predict the outcome of the election, represented by the random variable

$$O := \begin{cases} 1 & \text{if } \sum_{i=1}^{4} n_i\, S_i > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.159}$$

where n_i denotes the number of electors in state i (notice that the sum can never be zero).

From analyzing the poll data you conclude that the probability that candidate A wins each of the states is 0.15. If you assume that all the states are independent, this is enough to characterize the joint pmf. Table 3.2 lists the probability of all possible outcomes for this model. By (3.158) we only need to add up the outcomes for which $O = 1$. Under the full-independence assumption, the probability that candidate A wins is 6%.

You are not satisfied by the result because you suspect that the outcomes in different states are highly dependent. From past elections, you determine that the conditional probability of a

S_1	S_2	S_3	S_4	O	Prob. (indep.)	Prob. (Markov)
-1	-1	-1	-1	0	0.5220	0.6203
-1	-1	-1	1	0	0.0921	0.0687
-1	-1	1	-1	0	0.0921	0.0431
-1	-1	1	1	1	0.0163	0.0332
-1	1	-1	-1	0	0.0921	0.0431
-1	1	-1	1	0	0.0163	0.0048
-1	1	1	-1	1	0.0163	0.0208
-1	1	1	1	1	0.0029	0.0160
1	-1	-1	-1	0	0.0921	0.0687
1	-1	-1	1	0	0.0163	0.0077
1	-1	1	-1	1	0.0163	0.0048
1	-1	1	1	1	0.0029	0.0037
1	1	-1	-1	0	0.0163	0.0332
1	1	-1	1	1	0.0029	0.0037
1	1	1	-1	1	0.0029	0.0160
1	1	1	1	1	0.0005	0.0123

Table 3.2: Table of auxiliary values for Example 3.5.1.

candidate winning a state if they win an adjacent state is indeed very high. You incorporate your estimate of the conditional probabilities into a Markov-chain model described by (3.143):

$$p_{S_1}(1) = 0.15, \tag{3.160}$$

$$p_{S_{i+1}\,|\,S_i}(1\,|\,1) = 0.435, \quad 2 \leq i \leq 4, \tag{3.161}$$

$$p_{S_{i+1}\,|\,S_i}(-1\,|\,-1) = 0.900 \quad 2 \leq i \leq 4. \tag{3.162}$$

This means that if candidate B wins a state, they are very likely to win the adjacent one. If candidate A wins a state, their chance to win an adjacent state is significantly higher than if they don't (but still lower than candidate B). Under this model the marginal probability that candidate A wins each state is still 0.15. Table 3.2 lists the probability of all possible outcomes. The probability that candidate A wins is now 11%, almost double the probability than that obtained under the fully-independent model. This illustrates the danger of not accounting for dependencies between states, which for example may have been one of the reasons why many forecasts severely underestimated Donald Trump's chances in the 2016 election.

△

Section 2.5 explains how to derive the distribution of functions of univariate random variables by first computing their cdf and then differentiating it to obtain their pdf. This directly extends to multivariable random functions. Let X, Y be random variables defined on the same probability

space, and let $U = g(X,Y)$ and $V = h(X,Y)$ for two arbitrary functions $g, h : \mathbb{R}^2 \to \mathbb{R}$. Then,

$$F_{U,V}(u,v) = \mathrm{P}(U \leq u, V \leq v) \tag{3.163}$$
$$= \mathrm{P}(g(X,Y) \leq u, h(X,Y) \leq v) \tag{3.164}$$
$$= \int_{\{(x,y) \mid g(x,y) \leq u, h(x,y) \leq v\}} f_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y, \tag{3.165}$$

where the last equality only holds if the joint pdf of X and Y exists. The joint pdf can then be obtained by differentiation.

Theorem 3.5.2 (Pdf of the sum of two independent random variables)**.** *The pdf of $Z = X + Y$, where X and Y are independent random variables is equal to the* ***convolution*** *of their respective pdfs f_X and f_Y,*

$$f_Z(z) = \int_{u=-\infty}^{\infty} f_X(z-u)\, f_Y(u)\ du. \tag{3.166}$$

Proof. First we derive the cdf of Z

$$F_Z(z) = \mathrm{P}(X + Y \leq z) \tag{3.167}$$
$$= \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{z-y} f_X(x)\, f_Y(y)\, \mathrm{d}x\, \mathrm{d}y \tag{3.168}$$
$$= \int_{y=-\infty}^{\infty} F_X(z-y)\, f_Y(y)\, \mathrm{d}y. \tag{3.169}$$

Note that the joint pdf of X and Y is the product of the marginal pdfs because the random variables are independent. We now differentiate the cdf to obtain the pdf. Note that this requires an interchange of a limit operator with a differentiation operator and another interchange of an integral operator with a differentiation operator, which are justified because the functions involved are bounded and integrable.

$$f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z} \lim_{u \to \infty} \int_{y=-u}^{u} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.170}$$
$$= \lim_{u \to \infty} \frac{\mathrm{d}}{\mathrm{d}z} \int_{y=-u}^{u} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.171}$$
$$= \lim_{u \to \infty} \int_{y=-u}^{u} \frac{\mathrm{d}}{\mathrm{d}z} F_X(z-y)\, f_Y(y)\, \mathrm{d}y \tag{3.172}$$
$$= \lim_{u \to \infty} \int_{y=-u}^{u} f_X(z-y)\, f_Y(y)\, \mathrm{d}y. \tag{3.173}$$

□

Example 3.5.3 (Coffee beans)**.** A company that makes coffee buys beans from two small local producers in Colombia and Vietnam. The amount of beans they can buy from each producer varies depending on the weather. The company models these quantities C and V as independent random variables (assuming that the weather in Colombia is independent from the weather in Vietnam) which have uniform distributions in $[0,1]$ and $[0,2]$ (the unit is tons) respectively.

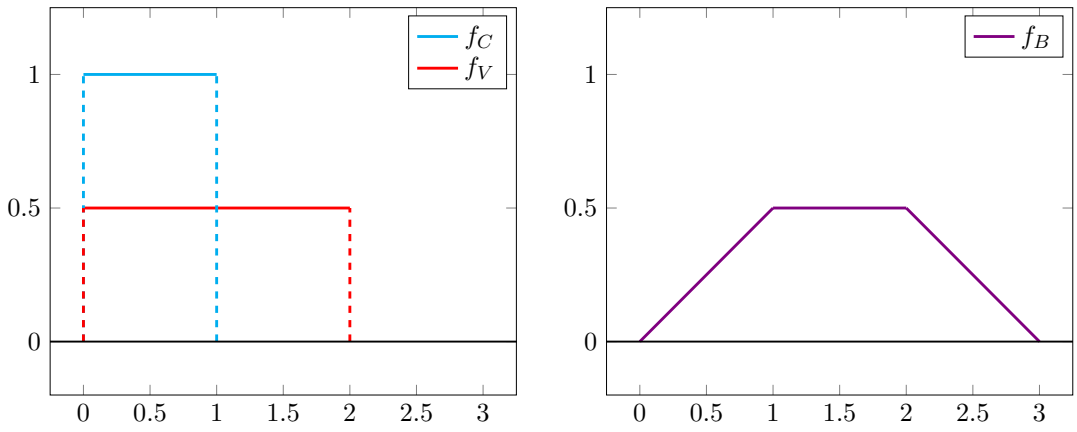


Figure 3.10: Probability density functions in Example 3.5.3.

We now compute the pdf of the total amount of coffee beans $B := E+V$ applying Theorem 3.5.2,

$$f_B(b) = \int_{u=-\infty}^{\infty} f_C(b-u)\, f_V(u)\, \mathrm{d}u \tag{3.174}$$

$$= \frac{1}{2}\int_{u=0}^{2} f_C(b-u)\, \mathrm{d}u \tag{3.175}$$

$$= \begin{cases} \frac{1}{2}\int_{u=0}^{b} \mathrm{d}u = \frac{b}{2} & \text{if } b \leq 1 \\ \frac{1}{2}\int_{u=b-1}^{b} \mathrm{d}u = \frac{1}{2} & \text{if } 1 \leq b \leq 2 \\ \frac{1}{2}\int_{u=b-1}^{2} \mathrm{d}u = \frac{3-b}{2} & \text{if } 2 \leq b \leq 3. \end{cases} \tag{3.176}$$

The pdf of B is shown in Figure 3.10.

△

3.6 Generating multivariate random variables

In Section 2.6 we consider the problem of generating independent samples from an arbitrary univariate distribution. Assuming that a procedure to achieve this is available, we can use it to sample from an arbitrary multivariate distribution by generating samples from the appropriate conditional distributions.

Algorithm 3.6.1 (Sampling from a multivariate distribution)**.** *Let X_1, X_2, ..., X_n be random variables belonging to the same probability space. To generate samples from their joint distribution we sequentially sample from their conditional distributions:*

1. *Obtain a sample x_1 of X_1.*
2. *For $i = 2, 3, \ldots, n$, obtain a sample x_i of X_i given the event $\{X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\}$ by sampling from $F_{X_i|X_1,\ldots,X_{i-1}}(\cdot|x_1, \ldots, x_{i-1})$.*

The chain rule implies that the output $x_1, \ldots, x_n$ of this procedure are samples from the joint distribution of the random variables. The following example considers the problem of sampling from a mixture of exponential random variables.

Example 3.6.2 (Mixture of exponentials)**.** Let B be a Bernoulli random variable with parameter p and X an exponential random variable with parameter 1 if $B = 0$ and 2 if $B = 1$. Assume that we have access to two independent samples u_1 and u_2 from a uniform distribution in $[0, 1]$. To obtain samples from B and X:

1. We set $b := 1$ if $u_1 \leq p$ and $b := 0$ otherwise. This ensures that b is a Bernoulli sample with the right parameter.
2. Then, we set

$$x := \frac{1}{\lambda} \log \left(\frac{1}{1 - u_2} \right) \tag{3.177}$$

where $\lambda := 1$ if $b = 0$ and $\lambda := 2$ if $b = 1$. By Example 2.6.4 x is distributed as an exponential with parameter λ.

△

3.7 Rejection sampling

We end the chapter by describing rejection sampling, also known as the accept-reject method, an alternative procedure for sampling from univariate distributions. The reason we have deferred it to this chapter is that analyzing this technique requires an understanding of multivariate random variables. Before presenting the method, we motivate it using discrete random variables.

3.7.1 Rejection sampling for discrete random variables

Our goal is to simulate a random variable Y using samples from another random variable X. To simplify the exposition, we assume that their pmfs p_X and p_Y have nonzero values in the set $\{1, 2, \ldots, n\}$ (generalizing to other discrete sets is straightforward). The idea behind rejection sampling is that we can choose a subset of the samples of X in a way that reshapes its distribution. When we obtain a sample of X we decide whether to accept it or reject with a certain probability. The probability depends on the value of the sample x, if $p_X(x)$ is much larger than $p_Y(x)$ we should probably reject it most of the time (but not always!). For each $x \in \{1, 2, \ldots, n\}$ we define the probability of accepting the sample by a_x.

We are interested in the distribution of only the accepted samples. Mathematically, the pmf of the accepted samples is equal to the conditional pmf of X, conditioned on the event that the sample is accepted,

$$p_{X \,|\, \text{Accepted}} \left(x \,|\, \text{Accepted} \right) = \frac{p_X \left(x \right) \mathrm{P} \left(\text{Accepted} \,|\, X = x \right)}{\sum_{i=1}^{n} p_X \left(i \right) \mathrm{P} \left(\text{Accepted} \,|\, X = i \right)} \quad \text{by Bayes' rule} \tag{3.178}$$

$$= \frac{p_X \left(x \right) a_x}{\sum_{i=1}^{n} p_X \left(i \right) a_i}. \tag{3.179}$$

We would like to fix the accept probabilities so that for all $x \in \{1, 2, \ldots, n\}$

$$p_{X \,|\, \text{Accepted}} \left(x \,|\, \text{Accepted} \right) = p_Y(x). \tag{3.180}$$

This can be achieved by fixing

$$a_x := \frac{p_Y(x)}{c p_X(x)}, \qquad x \in \{1, \ldots, n\}, \tag{3.181}$$

for any constant c. However, this will not yield a valid probability for any arbitrary c, because a_i could be larger than one! To avoid this issue, we need

$$c \geq \max_{x \in \{1,\ldots,n\}} \frac{p_Y(x)}{p_X(x)}, \quad \text{for all } x \in \{1, \ldots, n\}. \tag{3.182}$$

Finally, we can use a uniform random variable U between 0 and 1 to accept or reject, accepting each sample x if $U \leq a_x$. You might be wondering why we can't just generate Y directly from U. That would be indeed work and is much simpler; here we are just presenting the discrete case as a pedagogical introduction to the continuous case.

Algorithm 3.7.1 (Rejection sampling)**.** *Let X and Y be random variables with pmfs p_X and p_Y such that*

$$c \geq \max_{x \in \{1,\ldots,n\}} \frac{p_Y(x)}{p_X(x)} \tag{3.183}$$

for all x such that $p_Y(x)$ is nonzero, and U a random variable that is uniformly distributed in $[0, 1]$ and independent of X.

1. *Obtain a sample y of X.*
2. *Obtain a sample u of U.*
3. *Declare y to be a sample of Y if*

$$u \leq \frac{p_Y(y)}{c\, p_X(y)}. \tag{3.184}$$

3.7.2 Rejection sampling for continuous random variables

Here we show that the idea presented in the previous section can be applied in the continuous case. The goal is to obtain samples according to a target pdf f_Y by choosing samples obtained according to a different pdf f_X. As in the discrete case, we need

$$f_Y(y) \leq c\, f_X(y) \tag{3.185}$$

for all y, where c is a fixed positive constant. In words, the pdf of Y must be bounded by a scaled version of the pdf of X.

Algorithm 3.7.2 (Rejection sampling)**.** *Let X be a random variable with pdf f_X and U a random variable that is uniformly distributed in $[0, 1]$ and independent of X. We assume that (3.185) holds.*

1. *Obtain a sample y of X.*
2. *Obtain a sample u of U.*

3. Declare y to be a sample of Y if

$$u \leq \frac{f_Y(y)}{c\, f_X(y)}. \tag{3.186}$$

The following theorem establishes that the samples obtained by rejection sampling have the desired distribution.

Theorem 3.7.3 (Rejection sampling works)**.** *If assumption* (3.185) *holds, then the samples produced by rejection sampling are distributed according to f_Y.*

Proof. Let Z denote the random variable produced by rejection sampling. The cdf of Z is equal to

$$F_Z(y) = \mathrm{P}\left(X \leq y \,|\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) \tag{3.187}$$

$$= \frac{\mathrm{P}\left(X \leq y,\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right)}{\mathrm{P}\left(U \leq \frac{f_Y(X)}{c\, f_X(X)}\right)}. \tag{3.188}$$

To compute the numerator we integrate the joint pdf of U and X over the region of interest

$$\mathrm{P}\left(X \leq y,\, U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) = \int_{x=-\infty}^{y} \int_{u=0}^{\frac{f_Y(x)}{c\, f_X(x)}} f_X(x)\, \mathrm{d}u\, \mathrm{d}x \tag{3.189}$$

$$= \int_{x=-\infty}^{y} \frac{f_Y(x)}{c\, f_X(x)} f_X(x)\, \mathrm{d}x \tag{3.190}$$

$$= \frac{1}{c} \int_{x=-\infty}^{y} f_Y(x)\, \mathrm{d}x \tag{3.191}$$

$$= \frac{1}{c} F_Y(y). \tag{3.192}$$

The denominator is obtained in a similar way

$$\mathrm{P}\left(U \leq \frac{f_Y(X)}{c\, f_X(X)}\right) = \int_{x=-\infty}^{\infty} \int_{u=0}^{\frac{f_Y(x)}{c\, f_X(x)}} f_X(x)\, \mathrm{d}u\, \mathrm{d}x \tag{3.193}$$

$$= \int_{x=-\infty}^{\infty} \frac{f_Y(x)}{c\, f_X(x)} f_X(x)\, \mathrm{d}x \tag{3.194}$$

$$= \frac{1}{c} \int_{x=-\infty}^{\infty} f_Y(x)\, \mathrm{d}x \tag{3.195}$$

$$= \frac{1}{c}. \tag{3.196}$$

We conclude that

$$F_Z(y) = F_Y(y), \tag{3.197}$$

so the method produces samples from the distribution of Y. □

We now illustrate the method by applying it to produce a Gaussian random variable from an exponential and a uniform random variable.

Example 3.7.4 (Generating a Gaussian random variable)**.** In Example 2.6.4 we learned how to generate an exponential random variables using samples from a uniform distribution. In this example we will use samples from an exponential distribution to generate a standard Gaussian random variable applying rejection sampling.

The following lemma shows that we can generate a standard Gaussian random variable Y by:

1. Generating a random variable H with pdf
$$f_H(h) := \begin{cases} \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{h^2}{2}\right) & \text{if } h \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.198}$$
2. Generating a random variable S which is equal to 1 or -1 with probability 1/2, for example by applying the method described in Section 2.6.1.
3. Setting $Y := SH$.

Lemma 3.7.5. *Let H be a continuous random variable with pdf given by (3.198) and S a discrete random variable which equals 1 with probability 1/2 and -1 with probability 1/2. The random variable of $Y := SH$ is a standard Gaussian.*

Proof. The conditional pdf of Y given S is given by
$$f_{Y|S}(y|1) = \begin{cases} f_H(y) & \text{if } y \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.199}$$
$$f_{Y|S}(y|-1) = \begin{cases} f_H(-y) & \text{if } y < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.200}$$
By Lemma 3.3.5 we have
$$f_Y(y) = p_S(1) f_{Y|S}(y|1) + p_S(-1) f_{Y|S}(y|-1) \tag{3.201}$$
$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right). \tag{3.202}$$
△

The reason why we reduce the problem to generating H is that its pdf is only nonzero on the positive axis, which allows us to bound it with the exponential pdf of an exponential random variable X with parameter 1. If we set $c := \sqrt{2e/\pi}$ then $f_H(x) \leq c f_X(x)$ for all x, as illustrated in Figure 3.11. Indeed,
$$\frac{f_H(x)}{f_X(x)} = \frac{\frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}{\exp(-x)} \tag{3.203}$$
$$= \sqrt{\frac{2e}{\pi}} \exp\left(\frac{-(x-1)^2}{2}\right) \tag{3.204}$$
$$\leq \sqrt{\frac{2e}{\pi}}. \tag{3.205}$$

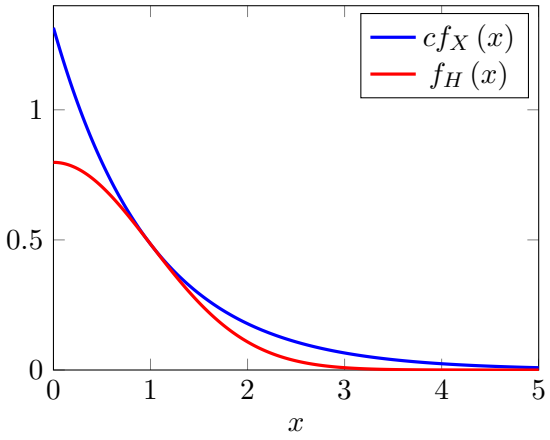


Figure 3.11: Bound on the pdf of the target distribution in Example 3.7.4.

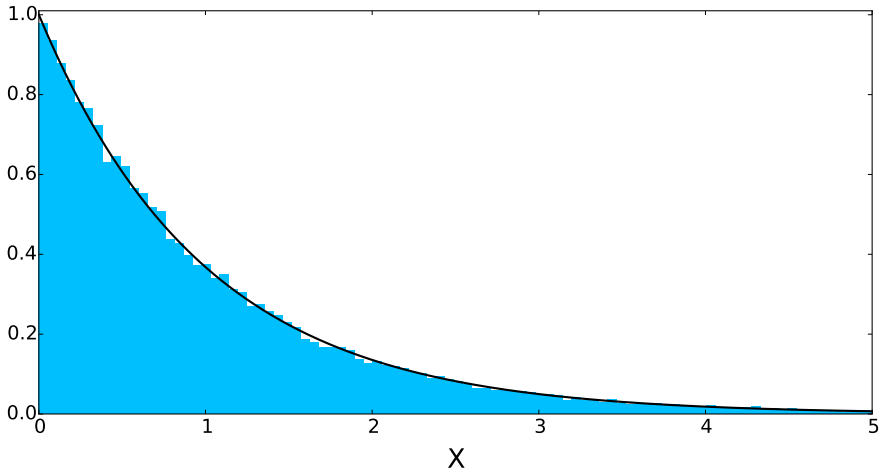
We can now apply rejection sampling to generate H. The steps are

1. Obtain a sample x from an exponential random variable X with parameter one
2. Obtain a sample u from U, which is uniformly distributed in $[0, 1]$.
3. Accept x as a sample of H if

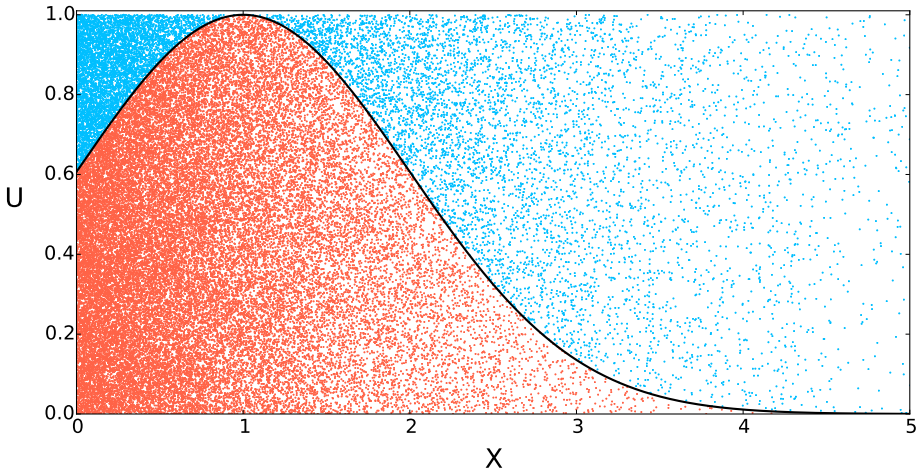
$$u \leq \exp\left(\frac{-(x-1)^2}{2}\right). \tag{3.206}$$

This procedure is illustrated in Figure 3.12. The rejection mechanism ensures that the accepted samples have the right distribution. △

Histogram of 50,000 iid samples from X (f_X is shown in black)



Scatterplot of samples from X and samples from U (accepted samples are colored red, $\exp\left(-(x-1)^2/2\right)$ is shown in black)



Histogram of accepted samples (f_H is shown in black)

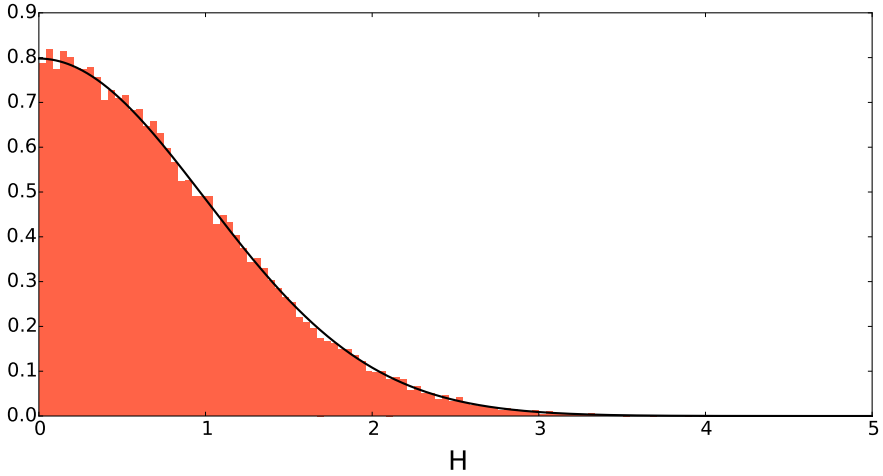


Figure 3.12: Illustration of how to generate 50,000 samples from the random variable H defined in Example 3.7.4 via rejection sampling.