ECE 302: Probability and Applications[1]

Week 7 Topics

- Entropy
  - entropy of a random variable
  - relative entropy
- Two Random Variables
  - Datasets & scattergrams
  - Joint, marginal, conditional pmfs for discrete RVs
- Joint, marginal, discrete RVs

1 Entropy

Entropy is a measure of the uncertainty in a random experiment. In this section, we introduce the notion of the entropy of a random variable and we develop several of its fundamental properties. We explain how entropy can be modified to obtain relative entropy, a measure of the similarity between two distributions.

1.1 The Entropy of a Random Variable

Let X be a discrete random variable with $S_X = \{1, 2, \ldots, K\}$ and pmf $p_k = P[X = k]$. We are interested in quantifying the uncertainty of the event $A_k = \{X = k\}$. Clearly, the uncertainty of A_k is low if the probability of A_k is close to one, and it is high if the probability of A_k is small. The following measure of uncertainty satisfies these two properties:

$$I(X = k) = \ln \frac{1}{P[X = k]} = -\ln P[X = k] \tag{1}$$

Note from Fig. 1 that $I(X = k) = 0$ if $P[X = k] = 1$, and $I(X = k)$ increases with decreasing $P[X = k]$. The **entropy of a random variable** $\boldsymbol{X}$ is defined as the expected value of the uncertainty of its outcomes:

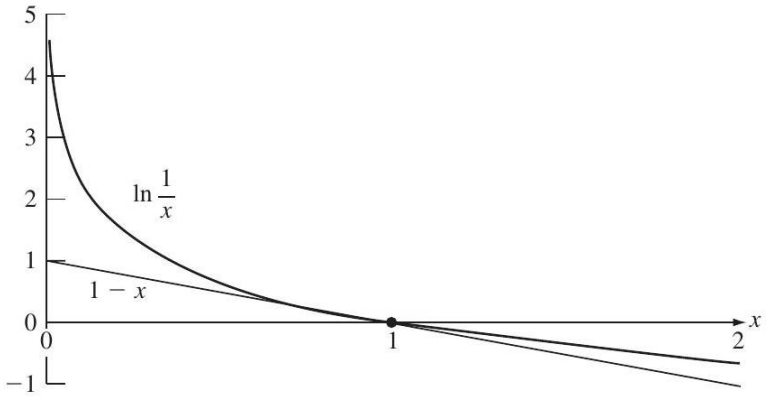

Figure 1: $\ln(1/x) \geq 1 - x$

$$H_X = E[I(X)] = \sum_{k=1}^{K} P[X = k] \ln \frac{1}{P[X = k]}$$

$$= -\sum_{k=1}^{K} P[X = k] \ln P[X = k]. \tag{2}$$

Note that $I(X)$ is simply a function of a random variable. We say that entropy is in units of "bits" when the logarithm is base 2. In the above expression we are using the natural logarithm, so we say the units are in "nats." Changing the base of the logarithm is equivalent to multiplying entropy by a constant, since $\ln(x) = \ln 2 \log_2 x$.

Example 1: Entropy of a Bernoulli Random Variable

Suppose that $S_X = \{0, 1\}$ and $p = P[X = 0] = 1 - P[X = 1]$. Figure 2 shows $-p \ln(p)$, $-(1-p)\ln(1-p)$, and the entropy of the binary random variable $H_X = h(p) = -p \ln(p) - (1-p)\ln(1-p)$ as functions of p. Note that $h(p)$ is symmetric about $p = 1/2$ and that it achieves its maximum at $p = 1/2$. Note also how the uncertainties of the events $\{X = 0\}$ and $\{X = 1\}$ vary together in complementary fashion: When $P[X = 0]$ is very small (i.e., highly uncertain), then $P[X = 1]$ is close to one (i.e., highly certain), and vice versa. Thus the highest average uncertainty occurs when $P[X = 0] = P[X = 1] = 1/2$. H_X can be viewed as the average uncertainty that

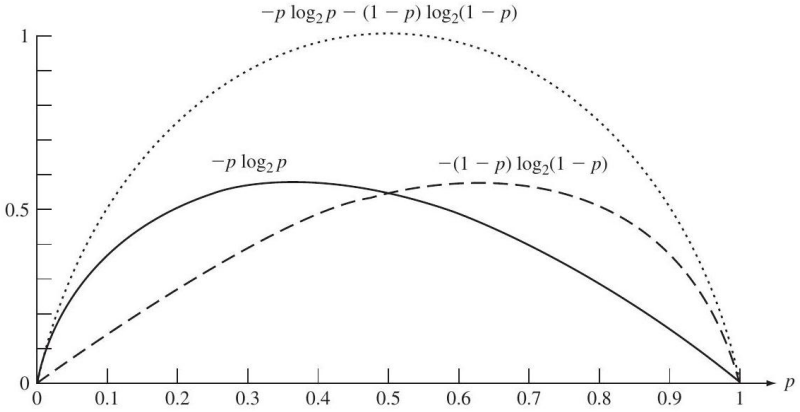

Figure 2: Entropy of binary random variable

is resolved by observing X. This suggests that if we are designing a binary experiment (for example, a yes/no question), then the average uncertainty that is resolved will be maximized if the two outcomes are designed to be equiprobable.

Example 2: Reduction of Entropy Through Partial Information

The binary representation of the random variable X takes on values from the set $\{000, 001, 010, \ldots, 111\}$ with equal probabilities. Find the reduction in the entropy of X given the event $A = \{X \text{ begins with a } 1\}$.

The entropy of X is

$$H_X = -\frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} - \cdots - \frac{1}{8}\log_2\frac{1}{8} = 3 \text{ bits.}$$

The event A implies that X is in the set $\{100, 101, 110, 111\}$, so the entropy of X given A is

$$H_{X|A} = -\frac{1}{4}\log_2\frac{1}{4} - \cdots - \frac{1}{4}\log_2\frac{1}{4} = 2 \text{ bits}$$

Thus the reduction in entropy is $H_X - H_{X|A} = 3 - 2 = 1$ bit.

Let $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$, and $\boldsymbol{q} = (q_1, q_2, \ldots, q_K)$ be two pmf's. The **relative entropy** of $\boldsymbol{q}$ with respect to $\boldsymbol{p}$ is defined by

$$H(p;q) = \sum_{k=1}^{K} p_k \ln\frac{1}{q_k} - H_X = \sum_{k=1}^{K} p_k \ln\frac{p_k}{q_k} \tag{3}$$

The relative entropy is nonnegative, and equal to zero if and only if $p_k = q_k$ *for all* k :

$$H(p;q) \geq 0 \quad \text{with equality iff} \quad p_k = q_k \quad \text{for } k = 1, \ldots, K \tag{4}$$

We will use this fact repeatedly in the remainder of this discussion. We note that the relative entropy is frequently used as a measure of the similarity between two random variables.

To show that the relative entropy is nonnegative, we use the inequality $\ln(1/x) \geq 1 - x$ with equality iff $x = 1$, as shown in Fig. 1. Equation (3) then becomes

$$H(p;q) = \sum_{k=1}^{K} p_k \ln\frac{p_k}{q_k} \geq \sum_{k=1}^{K} p_k \left(1 - \frac{q_k}{p_k}\right) = \sum_{k=1}^{K} p_k - \sum_{k=1}^{K} q_k = 0 \tag{5}$$

In order for equality to hold in the above expression, we must have $p_k = q_k$ for $k = 1, \ldots, K$.

Let X be any random variable with $S_X = \{1, 2, \ldots, K\}$ and pmf $\boldsymbol{p}$. If we let $q_k = 1/K$ in Eq. (4), then

$$H(p;q) = \ln K - H_X = \sum_{k=1}^{K} p_k \ln \frac{p_k}{1/K} \geq 0$$

which implies that for any random variable X with $S_X = \{1, 2, \ldots, K\}$,

$$H_X \leq \ln K \quad \text{with equality iff} \quad p_k = \frac{1}{K} \quad k = 1, \ldots, K \tag{6}$$

Thus *the maximum entropy attainable by the random variable X is $\ln K$, and this maximum is attained when all the outcomes are equiprobable.*

Equation (6) shows that the entropy of random variables with finite S_X is always finite. On the other hand, it also shows that as the size of S_X is increased, the entropy can increase without bound. The following example shows that some countably infinite random variables have finite entropy.

Example 3: Entropy of a Geometric Random Variable

The entropy of the geometric random variable with $S_X = \{0, 1, 2, \ldots\}$ is:

$$\begin{aligned} H_X &= -\sum_{k=0}^{\infty} p(1-p)^k \ln \left(p(1-p)^k\right) \\ &= -\ln p - \ln(1-p) \sum_{k=0}^{\infty} kp(1-p)^k \\ &= -\ln p - \frac{(1-p)\ln(1-p)}{p} \\ &= \frac{-p\ln p - (1-p)\ln(1-p)}{p} = \frac{h(p)}{p} \end{aligned}$$

where $h(p)$ is the entropy of a binary random variable. Note that $H_X = 2$ bits when $p = 1/2$.

For continuous random variables we have that $P[X = x] = 0$ for all x. Therefore by Eq. (1) the uncertainty for every event $\{X = x\}$ is infinite, and it follows from Eq. (2) that *the entropy of continuous random variables*

is infinite. The next example takes a look at how the notion of entropy may be applied to continuous random variables.

Example 4: Entropy of a Quantized Continuous Random Variable

Let X be a continuous random variable that takes on values in the interval $[a, b]$. Suppose that the interval $[a, b]$ is divided into a large number K of subintervals of length Δ. Let $Q(X)$ be the midpoint of the subinterval that contains X. Find the entropy of Q.

Let x_k be the midpoint of the k th subinterval, then $P\left[Q = x_k\right] = P[X$ is in k th subinterval $] = P\left[x_k - \Delta/2 < X < x_k + \Delta/2\right] \simeq f_X\left(x_k\right)\Delta$, and thus

$$\begin{aligned} H_Q &= \sum_{k=1}^{K} P\left[Q = x_k\right] \ln P\left[Q = x_k\right] \\ &\simeq -\sum_{k=1}^{K} f_X\left(x_k\right)\Delta \ln\left(f_X\left(x_k\right)\Delta\right) \\ &= -\ln(\Delta) - \sum_{k=1}^{K} f_X\left(x_k\right) \ln\left(f_X\left(x_k\right)\right)\Delta \end{aligned} \tag{7}$$

The above equation shows that there is a tradeoff between the entropy of Q and the quantization error $X - Q(X)$. As Δ is decreased the error decreases, but the entropy increases without bound, once again confirming the fact that the entropy of continuous random variables is infinite.

In the final expression for H_X in Eq. (7), as Δ approaches zero, the first expression approaches infinity, but the second expression approaches an integral which may be finite in some cases. The **differential entropy** is defined by this integral:

$$H_X = -\int_{-\infty}^{\infty} f_X(x) \ln f_X(x) dx = -E\left[\ln f_X(X)\right]$$

In the above expression, we reuse the term H_X with the understanding that we deal with differential entropy when dealing with continuous random variables.

Example 5: Differential Entropy of a Uniform Random Variable
The differential entropy for X uniform in $[a, b]$ is

$$H_X = -E\left[\ln\left(\frac{1}{b-a}\right)\right] = \ln(b-a)$$

Example 6: Differential Entropy of a Gaussian Random Variable
The differential entropy for X, a Gaussian random variable, is

$$\begin{aligned} H_X &= -E\left[\ln f_X(X)\right] \\ &= -E\left[\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(X-m)^2}{2\sigma^2}\right] \\ &= \frac{1}{2}\ln\left(2\pi\sigma^2\right) + \frac{1}{2} \\ &= \frac{1}{2}\ln\left(2\pi e\sigma^2\right) \end{aligned} \tag{8}$$

1.2 Entropy as a Measure of Information

Let X be a discrete random variable with $S_X = \{1, 2, \ldots, K\}$ and pmf $p_k = P[X = k]$. Suppose that the experiment that produces X is performed by John, and that he attempts to communicate the outcome to Mary by answering a series of yes/no questions. We are interested in characterizing the minimum average number of questions required to identify X.

Example 7
An urn contains 16 balls: 4 balls are labeled "1", 4 are labeled "2", 2 are labeled "3", 2 are labeled "4", and the remaining balls are labeled "5", "6", "7", and "8." John picks a ball from the urn at random, and he notes the number. Discuss what strategies Mary can use to find out the number of the ball through a series of yes/no questions. Compare the average number of questions asked to the entropy of X.

If we let X be the random variable denoting the number of the ball, then $S_X = \{1, 2, \ldots, 8\}$ and the pmf is $\boldsymbol{p} = (1/4, 1/4, 1/8, 1/8, 1/16, 1/16, 1/16, 1/16)$. We will compare the two strategies shown in Figs. 3(a) and (b).

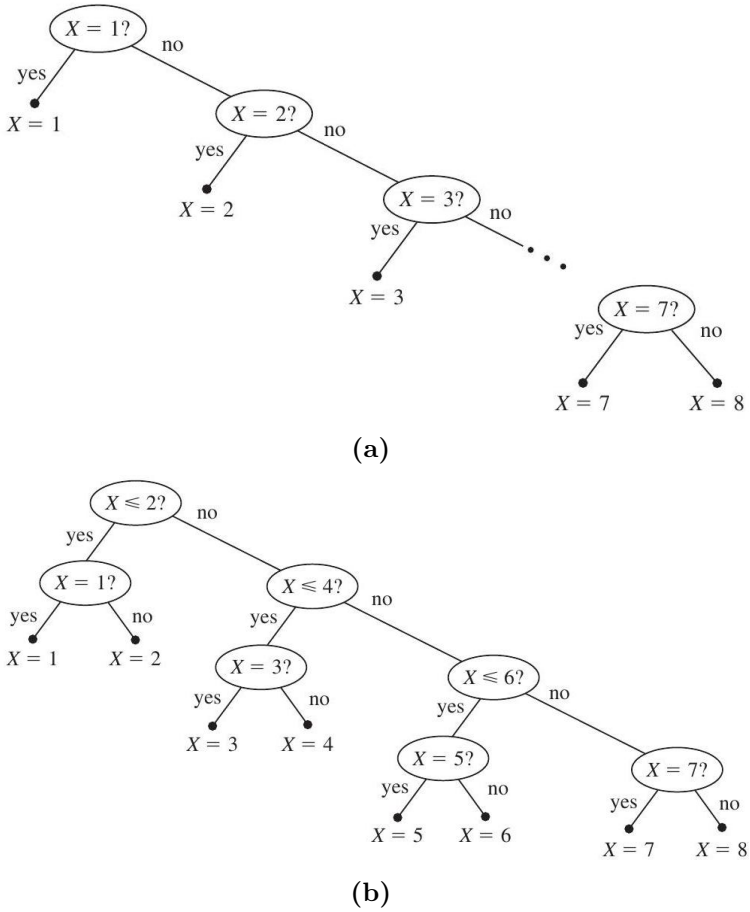

Figure 3: Two strategies for finding out the value of X through a series of yes/no questions

The series of questions in Fig. 3(a) uses the fact that the probability of $\{X = k\}$ decreases with k. Thus it is reasonable to ask the question {"Was X equal to 1?"}, {"Was X equal to 2?"}, and so on, until the answer is yes. Let L be the number of questions asked until the answer is yes, then

the average number of questions asked is

$$\begin{aligned} E[L] &= 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{8}\right) \\ &\quad + 5\left(\frac{1}{16}\right) + 6\left(\frac{1}{16}\right) + 7\left(\frac{1}{16}\right) + 7\left(\frac{1}{16}\right) \\ &= 51/16. \end{aligned}$$

The series of questions in Fig. 3(b) uses the observation made previously (see our discussion of the $m-$Erlang random variable) that yes/no questions should be designed so that the two answers are equiprobable. The questions in Fig. 3(b) meet this requirement. The average number of questions asked is

$$\begin{aligned} E[L] &= 2\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 3\left(\frac{1}{8}\right) \\ &\quad + 4\left(\frac{1}{16}\right) + 4\left(\frac{1}{16}\right) + 4\left(\frac{1}{16}\right) + 4\left(\frac{1}{16}\right) \\ &= 44/16. \end{aligned}$$

Thus the second series of questions has the better performance.

Finally, we find that the entropy of X is

$$H_X = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \cdots - \frac{1}{16}\log_2\frac{1}{16} = 44/16$$

which is equal to the performance of the second series of questions.

The problem of designing the series of questions to identify the random variable X is exactly the same as the problem of encoding the output of an information source. Each output of an information source is a random variable X, and the task of the encoder is to map each possible output into a unique string of binary digits. We can see this correspondence by taking the trees in Fig. 3 and identifying each yes/no answer with a 0/1. The sequence of 0's and 1's from the top node to each terminal node then defines the binary string ("codeword") for each outcome. It then follows that the problem of finding the best series of yes/no questions is the same as finding the binary tree code that minimizes the average codeword length.

2 Two Random Variables

Many random experiments involve several random variables. In some experiments a number of different quantities are measured. For example, the temperature at different locations at some specific time may be of interest. Other experiments involve the repeated measurement of a certain quantity such as the repeated measurement ("sampling") of the amplitude of an audio or video signal that varies with time. In this section, we extend the concepts from random variables to develop techniques for calculation the probabilities of events involving multiple random variables:

- We use the joint pmf, cdf, and pdf to calculate the probabilities of events that involve the joint behavior of two random variables;
- We use expected value to define joint moments that summarize the joint behavior of two random variables;
- We determine when two random variables are independent, and we quantify their degree of "correlation" when they are not independent;
- We obtain conditional probabilities involving a pair of random variables.

Basically we have already covered all the fundamental concepts of probability and random variables, and we now "simply" elaborate on the case of two or more random variables. However we will see that need more analytical tools, e.g., double summations of pmf's and double integration of pdf's, so we first discuss the case of two random variables in detail because we can draw on our geometric intuition. After this we will consider the general case of vector random variables.

2.1 Two Random Variables

A **scattergram** can be used to gain insight into the joint behavior of two random variables. A scattergram plot simply places a dot at every observation pair (x, y) that results from performing the experiment that generates (X, Y). Figure 4 shows the scattergram for 200 observations of four different pairs of random variables. The pairs in Fig. 4(a) appear to be uniformly distributed in the unit square. The pairs in Fig. 4(b) are clearly

confined to a disc of unit radius and appear to be more concentrated near the origin. The pairs in Fig. 4(c) are concentrated near the origin, and appear to have circular symmetry, but are not bounded to an enclosed region. The pairs in Fig. 4(d) again are concentrated near the origin and appear to have a clear linear relationship of some sort, that is, larger values of x tend to have linearly proportional increasing values of y. We later introduce various functions and moments to characterize the behavior of pairs of random variables illustrated in these examples.

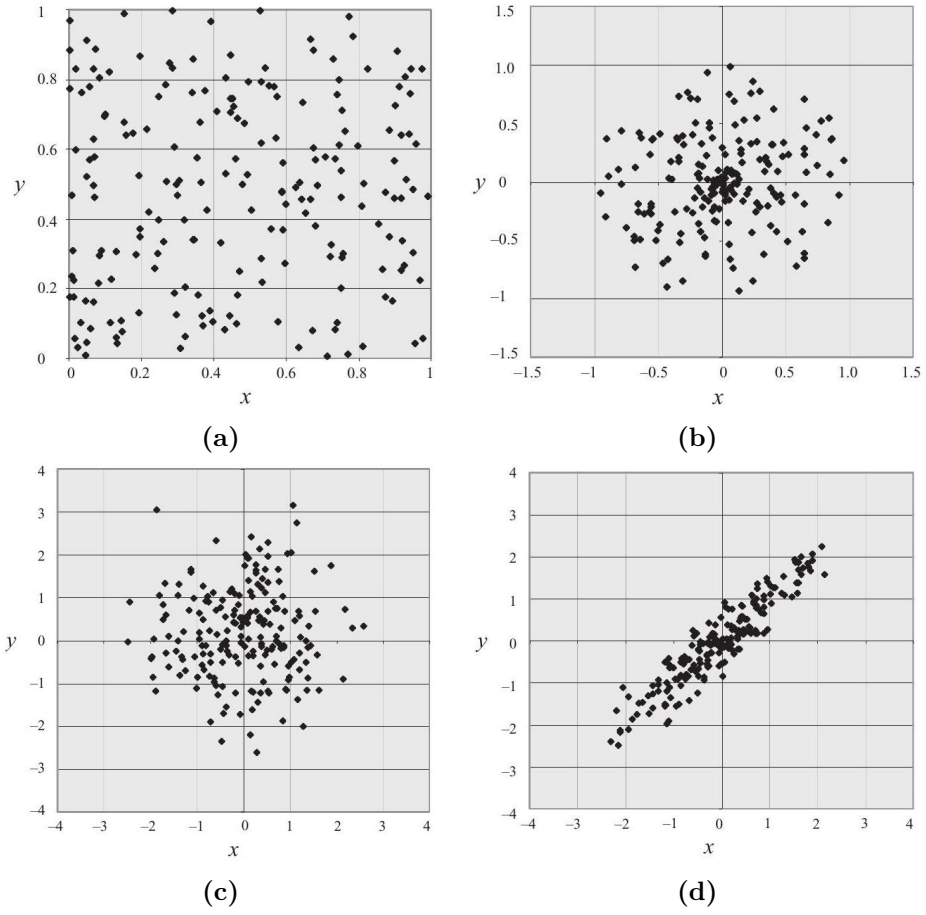

Figure 4: A scattergram for 200 observations of four different pairs of random variables.

The notion of a random variable as a mapping is easily generalized to the case where two or more quantities are of interest. Consider a random

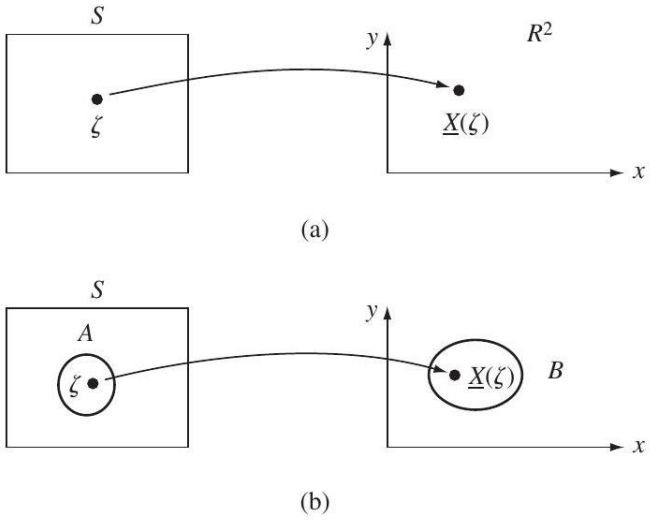

Figure 5: (a) A function assigns a pair of real numbers to each outcome in S. (b) Equivalent events for two random variables.

experiment with sample space S and event class $\mathcal{F}$. We are interested in a function that assigns a pair of real numbers $\mathbf{X}(\zeta) = (X(\zeta), Y(\zeta))$ to each outcome ζ in S. Basically we are dealing with a vector function that maps S into R^2, the real plane, as shown in Fig. 5(a). We are ultimately interested in events involving the pair (X, Y).

Example 8

Let a random experiment consist of selecting a student's name from an urn. Let ζ denote the outcome of this experiment, and define the following two functions:

$$H(\zeta) = \text{ height of student } \zeta \text{ in centimeters}$$
$$W(\zeta) = \text{ weight of student } \zeta \text{ in kilograms}$$

$(H(\zeta), W(\zeta))$ assigns a pair of numbers to each ζ in S. We are interested in events involving the pair (H, W). For example, the event $B = \{H \leq 183, W \leq 82\}$ represents students with height less that 183 cm (6 feet) and weight less than 82 kg(180 lb).

Example 9

A Web page provides the user with a choice either to watch a brief ad or to move directly to the requested page. Let ζ be the sequence of user arrivals and their choices (ad/no ad) in T seconds. Let $N_1(\zeta)$ be the number of users requesting the Web page directly and let $N_2(\zeta)$ be the number of times that the ad is chosen. An event of interest could be $A = \{N_1 \leq 1000, N_2 \geq 100\}$, that is, less than 1000 users reject the ad and more than 100 accept it.

Example 10

Let the outcome ζ in a random experiment be the length in bytes of a randomly selected message. Suppose that messages are broken into blocks of maximum length M=1000 bytes. Let Q be the number of full packets in a message and let R be the number of bytes left over. $(Q(\zeta), R(\zeta))$ assigns a pair of numbers to each ζ in S. Q takes on values in the range $0, 1, 2, \ldots$, and R takes on values in the range $0, 1, \ldots, M-1$. An event of interest may be $B = \{Q > 10, R < 500\}$, "the message is more than 10 blocks long, and the last block is less than half full."

Example 11

Let ζ be uniformly distributed in the interval $(0, 2\pi)$. Let

$$X = \cos\zeta \quad \text{and} \quad Y = \sin\zeta$$

The point (X, Y) then corresponds to the point on the unit circle specified by the angle ζ. An event of interest may be $B = \{X \leq 0, Y \leq 0\}$, that is, the point is in the lower quadrant of the plane.

We will show later that all of the events in the above example can be expressed in terms of the **joint cumulative distribution function of X and Y** which s defined as the probability of the event $\{X \leq x_1\} \cap \{Y \leq y_1\}$

$$F_{X,Y}(x_1, y_1) = P[X \leq x_1, Y \leq y_1] \tag{9}$$

Figure 6 shows three other examples of events:

$A = \{X + Y \leq 10\}$
$B = \{\min(X, Y) \leq 5\}$
$\mathrm{C} = \{X^2 + Y^2 \leq 100\}$

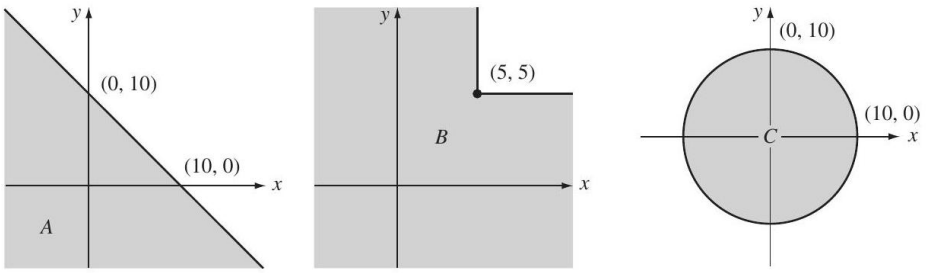

Figure 6: Examples of two-dimensional events

To determine the probability that the pair $\mathbf{X} = (X, Y)$ is in some region B in the plane, we find the equivalent event for B in the underlying sample space S:

$$A = \mathbf{X}^{-1}(B) = \{\zeta : (X(\zeta), Y(\zeta)) \text{ in } B\}$$

The relationship between $A = \mathbf{X}^{-1}(B)$ and B is shown in Fig. 5(b). If A is in $\mathcal{F}$, then it has a probability assigned to it, and we obtain:

$$P[X \text{ in } B] = P[A] = P[\{\zeta : (X(\zeta), Y(\zeta)) \text{ in } B\}] \tag{10}$$

The approach is identical to what we followed in the case of a single random variable. The only difference is that we are considering the *joint behavior of* X *and* Y that is induced by the underlying random experiment.

In the following we will introduce the joint probability mass function, joint cumulative distribution function, and joint probability density function which provide approaches to specifying the probability law that governs the behavior of the pair (X, Y).

2.2 Pairs of Discrete Random Variables

Let the vector random variable $\mathbf{X} = (X, Y)$ assume values from some countable set $S_{X,Y} = \{(x_j, y_k), j = 1, 2, \ldots, k = 1, 2, \ldots\}$. The joint probability mass function of $\mathbf{X}$ specifies the probabilities of the event $\{X = x\} \cap \{Y = y\}$:

$$\begin{aligned} p_{X,Y}(x, y) &= P[\{X = x\} \cap \{Y = y\}] \\ &\triangleq P[X = x, Y = y] \quad \text{for } (x, y) \in R^2 \end{aligned}$$

The values of the pmf on the set $S_{X,Y}$ provide the essential information:

$$\begin{aligned} p_{X,Y}(x_j, y_k) &= P[\{X = x_j\} \cap \{Y = y_k\}] \\ &\triangleq P[X = x_j, Y = y_k] \quad (x_j, y_k) \in S_{X,Y} \end{aligned}$$

There are several ways of showing the pmf graphically: (1) For small sample spaces we can present the pmf in the form of a table as shown in Fig. 7(a). (2) We can present the pmf using arrows of height $p_{X,Y}(x_j, y_k)$ placed at the points $\{(x_j, y_k)\}$ in the plane, as shown in Fig. 7(b), but this can be difficult to draw. (3) We can place dots at the points $\{(x_j, y_k)\}$ and label these with the corresponding pmf value as shown in Fig. (c).

The probability of any event B is the sum of the pmf over the outcomes in B :

$$P[\mathbf{X} \text{ in } B] = \sum_{(x_j, y_k) \text{ in } B} \sum p_{X,Y}(x_j, y_k) \tag{11}$$

Frequently it is helpful to sketch the region that contains the points in B as shown, for example, in Fig. 8. When the event B is the entire sample space $S_{X,Y}$, we have:

$$\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} p_{X,Y}(x_j, y_k) = 1 \tag{12}$$

Example 12

A packet switch has two input ports and two output ports. At a given time slot a packet arrives at an input port with probability 1/2, and each arriving packet is equally likely to be destined to output port 1 or 2 . Let X and Y be the number of packets destined for output ports 1 and 2, respectively. Find the pmf of X and Y, and show the pmf graphically.

The outcome I_j for an input port j can take the following values: "n", no packet arrival (with probability 1/2); "a1", packet arrival destined for output port 1 (with probability 1/4); "a2", packet arrival destined for output port 2 (with probability 1/4). The underlying sample space S consists of the pair of input outcomes $\zeta = (I_1, I_2)$. The mapping for (X, Y) is shown in the table below:

ζ	(n, n)	(n, a1)	(n, a2)	(a1, n)	(a1, a1)	(a1, a2)	(a2, n)	(a2, a1)	(a2, a2)
X, Y	(0, 0)	(1, 0)	(0, 1)	(1, 0)	(2, 0)	(1, 1)	(0, 1)	(1, 1)	(0, 2)

The pmf of (X, Y) is then:

$$p_{X,Y}(0,0) = P[\zeta = (\text{n}, \text{n})] = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$
$$p_{X,Y}(0,1) = P[\zeta \in \{(\text{n}, \text{a2}), (\text{a2}, \text{n})\}] = 2 * \frac{1}{8} = \frac{1}{4}$$
$$p_{X,Y}(1,0) = P[\zeta \in \{(\text{n}, \text{a1}), (\text{a1}, \text{n})\}] = \frac{1}{4}$$
$$p_{X,Y}(1,1) = P[\zeta \in \{(\text{a1}, \text{a2}), (\text{a2}, \text{a1})\}] = \frac{1}{8}$$
$$p_{X,Y}(0,2) = P[\zeta = (\text{a2}, \text{a2})] = \frac{1}{16}$$
$$p_{X,Y}(2,0) = P[\zeta = (\text{a1}, \text{a1})] = \frac{1}{16}$$

Figure 7(a) shows the pmf in tabular form where the number of rows and columns accommodate the range of X and Y respectively. Each entry in the table gives the pmf value for the corresponding x and y. Figure 7(b) shows the pmf using arrows in the plane. An arrow of height $p_{X,Y}(j,k)$ is placed at each of the points in $S_{X,Y} = \{(0,0), (0,1), (1,0), (1,1), (0,2), (2,0)\}$. Figure 7(c) shows the pmf using labeled dots in the plane. A dot with label $p_{X,Y}(j,k)$ is placed at each of the points in $S_{X,Y}$.

Example 13

A random experiment consists of tossing two "loaded" dice and noting the pair of numbers (X, Y) facing up. The joint pmf $p_{X,Y}(j,k)$ for $j = 1, \ldots, 6$ and $k = 1, \ldots, 6$ is given by the twodimensional table shown in Fig. 5.6. The (j,k) entry in the table contains the value $p_{X,Y}(j,k)$. Find the $P[\min(X,Y) = 3]$.

Figure 8 shows the region that corresponds to the set $\{\min(x,y) = 3\}$. The probability of this event is given by:

$$\begin{aligned} P[\min(X,Y) = 3] &= p_{X,Y}(6,3) + p_{X,Y}(5,3) + p_{X,Y}(4,3) \\ &\quad + p_{X,Y}(3,3) + p_{X,Y}(3,4) + p_{X,Y}(3,5) + p_{X,Y}(3,6) \\ &= 6\left(\frac{1}{42}\right) + \frac{2}{42} = \frac{8}{42} \end{aligned}$$

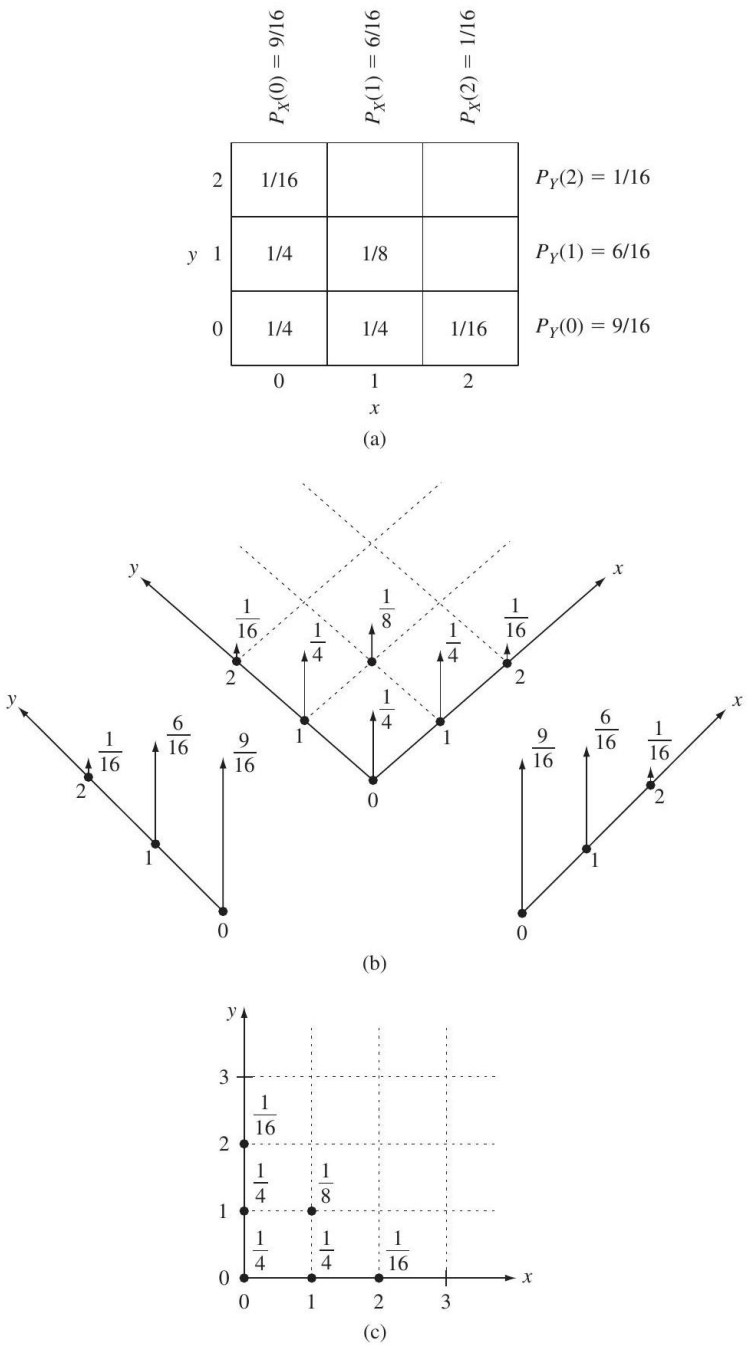

Figure 7: Graphical representations of pmf's: (a) in table format; (b) use of arrows to show height; (c) labeled dots corresponding to pmf value.

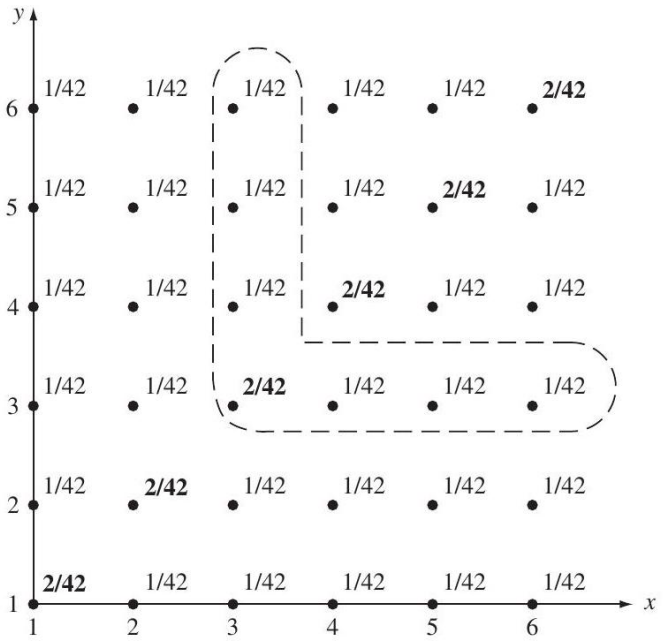

Figure 8: Showing the pmf via a sketch containing the points in B.

2.3 Marginal Probability Mass Function

The joint pmf of $\mathbf{X}$ provides the information about the joint behavior of X and Y. We are also interested in the probabilities of events involving each of the random variables in isolation. These can be found in terms of the **marginal probability mass functions**:

$$\begin{aligned} p_X(x_j) &= P[X = x_j] \\ &= P[X = x_j, Y = \text{ anything }] \\ &= P[\{X = x_j \text{ and } Y = y_1\} \cup \{X = x_j \text{ and } Y = y_2\} \cup \ldots] \\ &= \sum_{k=1}^{\infty} p_{X,Y}(x_j, y_k) \end{aligned} \tag{13}$$

and similarly,

$$\begin{aligned} p_Y(y_k) &= P[Y = y_k] \\ &= \sum_{j=1}^{\infty} p_{X,Y}(x_j, y_k) \end{aligned} \tag{14}$$

The marginal pmf's satisfy all the properties of one-dimensional pmf's, and they supply the information required to compute the probability of events involving the corresponding random variable.

The probability $p_{X,Y}(x_j, y_k)$ can be interpreted as the long-term relative frequency of the joint event $\{X = X_j\} \cap \{Y = Y_k\}$ in a sequence of repetitions of the random experiment. Equation (13) corresponds to the fact that the relative frequency of the event $\{X = X_j\}$ is found by adding the relative frequencies of all outcome pairs in which X_j appears. In general, it is impossible to deduce the relative frequencies of pairs of values X and Y from the relative frequencies of X and Y in isolation. The same is true for pmf's: In general, knowledge of the marginal pmf's is insufficient to specify the joint pmf.

Example 14

Find the marginal pmf for the output ports (X, Y) in Example 9.
Figure 7(a) shows that the marginal pmf is found by adding entries along a row or column in the table. For example, by adding along the $x = 1$ column we have:

$$p_X(1) = P[X = 1] = p_{X,Y}(1, 0) + p_{X,Y}(1, 1) = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$$

Similarly, by adding along the $y = 0$ row:

$$p_Y(0) = P[Y = 0] = p_{X,Y}(0, 0) + p_{X,Y}(1, 0) + p_{X,Y}(2, 0) = \frac{1}{4} + \frac{1}{4} + \frac{1}{16} = \frac{9}{16}$$

Figure 7(b) shows the marginal pmf using arrows on the real line.

Example 15

Find the marginal pmf's in the loaded dice experiment in Example 9.
The probability that $X = 1$ is found by summing over the first row:

$$P[X = 1] = \frac{2}{42} + \frac{1}{42} + \cdots + \frac{1}{42} = \frac{1}{6}$$

Similarly, we find that $P[X = j] = 1/6$ for $j = 2, \ldots, 6$. The probability that $Y = k$ is found by summing over the k th column. We then find that $P[Y = k] = 1/6$ for $k = 1, 2, \ldots, 6$. Thus each die, in isolation, appears to be fair in the sense that each face is equiprobable. If we knew only these marginal pmf's we would have no idea that the dice are loaded.

Example 16

In Example 10, let the number of bytes N in a message have a geometric distribution with parameter $1-p$ and range $S_N = \{0, 1, 2, \ldots\}$. Find the joint pmf and the marginal pmf's of Q and R.

If a message has N bytes, then the number of full packets is the quotient Q in the division of N by M, and the number of remaining bytes is the remainder R. The probability of the pair $\{(q, r)\}$ is given by

$$P[Q = q, R = r] = P[N = qM + r] = (1-p)p^{qM+r}$$

The marginal pmf of Q is

$$\begin{aligned} P[Q = q] &= P[N \text{ in} \{qM, qM+1, \ldots, qM + (M-1)\}] \\ &= \sum_{k=0}^{(M-1)} (1-p)p^{qM+k} \\ &= (1-p)p^{qM}\frac{1-p^M}{1-p} = \left(1-p^M\right)\left(p^M\right)^q \quad q = 0, 1, 2, \ldots \end{aligned}$$

The marginal pmf of Q is geometric with parameter p^M. The marginal pmf of R is:

$$\begin{aligned} &P[R = r] = P[N \text{ in } \{r, M+r, 2M+r, \ldots\}] \\ &= \sum_{q=0}^{\infty}(1-p)p^{qM+r} = \frac{(1-p)}{1-p^M}p^r \quad r = 0, 1, \ldots, M-1 \end{aligned}$$

R has a truncated geometric pmf. As an exercise, you should verify that all the above marginal pmf's add to 1.

3 The Joint cdf of X and Y

Previously we saw that semi-infinite intervals of the form $(-\infty, x]$ are a basic building block from which other one-dimensional events can be built. By defining the cdf $F_X(x)$ as the probability of $(-\infty, x]$, we were then able to express the probabilities of other events in terms of the cdf. In this section we repeat the above development for two-dimensional random variables. A basic building block for events involving two-dimensional

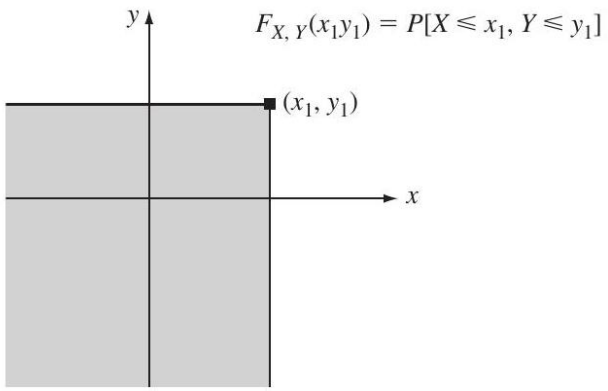

Figure 9: The joint cumulative distribution function is defined as the probability of the semi-infinite rectangle defined by the point (x_1, y_1).

random variables is the semi-infinite rectangle defined by $\{(x, y) : x \leq x_1$ and $y \leq y_1\}$, as shown in Fig. 9. We also use the more compact notation $\{x \leq x_1, y \leq y_1\}$ to refer to this region. The **joint cumulative distribution function of X and Y** is defined as the probability of the event $\{X \leq x_1\} \cap \{Y \leq y_1\}$

$$F_{X,Y}(x_1, y_1) = P[X \leq x_1, Y \leq y_1] \tag{15}$$

In terms of relative frequency, $F_{X,Y}(x_1, y_1)$ represents the long-term proportion of time in which the outcome of the random experiment yields a point X that falls in the rectangular region shown in Fig. 9. In terms of probability "mass," $F_{X,Y}(x_1, y_1)$ represents the amount of mass contained in the rectangular region.

The joint cdf satisfies the following properties.

(i) The joint cdf is a nondecreasing function of x and y :

$$F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2) \quad \text{if } x_1 \leq x_2 \text{ and } y_1 \leq y_2 \tag{16}$$

(ii) $F_{X,Y}(x_1, -\infty) = 0, \quad F_{X,Y}(-\infty, y_1) = 0, \quad F_{X,Y}(\infty, \infty) = 1.$

(iii) We obtain the marginal cumulative distribution functions by removing the constraint on one of the variables. The marginal cdf's are the probabilities of the regions shown in Fig. 10:

$$F_X(x_1) = F_{X,Y}(x_1, \infty) \quad \text{and} \quad F_Y(y_1) = F_{X,Y}(\infty, y_1) \tag{17}$$

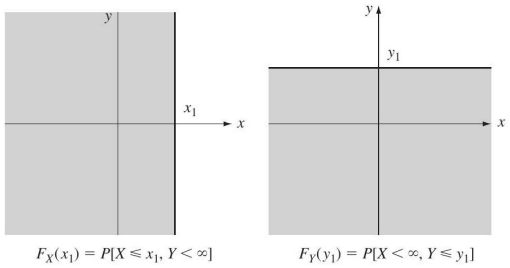

Figure 10: The marginal cdf's are the probabilities of these half-planes.

(iv) The joint cdf is continuous from the "north" and from the "east," that is,

$$\lim_{x \to a^+} F_{X,Y}(x,y) = F_{X,Y}(a,y) \quad \text{and} \quad \lim_{y \to b^+} F_{X,Y}(x,y) = F_{X,Y}(x,b) \tag{18}$$

(v) The probability of the rectangle $\{x_1 < x \leq x_2, y_1 < y \leq y_2\}$ is given by:

$$P\left[x_1 < X \leq x_2, y_1 < Y \leq y_2\right] = \tag{19}$$

$$F_{X,Y}\left(x_2,y_2\right) - F_{X,Y}\left(x_2,y_1\right) - F_{X,Y}\left(x_1,y_2\right) + F_{X,Y}\left(x_1,y_1\right) \tag{20}$$

Property (i) follows by noting that the semi-infinite rectangle defined by (x_1, y_1) is contained in that defined by (x_2, y_2) and applying Corollary 7. Properties (ii) to (iv) are obtained by limiting arguments. For example, the sequence $\{x \leq x_1 \text{ and } y \leq -n\}$ is decreasing and approaches the empty set $\varnothing$, so

$$F_{X,Y}\left(x_1,-\infty\right) = \lim_{n \to \infty} F_{X,Y}\left(x_1,-n\right) = P[\varnothing] = 0$$

For property (iii) we take the sequence $\{x \leq x_1 \text{ and } y \leq n\}$ which increases to $\{x \leq x_1\}$, so

$$\lim_{n \to \infty} F_{X,Y}\left(x_1,n\right) = P\left[X \leq x_1\right] = F_X\left(x_1\right)$$

For property (v) note in Fig. 11(a) that $B = \{x_1 < x \leq x_2, y \leq y_1\} = \{X \leq x_2,\ Y \leq y_1\} - \{X \leq x_1, Y \leq y_1\}$, so $P[B] = P\left[x_1 < X \leq x_2, Y \leq y_1\right] =$

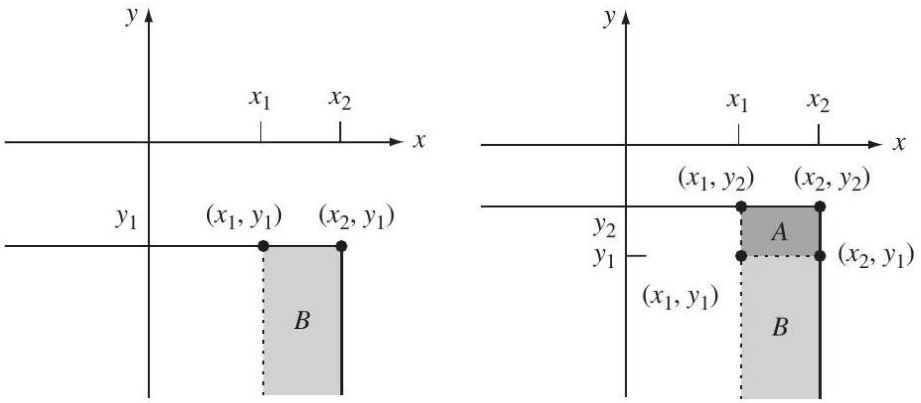

Figure 11: The joint cdf can be used to determine the probability of various events.

$F_{X,Y}(x_2, y_1)$ - $F_{X,Y}(x_1, y_1)$. In Fig. 11(b), note that $F_{X,Y}(x_2, y_2) = P[A] + P[B] + F_{X,Y}(x_1, y_2)$. Property (v) follows by solving for $P[A]$ and substituting the expression for $P[B]$.

Example 17

Plot the joint cdf of X and Y from Example 12. Find the marginal cdf of X.

To find the cdf of $\mathbf{X}$, we identify the regions in the plane according to which points in $S_{X,Y}$ are included in the rectangular region defined by (x, y). For example,

- The regions outside the first quadrant do not include any of the points, so $F_{X,Y}(x, y) = 0$.
- The region $\{0 \leq x < 1, 0 \leq y < 1\}$ contains the point $(0, 0)$, so $F_{X,Y}(x, y) = 1/4$.

Figure 12 shows the cdf after all possible regions are examined.

We need to consider several cases to find $F_X(x)$. For $x < 0$, we have $F_X(x) = 0$. For $0 \leq x < 1$, we have $F_X(x) = F_{X,Y}(x, \infty) = 9/16$. For $1 \leq x < 2$, we have $F_X(x) = F_{X,Y}(x, \infty) = 15/16$. Finally, for $x \geq 1$, we have $F_X(x) = F_{X,Y}(x, \infty) = 1$. Therefore $F_X(x)$ is a staircase function

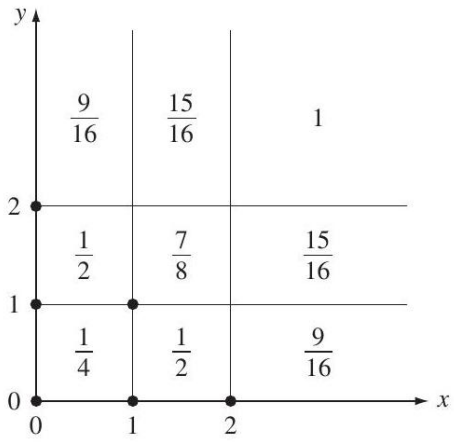

Figure 12: Joint cdf for packet switch example.

and X is a discrete random variable with $p_X(0) = 9/16, p_X(1) = 6/16$, and $p_X(2) = 1/16$.

Example 18

The joint cdf for the pair of random variables $\mathbf{X} = (X, Y)$ is given by

$$F_{X,Y}(x,y) = \begin{cases} 0 & x < 0 \text{ or } y < 0 \\ xy & 0 \le x \le 1, 0 \le y \le 1 \\ x & 0 \le x \le 1, y > 1 \\ y & 0 \le y \le 1, x > 1 \\ 1 & x \ge 1, y \ge 1 \end{cases}$$

Plot the joint cdf and find the marginal cdf of X.

Figure 13 shows a plot of the joint cdf of X and $Y \cdot F_{X,Y}(x,y)$ is continuous for all points in the plane. $F_{X,Y}(x,y) = 1$ for all $x \ge 1$ and $y \ge 1$, which implies that X and Y each assume values less than or equal to one.

The marginal cdf of X is:

$$F_X(x) = F_{X,Y}(x,\infty) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x \ge 1 \end{cases}$$

X is uniformly distributed in the unit interval.

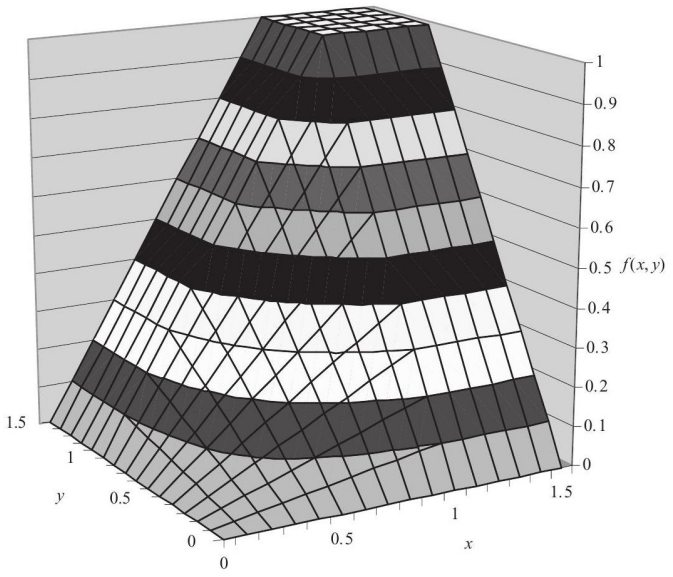

Figure 13: Joint cdf for two uniform random variables.

Example 19

The joint cdf for the vector of random variable $\mathbf{X} = (X, Y)$ is given by

$$F_{X,Y}(x,y) = \begin{cases} \left(1 - e^{-\alpha x}\right)\left(1 - e^{-\beta y}\right) & x \geq 0, y \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$

Find the marginal cdf's.

The marginal cdf's are obtained by letting one of the variables approach infinity:

$$\begin{aligned} F_X(x) &= \lim_{y\to\infty} F_{X,Y}(x,y) = 1 - e^{-\alpha x} \quad x \geq 0 \\ F_Y(y) &= \lim_{x\to\infty} F_{X,Y}(x,y) = 1 - e^{-\beta y} \quad y \geq 0. \end{aligned}$$

X and Y individually have exponential distributions with parameters α and β, respectively.

Example 20

Find the probability of the events $A = \{X \leq 1, Y \leq 1\}, B = \{X > x, Y > y\}$, where $x > 0$ and $y > 0$, and $D = \{1 < X \leq 2, 2 < Y \leq 5\}$ in Example 19.

The probability of A is given directly by the cdf:

$$P[A] = P[X \leq 1, Y \leq 1] = F_{X,Y}(1,1) = \left(1 - e^{-\alpha}\right)\left(1 - e^{-\beta}\right).$$

The probability of B requires more work. By DeMorgan's rule:

$$B^c = (\{X > x\} \cap \{Y > y\})^c = \{X \leq x\} \cup \{Y \leq y\}$$

Corollary 5 **in Section 2.2** gives the probability of the union of two events:

$$\begin{aligned} P\left[B^c\right] &= P[X \leq x] + P[Y \leq y] - P[X \leq x, Y \leq y] \\ &= \left(1 - e^{-\alpha x}\right) + \left(1 - e^{-\beta y}\right) - \left(1 - e^{-\alpha x}\right)\left(1 - e^{-\beta y}\right) \\ &= 1 - e^{-\alpha x}e^{-\beta y}. \end{aligned}$$

Finally we obtain the probability of B :

$$P[B] = 1 - P\left[B^c\right] = e^{-\alpha x}e^{-\beta y}$$

You should sketch the region B on the plane and identify the events involved in the calculation of the probability of B^c.

The probability of event D is found by applying property (vi) of the joint cdf:

$$\begin{aligned} P[1 <& X \leq 2, 2 < Y \leq 5] \\ =& F_{X,Y}(2,5) - F_{X,Y}(2,2) - F_{X,Y}(1,5) + F_{X,Y}(1,2) \\ =& \left(1 - e^{-2\alpha}\right)\left(1 - e^{-5\beta}\right) - \left(1 - e^{-2\alpha}\right)\left(1 - e^{-2\beta}\right) \\ & - \left(1 - e^{-\alpha}\right)\left(1 - e^{-5\beta}\right) + \left(1 - e^{-\alpha}\right)\left(1 - e^{-2\beta}\right). \end{aligned}$$

3.1 Random Variables That Differ in Type

In some problems it is necessary to work with joint random variables that differ in type, that is, one is discrete and the other is continuous. Usually it is rather clumsy to work with the joint cdf, and so it is preferable to work with either $P[X = k, Y \leq y]$ or $P\left[X = k, y_1 < Y \leq y_2\right]$. These probabilities are sufficient to compute the joint cdf should we have to.

Example 21: Communication Channel with Discrete Input and Continuous Output

The input X to a communication channel is +1 volt or -1 volt with equal probability. The output Y of the channel is the input plus a noise voltage N that is uniformly distributed in the interval from -2 volts to +2 volts. Find $P[X = +1, Y \leq 0]$.

This problem lends itself to the use of conditional probability:

$$P[X = +1, Y \leq y] = P[Y \leq y \mid X = +1]P[X = +1]$$

where $P[X = +1] = 1/2$. When the input $X = 1$, the output Y is uniformly distributed in the interval $[-1, 3]$; therefore

$$P[Y \leq y \mid X = +1] = \frac{y+1}{4} \quad \text{for } -1 \leq y \leq 3$$

Thus $P[X = +1, Y \leq 0] = P[Y \leq 0 \mid X = +1]P[X = +1] = (1/2)(1/4) = 1/8$.