Some Concepts in Probability and Information Theory

We begin this course with a condensed survey of basic concepts in probability theory and their applications in information theory. The notion of probability plays such a fundamental role in quantum mechanics that we must have some mathematical understanding of probability before beginning our study of quantum information.

Contents

1 Sample Spaces and Probability Distributions 1
2 Random Variables, Expectation Value, Entropy, and Variance 3
3 Joint Variables, Conditional Probabilities, and Bayes’ Theorem 6
4 The i.i.d. Paradigm, Typical Sequences, and Data Compression 7
5 Exercises 10

1 Sample Spaces and Probability Distributions

Any physical situation with an uncertain future can be thought of as an experiment, and the possible outcomes of a given experiment form a set \( \Omega \) called sample space. For example, the rolling of a six-sided die is an experiment, and the sample space for this experiment consists of the six possible numbers that the die can show when it stops:

\[
\Omega = \{1, 2, 3, 4, 5, 6\}.
\]

Suppose that in addition to rolling a die, the experiment also involves flipping a two-sided coin. In this case, sample space enumerates all possible outcomes of both the die and the coin. If we let \( H/T \) denote heads/tails of the coin’s final position, then sample space consists of twelve outcomes,

\[
\Omega = \{(1, H), (1, T), (2, H), (2, T), (3, H), (3, T), (4, H), (4, T), (5, H), (5, T), (6, H), (6, T)\}.
\]

As a final example, consider a Saluki basketball game. We can treat this as an experiment in which sample space consists of just two outcomes:

\[
\Omega = \{\text{SIU wins, SIU looses}\}.
\]

An event associated with a given experiment is any collection of outcomes. This collection can consist of just a single outcome, but it can also contain multiple outcomes. Returning to
the experiment of rolling a die and flipping a coin, one can consider the event that includes all outcomes obtaining tails on the coin flip. This event \( E \) is the subset of \( \Omega \) given by
\[
E = \{(1, T), (2, T), (3, T), (4, T), (5, T), (6, T)\}.
\]
It is helpful to think of outcomes as being individual points in an experiment and events as being groups of such points. In an experiment, we say that “event \( E \) occurs” if the outcome of experiment belongs to \( E \). For a sample space \( \Omega \), the collection of all events (i.e. the set of all subsets of \( \Omega \)) is denoted by \( 2^\Omega \), and it is called the power set of \( \Omega \)

One important event is the empty set \( \emptyset \), which is the event associated with no outcomes. The empty set arises when taking the intersection of disjoint events. In general, the union \( E_1 \cup E_2 \) of two events \( E_1 \) and \( E_2 \) is the event consisting of all outcomes contained in either \( E_1 \) or \( E_2 \), while their intersection \( E_1 \cap E_2 \) is the event consisting of all outcomes contained in both \( E_1 \) and \( E_2 \). If \( E_1 \) and \( E_2 \) are disjoint, then there is no outcome common to both, and we write \( E_1 \cap E_2 = \emptyset \).

All the previous examples describe experiments whose sample space contains a finite number of elements. But it is easy to conceive of experiments that have an infinite number of outcomes. For instance, suppose you repeatedly flip a coin and stop flipping only if the coin lands on the same side two times in a row. This whole process is an experiment with an unbounded sample space
\[
\Omega = \{(H, H), (T, T), (H, T, T), (T, H, H), (H, T, H, H), (T, H, T, T), \cdots \},
\]
where the sequence \((H, T, H, \cdots)\) describes the scenario of obtaining heads on the first flip, tails on the second flip, heads on the third flip, etc.

A sample space is called discrete if it contains either a finite or countably infinite number of elements. Recall that a set \( \Omega \) is called countably infinite if there exists a bijection between \( \Omega \) and the set of natural numbers \( \mathbb{N} = \{1, 2, 3, \cdots, \} \); in other words, every element in \( \Omega \) can be identified by some positive integer and vice-versa. To see that Eq. (2) is a countably infinite set, note that for every \( n \geq 2 \), there exists one sequence of length \( n \) that begins with \( H \) as well as one sequence of length \( n \) that begins with \( T \). Thus, we can numerically label the length-\( n \) sequence beginning with \( H \) as “\( 2n - 1 \)” and the length-\( n \) sequence as “\( 2n \)”;
this establishes the desired bijection. While we will be dealing primarily with discrete sample spaces in this course, important instances of non-discrete sample spaces arise in experiments that have numerical outcomes laying anywhere within some interval on the real number line. For example, if an electron is confined to a box with walls at coordinates \( x = 0 \) and \( x = 1 \), then the position of the electron can fall anywhere in the interval \((0, 1)\).

We next move to the formal definition of probability distributions.

**Definition 1.** A probability measure or probability distribution on some discrete sample space \( \Omega \) is a function \( p : 2^\Omega \to [0,1] \) such that

(i) Normalization: \( p(\Omega) = 1 \);

(ii) Additivity: \( p(E_1 \cup E_2) = p(E_1) + p(E_2) \) if \( E_1 \) and \( E_2 \) are two disjoint events.

For an event \( E \), we say that \( p(E) \) is the probability of event \( E \) occurring in an experiment with sample space \( \Omega \). When \( E = \{\omega\} \) consists of just a single outcome, we write its probability simply as \( p(\omega) \), or just \( p_\omega \). The combination \((\Omega, p)\) of a sample space and a probability measure is called a probability space.

The two mathematical properties in Defn. 1 fit consistently with our intuitive sense of probability. The normalization condition says that with probability one, some outcome of the experiment
will occur. The additive condition says that for two disjoint events, the probability that an outcome belongs to one of the events is just the total probability of either event occurring. For finite sample spaces, the uniform probability distribution is the distribution that assigns the same probability to each outcome. That is, \( p(\omega) = \frac{1}{|\Omega|} \) for every \( \omega \in \Omega \), where \( |\Omega| \) is the size of \( \Omega \). For any event \( E \), additivity then implies that \( p(E) = \frac{|E|}{|\Omega|} \).

Finally, note that for some probability measure \( p \), we must have that \( p(\emptyset) = 0 \). This follows immediately from properties (i) and (ii) since

\[
1 = p(\Omega) = p(\Omega \cup \emptyset) = p(\Omega) + p(\emptyset) = 1 + p(\emptyset).
\]

In general, any non-negative additive function with domain \( 2^\Omega \) and satisfying \( p(\emptyset) = 0 \) is called a measure on sample space \( \Omega \).

## 2 Random Variables, Expectation Value, Entropy, and Variance

Random variables are the basic objects studied in classical information theory. This is because random variables provide a context-independent way to analyze the statistics of some experiment. In other words, one can mathematically characterize the outcomes of an experiment in a way that does not depend on the particular physical details of the experiment. Formally, we define a random variable as follows.

**Definition 2.** For a discrete probability space \( (\Omega, p) \), a real-valued discrete random variable is a function \( X : \Omega \rightarrow \mathbb{R} \). Associated with every random variable is a probability space \( (\mathcal{X}, p_X) \), where \( p_X \) is the probability distribution given by

\[
p_X(x) = p(X^{-1}(x)).
\]

We say that \( X = x \) with probability \( p_X(x) \). When the underlying random variable is clear, we will sometimes write \( p_X(x) \) simply as \( p_X \).

To make the concept of a random variable concrete, let us examine a specific example. Consider a deck of 52 cards, and let \( X \) be the random variable given by

\[
X(c) = \begin{cases} 
1 & \text{if } c \text{ is an ace,} \\
 n & \text{if } c \text{ is a card of number } n, \\
10 & \text{if } c \text{ is a face card}
\end{cases}.
\]

If we assume that the cards in the deck are uniformly distribution, then the probability distribution \( p_X \) associated with \( X \) is given by

\[
\begin{align*}
p_X(1) &= p(X^{-1}(1)) = p(\text{ace}) = \frac{4}{52} \\
p_X(n) &= p(X^{-1}(n)) = p(\text{card } # n) = \frac{4}{52} \\
p_X(10) &= p(X^{-1}(10)) = p(\text{face card}) = \frac{12}{52}.
\end{align*}
\]

In practice, one does not need a specific experiment and sample space in mind to work with random variables. Indeed, since \( (\mathcal{X}, p_X) \) is itself a probability space, one can just focus on the random variable \( X \) without worrying about physical outcomes that the values of \( X \) represent. If desired, we can always later specify physical meaning to the values of \( X \), but it not necessary. This
is what is meant by saying that random variables allow for context-free mathematical analysis. Henceforth, whenever we deal with random variables $X, Y, \text{etc.}$, we will always assume that each of these has a probability space $(\mathcal{X}, p_X), (\mathcal{Y}, p_Y), \text{etc.}$ associated with it. Also, when the random variable is clear in context, we will often omit the variable subscript on the probability distribution; that is $p_X(x)$ will sometimes be denoted simply as $p(x)$.

For a real-valued random variable, we can consider events characterized by certain algebraic relationships. For instance, we may be interested in the event that $X \geq a$ for some constant $a$, which is just the union of all $x \in \mathcal{X}$ such that $x \geq a$. The probability of such an event is denoted by

$$Pr\{X \geq a\} := \sum_{x \geq a} p_X(x). \quad (6)$$

Starting from one random variable $X$, we can obtain another using any real-valued function whose domain is $\mathcal{X}$. That is, if $f : \mathcal{X} \to \mathcal{X}' \subset \mathbb{R}$, then we can define a new random variable $X' := f(X)$ whose sample space is $\mathcal{X}'$ and who takes on values $x \in \mathcal{X}'$ with probability $p_{X'}(x) = p_X(f^{-1}(x))$. One important example of such a function is the so-called self-information. For a distribution $p_X$ over $\mathcal{X}$, this is the function $J : \mathcal{X} \to (0, +\infty]$ given by

$$J(x) = -\log p_X(x). \quad (7)$$

In this course, we will always assume that the logarithm log is taken in base 2; i.e. $\log 2 = 1$. In contrast, the natural log $\ln$ is taken in base $e$.

The self-information can sometimes be interpreted as a function that quantifies how much “surprise” one would have if a given event occurs. That is, on a scale of “zero to infinity,” one would be surprised by an amount $J(x)$ should event $x$ occur in an experiment represented by the random variable $X$. Intuitively this seems plausible since $J(x)$ increases as the probability of $x$ decreases. At an extreme is a surprise of $+\infty$, which happens when an event occurs with probability 0. However, this interpretation of the self-information is only a heuristic, and it often fails to represent how our subjective experiences of “surprise” behave. For example, suppose we encounter an experiment with a large number of possible outcomes, say 1,000, each of them equally likely to occur. At the same time, consider the flipping of a highly biased coin in which heads lands upward only with probability $1/1,000$. In these two experiments, the same self-information is assigned to any outcome in the first experiment as it is assigned to the heads outcome in the second experiment. But clearly, we will be more surprised when obtaining heads in the second experiment than obtaining any one particular outcome in the first. Despite having inconsistencies like this in its interpretation, the self-information is still an important concept in information theory because of its relationship to entropy, which we define and study below.

For every real-valued random variable, we can define its expectation value. We all have an intuitive sense of what it means to say that an experiment generates some outcome on average. Using random variables, the notion of average or expectation is made precise.

**Definition 3.** The expectation value of a real-valued discrete random variable $X$ is given by

$$E[X] = \sum_{x \in \mathcal{X}} xp_X(x). \quad (8)$$

An important example is the expectation value of the self-information. This is called the Shannon entropy, and for random variable $X$, it is given by

$$H(X) := E[J(X)] = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x), \quad (9)$$
where we take as a definition that $0 := -0 \cdot \log 0$. This is called the Shannon entropy in honor of Claude Shannon, the father of information theory. Later in this course, we will encounter a generalization of the quantity in quantum systems known as the von Neumann entropy, named in honor of the mathematical physicist John von Neumann.

The expectation value of a random variable identifies a point on the real number line around which the values of $x$ are centered when “weighted” by their respective probabilities $p_X(x)$. However, $E[X]$ does not provide any indication on how close the individual values of $x$ are to this center value. For example, the random variable $X$ that takes on values 1 and $-1$ with equal probability has an expectation value of 0. However, the same is true for random variable $Y$ taking on values $10^6$ and $10^{-6}$ with equal probability. To distinguish between $X$ and $Y$, we would like to quantify how “spread out” a random variable is in the sense of how far its values lie from the expected value. One such indicator is given by the variance.

**Definition 4.** The variance of a real-valued discrete random variable $X$ is given by

$$\sigma^2(x) = E[X - E[X]]^2.$$  \hspace{1cm} (10)

In other words, the variance of $X$ is the squared average distance of each $x$ from the expected value $E[X]$. A large variance can be interpreted as a variable that is significantly “spread out” since its values will have a large average distance from its expected value. The square-root of the variance, $\sigma$, is often called the **standard deviation** of the random variable, and we will see one application of this quantity next.

Suppose we actually perform an experiment represented by random variable $X$. What is the probability that the outcome will like near the expectation value of $X$? The following theorem, originally given by Chebyshev, provides one bound on the probability of deviating from the average.

**Theorem 1 (Chebyshev’s Inequality).** Let $X$ be a random variable with expectation value $E[X]$ and non-zero variance $\sigma^2(X)$. For any $\kappa > 0$,

$$Pr\{|X - E[X]| \geq \kappa \sigma\} < \frac{1}{\kappa^2}.$$  \hspace{1cm} (11)

**Proof.** The follow easily from Markov’s Inequality, which says that for random variable $Z$,

$$Pr\{Z \geq a\} \leq \frac{E[Z]}{a}.$$  \hspace{1cm} (12)

The proof of Eq. (12) can be seen immediately from the chain of inequalities

$$E[Z] = \sum_{z \in Z} p_Z(z) \cdot z \geq \sum_{z < a} p_Z(z) \cdot z + \sum_{z \geq a} p_Z(z) \cdot a \geq a Pr\{Z \geq a\}.$$  

Now returning to Chebyshev’s Inequality, we have

$$Pr\{|X - E[X]| \geq \kappa \sigma\} = Pr\{(X - E[X])^2 \geq \kappa^2 \sigma^2\} \leq \frac{E[(X - E(X))^2]}{\kappa^2 \sigma^2} = \frac{1}{\kappa^2},$$  \hspace{1cm} (13)

where we have used Markov’s Inequality in the second line and the definition of variance in the last. \hfill \Box
3 Joint Variables, Conditional Probabilities, and Bayes’ Theorem

Consider again the experiment of rolling a die and flipping a coin. If we were to actually perform this experiment, we would not expect the outcome of the die roll to have any “effect” on the outcome of the coin flip. A bit more precisely, for any generic die and coin, the probability that the coin lands heads should not change based on the number obtained when rolling the die. We can describe this by saying the outcome of the coin flip is independent of die roll, and vice versa.

On the other hand, one can conceive of a die and coin that are cleverly engineered so that whenever the coin is heads the die lands on an even number, while whenever the coin is tails the die lands on an odd number. In this case, we say that the outcomes of the die roll and coin flip are correlated, and the sample space is given by

$$\Omega = \{(1, T), (3, T), (5, T), (2, H), (4, H), (6, H)\}.$$  \hspace{1cm} (14)

Now the probability that the coin lands heads definitely changes based on the number of the die roll, and we know with certainty (i.e. probability = 1) that the coin is heads. If instead we learned that the die roll is even, then the sample space shrinks to

$$\Omega_{\text{even}} = \{(2, H), (4, H), (6, H)\},$$

and we know with certainty (i.e. probability = 1) that the coin is heads. If instead we learned that the die roll is odd, then the sample space shrinks to

$$\Omega_{\text{odd}} = \{(1, T), (3, T), (5, T)\},$$

and the coin will land heads with probability zero. This example illustrates the basic idea of conditional probability. The conceptual lesson is that the probability distribution over outcomes in an experiment can change when we consider the outcomes of other experiments.

Let us discuss this in more detail using two random variables $X$ and $Y$. Associated with $X$ is the probability space $(\mathcal{X}, p_X)$ and with $Y$ the probability space $(\mathcal{Y}, p_Y)$. Thus, the probability that $X = x$ is $p_X(x)$ and the probability that $Y = y$ is $p_Y(y)$. But what is the probability that $X = x$ and $Y = y$? We cannot yet answer this question because it is asking about the joint random variable $XY$ whose range is the product sample space $\mathcal{X} \times \mathcal{Y}$, and we have not yet specified a distribution for the elements in $\mathcal{X} \times \mathcal{Y}$. The latter is the joint distribution $p_{XY}(x, y)$, and together with the set $\mathcal{X} \times \mathcal{Y}$ they form a probability space $(\mathcal{X} \times \mathcal{Y}, p_{XY})$ for the joint random variables $XY$. The distributions $p_X$ and $p_Y$ are called the marginal or reduced distributions of the joint distribution $p_{XY}$, and they can be obtained directly from the joint distribution using the formulas

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{XY}(x, y)$$  \hspace{1cm} (15)

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_{XY}(x, y).$$  \hspace{1cm} (16)

Two random variables $X$ and $Y$ are called independent or uncorrelated if

$$p_{XY}(x, y) = p_X(x) \cdot p_Y(y) \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \hspace{1cm} (17)$$

Otherwise the variables are correlated and we have $p_{XY}(x, y) \neq p_X(x) \cdot p_Y(y)$ for some pair of outcomes $(x, y)$. As we saw in the example above, for correlated variables, the probability that one variable obtains a particular outcome depends on the outcome of the other variable. The precise form of this dependence is given by the conditional probability.
Definition 5. For random variables $X$ and $Y$ with joint distribution $p_{XY}$, whenever $p_Y(y) \neq 0$, the conditional distribution of $X$ given $Y = y$ is

$$p_{X|Y=y}(x) := \frac{p_{XY}(x,y)}{p_Y(y)}. \quad (18)$$

Likewise, whenever $p_X(x) \neq 0$, the conditional distribution of $Y$ given $X = x$ is

$$p_{Y|X=x}(y) := \frac{p_{XY}(x,y)}{p_X(x)}. \quad (19)$$

You can easily verify that $p_{X|Y=y}$ and $p_{Y|X=x}$ indeed specify probability distributions over $\mathcal{X}$ and $\mathcal{Y}$ respectively by summing

$$\sum_{x \in \mathcal{X}} p_{X|Y=y}(x) = \sum_{x \in \mathcal{X}} \frac{p_{XY}(x,y)}{p_Y(y)} = p_Y(y) = 1,$$

and likewise for $p_{Y|X=x}$. Through simple algebra the conditional distribution of $X$ given $Y$ and the conditional distribution of $Y$ given $X$ can be related as

$$p_{X|Y=y}(x) = \frac{p_{Y|X=x}(y)p_X(x)}{p_Y(y)}. \quad (20)$$

This is sometimes called Bayes’ Theorem, and it has vast applications in applied statistics. Later in the course we will use Bayes’ Theorem when evaluating probabilities of certain events using the information obtained through quantum measurement.

For random variables $X$ and $Y$, the expectation value of their sum $X + Y$ is easily found to be

$$E[X + Y] = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y)(x + y)$$

$$= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y)x + \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y)y$$

$$= \sum_{x \in \mathcal{X}} p(x)x + \sum_{y \in \mathcal{Y}} p(y)y = E[X] + E[Y]. \quad (21)$$

Note that $E[X + Y] = E[X] + E[Y]$ even if $X$ and $Y$ are not independent. A similar relations for the variance $\sigma^2(X + Y)$ does not hold in general. However, if $X$ and $Y$ are independent, then it is not difficult to show that $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$.

4 The i.i.d. Paradigm, Typical Sequences, and Data Compression

We can generalize the notion of a joint probability distribution to more than two random variables. A sequence of random variables $(X_1, \cdots, X_n)$ takes on a sequence of values $(x_1, \cdots, x_n)$ from the product set $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with probability $p_{X_1 \cdots X_n}(x_1, \cdots, x_n)$. Here, $p_{X_1 \cdots X_n}$ is the joint distribution for the $n$ random variables.

An important scenario considered in information theory involves a sequence of identical and identically distributed (i.i.d) random variables. More precisely, variables $(X_1, \cdots, X_n)$ are said to be i.i.d. if

1. Identical: $p_X = p_{X_i}$ for all $i = 1, \cdots, n$, where $p_X$ is some fixed distribution;
2. Independent: \( p_{X_1 \cdots X_n} = \prod_{i=1}^{n} p_{X_i} \).

Physically, we use i.i.d. variables to model some process that generates a “stream” of random variables, each being the same and independent of one another. This physical process is sometimes called an i.i.d. source. The easiest example of an i.i.d. source is \( n \) independent coin flips. In this case, each coin flip generates heads or tails with probability 1/2 each, and the outcome of the \( i^{th} \) coin flip is independent of the \( j^{th} \) coin flip for any \( i \neq j \).

A sequence of i.i.d. variables \((X_1, \cdots, X_n)\) is usually denoted as \( X^n = (X_1, \cdots, X_n) \), and a generic sequence of values from \( \mathcal{X}^n \) is denote as \( x^n = (x_1, \cdots, x^n) \). Observe that \( X^n \) is itself a random variable, and its self-information is computed as

\[
J(x^n) = - \log p(x_1, \ldots, x_n) = - \log \prod_{i=1}^{n} p(x_i) = - \sum_{i=1}^{n} \log p(x_i) = \sum_{i=1}^{n} J(x_i),
\]

(22)

where the fact that \( p(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i) \) follows from the independence of the \( X_i \). Hence, the entropy of \( X^n \) is

\[
H(X^n) = E[J(X^n)] = E \left[ \sum_{i=1}^{n} J(X_i) \right] = \sum_{i=1}^{n} E[J(X_i)] = \sum_{i=1}^{n} H(X_i) = nH(X).
\]

(23)

The last equality uses the assumption that the \( X_i \) are identical with common distribution \( p_X \). Similarly, since the \( X_i \) are independent, the variance of \( X^n \) is given by

\[
\sigma^2(X^n) = n\sigma^2(X).
\]

(24)

Let us now turn to an actual experiment described by i.i.d. random variables \( X^n = (X_1, \cdots, X_n) \). Let \( \epsilon > 0 \) be any real number. What is the probability that the experiment should generate a sequence \( X^n \) whose self-information is at least \( n\epsilon \) away from the expected self-information \( nH(X) = E[J(X^n)] \)? Mathematically, this is phrased in terms of random variables and events as

\[
Pr\{[J(x^n) - E[J(X^n)]] \geq n\epsilon \} = “The probability that [\text{\textit{J}}(x^n) - \text{\textit{E}}[J(X^n)]] \geq n\epsilon \text{ for variable } X^n”. \]

Since \( \sigma(X^n) = \sqrt{n}\sigma(X) \), this probability can be computed from Chebyshev’s Inequality by taking \( \kappa = \frac{\sqrt{n}\epsilon}{\sigma(X)} \). Then Eq. (11) directly yields

\[
Pr\{[J(X^n) - nH(X)] \geq n\epsilon \} \leq \frac{\sigma(X)}{n\epsilon^2}.
\]

(25)

In other words, for an i.i.d. sequence of random variables \((X_1, \cdots, X_N)\),

\[
Pr\left\{ - \frac{1}{n} \sum_{i=1}^{n} \log P(X_i) - H(X) \right\} \geq \epsilon \right\} \leq \frac{\sigma(X)}{n\epsilon^2}.
\]

(26)

Notice that for any fixed \( \epsilon > 0 \), the RHS of Eq. (26) goes to zero as \( n \to \infty \). In particular, for arbitrary \( \delta > 0 \), \( n \) can be taken sufficiently large so that the RHS is less than \( \delta \). This says that for any \( \epsilon, \delta > 0 \), the probability that \( J(x^n) \) deviates from \( H(X) \) by more than \( \epsilon \) can be made less than \( \delta \) by taking \( n \) large enough. Usually \( \epsilon \) and \( \delta \) are both taken small so that with very high probability one obtains a sequence \( x^n \) with self-information very close to \( H(X) \). This conclusion motivates the following class of sequences.
Definition 6. Fix arbitrary $\epsilon > 0$ and integer $n$. For a random variable $X$ with distribution $p_X$ over set $\mathcal{X}$, an i.i.d. generated sequence $x^n \in \mathcal{X}^n$ is called $\epsilon$-typical if $|J(x^n) - H(X)| < \epsilon$. This is equivalent to the condition that
\[
2^{-n(H(X)+\epsilon)} \leq p_X^n(x^n) \leq 2^{-n(H(X)-\epsilon)}.
\] (27)

For every $\epsilon$ and integer $n$, the collection of all $\epsilon$-typical sequences is called the typical set, and it is denoted by $A_\epsilon^{(n)}$.

We now state and prove two fundamental properties of the typical set, which shows the importance of the entropy in analyzing i.i.d. variables. Later in the course, we will study a measure of entanglement called the entropy of entanglement. The physical meaning of this entanglement measure relies on the following theorem.

Theorem 2 (Typicality Theorem). For a random variable $X$, let $\epsilon > 0$ be any fixed integer.

1. The typical set is highly probable set. For any $\delta > 0$,
\[
Pr\{X^n \in A_\epsilon^{(n)}\} > 1 - \delta
\]
for all $n$ sufficiently large.

2. The number of typical sequences $|A_\epsilon^{(n)}|$ is bounded. For any $\delta > 0$,
\[
(1 - \delta)2^{n(H(X)+\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}
\]
for all $n$ sufficiently large.

Proof. Property 1. follows immediately from Eq. (26). For property 2., the upper bound is easily obtained by observing
\[
1 = \sum_{x^n \in \mathcal{X}^n} p_X^n(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)}} p_X^n(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \geq |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)}.
\]

Similarly, the lower bound follows from property 1. since
\[
1 - \delta < Pr\{X^n \in A_\epsilon^{(n)}\} = \sum_{x^n \in A_\epsilon^{(n)}} p_X^n(x^n) \leq |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)}.
\]

\[\square\]

We close this section by applying Theorem 2 to the information-theoretic task of data compression. Suppose you want to communicate the outcome of an i.i.d. experiment $X^n$ to your friend, but your ability to communicate is limited. Specifically, suppose that you can only send your friend a sequence of 0’s and 1’s consisting of $m < n$ digits; that is, your message will have the form $(0, 0, 1, 1, \ldots, 0)$. This is known as an $m$-bit message with each 0/1 in the sequence being called a bit of classical information. From this $m$-bit message, your friend wants to correctly determine which outcome sequence $x^n$ of your experiment actually occurred. This task is called data compression, and it leads to the following question: How small can $m$ be so that with high probability your friend will correctly learn the outcome of your experiment?
For example, suppose that $X = \{0, 1, \cdots, 9\}$ is a set of the first 10 non-negative integers. Then a sequence $x^n \in X^n$ will look like $x^n = (2, 7, 5, 1, 1, 0, \cdots, 8, 9)$, where there are a total of $n$ digits in this sequence. In total there are $|X|^n = 10^n$ such sequences in $X^n$. For each of these sequences, you must assign one of the $2^m$ $m$-bit sequence consisting of 0’s and 1’s.

Formally, we describe the task of data compression using encoding and decoding functions. The encoder $f$ is a mapping $f : X^n \rightarrow \{0, 1\}^m$ and the decoder $g$ is another mapping $g : \{0, 1\}^m \rightarrow X^n$. For a given error threshold $\delta > 0$, the goal is to find an encoder/decoder pair such that

$$Pr\{X^n = g(f(X^n))\} > 1 - \delta. \quad (29)$$

If such an encoder/decoder pair can be found, we say that $\delta$-good compression is achievable at rate $\frac{R}{n}$. The following theorem was first proved by Shannon in his 1948 seminal paper [Sha48].

**Theorem 3 (Data Compression).** For any $\delta > 0$ and $R > H(X)$, $\delta$-good compression can always be achieved at rate $R$.

**Proof.** Let $\epsilon > 0$ be arbitrarily chosen. By property 2. of Theorem 2, there are at most $2^{n(H(X) + \epsilon)}$ typical sequences in $A_{\epsilon}^{(n)}$. Thus, there exists an encoder $f$ that maps each element of $A_{\epsilon}^{(n)}$ to a unique $m$-bit sequence, where $m = \lceil n(H(X) + \epsilon) \rceil$. If $x^n \notin A_{\epsilon}^{(n)}$, then let $f(x^n) = f(\hat{x}^n)$ for some fixed $\hat{x}^n \in A_{\epsilon}^{(n)}$. In this way, every typical sequence is mapped to a different $m$-bit sequence while all atypical sequences are mapped to the same $m$-bit sequence; this is how the data is being compressed. We define the decoder $g$ to simply invert the encoder $f$, while ignoring the atypical sequences. That is, $g(f(x^n)) := f^{-1}(f(x^n)) \cap A_{\epsilon}^{(n)}$. Clearly, $f^{-1}(f(x^n)) \cap A_{\epsilon}^{(n)} = x^n$ whenever $x^n \in A_{\epsilon}^{(n)}$. Hence, this compression scheme fails only for atypical sequences. That is,

$$Pr\{X^n = g(f(X^n))\} = 1 - Pr\{X^n \in A_{\epsilon}^{(n)}\} > 1 - \delta \quad (30)$$

for all $n$ sufficiently large, where we use property 1. of Theorem 2. The rate of this $\delta$-good compression scheme is

$$R = \frac{m}{n} = \frac{\lceil n(H(X) + \epsilon) \rceil}{n} \geq \frac{n(H(X) + \epsilon)}{n} = H(X) + \epsilon. \quad (31)$$

Since $\epsilon$ is arbitrary, any $R > H(X)$ is achievable.

Theorem 3 says that any i.i.d. source can be reliably compressed to $H(X)$ bits per copy of $X$. It is also possible to prove the converse of this statement. That is, for any compression scheme with rate less than $H(X)$, the probability of a decoding error cannot be made arbitrary small (in fact the error probability converges to one as $n \rightarrow \infty$). Thus, in summary, the entropy $H(X)$ of a random variable characterizes precisely the optimal rate that it can be reliably compressed and restored when presented as an i.i.d. source.

## 5 Exercises

### Exercise 1

Consider the rolling of two six-sided dice. Let $E_1$ be the event that the dice land on the same number, and $E_2$ the event that their sum is greater than six.

1. What is the size of $E_2$ (i.e how many outcomes does it contain)?
2. What is $E_1 \cap E_2$?
3. What is the size of $E_1 \cup E_2$?

**Exercise 2**

Suppose you are given a well-shuffled standard deck of 52 cards (consisting of Aces, 2-10s, Jacks, Queens, and Kings, with 4 suits each). You select two cards from the deck. Assuming a uniform distribution over the cards,

1. What is the probability you choose two cards of the same color?
2. What is the probability you choose two cards of the same suit?
3. What is the probability you choose cards that do not form a pair?

**Exercise 3**

Assume that every human pregnancy yields a male or female with equal probability.

1. A woman has two children. One of them is a boy. What is the probability that this boy has a sister?
2. A woman has two children. The oldest one is a boy. What is the probability that this boy has a sister.

**Exercise 4**

A new disease is discovered that is found to be fatal 50% of the time when contracted. An experimental drug is developed to treat it. Among the survivors of the disease, 40% of them took the drug, while among the non-survivors, 10% of them also took the drug. Based on these findings, what is the probability of surviving the disease if the experimental drug is taken?

**Exercise 5**

From the definition of variance, prove that $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$ if $X$ and $Y$ are independent. Give a specific example of correlated random variables $X$ and $Y$ for which this equality does not hold.

**Exercise 6**

Consider a binary random variable $X$ with probabilities $p_X(0) = 3/4$ and $p_X(1) = 1/4$. Let $n = 4$ and $\epsilon = 1/100$.

1. What sequences of $\{0, 1\}^n$ belong to $A_\epsilon^{(n)}$?
2. For the sequence of i.i.d. random variables $X^n$, explicitly compute $Pr\{X^n \in A_\epsilon^{(n)}\}$.

**References**