# 2 Fundamentals of information theory

We begin with a brief overview of some of the fundamental concepts and mathematical tools of information theory. This allows us to establish notation and to set the stage for the results presented in subsequent chapters. For a comprehensive introduction to the fundamental concepts and methods of information theory, we refer the interested reader to the textbooks of Gallager [2], Cover and Thomas [3], Yeung [4], and Csiszár and Körner [5].

The rest of the chapter is organized as follows. Section 2.1 provides an overview of the basic mathematical tools and metrics that are relevant for subsequent chapters. Section 2.2 illustrates the fundamental proof techniques used in information theory by discussing the point-to-point communication problem and Shannon's coding theorems. Section 2.3 is entirely devoted to network information theory, with a special emphasis on distributed source coding and multi-user communications as they relate to information-theoretic security.

## 2.1 Mathematical tools of information theory

The following subsections describe a powerful set of metrics and tools that are useful to characterize the fundamental limits of communication systems. All results are stated without proof through a series of lemmas and theorems, and we refer the reader to standard textbooks [2, 3, 4] for details. Unless specified otherwise, all random variables and random vectors used throughout this book are real-valued random vectors.

### 2.1.1 Useful bounds

We start by recalling a few inequalities that are useful to bound the probabilities of rare events.

**Lemma 2.1** (Markov's inequality)**.** *Let* $\mathsf{X}$ *be a non-negative real-valued random variable. Then,*

$$\forall a > 0 \quad \mathbb{P}[\mathsf{X} \geqslant a] \leqslant \frac{\mathbb{E}_{\mathsf{X}}[\mathsf{X}]}{a}.$$

The following consequence of Markov's inequality is particularly useful.

**Lemma 2.2** (Selection lemma)**.** *Let $\mathsf{X}_n \in \mathcal{X}_n$ be a random variable and let $\mathcal{F}$ be a finite set of functions $f : \mathcal{X}_n \to \mathbb{R}^+$ such that $|\mathcal{F}|$ does not depend on $n$ and*

$$\forall f \in \mathcal{F} \quad \mathbb{E}_{\mathsf{X}_n}[f(\mathsf{X}_n)] \leqslant \delta(n).$$

*Then, there exists a specific realization $x_n$ of $\mathsf{X}_n$ such that*

$$\forall f \in \mathcal{F} \quad f(x_n) \leqslant \delta(n).$$

*Proof.* Let $\epsilon_n \triangleq \delta(n)$. Using the union bound and Markov's inequality, we obtain

$$\begin{aligned}
\mathbb{P}_{\mathsf{X}_n}\left[\bigcup_{f\in\mathcal{F}} \{f(\mathsf{X}_n) \geqslant (|\mathcal{F}|+1)\epsilon_n\}\right] &\leqslant \sum_{f\in\mathcal{F}} \mathbb{P}_{\mathsf{X}_n}[f(\mathsf{X}_n) \geqslant (|\mathcal{F}|+1)\epsilon_n] \\
&\leqslant \sum_{f\in\mathcal{F}} \frac{\mathbb{E}_{\mathsf{X}_n}[f(\mathsf{X}_n)]}{(|\mathcal{F}|+1)\epsilon_n} \\
&\leqslant \frac{|\mathcal{F}|}{|\mathcal{F}|+1} \\
&< 1.
\end{aligned}$$

Therefore, there exists *at least* one realization $x_n$ of $\mathsf{X}_n$ such that

$$\forall f \in \mathcal{F} \quad f(x_n) \leqslant (|\mathcal{F}|+1)\epsilon_n.$$

Since $\epsilon_n = \delta(n)$ and $|\mathcal{F}|$ is finite and independent of $n$, we can write $(|\mathcal{F}|+1)\epsilon_n$ as $\delta(n)$. □

We call Lemma 2.2 the "selection lemma" because it tells us that, if $\mathbb{E}_{\mathsf{X}_n}[f(X_n)] \leqslant \delta(n)$ for all $f \in \mathcal{F}$, we can select a specific realization $x_n$ such that $f(x_n) \leqslant \delta(n)$ for all $f \in \mathcal{F}$. Two other useful consequences of Markov's inequality are Chebyshev's inequality and Chernov bounds.

**Lemma 2.3** (Chebyshev's inequality)**.** *Let $\mathsf{X}$ be a real-valued random variable. Then,*

$$\forall a > 0 \quad \mathbb{P}[|\mathsf{X} - \mathbb{E}_{\mathsf{X}}[\mathsf{X}]| \geqslant a] \leqslant \frac{\mathrm{Var}(\mathsf{X})}{a^2}.$$

**Lemma 2.4** (Chernov bounds)**.** *Let $\mathsf{X}$ be a real-valued random variable. Then, for all $a > 0$,*

$$\begin{aligned}
\forall s > 0 \quad \mathbb{P}[\mathsf{X} \geqslant a] &\leqslant \mathbb{E}_{\mathsf{X}}\left[\mathrm{e}^{s\mathsf{X}}\right]\mathrm{e}^{-sa}, \\
\forall s < 0 \quad \mathbb{P}[\mathsf{X} \leqslant a] &\leqslant \mathbb{E}_{\mathsf{X}}\left[\mathrm{e}^{s\mathsf{X}}\right]\mathrm{e}^{-sa}.
\end{aligned}$$

### 2.1.2 Entropy and mutual information

In this section, we define a series of useful information-theoretic quantities, whose operational significance will become clear in the next section.

**Definition 2.1.** *Let* $\mathsf{X}$ *and* $\mathsf{X}'$ *be two discrete random variables defined on the same alphabet* $\mathcal{X}$*. The variational distance between* $\mathsf{X}$ *and* $\mathsf{X}'$ *is*

$$\mathbb{V}(\mathsf{X}, \mathsf{X}') \triangleq \sum_{x \in \mathcal{X}} |p_{\mathsf{X}}(x) - p_{\mathsf{X}'}(x)|.$$

**Definition 2.2.** *Let* $\mathsf{X} \in \mathcal{X}$ *be a discrete random variable with distribution* $p_{\mathsf{X}}$*. The Shannon entropy (or entropy for short) of* $\mathsf{X}$ *is defined as*

$$\mathbb{H}(\mathsf{X}) \triangleq -\sum_{x \in \mathcal{X}} p_{\mathsf{X}}(x) \log p_{\mathsf{X}}(x),$$

*with the convention that* $0 \log 0 \triangleq 0$*. Unless specified otherwise, all logarithms are taken to the base two and the unit for entropy is called a bit.*

If $\mathcal{X} = \{0, 1\}$, then $\mathsf{X}$ is a binary random variable and its entropy depends solely on the parameter $p = \mathbb{P}[\mathsf{X} = 0]$. The *binary entropy function* is defined as

$$\mathbb{H}_{\mathrm{b}}(p) \triangleq -p \log p - (1 - p) \log(1 - p).$$

**Lemma 2.5.** *For any discrete random variable* $\mathsf{X} \in \mathcal{X}$

$$0 \leqslant \mathbb{H}(\mathsf{X}) \leqslant \log|\mathcal{X}|.$$

*The equality* $\mathbb{H}(\mathsf{X}) = 0$ *holds if and only if* $\mathsf{X}$ *is a constant while the equality* $\mathbb{H}(\mathsf{X}) = \log|\mathcal{X}|$ *holds if and only if* $\mathsf{X}$ *is uniform on* $\mathcal{X}$*.*

Conceptually, $\mathbb{H}(\mathsf{X})$ can be viewed as a measure of the average amount of information contained in $\mathsf{X}$ or, equivalently, the amount of uncertainty that subsists until the outcome of $\mathsf{X}$ is revealed.

**Proposition 2.1** (Csiszár and Körner)**.** *Let* $\mathsf{X}$ *and* $\mathsf{X}'$ *be two discrete random variables defined on the same alphabet* $\mathcal{X}$*. Then,*

$$\left|\mathbb{H}(\mathsf{X}) - \mathbb{H}(\mathsf{X}')\right| \leqslant \mathbb{V}(\mathsf{X}, \mathsf{X}') \log\left(\frac{|\mathcal{X}|}{\mathbb{V}(\mathsf{X}, \mathsf{X}')}\right).$$

**Definition 2.3.** *Let* $\mathsf{X} \in \mathcal{X}$ *and* $\mathsf{Y} \in \mathcal{Y}$ *be two discrete random variables with joint distribution* $p_{\mathsf{XY}}$*. The joint entropy of* $\mathsf{X}$ *and* $\mathsf{Y}$ *is defined as*

$$\mathbb{H}(\mathsf{XY}) \triangleq -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{\mathsf{XY}}(x, y) \log p_{\mathsf{XY}}(x, y).$$

**Definition 2.4.** *Let* $\mathsf{X} \in \mathcal{X}$ *and* $\mathsf{Y} \in \mathcal{Y}$ *be two discrete random variables with joint distribution* $p_{\mathsf{XY}}$*. The conditional entropy of* $\mathsf{X}$ *given* $\mathsf{Y}$ *is defined as*

$$\mathbb{H}(\mathsf{Y}|\mathsf{X}) \triangleq -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{\mathsf{XY}}(x, y) \log p_{\mathsf{Y}|\mathsf{X}}(y|x).$$

By expanding $\mathbb{H}(\mathsf{XY})$ with Bayes' rule, one can verify that

$$\mathbb{H}(\mathsf{XY}) = \mathbb{H}(\mathsf{X}) + \mathbb{H}(\mathsf{Y}|\mathsf{X}).$$

This expansion generalizes to the entropy of a random vector $\mathsf{X}^n = (\mathsf{X}_1, \ldots, \mathsf{X}_n)$ as

$$\begin{aligned}\mathbb{H}(\mathsf{X}^n) &= \mathbb{H}(\mathsf{X}_1) + \mathbb{H}(\mathsf{X}_2|\mathsf{X}_1) + \cdots + \mathbb{H}\big(\mathsf{X}_n|\mathsf{X}^{n-1}\big) \\ &= \sum_{i=1}^{n} \mathbb{H}\big(\mathsf{X}_i|\mathsf{X}^{i-1}\big),\end{aligned}$$

with the convention that $\mathbb{H}\big(\mathsf{X}_1|\mathsf{X}^0\big) \triangleq \mathbb{H}(\mathsf{X}_1)$. This expansion is known as the *chain rule of entropy*.

**Lemma 2.6** ("Conditioning does not increase entropy")**.** *Let $\mathsf{X} \in \mathcal{X}$ and $\mathsf{Y} \in \mathcal{Y}$ be two discrete random variables with joint distribution $p_{\mathsf{XY}}$. Then,*

$$\mathbb{H}(\mathsf{X}|\mathsf{Y}) \leqslant \mathbb{H}(\mathsf{X}).$$

In other words, this lemma asserts that knowledge of $\mathsf{Y}$ cannot increase our uncertainty about $\mathsf{X}$.

**Definition 2.5.** *Let $\mathsf{X} \in \mathcal{X}$ and $\mathsf{Y} \in \mathcal{Y}$ be two discrete random variables with joint distribution $p_{\mathsf{XY}}$. The mutual information between $\mathsf{X}$ and $\mathsf{Y}$ is defined as*

$$\mathbb{I}(\mathsf{X};\mathsf{Y}) \triangleq \mathbb{H}(\mathsf{X}) - \mathbb{H}(\mathsf{X}|\mathsf{Y}).$$

*Let $\mathsf{X} \in \mathcal{X}$, $\mathsf{Y} \in \mathcal{Y}$, and $\mathsf{Z} \in \mathcal{Z}$ be discrete random variables with joint distribution $p_{\mathsf{XYZ}}$. The conditional mutual information between $\mathsf{X}$ and $\mathsf{Y}$ given $\mathsf{Z}$ is*

$$\mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) \triangleq \mathbb{H}(\mathsf{X}|\mathsf{Z}) - \mathbb{H}(\mathsf{X}|\mathsf{YZ}).$$

Intuitively, $\mathbb{I}(\mathsf{X};\mathsf{Y})$ represents the uncertainty about $\mathsf{X}$ that is not resolved by the observation of $\mathsf{Y}$. By using the chain rule of entropy, one can expand the mutual information between a random vector $\mathsf{X}^n = (\mathsf{X}_1, \ldots, \mathsf{X}_n)$ and a random variable $\mathsf{Y}$ as

$$\mathbb{I}(\mathsf{X}^n;\mathsf{Y}) = \sum_{i=1}^{n} \mathbb{I}\big(\mathsf{X}_i;\mathsf{Y}|\mathsf{X}^{i-1}\big),$$

with the convention that $\mathbb{I}\big(\mathsf{X}_1;\mathsf{Y}|\mathsf{X}^0\big) \triangleq \mathbb{I}(\mathsf{X}_1;\mathsf{Y})$. This expansion is known as the *chain rule of mutual information*.

**Lemma 2.7.** *Let $\mathsf{X} \in \mathcal{X}$ and $\mathsf{Y} \in \mathcal{Y}$ be two discrete random variables with joint distribution $p_{\mathsf{XY}}$. Then,*

$$0 \leqslant \mathbb{I}(\mathsf{X};\mathsf{Y}) \leqslant \min(\mathbb{H}(\mathsf{X}), \mathbb{H}(\mathsf{Y})).$$

*The equality $\mathbb{I}(\mathsf{X};\mathsf{Y}) = 0$ holds if and only if $\mathsf{X}$ and $\mathsf{Y}$ are independent. The equality $\mathbb{I}(\mathsf{X};\mathsf{Y}) = \mathbb{H}(\mathsf{X})$ ($\mathbb{I}(\mathsf{X};\mathsf{Y}) = \mathbb{H}(\mathsf{Y})$) holds if and only if $\mathsf{X}$ is a function of $\mathsf{Y}$ ($\mathsf{Y}$ is a function of $\mathsf{X}$).*

**Lemma 2.8.** *Let $\mathsf{X} \in \mathcal{X}$, $\mathsf{Y} \in \mathcal{Y}$, and $\mathsf{Z} \in \mathcal{Z}$ be three discrete random variables with joint distribution $p_{\mathsf{XYZ}}$. Then,*

$$0 \leqslant \mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) \leqslant \min(\mathbb{H}(\mathsf{X}|\mathsf{Z}), \mathbb{H}(\mathsf{Y}|\mathsf{Z})).$$

*The equality* $\mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) = 0$ *holds if and only if* $\mathsf{X}$ *and* $\mathsf{Y}$ *are conditionally independent given* $\mathsf{Z}$*. In this case, we say that* $\mathsf{X} \to \mathsf{Z} \to \mathsf{Y}$ *forms a Markov chain. The equality* $\mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) = \mathbb{H}(\mathsf{X}|\mathsf{Z})$ *(*$\mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{Z}) = \mathbb{H}(\mathsf{Y}|\mathsf{Z})$*) holds if and only if* $\mathsf{X}$ *is a function of* $\mathsf{Y}$ *and* $\mathsf{Z}$ *(*$\mathsf{Y}$ *is a function of* $\mathsf{X}$ *and* $\mathsf{Z}$*).*

**Lemma 2.9** (Data-processing inequality)**.** *Let* $\mathsf{X} \in \mathcal{X}$*,* $\mathsf{Y} \in \mathcal{Y}$*, and* $\mathsf{Z} \in \mathcal{Z}$ *be three discrete random variables such that* $\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$ *forms a Markov chain. Then,*

$$\mathbb{I}(\mathsf{X};\mathsf{Y}) \geqslant \mathbb{I}(\mathsf{X};\mathsf{Z}).$$

An equivalent form of the data-processing inequality is $\mathbb{H}(\mathsf{X}|\mathsf{Y}) \leqslant \mathbb{H}(\mathsf{X}|\mathsf{Z})$, which means that, on average, processing $\mathsf{Y}$ can only increase our uncertainty about $\mathsf{X}$.

**Lemma 2.10** (Fano's inequality)**.** *Let* $\mathsf{X} \in \mathcal{X}$ *be a discrete random variable and let* $\mathsf{X}'$ *be any estimate of* $\mathsf{X}$ *that takes values in the same alphabet* $\mathcal{X}$*. Let* $P_{\mathrm{e}} \triangleq \mathbb{P}[\mathsf{X} \neq \mathsf{X}']$ *be the probability of error obtained when estimating* $\mathsf{X}$ *with* $\mathsf{X}'$*. Then,*

$$\mathbb{H}\big(\mathsf{X}|\mathsf{X}'\big) \leqslant \mathbb{H}_{\mathrm{b}}(P_{\mathrm{e}}) + P_{\mathrm{e}} \log(|\mathcal{X}| - 1),$$

*where* $\mathbb{H}_{\mathrm{b}}(P_{\mathrm{e}})$ *is the binary entropy function defined earlier.*

Fano's inequality is the key ingredient of many proofs in this book because it relates an information-theoretic quantity (the conditional entropy $\mathbb{H}(\mathsf{X}|\mathsf{X}')$) to an operational quantity (the probability of error $P_{\mathrm{e}}$). In what follows, we often write Fano's inequality in the form $\mathbb{H}(\mathsf{X}|\mathsf{X}') \leqslant \delta(P_{\mathrm{e}})$ to emphasize that $\mathbb{H}(\mathsf{X}|\mathsf{X}')$ goes to zero if $P_{\mathrm{e}}$ goes to zero.

**Definition 2.6.** *A function* $f : \mathcal{I} \to \mathbb{R}$ *defined on a set* $\mathcal{I}$ *is convex on* $\mathcal{I}$ *if, for all* $(x_1, x_2) \in \mathcal{I}^2$ *and for all* $\lambda \in [0, 1]$*,*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leqslant \lambda f(x_1) + (1 - \lambda) f(x_2).$$

*If the equation above holds with strict inequality,* $f$ *is strictly convex on* $\mathcal{I}$*. A function* $f : \mathcal{I} \to \mathbb{R}$ *defined on a set* $\mathcal{I}$ *is (strictly) concave on* $\mathcal{I}$ *if the function* $-f$ *is (strictly) convex on* $\mathcal{I}$*.*

**Lemma 2.11** (Jensen's inequality)**.** *Let* $\mathsf{X} \in \mathcal{X}$ *be a random variable and let* $f : \mathcal{X} \to \mathbb{R}$ *be a convex function. Then,*

$$\mathbb{E}[f(\mathsf{X})] \geqslant f(\mathbb{E}[\mathsf{X}]).$$

*If* $f$ *is strictly convex, then equality holds if and only if* $\mathsf{X}$ *is a constant.*

**Lemma 2.12.** *Let* $\mathsf{X} \in \mathcal{X}$ *and* $\mathsf{Y} \in \mathcal{Y}$ *be two discrete random variables with joint distribution* $p_{\mathsf{XY}}$*. Then,*

- $\mathbb{H}(\mathsf{X})$ *is a concave function of* $p_{\mathsf{X}}$*;*
- $\mathbb{I}(\mathsf{X};\mathsf{Y})$ *is a concave function of* $p_{\mathsf{X}}$ *for* $p_{\mathsf{Y}|\mathsf{X}}$ *fixed;*
- $\mathbb{I}(\mathsf{X};\mathsf{Y})$ *is a convex function of* $p_{\mathsf{Y}|\mathsf{X}}$ *for* $p_{\mathsf{X}}$ *fixed.*

If $\mathsf{X}$ is a continuous random variable then, in general, $\mathbb{H}(\mathsf{X})$ is not well defined. It is convenient to define the differential entropy as follows.

**Definition 2.7.** *Let $\mathsf{X} \in \mathcal{X}$ be a continuous random variable with distribution $p_{\mathsf{X}}$. The differential entropy of $\mathsf{X}$ is defined as*

$$\mathbb{h}(\mathsf{X}) \triangleq -\int_{x\in\mathcal{X}} p_{\mathsf{X}}(x)\log p_{\mathsf{X}}(x)\mathrm{d}x.$$

The notions of joint entropy, conditional entropy, and mutual information for continuous random variables are identical to their discrete counterparts, but with the differential entropy $\mathbb{h}$ in place of the entropy $\mathbb{H}$. With the exception of Lemma 2.5 and Fano's inequality, all of the properties of entropy and mutual information stated above hold also for continuous random variables. In addition, the following properties will be useful in the next chapters.

**Lemma 2.13.** *If $\mathsf{X} \in \mathcal{X}$ is a continuous random variable with variance $\mathrm{Var}(\mathsf{X}) \leqslant \sigma^2$, then*

$$\mathbb{h}(\mathsf{X}) \leqslant \frac{1}{2}\log\left(2\pi e\sigma^2\right),$$

*with equality if and only if $\mathsf{X}$ has a Gaussian distribution with variance $\sigma^2$.*

**Lemma 2.14** (Entropy–power inequality)**.** *Let $\mathsf{X} \in \mathcal{X}$ and $\mathsf{Y} \in \mathcal{Y}$ be independent continuous random variables with entropy $\mathbb{h}(\mathsf{X})$ and $\mathbb{h}(\mathsf{Y})$, respectively. Let $\mathsf{X}'$ and $\mathsf{Y}'$ be independent Gaussian random variables such that $\mathbb{h}(\mathsf{X}') = \mathbb{h}(\mathsf{X})$ and $\mathbb{h}(\mathsf{Y}') = \mathbb{h}(\mathsf{Y})$. Then,*

$$\mathbb{h}(\mathsf{X}+\mathsf{Y}) \geqslant \mathbb{h}\left(\mathsf{X}'+\mathsf{Y}'\right),$$

*or, equivalently,*

$$2^{2\mathbb{h}(\mathsf{X}+\mathsf{Y})} \geqslant 2^{2\mathbb{h}(\mathsf{X}')} + 2^{2\mathbb{h}(\mathsf{Y}')}.$$

### 2.1.3 Strongly typical sequences

Let $x^n \in \mathcal{X}^n$ be a sequence whose $n$ elements are in a finite alphabet $\mathcal{X}$. The number of occurrences of a symbol $a \in \mathcal{X}$ in the sequence $x^n$ is denoted by $N(a; x^n)$, and the *empirical distribution* (or histogram) of $x^n$ is defined as the set $\{N(a; x^n)/n : a \in \mathcal{X}\}$.

**Definition 2.8** (Strong typical set)**.** *Let $p_{\mathsf{X}}$ be a distribution on a finite alphabet $\mathcal{X}$ and let $\epsilon > 0$. A sequence $x^n \in \mathcal{X}^n$ is (strongly) $\epsilon$-typical with respect to $p_{\mathsf{X}}$ if*

$$\forall a \in \mathcal{X} \quad \left|\frac{1}{n}N(a; x^n) - p_{\mathsf{X}}(a)\right| \leqslant \epsilon\, p_{\mathsf{X}}(a).$$

*The set of all $\epsilon$-typical sequences with respect to $p_{\mathsf{X}}$ is called the strong typical set and is denoted by $\mathcal{T}_\epsilon^n(\mathsf{X})$.*

In other words, the *typical set* $\mathcal{T}_\epsilon^n(\mathsf{X})$ contains all sequences $x^n$ whose empirical distribution is "close" to $p_{\mathsf{X}}$. The notion of typicality is particularly useful in information theory because of a result known as the *asymptotic equipartition property* (AEP).

**Theorem 2.1** (AEP). *Let $p_{\mathsf{X}}$ be a distribution on a finite alphabet $\mathcal{X}$ and let $0 < \epsilon < \min_{x\in\mathcal{X}} p_{\mathsf{X}}(x)$. Let $\mathsf{X}^n$ be a sequence of independent and identically distributed (i.i.d.) random variables with distribution $p_{\mathsf{X}}$. Then,*

$$\begin{aligned} 1-\delta_\epsilon(n) &\leqslant \mathbb{P}\left[\mathsf{X}^n \in \mathcal{T}_\epsilon^n(\mathsf{X})\right] \leqslant 1,\\ (1-\delta_\epsilon(n))2^{n(\mathbb{H}(\mathsf{X})-\delta(\epsilon))} &\leqslant \left|\mathcal{T}_\epsilon^n(\mathsf{X})\right| \leqslant 2^{n(\mathbb{H}(\mathsf{X})+\delta(\epsilon))},\\ \forall x^n \in \mathcal{T}_\epsilon^n(\mathsf{X}) \quad 2^{-n(\mathbb{H}(\mathsf{X})+\delta(\epsilon))} &\leqslant p_{\mathsf{X}^n}(x^n) \leqslant 2^{-n(\mathbb{H}(\mathsf{X})-\delta(\epsilon))}. \end{aligned}$$

In simple terms, the AEP states that, for sufficiently large $n$, the probability that the realization $x^n$ of a sequence of i.i.d. random variables belongs to the typical set is close to unity. Moreover, for practical purposes we may assume that the probability of any strongly typical sequence is about $2^{-n\mathbb{H}(\mathsf{X})}$ and the number of strongly typical sequences is approximately $2^{n\mathbb{H}(\mathsf{X})}$. In some sense, the AEP provides an operational interpretation of entropy.

**Remark 2.1.** *It is possible to provide explicit expressions for $\delta_\epsilon(n)$ and $\delta(\epsilon)$ [6], but the rough characterization used in Theorem 2.1 is sufficient for our purposes. In particular, it makes it easier to keep track of small terms that depend on $\epsilon$ or $n$ because we can write equations such as $\delta(\epsilon)+\delta(\epsilon)=\delta(\epsilon)$ without worrying about the exact dependence on $\epsilon$.*

The notion of typicality generalizes to multiple random variables. Assume that $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is a pair of sequences with elements in finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. The number of occurrences of a pair $(a, b) \in \mathcal{X} \times \mathcal{Y}$ in the pair of sequences $(x^n, y^n)$ is denoted by $N(a, b; x^n, y^n)$.

**Definition 2.9** (Jointly typical set). *Let $p_{\mathsf{XY}}$ be a joint distribution on the finite alphabets $\mathcal{X} \times \mathcal{Y}$ and let $\epsilon > 0$. Sequences $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$ are $\epsilon$-jointly typical with respect to $p_{\mathsf{XY}}$ if*

$$\forall (a, b) \in \mathcal{X} \times \mathcal{Y} \quad \left|\frac{1}{n}N(a, b; x^n, y^n) - p_{\mathsf{XY}}(a, b)\right| \leqslant \epsilon\, p_{\mathsf{XY}}(a, b).$$

*The set of all $\epsilon$-jointly typical sequences with respect to $p_{\mathsf{XY}}$ is called the jointly typical set and is denoted by $\mathcal{T}_\epsilon^n(\mathsf{XY})$.*

One can check that $\mathcal{T}_\epsilon^n(\mathsf{XY}) \subseteq \mathcal{T}_\epsilon^n(\mathsf{X}) \times \mathcal{T}_\epsilon^n(\mathsf{Y})$. In other words, $(x^n, y^n) \in \mathcal{T}_\epsilon^n(\mathsf{XY})$ implies that $x^n \in \mathcal{T}_\epsilon^n(\mathsf{X})$ and $y^n \in \mathcal{T}_\epsilon^n(\mathsf{Y})$. This property is known as the *consistency* of joint typicality. Notice that the jointly typical set $\mathcal{T}_\epsilon^n(\mathsf{XY})$ is the typical set $\mathcal{T}_\epsilon^n(\mathsf{Z})$ for the random variable $\mathsf{Z} = (\mathsf{X}, \mathsf{Y})$. Therefore, the result below follows directly from Theorem 2.1.

**Corollary 2.1** (Joint AEP). *Let $p_{\mathsf{XY}}$ be a joint distribution on the finite alphabets $\mathcal{X} \times \mathcal{Y}$ and let $0 < \epsilon < \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{\mathsf{XY}}(x, y)$. Let $(\mathsf{X}^n, \mathsf{Y}^n)$ be a sequence of i.i.d. random variables with joint distribution $p_{\mathsf{XY}}$. Then,*

$$\begin{aligned} 1-\delta_\epsilon(n) &\leqslant \mathbb{P}\left[(\mathsf{X}^n, \mathsf{Y}^n) \in \mathcal{T}_\epsilon^n(\mathsf{XY})\right] \leqslant 1,\\ (1-\delta_\epsilon(n))2^{n(\mathbb{H}(\mathsf{XY})-\delta(\epsilon))} &\leqslant \left|\mathcal{T}_\epsilon^n(\mathsf{XY})\right| \leqslant 2^{n(\mathbb{H}(\mathsf{XY})+\delta(\epsilon))},\\ \forall (x^n, y^n) \in \mathcal{T}_\epsilon^n(\mathsf{XY}) \quad 2^{-n(\mathbb{H}(\mathsf{XY})+\delta(\epsilon))} &\leqslant p_{\mathsf{X}^n\mathsf{Y}^n}(x^n, y^n) \leqslant 2^{-n(\mathbb{H}(\mathsf{XY})-\delta(\epsilon))}. \end{aligned}$$

It is also useful to introduce a conditional typical set, for which we can establish a conditional version of the AEP.

**Definition 2.10.** *Let $p_{\mathsf{XY}}$ be a joint distribution on the finite alphabets $\mathcal{X} \times \mathcal{Y}$ and let $\epsilon > 0$. Let $x^n \in \mathcal{T}_\epsilon^n(\mathsf{X})$. The set*

$$\mathcal{T}_\epsilon^n(\mathsf{XY}|x^n) \triangleq \left\{ y^n \in \mathcal{Y}^n : (x^n, y^n) \in \mathcal{T}_\epsilon^n(\mathsf{XY}) \right\}$$

*is called the conditional typical set with respect to $x^n$.*

**Theorem 2.2** (Conditional AEP)**.** *Let $p_{\mathsf{XY}}$ be a joint distribution on the finite alphabets $\mathcal{X} \times \mathcal{Y}$ and suppose that $0 < \epsilon' < \epsilon \leqslant \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{\mathsf{XY}}(x, y)$. Let $x^n \in \mathcal{T}_{\epsilon'}^n(\mathsf{X})$ and let $\tilde{\mathsf{Y}}^n$ be a sequence of random variables such that*

$$\forall y^n \in \mathcal{Y}^n \quad p_{\tilde{\mathsf{Y}}^n}(y^n) = \prod_{i=1}^n p_{\mathsf{Y}|\mathsf{X}}(y_i|x_i).$$

*Then,*

$$\begin{aligned} 1 - \delta_{\epsilon\epsilon'}(n) &\leqslant \mathbb{P}\left[\tilde{\mathsf{Y}}^n \in \mathcal{T}_\epsilon^n(\mathsf{XY}|x^n)\right] \leqslant 1, \\ (1 - \delta_{\epsilon\epsilon'}(n))2^{n(\mathbb{H}(\mathsf{Y}|\mathsf{X})-\delta(\epsilon))} &\leqslant \left|\mathcal{T}_\epsilon^n(\mathsf{XY}|x^n)\right| \leqslant 2^{n(\mathbb{H}(\mathsf{Y}|\mathsf{X})+\delta(\epsilon))}, \\ \forall y^n \in \mathcal{T}_\epsilon^n(\mathsf{XY}|x^n) \quad 2^{-n(\mathbb{H}(\mathsf{Y}|\mathsf{X})+\delta(\epsilon))} &\leqslant p_{\mathsf{Y}^n|\mathsf{X}^n}(y^n|x^n) \leqslant 2^{-n(\mathbb{H}(\mathsf{Y}|\mathsf{X})-\delta(\epsilon))}. \end{aligned}$$

The conditional AEP means that, if $x^n$ is a typical sequence and if $\tilde{\mathsf{Y}}^n$ is distributed according to $\prod_{i=1}^n p_{\mathsf{Y}|\mathsf{X}}(y_i|x_i)$, then $\tilde{\mathsf{Y}}^n$ is jointly typical with $x^n$ with high probability for $n$ large enough. In addition, the number of sequences $y^n$ that are jointly typical with $x^n$ is approximately $2^{n\mathbb{H}(\mathsf{Y}|\mathsf{X})}$, and their probability is on the order of $2^{-n\mathbb{H}(\mathsf{Y}|\mathsf{X})}$. The following corollary of the conditional AEP will be useful.

**Corollary 2.2.** *Let $p_{\mathsf{XY}}$ be a joint distribution on the finite alphabets $\mathcal{X} \times \mathcal{Y}$ and let $0 < \epsilon < \mu_{\mathsf{XY}}$ with $\mu_{\mathsf{XY}} \triangleq \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{\mathsf{XY}}(x, y)$. Let $\tilde{\mathsf{Y}}^n$ be a sequence of i.i.d. random variables with distribution $p_{\mathsf{Y}}$. Then,*

- *if $x^n \in \mathcal{T}_\epsilon^n(\mathsf{X})$,*

$$(1 - \delta_\epsilon(n))2^{-n(\mathbb{I}(\mathsf{X};\mathsf{Y})+\delta(\epsilon))} \leqslant \mathbb{P}\left[\tilde{\mathsf{Y}}^n \in \mathcal{T}_\epsilon^n(\mathsf{XY}|x^n)\right] \leqslant 2^{-n(\mathbb{I}(\mathsf{X};\mathsf{Y})-\delta(\epsilon))};$$

- *if $\tilde{\mathsf{X}}^n$ is a sequence of random variables independent of $\tilde{\mathsf{Y}}^n$ and with arbitrary distribution $p_{\tilde{\mathsf{X}}^n}$ on $\mathcal{X}^n$,*

$$\mathbb{P}\left[(\tilde{\mathsf{X}}^n, \tilde{\mathsf{Y}}^n) \in \mathcal{T}_\epsilon^n(\mathsf{XY})\right] \leqslant 2^{-n(\mathbb{I}(\mathsf{X};\mathsf{Y})-\delta(\epsilon))}.$$

In other words, if $\tilde{\mathsf{Y}}^n$ is generated independently of $x^n$, the probability that $\tilde{\mathsf{Y}}^n$ is jointly typical with $x^n$ is small and on the order of $2^{-n\mathbb{I}(\mathsf{X};\mathsf{Y})}$. Corollary 2.2 generalizes to more than two random variables; in particular, we make extensive use of the following result.

**Corollary 2.3.** *Let $p_{\mathsf{UXY}}$ be a joint distribution on the finite alphabets $\mathcal{U} \times \mathcal{X} \times \mathcal{Y}$ and suppose $0 < \epsilon < \mu_{\mathsf{UXY}}$ with $\mu_{\mathsf{UXY}} \triangleq \min_{(u,x,y)\in\mathcal{U}\times\mathcal{X}\times\mathcal{Y}} p_{\mathsf{UXY}}(u, x, y)$. Let $(\tilde{\mathsf{U}}^n, \tilde{\mathsf{X}}^n)$ be*

*a sequence of random variables with arbitrary distribution* $p_{\tilde{\mathsf{U}}^n\tilde{\mathsf{X}}^n}$ *on* $\mathcal{U}^n \times \mathcal{X}^n$. *Let* $\tilde{\mathsf{Y}}^n$ *be a sequence of random variables conditionally independent of* $\tilde{\mathsf{X}}^n$ *given* $\tilde{\mathsf{U}}^n$ *such that*

$$\forall (u^n, x^n, y^n) \in \mathcal{U}^n \times \mathcal{X}^n \times \mathcal{Y}^n$$
$$p_{\tilde{\mathsf{U}}^n\tilde{\mathsf{X}}^n\tilde{\mathsf{Y}}^n}(u^n, x^n, y^n) = \left(\prod_{i=1}^{n} p_{\mathsf{Y}|\mathsf{U}}(y_i|u_i)\right) p_{\tilde{\mathsf{U}}^n\tilde{\mathsf{X}}^n}(u^n, x^n).$$

*Then,*

$$\mathbb{P}_{\tilde{\mathsf{U}}^n\tilde{\mathsf{X}}^n\tilde{\mathsf{Y}}^n}\left[(\tilde{\mathsf{U}}^n, \tilde{\mathsf{X}}^n, \tilde{\mathsf{Y}}^n) \in \mathcal{T}_\epsilon^n(\mathsf{UXY})\right] \leqslant 2^{-n(\mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{U})-\delta(\epsilon))}.$$

### 2.1.4 Weakly typical sequences

Strong typicality requires the relative frequency of each possible symbol to be close to the corresponding probability; however, the notion of strong typicality does not apply to continuous random variables and it is sometimes convenient to use a *weaker* notion of typicality, which merely requires the empirical entropy of a sequence to be close to the true entropy of the corresponding random variable. All definitions and results in this section are stated for discrete random variables but hold also for continuous random variables on replacing the entropy $\mathbb{H}$ with the differential entropy $\mathbb{h}$.

**Definition 2.11** (Weakly typical set)**.** *Let* $p_{\mathsf{X}}$ *be a distribution on a finite alphabet* $\mathcal{X}$ *and let* $\epsilon > 0$. *A sequence* $x^n \in \mathcal{X}^n$ *is (weakly)* $\epsilon$*-typical with respect to* $p_{\mathsf{X}}$ *if*

$$\left| -\frac{1}{n} \log p_{\mathsf{X}^n}(x^n) - \mathbb{H}(\mathsf{X}) \right| \leqslant \epsilon.$$

*The set of all weakly* $\epsilon$*-typical sequences with respect to* $p_{\mathsf{X}}$ *is called the weakly typical set and is denoted* $\mathcal{A}_\epsilon^n(\mathsf{X})$.

The weak version of the AEP then follows from the weak law of large numbers.

**Theorem 2.3** (AEP)**.** *Let* $p_{\mathsf{X}}$ *be a distribution on a finite alphabet* $\mathcal{X}$ *and let* $\epsilon > 0$. *Let* $\mathsf{X}^n$ *be a sequence of i.i.d. random variables with distribution* $p_{\mathsf{X}}$. *Then,*

- *for n sufficiently large,* $\mathbb{P}\left[\mathsf{X}^n \in \mathcal{A}_\epsilon^n(\mathsf{X})\right] > 1 - \epsilon$;
- *if* $x^n \in \mathcal{A}_\epsilon^n(\mathsf{X})$, *then* $2^{-n(\mathbb{H}(\mathsf{X})+\epsilon)} \leqslant p_{\mathsf{X}^n}(x^n) \leqslant 2^{-n(\mathbb{H}(\mathsf{X})-\epsilon)}$;
- *for n sufficiently large,* $(1-\epsilon)2^{n(\mathbb{H}(\mathsf{X})-\epsilon)} \leqslant \left|\mathcal{A}_\epsilon^n(\mathsf{X})\right| \leqslant 2^{n(\mathbb{H}(\mathsf{X})+\epsilon)}$.

**Definition 2.12** (Jointly weak typical set)**.** *Let* $p_{\mathsf{XY}}$ *be a joint distribution on the finite alphabets* $\mathcal{X} \times \mathcal{Y}$ *and let* $\epsilon > 0$. *Sequences* $x^n \in \mathcal{X}^n$ *and* $y^n \in \mathcal{Y}^n$ *are jointly (weakly)* $\epsilon$*-typical with respect to* $p_{\mathsf{XY}}$ *if*

$$\left| -\frac{1}{n} \log p_{\mathsf{X}^n\mathsf{Y}^n}(x^n, y^n) - \mathbb{H}(\mathsf{XY}) \right| \leqslant \epsilon,$$
$$\left| -\frac{1}{n} \log p_{\mathsf{X}^n}(x^n) - \mathbb{H}(\mathsf{X}) \right| \leqslant \epsilon,$$
$$\left| -\frac{1}{n} \log p_{\mathsf{Y}^n}(y^n) - \mathbb{H}(\mathsf{Y}) \right| \leqslant \epsilon.$$

*The set of all jointly weakly $\epsilon$-typical sequences with respect to $p_{XY}$ is called the jointly weakly typical set and is denoted $\mathcal{A}_\epsilon^n(XY)$.*

**Theorem 2.4** (joint AEP)**.** *Let $p_{XY}$ be a joint distribution on the finite alphabets $\mathcal{X} \times \mathcal{Y}$ and let $\epsilon > 0$. Let $(X^n, Y^n)$ be a sequence of i.i.d. random variables with joint distribution $p_{XY}$. Then,*

- *for n sufficiently large, $\mathbb{P}\left[(X^n, Y^n) \in \mathcal{A}_\epsilon^n(XY)\right] > 1 - \epsilon$;*
- *if $(x^n, y^n) \in \mathcal{A}_\epsilon^n(XY)$, then $2^{-n(\mathbb{H}(XY)+\epsilon)} \leqslant p_{X^nY^n}(x^n, y^n) \leqslant 2^{-n(\mathbb{H}(XY)-\epsilon)}$;*
- *for n sufficiently large, $(1-\epsilon)2^{n(\mathbb{H}(XY)-\epsilon)} \leqslant \left|\mathcal{A}_\epsilon^n(XY)\right| \leqslant 2^{n(\mathbb{H}(XY)+\epsilon)}$.*

With weak typicality, there is no exact counterpart to the conditional AEP given in Corollary 2.2 but the following result holds nevertheless.

**Theorem 2.5.** *Let $p_{XY}$ be a joint distribution on the finite alphabets $\mathcal{X} \times \mathcal{Y}$ and let $\epsilon > 0$. Let $\tilde{Y}^n$ be a sequence of i.i.d. random variables with distribution $p_Y$, and let $\tilde{X}^n$ be an independent sequence of i.i.d. random variables with distribution $p_X$. Then,*

$$\mathbb{P}\left[(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon^n(XY)\right] \leqslant 2^{-n(\mathbb{I}(X;Y)-\delta(\epsilon))}.$$

In subsequent chapters, we use the term AEP for both strong and weak typicality; however, it will be clear from the context whether we refer to the theorems of Section 2.1.3 or those of Section 2.1.4.

### 2.1.5 Markov chains and functional dependence graphs

The identification of Markov chains among random variables that depend on each other via complicated relations is a recurrent problem in information theory. In principle, Markov chains can be identified by manipulating the joint probability distribution of random variables, but this is often a tedious task. In this short section, we describe a graphical yet correct method for identifying Markov chains that is based on the *functional dependence graph* of random variables.

**Definition 2.13** (Functional dependence graph)**.** *Consider m independent random variables and n functions of these variables. A functional dependence graph is a directed graph having $m + n$ vertices, and in which edges are drawn from one vertex to another if the random variable of the former vertex is an argument in the function defining the latter.*

---

**Example 2.1.** Let $M \in \mathcal{M}$ and $Z^n \in \mathbb{R}^n$ be independent random variables. Let $\{f_i\}_n$ be a set of functions from $\mathcal{M}$ to $\mathbb{R}^n$. For $i \in [\![1, n]\!]$ define the random variables $X_i = f_i(M)$ and $Y_i = X_i + Z_i$. The functional dependence graph of the random variables $M$, $X^n$, $Y^n$, and $Z^n$ is shown in Figure 2.1.

---

**Definition 2.14** (d-separation)**.** *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be disjoint subsets of vertices in a functional dependence graph $\mathcal{G}$. The subset $\mathcal{Z}$ is said to* d-separate *$\mathcal{X}$ from $\mathcal{Y}$ if there*

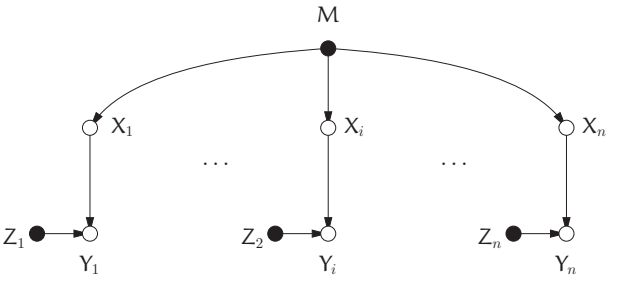


**Figure 2.1** Functional dependence graph of variables in Example 2.1. For clarity, the independent random variables are indicated by filled circles (●) whereas the functions of these random variables are indicated by empty circles (○).

*exists no path between a vertex of $\mathcal{X}$ and a vertex of $\mathcal{Y}$ after the following operations have been performed:*

- *construct the subgraph $\mathcal{G}'$ consisting of all vertices in $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, as well as the edges and vertices encountered when moving backward starting from any of the vertices in $\mathcal{X}$, $\mathcal{Y}$, or $\mathcal{Z}$;*
- *in the subgraph $\mathcal{G}'$, delete all edges coming out of $\mathcal{Z}$;*
- *remove all arrows in $\mathcal{G}'$ to obtain an undirected graph.*

The usefulness of d-separation is justified by the following theorem.

**Theorem 2.6.** *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be disjoint subsets of the vertices in a functional dependence graph. If $\mathcal{Z}$ d-separates $\mathcal{X}$ from $\mathcal{Y}$, and if we collect the random variables in $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ in the random vectors $\mathsf{X}$, $\mathsf{Y}$, and $\mathsf{Z}$, respectively, then $\mathsf{X} \to \mathsf{Z} \to \mathsf{Y}$ forms a Markov chain.*

Theorem 2.6 is particularly useful in the converse proofs of channel coding theorems.

---

**Example 2.2.** On the basis of the functional dependence graph of Figure 2.1, one can check that, for any $i \neq j$, $\mathsf{X}_i \to \mathsf{X}_j \to \mathsf{Y}_j$.

---

## 2.2 The point-to-point communication problem

The foundations of information theory were laid by Claude E. Shannon in his 1948 paper "A mathematical theory of communication" [7]. In his own words, *the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point*. If the message – for example, a letter from the alphabet, the gray level of a pixel or some physical quantity measured by a sensor – is to be reproduced at a remote location with a certain fidelity, some amount of information must be transmitted over a physical channel. This observation is the basis of Shannon's general model for *point-to-point* communication reproduced in Figure 2.2. It consists of the following elements.

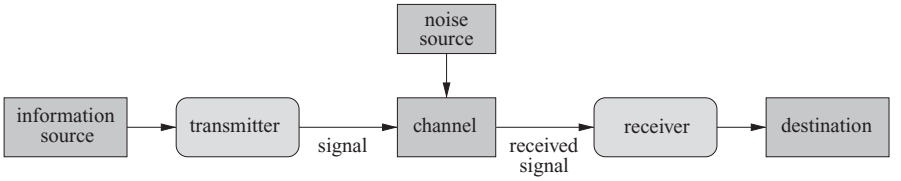


**Figure 2.2** Shannon's communication model (from [7]).

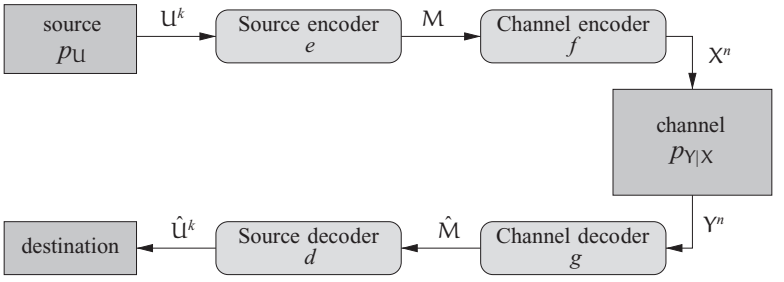


**Figure 2.3** Mathematical model of a two-stage communication system.

- The *information source* generates messages according to some random process.
- The *transmitter* observes the messages and forms a signal to be sent over the channel.
- The *channel* is governed by a *noise source*, which corrupts the original input signal; this models the physical constraints of a communication system, for instance thermal noise in electronic circuits or multipath fading in a wireless medium.
- The *receiver* takes the received signal, forms a reconstructed version of the original message, and delivers the result to the *destination*.

Given the statistical properties of the information source and the noisy channel, the goal of the communication engineer is to design the transmitter and the receiver in a way that allows the information sent by the source to reach its destination in a *reliable* way. Information theory can help us achieve this goal by characterizing the fundamental mechanisms behind communication systems and providing us with precise mathematical conditions under which reliable communication is possible.

### 2.2.1 Point-to-point communication model

To give a precise formulation of the point-to-point communication problem, we require definitions for each of its constituent modules. We assume that the *source* and the *channel* are described by discrete-time random processes, and we determine that the receiver and the transmitter agree on a common *code*, specified by an *encoder* and *decoder* pair. As illustrated in Figure 2.3, we consider a two-stage system in which the source is compressed before being encoded for channel transmission, and channel outputs are decoded before being decompressed. The basic relationships among the components in Figure 2.3 are described in the following lines.

**Definition 2.15.** *A discrete memoryless source (DMS)* $(\mathcal{U}, p_{\mathsf{U}})$ *generates a sequence of i.i.d. symbols (or letters) from the finite alphabet* $\mathcal{U}$ *according to the probability distribution* $p_{\mathsf{U}}$. *The random variable representing a source symbol is denoted by* $\mathsf{U}$.

**Definition 2.16.** *A discrete memoryless channel (DMC)* $\left(\mathcal{X}, p_{\mathsf{Y}|\mathsf{X}}, \mathcal{Y}\right)$ *is described by a finite input alphabet* $\mathcal{X}$, *a finite output alphabet* $\mathcal{Y}$, *and a conditional probability distribution* $p_{\mathsf{Y}|\mathsf{X}}$, *such that* $\mathsf{X}$ *and* $\mathsf{Y}$ *denote the channel input and the channel output, respectively. The set of conditional probabilities (also called transition probabilities) can be represented by a channel transition probability matrix* $\left(p_{\mathsf{Y}|\mathsf{X}}(y|x)\right)_{\mathcal{X},\mathcal{Y}}$.

In what follows, we illustrate many results numerically with the following DMCs.

---

**Example 2.3.** A binary symmetric channel with cross-over probability $p \in [0, 1]$, denoted by BSC($p$), is a DMC $\left(\{0, 1\}, p_{\mathsf{Y}|\mathsf{X}}, \{0, 1\}\right)$ characterized by the transition probability matrix

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

---

**Example 2.4.** A binary erasure channel with erasure probability $\epsilon \in [0, 1]$, denoted by BEC($\epsilon$), is a DMC $\left(\{0, 1\}, p_{\mathsf{Y}|\mathsf{X}}, \{0, ?, 1\}\right)$ characterized by the transition probability matrix

$$\begin{pmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & \epsilon & 1-\epsilon \end{pmatrix}.$$

---

**Definition 2.17.** *A* $(2^{kR}, k)$ source code $\mathcal{C}_k$ *for a DMS* $(\mathcal{U}, p_{\mathsf{U}})$ *consists of*

- *a message set* $\mathcal{M} = [\![1, 2^{kR}]\!]$;
- *an encoding function* $e : \mathcal{U}^k \to \mathcal{M}$, *which maps a sequence of* $k$ *source symbols* $u^k$ *to a message* $m$;
- *a decoding function* $d : \mathcal{M} \to \mathcal{U}^k \cup \{?\}$, *which maps a message* $m$ *to a sequence of source symbols* $\hat{u}^k \in \mathcal{U}^k$ *or an error message* **?**.

The *compression rate* of the source code is defined as $(1/k)\log\lceil 2^{kR}\rceil$ in bits[1] per source symbol, and its probability of error is

$$\mathbf{P}_{\mathrm{e}}(\mathcal{C}_k) \triangleq \mathbb{P}\left[\hat{\mathsf{U}}^k \neq \mathsf{U}^k \mid \mathcal{C}_k\right].$$

**Definition 2.18.** *A rate* $R$ *is an achievable compression rate for the source* $(\mathcal{U}, p_{\mathsf{U}})$ *if there exists a sequence of* $\left(2^{kR}, k\right)$ *source codes* $\{\mathcal{C}_k\}_{k\geqslant 1}$, *such that*

$$\lim_{k\to\infty} \mathbf{P}_{\mathrm{e}}(\mathcal{C}_k) = 0,$$

*that is, the source sequences can be reconstructed with arbitrarily small probability of error with compression rates arbitrarily close to* $R$.

[1] Unless specified otherwise, all logarithms are taken to the base two.

**Definition 2.19.** *A* $\left(2^{nR}, n\right)$ channel code $\mathcal{C}_n$ *for a DMC* $\left(\mathcal{X}, p_{\mathsf{Y}|\mathsf{X}}, \mathcal{Y}\right)$ *consists of*

- *a message set* $\mathcal{M} = [\![1, 2^{nR}]\!]$*;*
- *an encoding function* $f : \mathcal{M} \to \mathcal{X}^n$*, which maps a message* $m$ *to a* codeword $x^n$ *with* $n$ *symbols;*
- *a decoding function* $g : \mathcal{Y}^n \to \mathcal{M} \cup \{\textbf{?}\}$*, which maps a block of* $n$ *channel outputs* $y^n$ *to a message* $\hat{m} \in \mathcal{M}$ *or an error message* **?**.

The set of codewords $\{f(m) : m \in [\![1, 2^{nR}]\!]\}$ is called the *codebook* of $\mathcal{C}_n$. With a slight abuse of notation, we denote the codebook itself by $\mathcal{C}_n$ as well. Unless specified otherwise, messages are represented by a random variable $\mathsf{M}$ uniformly distributed in $\mathcal{M}$, and the rate of the channel code is defined as $(1/n)\log\lceil 2^{nR}\rceil$ in bits per channel use. The average probability of error is defined as

$$\mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) \triangleq \mathbb{P}\left[\hat{\mathsf{M}} \neq \mathsf{M} \mid \mathcal{C}_n\right].$$

**Definition 2.20.** *A rate* $R$ *is an achievable transmission rate for the DMC* $\left(\mathcal{X}, p_{\mathsf{Y}|\mathsf{X}}, \mathcal{Y}\right)$ *if there exists a sequence of* $\left(2^{nR}, n\right)$ *codes* $\{\mathcal{C}_n\}_{n\geqslant 1}$ *such that*

$$\lim_{n\to\infty} \mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) = 0;$$

*that is, messages can be transmitted at a rate arbitrarily close to* $R$ *and decoded with arbitrarily small probability of error. The channel capacity of the DMC is defined as*

$$C \triangleq \sup\{R : R \text{ is an achievable transmission rate}\}.$$

The typical goal of information theory is to characterize achievable rates on the basis of information-theoretic quantities that depend only on the given probability distributions and not on the block lengths $k$ or $n$. A theorem that confirms the existence of codes for a class of achievable rates is often referred to as a *direct result* and the arguments that lead to this result constitute an *achievability proof*. On the other hand, when a theorem asserts that codes with certain properties do not exist, we speak of a *converse result* and a *converse proof*. A fundamental result that includes both the achievability and the converse parts is called a *coding theorem*. The mathematical tools that enable this characterization are those presented in Section 2.1, and we illustrate their use by discussing two of Shannon's fundamental coding theorems. These results form the basis of information theory and are of great use in several of the proofs developed in subsequent chapters.

**Remark 2.2.** *Notice that the formulation of the point-to-point communication problem does not put any constraints either on the computational complexity or on the delay of the encoding and decoding procedures. In other words, the goal is to describe the fundamental limits of communications systems irrespective of their technological limitations.*

### 2.2.2 The source coding theorem

The *source coding theorem* gives a complete solution (achievability and converse) for the point-to-point communication problem stated in Section 2.2.1 when the channel

is noiseless, that is $Y = X$. In that case, it is not necessary to use a channel code to compensate for the impairments caused by the channel, but it is still useful to encode the messages produced by the source to achieve a more efficient representation of the source information in bits per source symbol. This procedure is called *source coding* or *data compression*. The main idea is to consider only a subset $\mathcal{A}$ of all possible source sequences $\mathcal{U}^k$, and assign a different index $i \in [\![1, |\mathcal{A}|]\!]$ to each of the sequences $u^k \in \mathcal{A}$. If the source produces a sequence $u^k \in \mathcal{A}$, then the encoder outputs the corresponding index $i$, otherwise it outputs some predefined constant. The decoder receives the index $i$ and outputs the corresponding sequence in $\mathcal{A}$.

Since information theory is primarily concerned with the fundamental limits of reliable communication, it is possible to prove the existence of codes without having to search for explicit code constructions. One technique, which is particularly useful in information-theoretic problems related to source coding, consists of *throwing* sequences $u^n \in \mathcal{U}^n$ randomly into a finite set of bins, such that the sequences that land in the same bin share a common bin index. If each sequence is assigned a bin at random according to a uniform distribution, then we refer to this procedure as *random binning*. If we want to prove that there exists a code such that the error probability goes to zero, it suffices to show that the average of the probability of error taken over all possible bin assignments goes to zero and to use the selection lemma. The following theorem exploits random binning to characterize the set of achievable compression rates.

**Theorem 2.7** (Source coding theorem)**.** *For a discrete memoryless source* $(\mathcal{U}, p_U)$*,*

$$\inf\{R : R \text{ is an achievable compression rate}\} = \mathbb{H}(U).$$

*In other words, if a compression rate* $R$ *satisfies* $R > \mathbb{H}(U)$ *then* $R$ *is achievable and any achievable compression rate* $R$ *must satisfy* $R \geqslant \mathbb{H}(U)$*.*

*Proof.* We start with the achievability part of the proof, which is based on random binning. The idea is to randomly assign each source sequence to one of a finite number of bins; then, as long as the number of bins is larger than $2^{k\mathbb{H}(U)}$, the probability of finding more than one typical sequence in the same bin is very small. If each typical sequence is mapped to a different bin index, an arbitrarily small probability of error can be achieved by letting the decoder output the typical sequence that corresponds to the received index. Formally, let $\epsilon > 0$ and $k \in \mathbb{N}^*$. Let $R > 0$ be a rate to be specified later. We construct a $(2^{kR}, k)$ source code $\mathcal{C}_k$ as follows.

- *Binning.* For each sequence $u^k \in \mathcal{T}_\epsilon^k(U)$, draw an index uniformly at random in the set $[\![1, 2^{kR}]\!]$. The index assignment defines the encoding function

$$e : \mathcal{U}^k \to [\![1, 2^{kR}]\!],$$

  which is revealed to the encoder and decoder.
- *Encoder.* Given an observation $u^k$, output $m = e(u^k)$ if $u^k \in \mathcal{T}_\epsilon^k(U)$; otherwise output $m = 1$.
- *Decoder.* Given message $m$, output $\hat{u}^k$ if it is the unique sequence such that $\hat{u}^k \in \mathcal{T}_\epsilon^k(U)$ and $e(\hat{u}^k) = m$; otherwise output an error **?**.

The random variable that represents the randomly generated encoding function $e$ is denoted by $\mathsf{E}$ while the random variable that represents the randomly generated code $\mathcal{C}_k$ is denoted by $\mathsf{C}_k$. We proceed to bound $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)]$. First, note that $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)]$ can be expressed in terms of the events

$$\begin{aligned}\mathcal{E}_0 &= \{\mathsf{U}^k \notin \mathcal{T}_\epsilon^k(\mathsf{U})\},\\ \mathcal{E}_1 &= \{\exists \hat{u}^k \neq \mathsf{U}^k : \mathsf{E}(\hat{u}^k) = \mathsf{E}(\mathsf{U}^k) \text{ and } \hat{u}^k \in \mathcal{T}_\epsilon^k(\mathsf{U})\}\end{aligned}$$

as $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)] = \mathbb{P}[\mathcal{E}_0 \cup \mathcal{E}_1]$. By the union bound,

$$\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)] \leqslant \mathbb{P}[\mathcal{E}_0] + \mathbb{P}[\mathcal{E}_1]. \tag{2.1}$$

By the AEP,

$$\mathbb{P}[\mathcal{E}_0] \leqslant \delta_\epsilon(k) \tag{2.2}$$

and we can upper bound $\mathbb{P}[\mathcal{E}_1]$ as

$$\begin{aligned}\mathbb{P}[\mathcal{E}_1] &= \sum_{u^k} p_{\mathsf{U}^k}\left(u^k\right) \mathbb{P}\left[\exists \hat{u}^k \neq u^k : \mathsf{E}(\hat{u}^k) = \mathsf{E}(u^k) \text{ and } \hat{u}^k \in \mathcal{T}_\epsilon^k(\mathsf{U})\right]\\ &\leqslant \sum_{u^k} p_{\mathsf{U}^k}\left(u^k\right) \sum_{\substack{\hat{u}^k \in \mathcal{T}_\epsilon^k(\mathsf{U})\\ \hat{u}^k \neq u^k}} \mathbb{P}\left[\mathsf{E}(\hat{u}^k) = \mathsf{E}(u^k)\right]\\ &= \sum_{u^k} p_{\mathsf{U}^k}\left(u^k\right) \sum_{\substack{\hat{u}^k \in \mathcal{T}_\epsilon^k(\mathsf{U})\\ \hat{u}^k \neq u^k}} \frac{1}{\lceil 2^{kR} \rceil}\\ &\leqslant \sum_{u^k} p_{\mathsf{U}^k}\left(u^k\right) \frac{1}{\lceil 2^{kR} \rceil} \left|\mathcal{T}_\epsilon^k(\mathsf{U})\right|\\ &\leqslant \sum_{u^k} p_{\mathsf{U}^k}\left(u^k\right) \frac{1}{\lceil 2^{kR} \rceil} 2^{k(\mathbb{H}(\mathsf{U})+\delta(\epsilon))}\\ &\leqslant 2^{k(\mathbb{H}(\mathsf{U})+\delta(\epsilon)-R)}.\end{aligned}$$

Hence, if we choose $R > \mathbb{H}(\mathsf{U}) + \delta(\epsilon)$, we have

$$\mathbb{P}[\mathcal{E}_1] \leqslant \delta_\epsilon(k). \tag{2.3}$$

On substituting (2.2) and (2.3) into (2.1), we obtain $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)] \leqslant \delta_\epsilon(k)$. By applying the selection lemma to the random variable $\mathsf{C}_k$ and the function $\mathbf{P}_{\mathrm{e}}$, we conclude that there exists at least one source code $\mathcal{C}_k$ such that $\mathbf{P}_{\mathrm{e}}(\mathcal{C}_k) \leqslant \delta_\epsilon(k)$. Since $\epsilon$ can be chosen arbitrarily small, all rates $R > \mathbb{H}(\mathsf{U})$ are achievable.

We now establish the converse result and show that any achievable rate must satisfy $R \geqslant \mathbb{H}(\mathsf{U})$. Let $R$ be an achievable rate and let $\epsilon > 0$. By definition, there exists a source code $\mathcal{C}_k$ such that $\mathbf{P}_{\mathrm{e}}(\mathcal{C}_k) \leqslant \delta(\epsilon)$. If we let $\mathsf{M}$ denote the message output by the encoder,

then Fano's inequality guarantees that

$$\frac{1}{k}\mathbb{H}\big(\mathsf{U}^k|\mathsf{M}\mathcal{C}_k\big) \leqslant \delta(\mathbf{P}_{\mathrm{e}}(\mathcal{C}_k)) \leqslant \delta(\epsilon).$$

We drop the conditioning on $\mathcal{C}_k$ in subsequent calculations to simplify the notation. Note that

$$\begin{aligned}\mathbb{H}(\mathsf{U}) &= \frac{1}{k}\mathbb{H}\big(\mathsf{U}^k\big)\\ &= \frac{1}{k}\mathbb{I}\big(\mathsf{U}^k;\mathsf{M}\big) + \frac{1}{k}\mathbb{H}\big(\mathsf{U}^k|\mathsf{M}\big)\\ &\leqslant \frac{1}{k}\mathbb{I}\big(\mathsf{U}^k;\mathsf{M}\big) + \delta(\epsilon)\\ &\leqslant \frac{1}{k}\mathbb{H}(\mathsf{M}) + \delta(\epsilon)\\ &\leqslant R + \delta(k) + \delta(\epsilon).\end{aligned}$$

Since $\epsilon$ can be chosen arbitrarily small and $k$ can be chosen arbitrarily large, we obtain $R \geqslant \mathbb{H}(\mathsf{U})$. □

**Remark 2.3.** *Alternatively, the achievability part of the source coding theorem can be established on the basis of the AEP alone. In fact, for large k the AEP guarantees that any sequence $u^k$ produced by the source $(\mathcal{U}, p_{\mathsf{U}})$ belongs with high probability to the typical set $\mathcal{T}_\epsilon^k(\mathsf{U})$; hence, we need only index the approximately $2^{k\mathbb{H}(\mathsf{U})}$ typical sequences to achieve arbitrarily small probability of error and the corresponding rate is on the order of $\mathbb{H}(\mathsf{U})$.*

### 2.2.3 The channel coding theorem

The *channel coding theorem* gives a complete solution (achievability and converse) for the point-to-point communication problem stated in Section 2.2.1 when the source $(\mathcal{U}, p_{\mathsf{U}})$ is uniform over $\mathcal{U}$. According to the source coding theorem, there is no need to encode the source since $\mathbb{H}(\mathsf{U}) = \log|\mathcal{U}|$ is maximal. We simply group the source symbols in sequences of length $k$. Letting $M \triangleq |\mathcal{U}|^k$ and $\mathcal{M} = [\![1, M]\!]$, we index each sequence of length $k$ with an integer $m \in \mathcal{M}$. We use a channel code of rate $(1/n)\log M$ to transmit the messages produced by source $\mathsf{U}$ over a discrete memoryless channel $\big(\mathcal{X}, p_{\mathsf{Y}|\mathsf{X}}, \mathcal{Y}\big)$.

As was done for the source coding theorem, it is possible to prove the existence of codes without having to search for explicit code constructions. The idea is to construct a random code by drawing the symbols of codewords independently at random according to a fixed probability distribution $p_{\mathsf{X}}$ on $\mathcal{X}$. Then, if we want to prove that there exists a code such that the error probability goes to zero for $n$ sufficiently large, it suffices to show that the average of the probability of error taken over all possible random codebooks goes to zero for $n$ sufficiently large and use the selection lemma. This technique is

referred to as *random coding* and is used in the proof of the following theorem to characterize the set of achievable rates.

**Theorem 2.8** (Channel coding theorem)**.** *The capacity of a DMC* $\left(\mathcal{X}, p_{\mathsf{Y}|\mathsf{X}}, \mathcal{Y}\right)$ *is* $C = \max_{p_{\mathsf{X}}} \mathbb{I}(\mathsf{X};\mathsf{Y})$*. In other words, if* $R < C$ *then* $R$ *is an achievable transmission rate and an achievable transmission rate must satisfy* $R \leqslant C$*.*

*Proof.* We begin with the achievability part based on random coding. We choose a probability distribution $p_{\mathsf{X}}$ on $\mathcal{X}$ and, without loss of generality, we assume that $p_{\mathsf{X}}$ is such that $\mathbb{I}(\mathsf{X};\mathsf{Y}) > 0$. Let $0 < \epsilon < \mu_{\mathsf{XY}}$, where $\mu_{\mathsf{XY}} = \min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{\mathsf{XY}}(x, y)$, and let $n \in \mathbb{N}^*$. Let $R > 0$ be a rate to be specified later. We construct a $(2^{nR}, n)$ code $\mathcal{C}_n$ as follows.

- *Codebook construction.* Construct a codebook with $\lceil 2^{nR} \rceil$ codewords, labeled $x^n(m)$ with $m \in [\![1, 2^{nR}]\!]$, by generating the symbols $x_i(m)$ for $i \in [\![1, n]\!]$ and $m \in [\![1, 2^{nR}]\!]$ independently according to $p_{\mathsf{X}}$. The codebook is revealed both to the encoder and to the decoder.
- *Encoder f.* Given $m$, transmit $x^n(m)$.
- *Decoder g.* Given $y^n$, output $\hat{m}$ if it is the unique message such that $(x^n(\hat{m}), y^n) \in \mathcal{T}_\epsilon^n(\mathsf{XY})$; otherwise, output an error **?**.

The random variable that represents the randomly generated codebook $\mathcal{C}_n$ is denoted by $\mathsf{C}_n$. We first develop an upper bound for $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)]$. Notice that

$$\begin{aligned}\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] &= \mathbb{E}_{\mathsf{C}_n}\left[\mathbb{P}\left[\mathsf{M} \neq \hat{\mathsf{M}} \mid \mathsf{C}_n\right]\right] \\ &= \sum_m \mathbb{E}_{\mathsf{C}_n}\left[\mathbb{P}\left[\mathsf{M} \neq \hat{\mathsf{M}} \mid \mathsf{M} = m, \mathsf{C}_n\right]\right] p_{\mathsf{M}}(m).\end{aligned}$$

By virtue of the symmetry of the random code construction, we have that $\mathbb{E}_{\mathsf{C}_n}\left[\mathbb{P}\left[\mathsf{M} \neq \hat{\mathsf{M}} \mid \mathsf{M} = m, \mathsf{C}_n\right]\right]$ is independent of $m$. Therefore, we can assume without losing generality that message $m = 1$ was sent and write

$$\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] = \mathbb{E}_{\mathsf{C}_n}\left[\mathbb{P}\left[\mathsf{M} \neq \hat{\mathsf{M}} \mid \mathsf{M} = 1, \mathsf{C}_n\right]\right].$$

Notice that $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)]$ can be expressed in terms of the events

$$\mathcal{E}_i = \{(\mathsf{X}^n(i), \mathsf{Y}^n) \in \mathcal{T}_\epsilon^n(\mathsf{XY})\} \quad \text{for } i \in [\![1, 2^{nR}]\!]$$

as $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] = \mathbb{P}\left[\mathcal{E}_1^{\mathrm{c}} \cup \bigcup_{i\neq 1} \mathcal{E}_i\right]$. By the union bound,

$$\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] \leqslant \mathbb{P}\left[\mathcal{E}_1^{\mathrm{c}}\right] + \sum_{i\neq 1} \mathbb{P}[\mathcal{E}_i]. \tag{2.4}$$

By the AEP,

$$\mathbb{P}\left[\mathcal{E}_1^{\mathrm{c}}\right] \leqslant \delta_\epsilon(n). \tag{2.5}$$

Since $\mathsf{Y}^n$ is the output of the channel when $\mathsf{X}^n(1)$ is transmitted and since $\mathsf{X}^n(1)$ is independent of $\mathsf{X}^n(i)$ for $i \neq 1$, note that $\mathsf{Y}^n$ is independent of $\mathsf{X}^n(i)$ for $i \neq 1$; hence,

Corollary 2.2 applies and

$$\mathbb{P}[\mathcal{E}_i] \leqslant 2^{-n(\mathbb{I}(\mathsf{X};\mathsf{Y})-\delta(\epsilon))} \quad \text{for } i \neq 1. \tag{2.6}$$

On substituting (2.5) and (2.6) into (2.4), we obtain

$$\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] \leqslant \delta_\epsilon(n) + \lceil 2^{nR} \rceil 2^{-n(\mathbb{I}(\mathsf{X};\mathsf{Y})-\delta(\epsilon))}.$$

Hence, if we choose the rate $R$ such that $R < \mathbb{I}(\mathsf{X};\mathsf{Y}) - \delta(\epsilon)$, then

$$\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] \leqslant \delta_\epsilon(n).$$

By applying the selection lemma to the random variable $\mathsf{C}_n$ and the function $\mathbf{P}_{\mathrm{e}}$, we conclude that there exists a $(2^{nR}, n)$ code $\mathcal{C}_n$ such that $\mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) \leqslant \delta_\epsilon(n)$. Since $\epsilon$ can be chosen arbitrarily small and since the distribution $p_{\mathsf{X}}$ is arbitrary, we conclude that all rates $R < \max_{p_{\mathsf{X}}} \mathbb{I}(\mathsf{X};\mathsf{Y})$ are achievable.

We now establish the converse part of the proof. Let $R$ be an achievable rate and let $\epsilon > 0$. For $n$ sufficiently large, there exists a $(2^{nR}, n)$ code $\mathcal{C}_n$ such that

$$\frac{1}{n}\mathbb{H}(\mathsf{M}|\mathcal{C}_n) \geqslant R \quad \text{and} \quad \mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) \leqslant \delta(\epsilon).$$

In the remaining part of the proof we drop the conditioning on $\mathcal{C}_n$ to simplify the notation. By virtue of Fano's inequality, it also holds that

$$\frac{1}{n}\mathbb{H}(\mathsf{M}|\mathsf{Y}^n) \leqslant \delta(\mathbf{P}_{\mathrm{e}}(\mathcal{C}_n)) = \delta(\epsilon).$$

Therefore,

$$\begin{aligned}
R &\leqslant \frac{1}{n}\mathbb{H}(\mathsf{M}) \\
&\leqslant \frac{1}{n}\mathbb{I}(\mathsf{M};\mathsf{Y}^n) + \frac{1}{n}\mathbb{H}(\mathsf{M}|\mathsf{Y}^n) \\
&\leqslant \frac{1}{n}\mathbb{I}(\mathsf{M};\mathsf{Y}^n) + \delta(\epsilon) \\
&\overset{(a)}{\leqslant} \frac{1}{n}\mathbb{I}(\mathsf{X}^n;\mathsf{Y}^n) + \delta(\epsilon) \\
&= \frac{1}{n}\mathbb{H}(\mathsf{Y}^n) - \frac{1}{n}\mathbb{H}(\mathsf{Y}^n|\mathsf{X}^n) + \delta(\epsilon) \\
&\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{H}\big(\mathsf{Y}_i|\mathsf{Y}^{i-1}\big) - \frac{1}{n}\mathbb{H}(\mathsf{Y}_i|\mathsf{X}_i)\right) + \delta(\epsilon) \\
&\overset{(c)}{\leqslant} \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{H}(\mathsf{Y}_i) - \frac{1}{n}\mathbb{H}(\mathsf{Y}_i|\mathsf{X}_i)\right) + \delta(\epsilon) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(\mathsf{X}_i;\mathsf{Y}_i) + \delta(\epsilon) \\
&\leqslant \max_{p_{\mathsf{X}}} \mathbb{I}(\mathsf{X};\mathsf{Y}) + \delta(\epsilon),
\end{aligned}$$

where (a) follows from the data-processing inequality applied to the Markov chain $M \to X^n \to Y^n$, (b) follows because the channel is memoryless, and (c) follows because conditioning does not increase entropy. Since $\epsilon$ can be chosen arbitrarily small, we obtain $R \leqslant \max_{p_X} \mathbb{I}(X;Y)$. □

The channel coding theorem shows that the channel capacity is equal to the maximum mutual information between the channel input X and the channel output Y, where the maximization is carried out over all possible input probability distributions $p_X$. The proof technique and structure are common to most proofs in subsequent chapters.

---

**Example 2.5.** The capacity of a binary symmetric channel BSC($p$) is $1 - \mathbb{H}_b(p)$. The capacity of a binary erasure channel BEC($\epsilon$) is $1 - \epsilon$.

---

Among the many channel models, the additive white Gaussian noise (AWGN) channel (Gaussian channel for short) takes a particularly prominent role in information and communication theory, because it captures the impact of thermal noise and interference on wired and wireless communications. The channel output at each time $i \geqslant 1$ is given by $Y_i = X_i + N_i$, where $X_i$ denotes the transmitted symbol and $\{N_i\}_{i \geqslant 1}$ are i.i.d. random variables with distribution $\mathcal{N}(0, \sigma^2)$. Since the channel capacity of the Gaussian channel can be infinite without further restrictions, we add an average power constraint in the form of

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] \leqslant P.$$

**Theorem 2.9.** *The capacity of a Gaussian channel is given by*

$$C = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2}\right),$$

*where $P$ denotes the power constraint and $\sigma^2$ is the variance of the noise.*

*Sketch of proof.* The proof developed for Theorem 2.8 does not apply directly to the Gaussian channel because of the power constraint imposed on channel inputs and the continuous nature of the channel. Nevertheless, it is possible to develop a similar proof by using weakly typical sequences and the weak AEP (see for instance [3, Chapter 9]). The power constraint can be dealt with by introducing an error event that accounts for the sequences violating the power constraint in the codebook generation. □

## 2.3 Network information theory

Shannon's coding theorems characterize the fundamental limits of communication between two users. However, in many communication scenarios – for example, satellite broadcasting, cellular telephony, the Internet, and wireless sensor networks – the information is sent by one or more transmitting nodes to one or more receiving nodes

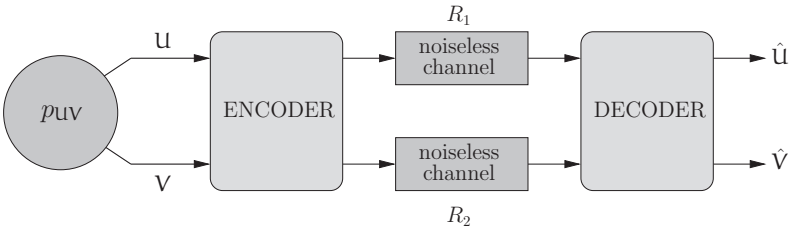


**Figure 2.4** Joint encoding of correlated sources.

over more or less intricate communication networks. The interactions between the users of said networks introduce a whole new range of fundamental communication aspects that are not present in the classical point-to-point problem, such as *interference*, *user cooperation*, and *feedback*. The central goal of *network information theory* is to provide a thorough understanding of these basic mechanisms, by characterizing the fundamental limits of communication systems with multiple users. In this section, we discuss some results of network information theory that are useful for understanding information-theoretic security in subsequent chapters.

### 2.3.1 Distributed source coding

Consider a DMS $(\mathcal{UV}, p_{UV})$ that consists of two components U and V with joint distribution $p_{UV}$. As shown in Figure 2.4, the two components are to be processed by a joint encoder and transmitted to a common destination over two noiseless channels. The joint distribution $p_{UV}$ can be arbitrary and the symbols produced by U and V at any given point in time are statistically dependent; therefore, we refer to U and V as *correlated sources*. Since the channels to the destination do not introduce any errors, we may ask the following question: at what rates $R_1$ and $R_2$ can we transmit information generated by U and V with an arbitrarily small probability of error? Since there is a common encoder and a common decoder, this problem reduces to the classical point-to-point problem and the solution follows naturally from Shannon's source coding theorem: the messages can be reconstructed with an arbitrarily small probability of error at the receiver if and only if

$$R_1 + R_2 > \mathbb{H}(UV);$$

that is, the sum rate must be greater than the joint entropy of U and V.

As illustrated in Figure 2.5, the problem becomes more challenging if instead of a joint encoder we consider two *separate* encoders. Here, each encoder observes only the realizations of the one source it is assigned to and does not know the output symbols of the other source. In this case, it is not immediately clear which encoding rates guarantee reconstruction with an arbitrarily small probability of error at the receiver. If we encode U at rate $R_1 > \mathbb{H}(U)$ and V at rate $R_2 > \mathbb{H}(V)$, then the source coding theorem guarantees once again that an arbitrarily small probability of error is possible. But, in this case, the sum rate satisfies $R_1 + R_2 > \mathbb{H}(U) + \mathbb{H}(V)$, which, in general, is greater than the joint entropy $\mathbb{H}(UV)$.

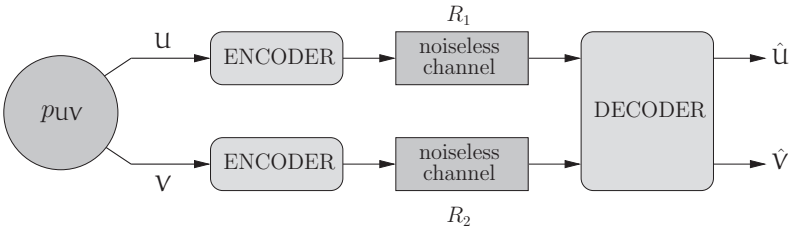


**Figure 2.5** Separate encoding of correlated sources (the *Slepian–Wolf problem*).

Surprisingly, it turns out that the sum rate required by two separate encoders is the same as that required by a joint encoder, that is $R_1 + R_2 > \mathbb{H}(\mathsf{UV})$ is sufficient to reconstruct $\mathsf{U}$ and $\mathsf{V}$ with an arbitrarily small probability of error. In other words, there is no penalty in overall compression rate due to the fact that the encoders can observe only the realizations of the one source they have been assigned to. However, it is important to point out that the decoder does require a minimum amount of rate from each encoder; specifically, the average remaining uncertainty about the messages of one source given the messages of the other source, $\mathbb{H}(\mathsf{U}|\mathsf{V})$ and $\mathbb{H}(\mathsf{V}|\mathsf{U})$. Formally, a code for the distributed source coding problem is defined as follows.

**Definition 2.21.** *A* $(2^{kR_1}, 2^{kR_2}, k)$ *source code* $\mathcal{C}_k$ *for the DMS* $(\mathcal{UV}, p_{\mathsf{UV}})$ *consists of*

- *two message sets* $\mathcal{M}_1 = [\![1, 2^{kR_1}]\!]$ *and* $\mathcal{M}_2 = [\![1, 2^{kR_2}]\!]$*;*
- *an encoding function* $e_1 : \mathcal{U}^k \to \mathcal{M}_1$*, which maps a sequence of* $k$ *source symbols* $u^k$ *to a message* $m_1$*;*
- *an encoding function* $e_2 : \mathcal{V}^k \to \mathcal{M}_2$*, which maps a sequence of* $k$ *source symbols* $v^k$ *to a message* $m_2$*;*
- *a decoding function* $d : \mathcal{M}_1 \times \mathcal{M}_2 \to (\mathcal{U}^k \times \mathcal{V}^k) \cup \{\textbf{?}\}$*, which maps a message pair* $(m_1, m_2)$ *to a pair of source sequences* $(\hat{u}^k, \hat{v}^k) \in \mathcal{U}^k \times \mathcal{V}^k$ *or an error message* **?**.

The performance of a code $\mathcal{C}_k$ is measured in terms of the average probability of error

$$\mathbf{P}_{\mathrm{e}}(\mathcal{C}_k) \triangleq \mathbb{P}\left[(\hat{\mathsf{U}}^k, \hat{\mathsf{V}}^k) \neq (\mathsf{U}^k, \mathsf{V}^k) \mid \mathcal{C}_k\right].$$

**Definition 2.22.** *A rate pair* $(R_1, R_2)$ *is achievable if there exists a sequence of* $(2^{kR_1}, 2^{kR_2}, k)$ *codes* $\{\mathcal{C}_k\}_{k \geqslant 1}$ *such that*

$$\lim_{k \to \infty} \mathbf{P}_{\mathrm{e}}(\mathcal{C}_k) = 0.$$

*The achievable rate region is defined as*

$$\mathcal{R}^{\mathrm{SW}} \triangleq \mathrm{cl}(\{(R_1, R_2) : (R_1, R_2) \text{ is achievable}\}).$$

The achievable rate region with separate encoding was first characterized by Slepian and Wolf; hence, the region is often called the Slepian–Wolf region and codes for the distributed source coding problem are often referred to as Slepian–Wolf codes.

**Theorem 2.10** (Slepian–Wolf theorem)**.** *The achievable rate region with separate encoding for a source* $(\mathcal{U}\mathcal{V}, p_{\mathsf{U}\mathsf{V}})$ *is*

$$\mathcal{R}^{\mathrm{SW}} \triangleq \left\{ (R_1, R_2): \begin{array}{l} R_1 \geqslant \mathbb{H}(\mathsf{U}|\mathsf{V}) \\ R_2 \geqslant \mathbb{H}(\mathsf{V}|\mathsf{U}) \\ R_1 + R_2 \geqslant \mathbb{H}(\mathsf{U}\mathsf{V}) \end{array} \right\}.$$

*Proof.* We begin with the achievability part of the proof, which is based on joint typicality and random binning. Let $\epsilon > 0$ and $k \in \mathbb{N}^*$. Let $R_1 > 0$ and $R_2 > 0$ be rates to be specified later. We construct a $(2^{kR_1}, 2^{kR_2}, k)$ code $\mathcal{C}_k$ as follows.

- *Binning.* For each sequence $u^k \in \mathcal{T}_\epsilon^k(\mathsf{U})$, draw an index uniformly at random in the set $[\![1, 2^{kR_1}]\!]$. For each sequence $v^k \in \mathcal{T}_\epsilon^k(\mathsf{V})$, draw an index uniformly at random in the set $[\![1, 2^{kR_2}]\!]$. The index assignments define the encoding functions

$$e_1 : \mathcal{U}^k \to [\![1, 2^{kR_1}]\!] \quad \text{and} \quad e_2 : \mathcal{V}^k \to [\![1, 2^{kR_2}]\!],$$

  which are revealed to all parties.
- *Encoder 1.* Given the observation $u^k$, if $u^k \in \mathcal{T}_\epsilon^k(\mathsf{U})$, output $m_1 = e_1(u^k)$; otherwise output $m_1 = 1$.
- *Encoder 2.* Given the observation $v^k$, if $v^k \in \mathcal{T}_\epsilon^k(\mathsf{V})$, output $m_2 = e_2(v^k)$; otherwise output $m_2 = 1$.
- *Decoder.* Given messages $m_1$ and $m_2$, output $\hat{u}^k$ and $\hat{v}^k$ if they are the unique sequences such that $(\hat{u}^k, \hat{v}^k) \in \mathcal{T}_\epsilon^k(\mathsf{U}\mathsf{V})$ and $e_1(\hat{u}^k) = m_1$, $e_2(\hat{v}^k) = m_2$; otherwise, output **?**.

The random variables that represent the randomly generated functions $e_1$ and $e_2$ are denoted by $\mathsf{E}_1$ and $\mathsf{E}_2$, and the random variable that represents the randomly generated code $\mathcal{C}_k$ is denoted by $\mathsf{C}_k$. We proceed to bound $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)]$, which can be expressed in terms of the following events:

$$\begin{aligned}
\mathcal{E}_0 &= \{(\mathsf{U}^k, \mathsf{V}^k) \notin \mathcal{T}_\epsilon^k(\mathsf{U}\mathsf{V})\}, \\
\mathcal{E}_1 &= \{\exists \hat{u}^k \neq \mathsf{U}^k : \mathsf{E}_1(\hat{u}^k) = \mathsf{E}_1(\mathsf{U}^k) \text{ and } (\hat{u}^k, \mathsf{V}^k) \in \mathcal{T}_\epsilon^k(\mathsf{U}\mathsf{V})\}, \\
\mathcal{E}_2 &= \{\exists \hat{v}^k \neq \mathsf{V}^k : \mathsf{E}_2(\hat{v}^k) = \mathsf{E}_2(\mathsf{V}^k) \text{ and } (\hat{u}^k, \mathsf{V}^k) \in \mathcal{T}_\epsilon^k(\mathsf{U}\mathsf{V})\}, \\
\mathcal{E}_{12} &= \{\exists \hat{v}^k \neq \mathsf{V}^k, \hat{u}^k \neq \mathsf{U}^k, : \mathsf{E}_1(\hat{u}^k) = \mathsf{E}_1(\mathsf{U}^k),\ \mathsf{E}_2(\hat{v}^k) = \mathsf{E}_2(\mathsf{V}^k) \\
&\qquad \text{and } (\hat{u}^k, \hat{v}^k) \in \mathcal{T}_\epsilon^k(\mathsf{U}\mathsf{V})\},
\end{aligned}$$

since $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)] = \mathbb{P}[\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_{12}]$. By the union bound,

$$\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)] \leqslant \mathbb{P}[\mathcal{E}_0] + \mathbb{P}[\mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2] + \mathbb{P}[\mathcal{E}_{12}]. \tag{2.7}$$

By the AEP,

$$\mathbb{P}[\mathcal{E}_0] \leqslant \delta_\epsilon(k). \tag{2.8}$$

Using Theorem 2.2,

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_1] &= \sum_{u^k, v^k} p_{\mathsf{U}^k\mathsf{V}^k}(u^k, v^k)\, \mathbb{P}\left[\exists \hat{u}^k \neq u^k : \mathsf{E}_1(\hat{u}^k) = \mathsf{E}_1(u^k) \text{ and } (\hat{u}^k, v^k) \in \mathcal{T}_\epsilon^k(\mathsf{U}\mathsf{V})\right] \\
&\leqslant \sum_{u^k, v^k} p_{\mathsf{U}^k\mathsf{V}^k}(u^k, v^k) \sum_{\substack{\hat{u}^k \in \mathcal{T}_\epsilon^k(\mathsf{U}|v^k) \\ \hat{u}^k \neq u^k}} \mathbb{P}\left[\mathsf{E}_1(\hat{u}^k) = \mathsf{E}_1(u^k)\right] \\
&= \sum_{u^k, v^k} p_{\mathsf{U}^k\mathsf{V}^k}(u^k, v^k) \sum_{\substack{\hat{u}^k \in \mathcal{T}_\epsilon^k(\mathsf{U}|v^k) \\ \hat{u}^k \neq u^k}} \frac{1}{\lceil 2^{kR_1} \rceil} \\
&\leqslant \sum_{u^k, v^k} p_{\mathsf{U}^k\mathsf{V}^k}(u^k, v^k) \frac{1}{\lceil 2^{kR_1} \rceil} \left|\mathcal{T}_\epsilon^k(\mathsf{U}|v^k)\right| \\
&\leqslant \sum_{u^k, v^k} p_{\mathsf{U}^k\mathsf{V}^k}(u^k, v^k) \frac{1}{\lceil 2^{kR_1} \rceil} 2^{k(\mathbb{H}(\mathsf{U}|\mathsf{V})+\delta(\epsilon))} \\
&\leqslant 2^{k(\mathbb{H}(\mathsf{U}|\mathsf{V})+\delta(\epsilon)-R_1)}. \qquad (2.9)
\end{aligned}$$

Similarly, we obtain the following bounds for $\mathbb{P}[\mathcal{E}_2]$ and $\mathbb{P}[\mathcal{E}_{12}]$:

$$\mathbb{P}[\mathcal{E}_2] \leqslant 2^{k(\mathbb{H}(\mathsf{V}|\mathsf{U})+\delta(\epsilon)-R_2)}, \qquad (2.10)$$

$$\mathbb{P}[\mathcal{E}_{12}] \leqslant 2^{k(\mathbb{H}(\mathsf{U}\mathsf{V})+\delta(\epsilon)-(R_1+R_2))}. \qquad (2.11)$$

Hence, if we choose the rates $R_1$ and $R_2$ such that

$$\begin{aligned}
R_1 &> \mathbb{H}(\mathsf{U}|\mathsf{V}) + \delta(\epsilon), \\
R_2 &> \mathbb{H}(\mathsf{V}|\mathsf{U}) + \delta(\epsilon), \\
R_1 + R_2 &> \mathbb{H}(\mathsf{U}\mathsf{V}) + \delta(\epsilon),
\end{aligned}$$

and substitute (2.8)–(2.11) into (2.7), we obtain $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_k)] \leqslant \delta_\epsilon(k)$. By applying the selection lemma to the random variable $\mathsf{C}_k$ and the function $\mathbf{P}_{\mathrm{e}}$, we conclude that there exists a specific code $\mathcal{C}_k$ such that $\mathbf{P}_{\mathrm{e}}(\mathcal{C}_k) \leqslant \delta_\epsilon(k)$. Since $\epsilon$ can be chosen arbitrarily small, we conclude that

$$\left\{ (R_1, R_2): \begin{array}{l} R_1 \geqslant \mathbb{H}(\mathsf{U}|\mathsf{V}) \\ R_2 \geqslant \mathbb{H}(\mathsf{V}|\mathsf{U}) \\ R_1 + R_2 \geqslant \mathbb{H}(\mathsf{U}\mathsf{V}) \end{array} \right\} \subseteq \mathcal{R}^{\mathrm{SW}}.$$

The converse part of the proof follows from the converse of the source coding theorem and is omitted. □

Figure 2.6 illustrates the typical shape of the Slepian–Wolf region $\mathcal{R}^{\mathrm{SW}}$. A special case of the Slepian–Wolf problem is when one of the components of the DMS $(\mathcal{U}\mathcal{V}, p_{\mathsf{U}\mathsf{V}})$, say $\mathsf{V}$, is directly available at the decoder as side information and only $\mathsf{U}$ should be compressed. This problem is known as source coding with side information. The characterization of the minimum compression rate required to reconstruct $\mathsf{U}$ reliably at the decoder follows from Theorem 2.10.

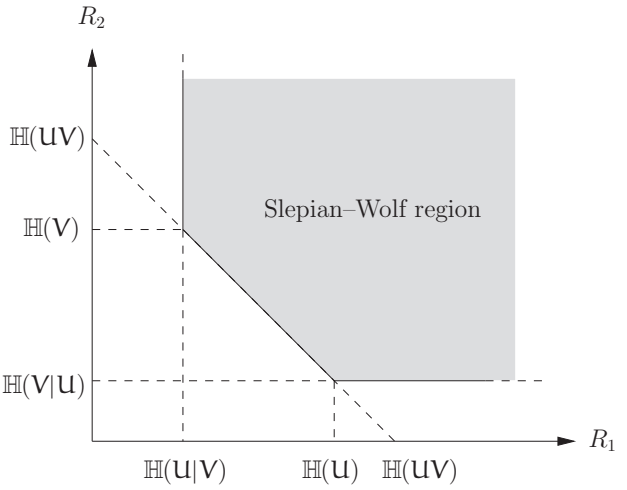


**Figure 2.6** The Slepian–Wolf region for a DMS $(\mathcal{U}\mathcal{V}, p_{\mathsf{U}\mathsf{V}})$.

**Corollary 2.4** (Source coding with side information)**.** *Consider a DMS* $(\mathcal{U}\mathcal{V}, p_{\mathsf{U}\mathsf{V}})$ *and assume that* $(\mathcal{U}, p_{\mathsf{U}})$ *should be compressed knowing that* $(\mathcal{V}, p_{\mathsf{V}})$ *is available as side information at the decoder. Then,*

$$\inf\{R : R \text{ is an achievable compression rate}\} = \mathbb{H}(\mathsf{U}|\mathsf{V}).$$

Corollary 2.4 plays a fundamental role for secret-key agreement in Chapters 4 and 6.

### 2.3.2 The multiple-access channel

In the previous problem, we assumed that the information generated by multiple sources is transmitted over noiseless channels. If these data are to be communicated over a common noisy channel to a single destination, we call this type of channel a *multiple-access channel* (MAC). As illustrated in Figure 2.7, a discrete memoryless multiple access channel $\left(\mathcal{X}_1, \mathcal{X}_2, p_{\mathsf{Y}|\mathsf{X}_1\mathsf{X}_2}, \mathcal{Y}\right)$ consists of two finite input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$, one finite output alphabet $\mathcal{Y}$, and transition probabilities $p_{\mathsf{Y}|\mathsf{X}_1\mathsf{X}_2}$ such that

$$\forall n \geqslant 1 \quad \forall (x_1^n, x_2^n, y^n) \in \mathcal{X}_1^n \times \mathcal{X}_2^n \times \mathcal{Y}^n$$
$$p_{\mathsf{Y}^n|\mathsf{X}_1^n\mathsf{X}_2^n}\left(y^n|x_1^n x_2^n\right) = \prod_{i=1}^{n} p_{\mathsf{Y}|\mathsf{X}_1\mathsf{X}_2}(y_i|x_{1,i}, x_{2,i}).$$

**Definition 2.23.** *A* $(2^{nR_1}, 2^{nR_2}, n)$ *code* $\mathcal{C}_n$ *for the MAC consists of*

- *two message sets* $\mathcal{M}_1 = [\![1, 2^{nR_1}]\!]$ *and* $\mathcal{M}_2 = [\![1, 2^{nR_2}]\!]$*;*
- *two encoding functions,* $f_1 : \mathcal{M}_1 \to \mathcal{X}_1^n$ *and* $f_2 : \mathcal{M}_2 \to \mathcal{X}_2^n$*, which map a message* $m_1$ *or* $m_2$ *to a codeword* $x_1^n$ *or* $x_2^n$*;*

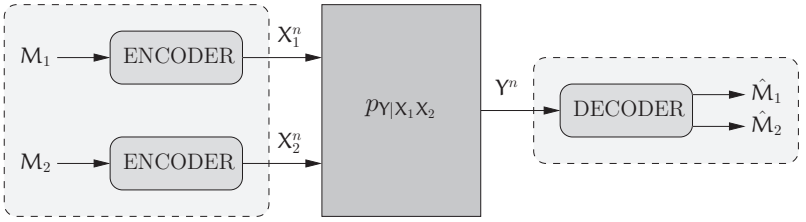


**Figure 2.7** Communication over a two-user multiple-access channel.

- *a decoding function* $g : \mathcal{Y}^n \to \mathcal{M}_1 \times \mathcal{M}_2 \cup \{?\}$*, which maps each channel observation* $y^n$ *to a message pair* $(\hat{m}_1, \hat{m}_2) \in \mathcal{M}_1 \times \mathcal{M}_2$ *or an error message* **?**.

The messages $\mathsf{M}_1$ and $\mathsf{M}_2$ are assumed uniformly distributed in their respective sets, and the performance of a code $\mathcal{C}_n$ is measured in terms of the average probability of error

$$\mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) \triangleq \mathbb{P}\left[(\hat{\mathsf{M}}_1, \hat{\mathsf{M}}_2) \neq (\mathsf{M}_1, \mathsf{M}_2)) \middle| \mathcal{C}_n\right].$$

**Definition 2.24.** *A rate pair* $(R_1, R_2)$ *is achievable for the MAC if there exists a sequence of* $(2^{nR_1}, 2^{nR_2}, n)$ *codes* $\{\mathcal{C}_n\}_{n \geqslant 1}$ *such that*

$$\lim_{n\to\infty} \mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) = 0.$$

*The capacity region of a MAC is defined as*

$$\mathcal{C}^{\mathsf{MAC}} \triangleq \mathrm{cl}(\{(R_1, R_2) : (R_1, R_2) \text{ is achievable}\}).$$

The characterization of the capacity region requires the notion of a *convex hull*, which we define below.

**Definition 2.25.** *The convex hull of a set* $\mathcal{S} \subseteq \mathbb{R}^n$ *is the set*

$$\mathrm{co}(\mathcal{S}) \triangleq \left\{ \sum_{i=1}^{k} \lambda_i x_i^n : k \geqslant 1,\ \{\lambda_i\}_k \in [0,1]^k,\ \sum_{i=1}^{k} \lambda_i = 1,\ \{x_i^n\}_k \in \mathcal{S}^k \right\}.$$

**Theorem 2.11** (Ahlswede and Liao)**.** *Consider a MAC* $\left(\mathcal{X}_1, \mathcal{X}_2, p_{\mathsf{Y}|\mathsf{X}_1\mathsf{X}_2}, \mathcal{Y}\right)$*. For any independent distributions* $p_{\mathsf{X}_1}$ *on* $\mathcal{X}_1$ *and* $p_{\mathsf{X}_2}$ *on* $\mathcal{X}_2$*, define the set* $\mathcal{R}(p_{\mathsf{X}_1} p_{\mathsf{X}_2})$ *as*

$$\mathcal{R}(p_{\mathsf{X}_1} p_{\mathsf{X}_2}) \triangleq \left\{ (R_1, R_2): \begin{array}{c} 0 \leqslant R_1 \leqslant \mathbb{I}(\mathsf{X}_1; \mathsf{Y}|\mathsf{X}_2) \\ 0 \leqslant R_2 \leqslant \mathbb{I}(\mathsf{X}_2; \mathsf{Y}|\mathsf{X}_1) \\ 0 \leqslant R_1 + R_2 \leqslant \mathbb{I}(\mathsf{X}_1\mathsf{X}_2; \mathsf{Y}) \end{array} \right\},$$

*where the joint distribution of* $\mathsf{X}_1$*,* $\mathsf{X}_2$*, and* $\mathsf{Y}$ *factorizes as* $p_{\mathsf{X}_1} p_{\mathsf{X}_2} p_{\mathsf{Y}|\mathsf{X}_1\mathsf{X}_2}$*. Then, the capacity region of a MAC is*

$$\mathcal{C}^{\mathsf{MAC}} \triangleq \mathrm{co}\left( \bigcup_{p_{\mathsf{X}_1} p_{\mathsf{X}_2}} \mathcal{R}(p_{\mathsf{X}_1} p_{\mathsf{X}_2}) \right).$$

*Proof.* We provide only the achievability part of the proof, which is similar to the proof of Shannon's channel coding theorem and is based on joint typicality and random

coding. Fix two independent probability distributions, $p_{X_1}$ on $\mathcal{X}_1$ and $p_{X_2}$ on $\mathcal{X}_2$. Let $0 < \epsilon < \mu_{X_1X_2Y}$, where

$$\mu_{X_1X_2Y} \triangleq \min_{(x_1,x_2,y)\in\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{Y}} p_{X_1}(x_1)p_{X_2}(x_2)p_{Y|X_1X_2}(y|x_1,x_2),$$

and let $n \in \mathbb{N}^*$. Let $R_1 > 0$ and $R_2 > 0$ be rates to be specified later. We construct a $(2^{nR_1}, 2^{nR_2}, n)$ code $\mathcal{C}_n$ as follows.

- *Codebook construction.* Construct a codebook for user 1 with $\lceil 2^{nR_1} \rceil$ codewords, labeled $x_1^n(m_1)$ with $m_1 \in [\![1, 2^{nR_1}]\!]$, by generating the symbols $x_{1,i}(m_1)$ for $i \in [\![1, n]\!]$ and $m_1 \in [\![1, 2^{nR_1}]\!]$ independently according to $p_{X_1}$. Similarly, construct a codebook for user 2 with $\lceil 2^{nR_2} \rceil$ codewords, labeled $x_2^n(m_2)$ with $m_2 \in [\![1, 2^{nR_2}]\!]$, by generating the symbols $x_{2,i}(m_2)$ for $i \in [\![1, n]\!]$ and $m_2 \in [\![1, 2^{nR_2}]\!]$ independently according to $p_{X_2}$. The codebooks are revealed to all encoders and decoders.
- *Encoder 1.* Given $m_1$, transmit $x_1^n(m_1)$.
- *Encoder 2.* Given $m_2$, transmit $x_2^n(m_2)$.
- *Decoder.* Given $y^n$, output $(\hat{m}_1, \hat{m}_2)$ if it is the unique message pair such that $(x_1^n(\hat{m}_1), x_2^n(\hat{m}_2), y^n) \in \mathcal{T}_\epsilon^n(X_1X_2Y)$; otherwise, output an error **?**.

The random variable that represents the randomly generated code $\mathcal{C}_n$ is denoted by $C_n$ and we proceed to bound $\mathbb{E}[\mathbf{P}_e(C_n)]$. By virtue of the symmetry of the random code construction

$$\begin{aligned}\mathbb{E}[\mathbf{P}_e(C_n)] &= \mathbb{E}_{C_n}\left[\mathbb{P}\left[(\hat{M}_1, \hat{M}_2) \neq (M_1, M_2)\middle| C_n\right]\right] \\ &= \mathbb{E}_{C_n}\left[\mathbb{P}\left[(\hat{M}_1, \hat{M}_2) \neq (M_1, M_2)\middle| M_1 = 1, M_2 = 1, C_n\right]\right].\end{aligned}$$

Therefore, the probability of error can be expressed in terms of the error events

$$\mathcal{E}_{ij} \triangleq \{(X_1^n(i), X_2^n(j), Y^n) \in \mathcal{T}_\epsilon^n(X_1X_2Y)\} \quad \text{for } i \in [\![1, 2^{nR_1}]\!] \text{ and } j \in [\![1, 2^{nR_2}]\!]$$

as

$$\mathbb{E}[\mathbf{P}_e(C_n)] = \mathbb{P}\left[\mathcal{E}_{11}^c \cup \bigcup_{i\neq 1}\mathcal{E}_{i1} \cup \bigcup_{j\neq 1}\mathcal{E}_{1j} \cup \bigcup_{(i,j)\neq(1,1)}\mathcal{E}_{ij}\right].$$

By the union bound,

$$\mathbb{E}[\mathbf{P}_e(C_n)] \leqslant \mathbb{P}\left[\mathcal{E}_{11}^c\right] + \sum_{i\neq 1}\mathbb{P}[\mathcal{E}_{i1}] + \sum_{j\neq 1}\mathbb{P}\left[\mathcal{E}_{1j}\right] + \sum_{(i,j)\neq(1,1)}\mathbb{P}\left[\mathcal{E}_{ij}\right]. \tag{2.12}$$

By the joint AEP,

$$\mathbb{P}\left[\mathcal{E}_{11}^c\right] \leqslant \delta_\epsilon(n). \tag{2.13}$$

For $i \neq 1$, $X_1^n(i)$ is conditionally independent of $Y^n$ given $X_2^n(1)$; therefore, by Corollary 2.3,

$$\mathbb{P}[\mathcal{E}_{i1}] \leqslant 2^{-n(\mathbb{I}(X_1;Y|X_2)-\delta(\epsilon))} \quad \text{for } i \neq 1. \tag{2.14}$$

Similarly, we can show

$$\mathbb{P}\left[\mathcal{E}_{1j}\right] \leqslant 2^{-n(\mathbb{I}(\mathsf{X}_2;\mathsf{Y}|\mathsf{X}_1)-\delta(\epsilon))} \quad \text{for } j \neq 1, \tag{2.15}$$

$$\mathbb{P}\left[\mathcal{E}_{ij}\right] \leqslant 2^{-n(\mathbb{I}(\mathsf{X}_1\mathsf{X}_2;\mathsf{Y})-\delta(\epsilon))} \quad \text{for } i \neq 1 \text{ and } j \neq 1. \tag{2.16}$$

On substituting (2.13)–(2.16) into (2.12), we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] \leqslant \delta_\epsilon(n) &+ \lceil 2^{nR_1}\rceil 2^{-n(\mathbb{I}(\mathsf{X}_1;\mathsf{Y}|\mathsf{X}_2)-\delta(\epsilon))} + \lceil 2^{nR_2}\rceil 2^{-n(\mathbb{I}(\mathsf{X}_2;\mathsf{Y}|\mathsf{X}_1)-\delta(\epsilon))} \\ &+ \lceil 2^{nR_1}\rceil\lceil 2^{nR_2}\rceil 2^{-n(\mathbb{I}(\mathsf{X}_1\mathsf{X}_2;\mathsf{Y})-\delta(\epsilon))}.\end{aligned}$$

Hence, if we choose $R_1$ and $R_2$ to satisfy

$$\begin{aligned}R_1 &< \mathbb{I}(\mathsf{X}_1;\mathsf{Y}|\mathsf{X}_2) - \delta(\epsilon),\\ R_2 &< \mathbb{I}(\mathsf{X}_2;\mathsf{Y}|\mathsf{X}_1) - \delta(\epsilon),\\ R_1 + R_2 &< \mathbb{I}(\mathsf{X}_1\mathsf{X}_2;\mathsf{Y}) - \delta(\epsilon),\end{aligned}$$

we obtain $\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathsf{C}_n)] \leqslant \delta_\epsilon(n)$. By applying the selection lemma to the random variable $\mathsf{C}_n$ and the function $\mathbf{P}_{\mathrm{e}}$, we conclude that there exists a $(2^{nR_1}, 2^{nR_2}, n)$ code $\mathcal{C}_n$ such that $\mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) \leqslant \delta_\epsilon(n)$. Since $\epsilon$ can be chosen arbitrarily small and since the distributions $p_{\mathsf{X}_1}$ and $p_{\mathsf{X}_2}$ are arbitrary, we conclude that

$$\bigcup_{p_{\mathsf{X}_1}p_{\mathsf{X}_2}} \left\{ (R_1, R_2): \begin{array}{c} 0 \leqslant R_1 \leqslant \mathbb{I}(\mathsf{X}_1;\mathsf{Y}|\mathsf{X}_2) \\ 0 \leqslant R_2 \leqslant \mathbb{I}(\mathsf{X}_2;\mathsf{Y}|\mathsf{X}_1) \\ 0 \leqslant R_1 + R_2 \leqslant \mathbb{I}(\mathsf{X}_1\mathsf{X}_2;\mathsf{Y}) \end{array} \right\} \subseteq \mathcal{C}^{\mathsf{MAC}}$$

is achievable. Finally, it can be shown that time-sharing between different codes achieves the entire convex hull [3, Section 15.3]. We refer the reader to [3, 6] for the converse part of the proof. □

The typical shape of the region $\mathcal{R}(p_{\mathsf{X}_1}p_{\mathsf{X}_2})$ is illustrated in Figure 2.8. The boundaries of the capacity region can be explained in a very intuitive way. When encoder 1 views the signals sent by encoder 2 as noise, its maximum achievable rate is on the order of $\mathbb{I}(\mathsf{X}_1;\mathsf{Y})$, which is a direct consequence of the channel coding theorem. Then, the decoder can estimate the codeword $x_1^n$ and subtract it from the channel output sequence $y_1^n$, thus allowing encoder 2 to achieve a maximum rate on the order of $\mathbb{I}(\mathsf{X}_2;\mathsf{Y}|\mathsf{X}_1)$. This procedure is sometimes called *successive cancellation* and leads to the upper corner point of the region. The lower corner point corresponds to the symmetric case, in which encoder 2 views the signals sent by encoder 1 as noise.

### 2.3.3 The broadcast channel

While a multiple-access channel considers multiple sources and one destination, the *broadcast channel* (BC for short) considers a single information source that transmits to multiple users. Applications of the BC include the downlink channel of a satellite or of a base station in a mobile communication network, and the wiretap channel which is studied in detail in Chapter 3 and Chapter 5. As illustrated in Figure 2.9, a discrete

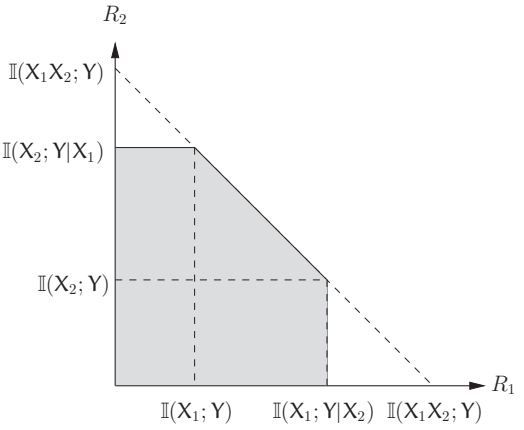


**Figure 2.8** Typical shape of the rate region $\mathcal{R}(p_{X_1}p_{X_2})$ of the multiple-access channel for fixed input distributions $p_{X_1}$ and $p_{X_2}$.

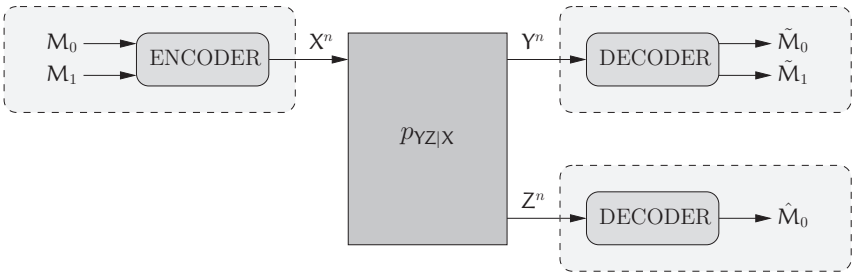


**Figure 2.9** Communication over a two-user broadcast channel.

memoryless two-user broadcast channel $(\mathcal{X}, p_{YZ|X}, \mathcal{Y}, \mathcal{Z})$ consists of a finite input alphabet $\mathcal{X}$, two finite output alphabets $\mathcal{Y}$ and $\mathcal{Z}$, and transition probabilities $p_{YZ|X}$ such that

$$\forall n \geqslant 1 \quad \forall (x^n, y^n, z^n) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n$$

$$p_{Y^nZ^n|X^n}(y^n, z^n|x^n) = \prod_{i=1}^{n} p_{YZ|X}(y_i, z_i|x_i).$$

We assume that the transmitter wants to send a *common message* $M_0$ to both receivers and a *private message* $M_1$ to the receiver observing $Y^n$. The receiver observing $Z^n$ is called a "weak" user, while the receiver observing $Y^n$ is called the "strong" user.

**Definition 2.26.** *A* $\left(2^{nR_0}, 2^{nR_1}, n\right)$ *code* $\mathcal{C}_n$ *for the BC consists of*

- *two message sets* $\mathcal{M}_0 = [\![1, 2^{nR_0}]\!]$ *and* $\mathcal{M}_1 = [\![1, 2^{nR_1}]\!]$*;*
- *an encoding function* $f : \mathcal{M}_0 \times \mathcal{M}_1 \to \mathcal{X}^n$*, which maps a message pair* $(m_0, m_1)$ *to a codeword* $x^n$*;*

- *a decoding function* $g : \mathcal{Y}^n \to (\mathcal{M}_0 \times \mathcal{M}_1) \cup \{?\}$*, which maps each channel observation* $y^n$ *to a message pair* $(\tilde{m}_0, \tilde{m}_1) \in \mathcal{M}_0 \times \mathcal{M}_1$ *or an error message* **?***;*
- *a decoding function* $h : \mathcal{Z}^n \to \mathcal{M}_0 \cup \{?\}$*, which maps each channel observation* $z^n$ *to a message* $\hat{m}_0 \in \mathcal{M}_0$ *or an error message* **?***.*

Messages $\mathsf{M}_0$ and $\mathsf{M}_1$ are assumed uniformly distributed in their respective sets and the performance of a code $\mathcal{C}_n$ is measured in terms of the average probability of error

$$\mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) \triangleq \mathbb{P}\left[\hat{\mathsf{M}}_0 \neq \mathsf{M}_0 \text{ or } (\tilde{\mathsf{M}}_0, \tilde{\mathsf{M}}_1) \neq (\mathsf{M}_0, \mathsf{M}_1) \mid \mathcal{C}_n\right].$$

**Definition 2.27.** *A rate pair* $(R_0, R_1)$ *is achievable for the BC if there exists a sequence of* $\left(2^{nR_0}, 2^{nR_1}, n\right)$ *codes* $\{\mathcal{C}_n\}_{n \geqslant 1}$ *such that*

$$\lim_{n \to \infty} \mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) = 0.$$

*The capacity region of a BC is defined as*

$$\mathcal{C}^{\mathrm{BC}} \triangleq \mathrm{cl}(\{(R_0, R_1) : (R_0, R_1) \text{ is achievable}\}).$$

As in many other fundamental problems of network information theory, determining the capacity region of the broadcast channel turns out to be a very difficult task. Therefore, we provide only an achievable rate region, which, in general, is strictly smaller than the capacity region.

**Theorem 2.12** (Bergsman and Gallager)**.** *Consider a BC* $(\mathcal{X}, p_{\mathsf{YZ}|\mathsf{X}}, \mathcal{Y}, \mathcal{Z})$*. For any joint distribution* $p_{\mathsf{UX}}$ *on* $\mathcal{U} \times \mathcal{X}$*, define the set* $\mathcal{R}(p_{\mathsf{UX}})$ *as*

$$\mathcal{R}(p_{\mathsf{UX}}) \triangleq \left\{(R_0, R_1): \begin{array}{l} 0 \leqslant R_0 \leqslant \min(\mathbb{I}(\mathsf{U};\mathsf{Y}), \mathbb{I}(\mathsf{U};\mathsf{Z})) \\ 0 \leqslant R_1 \leqslant \mathbb{I}(\mathsf{X};\mathsf{Y}|\mathsf{U}) \end{array}\right\},$$

*where the joint distribution of* $\mathsf{U}$*,* $\mathsf{X}$*,* $\mathsf{Y}$*, and* $\mathsf{Z}$ *factorizes as* $p_{\mathsf{UX}} p_{\mathsf{YZ}|\mathsf{X}}$*. Then,*

$$\mathcal{R}^{\mathrm{BC}} \triangleq \mathrm{co}\left(\bigcup_{p_{\mathsf{UX}}} \mathcal{R}(p_{\mathsf{UX}})\right) \subseteq \mathcal{C}^{\mathrm{BC}}.$$

*In addition, the cardinality of the auxiliary random variable* $\mathsf{U}$ *can be limited to* $|\mathcal{U}| \leqslant \min(|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{Z}|)$*.*

*Proof.* The proof that $\mathcal{R}^{\mathrm{BC}} \subseteq \mathcal{C}^{\mathrm{BC}}$ is based on joint typicality, random coding, and a code construction called *superposition coding*. As illustrated in Figure 2.10, the idea of superposition coding is to create a codebook with $\lceil 2^{nR_0} \rceil$ codewords for the weakest user and to superpose a codebook with $\lceil 2^{nR_1} \rceil$ codewords for the strongest user to *every* codeword. The codewords $u^n$ are often called "cloud centers," while the codewords $x^n$ are called "satellite codewords." Formally, fix a joint probability distribution $p_{\mathsf{UX}}$ on $\mathcal{U} \times \mathcal{X}$. Let $0 < \epsilon < \mu_{\mathsf{XYU}}$, where $\mu_{\mathsf{XYU}} \triangleq \min p_{\mathsf{XU}}(x, u) p_{\mathsf{Y}|\mathsf{XU}}(y|x, u)$ and let $n \in \mathbb{N}^*$. Let $R_0 > 0$ and $R_1 > 0$ be rates to be specified later. We construct a $\left(2^{nR_0}, 2^{nR_1}, n\right)$ code $\mathcal{C}_n$ as follows.

- *Codebook construction.* Construct a codebook with $\lceil 2^{nR_0} \rceil$ codewords, labeled $u^n(m_0)$ with $m_0 \in [\![1, 2^{nR_0}]\!]$, by generating the symbols $u_i(m_0)$ for $i \in [\![1, n]\!]$ and

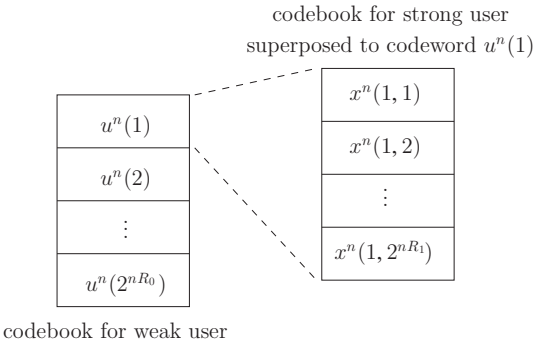


**Figure 2.10** Superposition coding. A codebook for the strong user is superposed to every codeword of the codebook for the weak user.

$m_0 \in [\![1, 2^{nR_0}]\!]$ independently according to $p_U$. For each $u^n(m_0)$ with $m_0 \in [\![1, 2^{nR_0}]\!]$, generate another codebook with $\lceil 2^{nR_1} \rceil$ codewords, labeled $x^n(m_0, m_1)$ with $m_1 \in [\![1, 2^{nR_1}]\!]$, by generating the symbols $x_i(m_0, m_1)$ for $i \in [\![1, n]\!]$ and $m_1 \in [\![1, 2^{nR_1}]\!]$ independently according to $p_{X|U=u_i(m_0)}$. The codebooks are revealed to the encoder and both decoders.

- *Encoder.* Given $(m_0, m_1)$, transmit $x^n(m_0, m_1)$.
- *Decoder for weak user.* Given $z^n$, output $\hat{m}_0$ if it is the unique message such that $(u^n(\hat{m}_0), z^n) \in \mathcal{T}_\epsilon^n(UZ)$; otherwise, output an error **?**.
- *Decoder for strong user.* Given $y^n$, output $(\tilde{m}_0, \tilde{m}_1)$ if it is the unique message pair such that $(u^n(\tilde{m}_0), y^n) \in \mathcal{T}_\epsilon^n(UY)$ and $(u^n(\tilde{m}_0), x^n(\tilde{m}_0, \tilde{m}_1), y^n) \in \mathcal{T}_\epsilon^n(UXY)$; otherwise, output an error **?**.

The random variable that denotes the randomly generated codebook $\mathcal{C}_n$ is denoted by $C_n$, and we proceed to bound $\mathbb{E}[\mathbf{P}_e(C_n)]$. From the symmetry of the random code construction, notice that

$$\begin{aligned}\mathbb{E}[\mathbf{P}_e(C_n)] &= \mathbb{E}_{C_n}\left[\mathbb{P}\left[\hat{M}_0 \neq M_0 \text{ or } (\tilde{M}_0, \tilde{M}_1) \neq (M_0, M_1) \mid C_n\right]\right] \\ &= \mathbb{E}_{C_n}\left[\mathbb{P}\left[\hat{M}_0 \neq M_0 \text{ or } (\tilde{M}_0, \tilde{M}_1) \neq (M_0, M_1) \mid M_0 = 1, M_1 = 1, C_n\right]\right].\end{aligned}$$

Therefore, $\mathbb{E}[\mathbf{P}_e(C_n)]$ can be expressed in terms of the events

$$\begin{aligned}\mathcal{E}_i &= \left\{(U^n(i), Y^n) \in \mathcal{T}_\epsilon^n(UY)\right\} \quad \text{for } i \in [\![1, 2^{nR_0}]\!], \\ \mathcal{F}_i &= \left\{(U^n(i), Z^n) \in \mathcal{T}_\epsilon^n(UZ)\right\} \quad \text{for } i \in [\![1, 2^{nR_0}]\!], \\ \mathcal{G}_{ij} &= \left\{(U^n(i), X^n(j), Y^n) \in \mathcal{T}_\epsilon^n(UXY)\right\} \quad \text{for } i \in [\![1, 2^{nR_0}]\!] \text{ and } j \in [\![1, 2^{nR_1}]\!]\end{aligned}$$

as

$$\mathbb{E}[\mathbf{P}_e(C_n)] = \mathbb{P}\left[\mathcal{E}_1^c \cup \bigcup_{i \neq 1} \mathcal{E}_i \cup \mathcal{F}_1^c \cup \bigcup_{i \neq 1} \mathcal{F}_i \cup \bigcup_{j \neq 1} \mathcal{G}_{1j}\right]$$

and, by the union bound,

$$\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathrm{C}_n)] \leqslant \mathbb{P}\left[\mathcal{E}_1^{\mathrm{c}}\right] + \sum_{i\neq 1} \mathbb{P}[\mathcal{E}_i] + \mathbb{P}\left[\mathcal{F}_1^{\mathrm{c}}\right] + \sum_{i\neq 1} \mathbb{P}[\mathcal{F}_i] + \sum_{j\neq 1} \mathbb{P}\left[\mathcal{G}_{1j}\right]. \tag{2.17}$$

By the joint AEP,

$$\mathbb{P}\left[\mathcal{E}_1^{\mathrm{c}}\right] \leqslant \delta_\epsilon(n) \quad \text{and} \quad \mathbb{P}\left[\mathcal{F}_1^{\mathrm{c}}\right] \leqslant \delta_\epsilon(n). \tag{2.18}$$

For $i \neq 1$, notice that $\mathrm{U}^n(i)$ is independent of $\mathrm{Y}^n$ and $\mathrm{Z}^n$; therefore, by Corollary 2.2,

$$\mathbb{P}[\mathcal{E}_i] \leqslant 2^{-n(\mathbb{I}(\mathrm{U};\mathrm{Y})-\delta(\epsilon))} \quad \text{and} \quad \mathbb{P}[\mathcal{F}_i] \leqslant 2^{-n(\mathbb{I}(\mathrm{U};\mathrm{Z})-\delta(\epsilon))} \quad \text{for } i \neq 1. \tag{2.19}$$

For $j \neq 1$, $\mathrm{X}^n(1, j)$ is conditionally independent of $\mathrm{Z}^n$ given $\mathrm{U}^n(1)$; therefore, by Corollary 2.3,

$$\mathbb{P}\left[\mathcal{G}_{1j}\right] \leqslant 2^{-n(\mathbb{I}(\mathrm{X};\mathrm{Y}|\mathrm{U})-\delta(\epsilon))} \quad \text{for } j \neq 1. \tag{2.20}$$

On substituting (2.18), (2.19), and (2.20) into (2.17), we obtain

$$\begin{aligned}\mathbb{E}[\mathbf{P}_{\mathrm{e}}(\mathrm{C}_n)] \leqslant \delta_\epsilon(n) &+ \lceil 2^{nR_0} \rceil 2^{-n(\mathbb{I}(\mathrm{U};\mathrm{Y})-\delta(\epsilon))} + \lceil 2^{nR_0} \rceil 2^{-n(\mathbb{I}(\mathrm{U};\mathrm{Z})-\delta(\epsilon))} \\ &+ \lceil 2^{nR_1} \rceil 2^{-n(\mathbb{I}(\mathrm{X};\mathrm{Y}|\mathrm{U})-\delta(\epsilon))}.\end{aligned}$$

Hence, if we choose the rates $R_0$ and $R_1$ to satisfy

$$\begin{aligned}R_0 &< \min(\mathbb{I}(\mathrm{U};\mathrm{Y}), \mathbb{I}(\mathrm{U};\mathrm{Z})) - \delta(\epsilon), \\ R_1 &< \mathbb{I}(\mathrm{X};\mathrm{Y}|\mathrm{U}) - \delta(\epsilon),\end{aligned}$$

we obtain $\mathbb{E}[\mathrm{C}_n] \leqslant \delta_\epsilon(n)$. By applying the selection lemma to the random variable $\mathrm{C}_n$ and the function $\mathbf{P}_{\mathrm{e}}$, we conclude that there exists a $\left(2^{nR_0}, 2^{nR_1}, n\right)$ code $\mathcal{C}_n$ such that $\mathbf{P}_{\mathrm{e}}(\mathcal{C}_n) \leqslant \delta_\epsilon(n)$. Since $\epsilon$ can be chosen arbitrarily small and since the distribution $p_{\mathrm{UX}}$ is arbitrary, we conclude that

$$\left\{(R_0, R_1): \begin{array}{l} 0 \leqslant R_0 \leqslant \min(\mathbb{I}(\mathrm{U};\mathrm{Y}), \mathbb{I}(\mathrm{U};\mathrm{Z})) \\ 0 \leqslant R_1 \leqslant \mathbb{I}(\mathrm{X};\mathrm{Y}|\mathrm{U}) \end{array}\right\} \subseteq \mathcal{C}^{\mathrm{BC}}.$$

Since $p_{\mathrm{UX}}$ is arbitrary and since it is possible to perform time-sharing, the theorem follows. The bound for the cardinality of the random variable $\mathrm{U}$ follows from Caratheodory's theorem, and we refer the reader to [3] for details. □

## 2.4 Bibliographical notes

The definitions of typical sequences and their properties are those described in the textbooks [3, 4, 6]. The notion of d-separation is a known result in statistical inference, and we have used the definition provided by Kramer [8].

There are several ways of proving the channel coding theorem, see for instance [2, 3, 4, 5], and our presentation is based on the approach in [3, 4]. The Slepian–Wolf theorem was established by Slepian and Wolf in [9]. Additional examples of results proved with random binning with side information can be found in [10] and [11].

The capacity of the two-user multiple-access channel with independent messages was obtained independently by Ahlswede [12] and Liao [13]. The broadcast channel model was proposed by Cover [14]. Bergmans [15] and Gallager [16] proved that the achievable region obtained in this chapter is the capacity region of a subclass of broadcast channels called *physically degraded* broadcast channels. Surveys of known results about network information theory can be found in Cover's survey paper [17] or Kramer's monograph [6].