Chapter 6

Elements of Information Theory

Entropy and *mutual information* were introduced by Shannon in 1948, and this began a remarkable development of *information theory.* Over the last fifty years information theory underwent many changes and found new applications. Entropy, the Shannon-McMillan-Breiman theorem and random coding technique nowadays are standard tools of analysis of algorithms. In this chapter, we discuss elements of information theory and illustrate its applications to analysis of algorithms. In particular, we prove three main results of Shannon, that of source coding, channel coding and rate distortion. In the application section, we discuss a variety of data compression schemes based on exact and approximate pattern matching.

ENTROPY of a single random variable X with probability mass $P(a) = \Pr\{X = a\}$ is defined as

$$h(X) = -\sum_{a \in \mathcal{A}} P(a) \log P(a)$$

where by convention $0 \log 0 = 0$. We left intentionally the base of the logarithm unspecified. In this chapter it is convenient to use the logarithm to base 2. If the base of the logarithm is 2, we measure the entropy in *bits*; with the natural logarithm we measure the entropy in *nats*. As we shall see, the entropy represents average uncertainty of X, and in fact it is the number of bits on the average required to describe the random variable (cf. Exercise 1). For example, one needs 5 bits to identify a value of a random variable distributed uniformly over 32 outcomes, and the entropy of X is $h(X) = 5$ bits.

The real virtue of entropy is manifested the best when investigating entropy rates of stochastic processes. The *Asymptotic Equipartition Property* (AEP) that follows from the most celebrated result of information theory, namely that of **Shannon-McMillan-Breiman** theorem, is the key result. Let us start with a simple observation: Consider an i.i.d. binary sequence of symbols of length n, say $(X_1, \ldots, X_n)$, with p denoting the probability of occurrence of one symbol and $q = 1 - p$ the probability of the other symbol. When $p = q = 1/2$, then $\Pr\{X_1, \ldots, X_n\} = 2^{-n}$ and it does not matter what are the actual values of $X_1, \ldots, X_n$. In general, $\Pr\{X_1, \ldots, X_n\}$ is not the same for all possible values of $X_1, \ldots, X_n$, however, we shall see that all **typical** sequences $(X_1, \ldots, X_n)$ have "asymptotically" the same probability. Consider $p \neq q$ in the example above. In a typical binary sequence of length n there are approximately $np + O(\sqrt{n})$ ones and $nq + O(\sqrt{n})$ zeros. We expect expect that the probability of such a sequence is well approximated by

$$p^{np+O(\sqrt{n})}q^{nq+O(\sqrt{n})} = 2^{-n(-p\log p - q\log q)+O(\sqrt{n})} \sim 2^{-nh}$$

where $h = -p\log p - q\log q$ is the *entropy* of the underlying Bernoulli random variable. We conclude that typical sequences have approximately the same probability equal to 2^{-nh}. We make this statement rigorous in this chapter.

Information theory was born in October 1948 when Shannon published his classic paper "A Mathematical Theory of Communication". He established three main results of information theory, namely that of source coding, channel coding and rate distortion. In source coding one wants to find the shortest description of a message. It turns out that you cannot beat a certain limit which is equal to entropy times the length of the original message. In channel coding, the goal is the opposite to source coding, namely, we want to send *reliably* as much information as possible. Finally, there are sources of information that cannot be described by a finite number of bits (e.g., an arbitrary real number, a continuous function, etc.). In such cases, we can only describe *approximately* the source of information. How well we can do? To answer this question we enter the realm of *rate distortion theory.* We touch in this chapter some of these problems. The reader is referred to the excellent books of Cover and Thomas [59], and Csiszár and Körner [62] for further information. In fact, a large part of this chapter is based on [59].

In the analysis of algorithms, which is the main topic of this book, entropy appears quite often in disparate problems as we shall see in the application section of this chapter. Actually, in the analysis of algorithm the so called Rényi's entropy of order b occurs almost as frequently as Shannon entropy. We study some properties of Rényi entropy and illustrate its applications for problems on strings. In particular, we study the phrase length in the Lempel-Ziv scheme and its lossy extension, the shortest common superstring problem, and the fixed database model for data compression.

6.1 Entropy, Relative Entropy, and Mutual Information

In this section we present basic definitions of entropy, relative entropy, Rényi entropy, and mutual information for single random variables. We shall follow the presentation of Cover and Thomas [59].

Let us consider a random variable X defined over a finite alphabet $\mathcal{A}$. We write $P(X)$ for a random function of X such that $P(a) = \Pr\{X = a\}$ for every $a \in \mathcal{A}$. **Shannon entropy** is defined as

$$h(X) = -\mathbf{E}[\log P(X)] = -\sum_{a\in\mathcal{A}} P(a)\log P(a) \tag{6.1}$$

where throughout this chapter we assume that the logarithm base is 2 and $0\log 0 = 0$. We sometimes write $h(P)$ or simply h for the entropy. In some situations, we specifically show that the average in (6.1) is taken over the measure P, that is, we write $h(X) = \mathbf{E}_P[-\log P(X)]$.

As mentioned in the introduction, the entropy is a measure of uncertainty of X. It is also the minimum average number of bits required to describe X, as illustrated in the following example.

Example 6.1: *Entropy and Random Variable Representation*

Let X be a random variable taken eight values with the corresponding probabilities $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. The entropy of X is

$$h(X) = 2 \text{ bits.}$$

Suppose we want to rely to another person the outcome of X. We can use 3 bits to encode 8 messages, however, we can do much better by encoding more likely outcomes with shorter codes and less likely with longer. For example, if we assign the following codes to eight messages: 0, 10, 110, 1110, 111100, 111101, 111110, 111111, then we need on the average only 2 bits. Notice that the description length in this case is equal to the entropy. In Exercise 1 the reader is asked to extend it to general discrete random variable. □

The entropy $h(X)$ leads to several new definitions that we discuss in the sequel.

Definition 6.1 (i) **Joint Entropy** $h(X,Y)$ *is defined as*

$$h(X,Y) = -\mathbf{E}[\log P(X,Y)] = -\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}'} P(a,a')\log P(a,a'). \tag{6.2}$$

(ii) **Conditional Entropy** $h(Y|X)$ *is*

$$h(Y|X) = -\mathbf{E}_{P(X,Y)}[\log P(Y|X)] = \sum_{a\in\mathcal{A}} P(a)h(Y|\ X=a) = -\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}'} P(a,a')\log P(a'|a). \tag{6.3}$$

(iii) **Relative Entropy** *or* **Kullback Leibler Distance** *between two distributions P and Q defined on the same probability space is*

$$D(P \parallel Q) = \mathbf{E}_P\left[\log\frac{P(X)}{Q(X)}\right] = \sum_{a\in\mathcal{A}} P(a)\log\frac{P(a)}{Q(a)} \tag{6.4}$$

where by convention $0\log 0/Q = 0$ and $P\log P/0 = \infty$.

(iv) **Mutual Information** *of X and Y is the relative entropy between the joint distribution of X and Y and the product distribution $P(X)P(Y)$, that is,*

$$I(X;Y) = \mathbf{E}_{P(X,Y)}\left[\log\frac{P(X,Y)}{P(X)P(Y)}\right] = \sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}'} P(a,a')\log\frac{P(a,a')}{P(a)P(a')}. \tag{6.5}$$

(v) **Rényi's Entropy of order** b *($-\infty \le b \le \infty$) is defined as*

$$h_b(X) = -\frac{\log \mathbf{E}[P^b(X)]}{b} = -\frac{1}{b}\log\sum_{a\in\mathcal{A}} P^{b+1}(a) \tag{6.6}$$

provided $b \neq 0$ where

$$
\begin{aligned}
h_{-\infty} &= \min_{i \in \mathcal{A}} \{P(i)\}, & (6.7)\\
h_{\infty} &= \max_{i \in \mathcal{A}} \{P(i)\} & (6.8)
\end{aligned}
$$

by Hilbert's inequalities on means (2.12)–(2.13). Observe that $h(X) = \lim_{b \to 0} h_b(X)$.

Before we justify these definitions and provide an interpretation, let us explore some mathematical relationships between them. These properties are easy to prove by elementary algebra, and the reader is asked in Exercise 3 to verify them (cf. Exercise 4 for more properties of the entropy).

Theorem 6.2 *The following holds*

$$
\begin{aligned}
h(X, Y) &= h(X) + h(Y|X), & (6.9)\\
h(X_1, X_2, \ldots, X_n) &= \sum_{i=1}^{n} h(X_i|X_{i-1}, \ldots, X_1), & (6.10)\\
I(X;Y) &= h(X) - h(X|Y) = h(Y) - h(Y|X), & (6.11)\\
I(X;X) &= h(X). & (6.12)
\end{aligned}
$$

It is intuitively clear that the conditional entropy $h(X|Y)$ should not be bigger than the entropy $h(X)$ since when Y is correlated with X we learn something about X knowing Y, hence its uncertainty decreases. We prove this fact, and some others, formally below.

Theorem 6.3 *The following inequalities hold*

$$
\begin{aligned}
D(P \parallel Q) &\geq 0, & (6.13)\\
I(X;Y) &\geq 0, & (6.14)\\
h(X) &\geq h(X|Y), & (6.15)\\
h(X_1, \ldots, X_n) &\leq \sum_{i=1}^{n} h(X_i), & (6.16)
\end{aligned}
$$

with equality in (6.13) if an only if $P(a) = Q(a)$ for all $a \in \mathcal{A}$, equality in (6.14)–(6.15) if and only if X and Y are independent, and equality in (6.16) if and only if $X_1, \ldots, X_n$ are independent.

Proof. Since $\log x \leq x - 1$ for $x > 0$, we obtain

$$
\begin{aligned}
-D(P \parallel Q) &= \mathbf{E}_P \left[\log \frac{Q}{P} \right]\\
&\leq \sum_{a \in \mathcal{A}} Q(a) - \sum_{a \in \mathcal{A}} P(a) = 1 - 1 = 0
\end{aligned}
$$

which proves (6.13). Then (6.14) follows from $I(X;Y) = D(P(X,Y) \parallel P(X)P(Y))$, and (6.15) is a consequence of the above and (6.11). Finally, (6.16) follows from (6.10) and (6.15). ∎

6.2 Entropy Rate and Rényi's Entropy Rates

We first define precisely typical sequences and then prove the Shannon-McMillan-Breiman theorem for memoryless, mixing, and stationary and ergodic sources. In the second part of this section, we introduce Rényi's entropy rates that are widely used in analysis of algorithms.

6.2.1 The Shannon-McMillan-Breiman Theorem

Let $\{X_k\}_{k=1}^{\infty}$ be a one-sided stationary sequence, and let $P(X_1^n)$ be a random variable equal to $\Pr\{X_1^n = (x_1, \ldots, x_n)\}$ whenever $X_1^n = (x_1, \ldots, x_n) \in \mathcal{A}^n$. We can define *entropy rate* of X_1^n in two different ways, namely:

$$h := -\lim_{n\to\infty} \frac{\mathbf{E}[\log P(X_1^n)]}{n} = \lim_{n\to\infty} \frac{h(X_1, \ldots, X_n)}{n}, \tag{6.17}$$

$$h' := \lim_{n\to\infty} h(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1) \tag{6.18}$$

provided the limits above exist. We shall prove that for stationary processes the limits exist and are equal.

First of all, observe that the conditional entropy $h(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$ is non-increasing with n and hence has a limit. Indeed, by (6.15) and stationarity

$$\begin{aligned} h(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_1) &\le h(X_{n+1} \mid X_n, X_{n-1}, \ldots, X_2) \\ &= h(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1). \end{aligned}$$

Now, we are in the position to prove the following statement.

Theorem 6.4 *For a stationary process, the limits in (6.17) and (6.18) exist and are equal, that is,*

$$h = h'.$$

Proof. By the chain rule (6.10),

$$\frac{h(X_1, \ldots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} h(X_i \mid X_{i-1}, \ldots, X_1).$$

But, we just proved that the conditional entropies $h(X_i \mid X_{i-1}, \ldots, X_1)$ tend to a limit h', hence by the *Cesáro Mean* theorem, the right-hand side converges to h'. We must then conclude that the limit of the left-hand side exist, and is equal to $h = h'$. ∎

Example 6.2: *Entropy Rates for Memoryless and Markovian Sources*

Let us first consider a memoryless source. Then $P(X_1^n) = P^n(X_1)$ and we have $h(X_1^n) = n\mathbf{E}[-\log P(X_1)] = nh(X_1)$.

For a Markovian source, we appeal to the definition (6.18). Let us consider a stationary Markov source of order one. Then

$$h = \lim_{n\to\infty} h(X_n|\ X_{n-1},\ldots,X_1) = \lim_{n\to\infty} h(X_n|\ X_{n-1}) = h(X_2|\ X_1).$$

Thus, by definition (6.3), we have

$$h = h(X_2|\ X_1) = -\sum_{i,j\in A} \pi_i p_{ij} \log p_{ij} \tag{6.19}$$

where π_i is the stationary distribution and $\mathsf{P} = \{p_{ij}\}_{i,j\in A}$ the transition matrix of the stationary Markov chain. □

We are now well prepared to derive the most important theorem of this chapter, namely the Shannon-McMillan-Breiman theorem that roughly says that the the average operator $\mathbf{E}[\cdot]$ can be taken away from the definition of the entropy (6.17) provided the deterministic limit is replaced by almost sure convergence of the random variable $\log P(X_1^n)$. The main consequence of this theorem is the so called *Asymptotic Equipartition Property* discussed in the next section.

Let us put some rigor into the above rough description. We first show how to derive the Shannon-McMillan-Breiman theorem for memoryless sources and then for mixing sources. Finally, we shall formulate the theorem in its full generality and prove it for stationary ergodic sources using the "sandwich" approach of Algoet and Cover [8]. Consider first a memoryless source. Then

$$\begin{aligned} -\frac{\log P(X_1^n)}{n} &= -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i) \\ &\to \mathbf{E}[-\log P(X_1)] = h \qquad \text{(a.s.)}\ , \end{aligned}$$

where the last implication follows from the *Strong Law of Large Numbers* applied to the sequence $(-\log P(X_1),\ldots,-\log P(X_n))$. One *should* notice a difference between the definition of the entropy (6.17) and the result above. In (6.17) we take the *average* of $\log P(X_1^n)$ while in the above we proved that *almost surely* the probability $P(X_1^n)$ can be well approximated by 2^{-nh}. We are aiming now at showing that the above conclusion is true for much more general sources.

As the next step, let us consider mixing sources that include as a special case Markovian sources (cf. Chapter 2). In fact, we consider a weak mixing since we only assume that there exists a constant c such that

$$P(X_1^{n+m}) \le cP(X_1^n)P(X_{n+1}^{n+m})$$

for integers $n, m \geq 0$. Taking logarithm we obtain

$$\log P(X_1^{n+m}) \leq \log P(X_1^n) + \log P(X_{n+1}^{n+m}) + \log c$$

which satisfies conditions of the subadditive ergodic theorem Theorem 5.8. Thus

$$h = -\lim_{n\to\infty} \frac{\log P(X_1^n)}{n} \qquad \text{(a.s.)} \ .$$

Again, the reader should notice the difference between this result and the definition of the entropy.

We are finally ready to state the Shannon-McMillan-Breiman theorem in its full generality. We prove it using a Markov approximation discussed in Lemma 4.2 (cf. also Theorem 6.33 in Exercise 10 of Chapter 5).

Theorem 6.5 (Shannon-McMillan-Breiman) *For a stationary and ergodic sequence $\{X_k\}$ the following holds*

$$h = -\lim_{n\to\infty} \frac{\log P(X_1^n)}{n} \qquad \text{(a.s.)} \ .$$

where h is the entropy rate of the process $\{X_k\}$.

Proof. We present here the Algoet and Cover "sandwich" proof. The main idea is to use the kth order Markov approximation of a stationary distribution. Let P be the stationary distribution, and

$$P^k(X_0^{n-1}) := P(X_0^{k-1}) \prod_{i=k}^{n-1} P(X_i|X_{i-k}^{i-1})$$

be its kth order Markov approximation. Using (4.6) of Lemma 4.2, we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P(X_0^{n-1})} \leq \limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P^k(X_0^{n-1})} = h^k$$

for any $k \geq 1$. In the above, h^k is the entropy of the kth order Markov chain, which we know to exist.

Now, we consider a lower bound by using (4.7) of Lemma 4.2, that is,

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{P(X_0^{n-1})} \geq \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P(X_0^{n-1}|X_{-\infty}^{-1})} = h^\infty$$

where h^∞ exists by the ergodic theorem. Putting everything together, we have

$$h^\infty \leq \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{P(X_0^{n-1})} \leq \limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P(X_0^{n-1})} \leq h^k.$$

We shall prove now that $h^k \downarrow h^\infty$ completing the proof since by Theorem 6.4 $h = h^\infty$. By the Doob–Levy martingale construction (cf. Exercise 6) of Chapter 5, we know that $P(x_0|X_{-k}^{-1})$ is a martingale. Therefore, by the martingale convergence Theorem 5.12

$$P(x_0|X_{-k}^{-1}) \to P(x_0|X_{-\infty}^{-1}) \quad \text{(a.s.)}$$

for all $x_0 \in \mathcal{A}$. Since $x \log x$ is bounded and continuous, the *bounded convergence theorem* allows to interchange of expectation and limit, yielding

$$\begin{aligned}\lim_{k\to\infty} h^k &= \lim_{k\to\infty} \mathbf{E}\left[-\sum_{x_0\in\mathcal{A}} P(x_0|X_{-k}^{-1}) \log P(x_0|X_{-k}^{-1})\right] \\ &= \mathbf{E}\left[-\sum_{x_0\in\mathcal{A}} P(x_0|X_{-\infty}^{-1}) \log P(x_0|X_{-\infty}^{-1})\right] = h^\infty\end{aligned}$$

which completes the proof. ∎

6.2.2 Rényi's Entropy Rates

In many problems on words **Rényi entropy rates** are widely used (cf. Section 6.5 and [22, 23, 25, 324, 325]). For $-\infty \le b \le \infty$, the bth *order Rényi entropy* is defined as

$$h_b = \lim_{n\to\infty} \frac{-\log \mathbf{E}[P^b(X_1^n)]}{bn} = \lim_{n\to\infty} \frac{-\log\left(\sum_{w\in\mathcal{A}^n} P^{b+1}(w)\right)^{-1/b}}{n}, \tag{6.20}$$

provided the above limit exists. In the above, we write $P(w) = \Pr\{X_1^n = w\}$ for $w \in \mathcal{A}^n$. We will prove below that for mixing processes generalized Rényi's entropies exit. For now, we observe that

$$h = \lim_{b\to 0} h_b,$$

and by Hilbert's inequalities on means (2.12)–(2.13)

$$\begin{aligned} h_{-\infty} &= \lim_{n\to\infty} \frac{\max\{-\log P(X_1^n)\ , P(X_1^n) > 0\}}{n}, \\ h_\infty &= \lim_{n\to\infty} \frac{\min\{-\log P(X_1^n)\ , P(X_1^n) > 0\}}{n}. \end{aligned}$$

Example 6.3 *Rényi's Entropy for Memoryless and Markovian Sources*

We compute here the Rényi's entropies for memoryless and Markovian sources.

1. Memoryless Source. In this case, for $b > 0$ and a binary alphabet

$$\sum_{w\in\mathcal{A}^n} P^{b+1}(w) = \sum_{k=0}^{n} \binom{n}{k} p^{(b+1)k} q^{(b+1)(n-k)} = (p^{b+1} + q^{b+1})^n$$

where, as always, $q = 1 - p$. For a general alphabet $\mathcal{A}$ of size V and probabilities $p_1, \ldots, p_V$ we obtain

$$
\begin{aligned}
h_b &= -\frac{1}{b}\log(p_1^{b+1} + \cdots + p_V^{b+1}) \qquad b > 0, && (6.21)\\
h_{-\infty} &= -\log p_{\max}, && (6.22)\\
h_{\infty} &= -\log p_{\min}, && (6.23)
\end{aligned}
$$

where $p_{\max} = \max\{p_1, \ldots, p_V\}$ and $p_{\min} = \min\{p_1, \ldots, p_V\}$.

2. Markovian Source. In Section 4.2.3 we proved that

$$\sum_{w \in \mathcal{A}^n} P^{b+1}(w) = \beta \lambda_{[b+1]}^n (1 + O(\rho^n))$$

for constants $b > 0$ and $\rho < 1$, where $\lambda_{[b]}$ is the largest eigenvalue of the Schur product (i.e., element-wise product) $\mathsf{P}_{[b]} = \mathsf{P} \circ \mathsf{P} \circ \cdots \circ \mathsf{P}$ of b transition matrices P. In view of this,

$$h_b = -\frac{1}{b}\log \lambda_{[b+1]}, \qquad b > 0. \tag{6.24}$$

With respect to $h_{-\infty}$ and h_∞, we need a result from digraphs (cf. [191, 290]). Consider a digraph on $\mathcal{A}$ with weights equal to $-\log p_{ij}$ where $i, j \in \mathcal{A}$. Define a cycle $\mathcal{C} = \{i_1, i_2, \ldots, i_v, i_1\}$ for some $v \le V$ such that $i_j \in \mathcal{A}$, and let $\ell(\mathcal{C}) = -\sum_{k=1}^{v} \log(p_{i_k, i_{k+1}})$ (with $i_{v+1} = i_1$) be the total weight of the cycle $\mathcal{C}$. Then

$$
\begin{aligned}
h_{-\infty} &= \min_{\mathcal{C}} \left\{ \frac{\ell(\mathcal{C})}{|\mathcal{C}|} \right\},\\
h_{\infty} &= \max_{\mathcal{C}} \left\{ \frac{\ell(\mathcal{C})}{|\mathcal{C}|} \right\}.
\end{aligned}
$$

Karp [191] showed how to compute these quantities efficiently. □

We should point out that we have already encountered Rényi's entropies on pages of this book. In Section 4.2.3 the entropies h_b and $h_{-\infty}$ appeared in the analysis of the height of digital trees, while $h_{-\infty}$ popped up in the study of the shortest path and fill-up level in a digital tree (cf. Exercise 10 of Chapter 4). The first Rényi's entropy h_1 plays a role in the analysis of molecular sequences as discussed in Arratia and Waterman [23, 25, 350].

6.3 Asymptotic Equipartition Property

An important conclusion from the Shannon-McMillan-Breiman theorem is the **Asymptotic Equipartition Property** (AEP) which asserts that asymptotically all *typical* sequences have the same probability approximately equal to 2^{-nh}. We first define precisely the basic AEP property, and then extend it to jointly typical sequences, formulate AEP for relative entropy (divergence), and finally discuss a lossy generalization of AEP.

6.3.1 Typical Sequences

Let us start with a definition.

Definition 6.6 *The typical set $\mathcal{G}_n^\varepsilon$ with respect to the probability measure $P = \mathsf{Pr}\{\cdot\}$ is the set of all sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{A}^n$ such that for given $\varepsilon > 0$*

$$2^{-n(h(P)+\varepsilon)} \leq P(x_1, \ldots, x_n) \leq 2^{-n(h(P)-\varepsilon)} \tag{6.25}$$

where $P(x_1, \ldots, x_n) := \mathsf{Pr}\{X_1^n = (x_1, \ldots, x_n)\}$.

The Shannon-McMillan-Breiman theorem implies the following property that is at the heart of information theory.

Theorem 6.7 (Asymptotic Equipartition Property) *For a stationary and ergodic sequence X_1^n, the state space $\mathcal{A}^n$ can be partitioned into two subsets $\mathcal{B}_n^\varepsilon$ ("bad set") and $\mathcal{G}_n^\varepsilon$ ("good set") such that for given $\varepsilon > 0$ there is N_ε so that for $n \geq N_\varepsilon$ we have*

$$\begin{aligned} 2^{-nh(1+\varepsilon)} &\leq P(x_1^n) \leq 2^{-nh(1-\varepsilon)} \qquad \text{for} \qquad x_1^n \in \mathcal{G}_n^\varepsilon, &(6.26)\\ \mathsf{Pr}\{\mathcal{B}_n^\varepsilon\} &\leq \varepsilon. &(6.27)\end{aligned}$$

Moreover,

$$(1-\varepsilon)2^{nh(1-\varepsilon)} \leq |\mathcal{G}_n^\varepsilon| \leq 2^{nh(1+\varepsilon)} \tag{6.28}$$

for sufficiently large n.

Proof. By Shannon-McMillan-Breiman theorem we have for any $\varepsilon > 0$

$$\mathsf{Pr}\left\{\left|-\frac{1}{n}\log P(X_1, \ldots, X_n) - h\right| < \varepsilon\right\} > 1 - \varepsilon$$

for, say $n > N_\varepsilon$. Denote by $\mathcal{B}_n^\varepsilon$ the set of $x_1^n = (x_1, \ldots, x_n)$ that does not satisfy the above. Its probability is smaller than ε. Call those x_1^n that fulfills the above inequality as $\mathcal{G}_n^\varepsilon$. Then for $x_1^n \in \mathcal{G}_n^\varepsilon$

$$h - \varepsilon \leq -\frac{1}{n}\log P(x_1, \ldots, x_n) \leq h + \varepsilon$$

which proves (6.26) and (6.27). Certainly, $\mathcal{G}_n^\varepsilon$ is the typical set according to Definition 6.6.

To prove the second part of the theorem, we proceed as follows

$$\begin{aligned} 1 &= \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \geq \sum_{x_1^n \in \mathcal{G}_n^\varepsilon} P(x_1^n) \geq \sum_{x_1^n \in \mathcal{G}_n^\varepsilon} 2^{-nh(1+\varepsilon)} \\ &= 2^{-nh(1+\varepsilon)}|\mathcal{G}_n^\varepsilon|, \end{aligned}$$

and

$$
\begin{aligned}
1-\varepsilon &< \Pr\{\mathcal{G}_n^\varepsilon\} \le \sum_{x_1^n \in \mathcal{G}_n^\varepsilon} 2^{-nh(1-\varepsilon)} \\
&= 2^{-nh(1-\varepsilon)} |\mathcal{G}_n^\varepsilon|.
\end{aligned}
$$

This completes the proof. ∎

6.3.2 Jointly Typical Sequences

In some applications (cf. Section 6.4.2), we need a generalization of AEP to a pair of sequences (X_1^n, Y_1^n). Let us start with a definition of *jointly typical set.*

Definition 6.8 *The typical set $\mathcal{G}_n^\varepsilon$ of sequences (x_1^n, y_1^n) with respect to the probability measure $P = \Pr\{\cdot,\cdot\}$ is the set of all n–sequences such that for given $\varepsilon > 0$ the sequence x_1^n is $P(x)$ ε–typical, y_1^n is $P(y)$ ε– typical and*

$$
2^{-n(h(X,Y)+\varepsilon)} \le P(x_1^n, y_1^n) \le 2^{-n(h(X,Y)-\varepsilon)} \tag{6.29}
$$

where $P(x_1^n, y_1^n) := \Pr\{X_1^n = (x_1, \ldots, x_n), Y_1^n = (y_1, \ldots, y_n)\}$. In short

$$
\begin{aligned}
\mathcal{G}_n^\varepsilon = \Big\{(x_1^n, y_1^n) :\ & 2^{-nh(1+\varepsilon)} \le P(x_1^n) \le 2^{-nh(1-\varepsilon)}, \\
& 2^{-nh(1+\varepsilon)} \le P(y_1^n) \le 2^{-nh(1-\varepsilon)}, \\
& 2^{-n(h(X,Y)+\varepsilon)} \le P(x_1^n, y_1^n) \le 2^{-n(h(X,Y)-\varepsilon)}\Big\}
\end{aligned}
$$

for any $\varepsilon > 0$.

The following is a generalization of AEP to jointly typical sequences.

Theorem 6.9 (Joint AEP) *Let $\{X_k\}$, $\{Y_k\}$ and $\{X_k, Y_k\}$ be mixing processes. Then for given $\varepsilon > 0$ there is N_ε so that for $n \ge N_\varepsilon$ we have*

$$
\Pr\{(X_1^n, Y_1^n) \in \mathcal{G}_n^\varepsilon\} \ \ge\ 1-\varepsilon, \tag{6.30}
$$

$$
(1-\varepsilon) 2^{n(h(X,Y)-\varepsilon)} \le |\mathcal{G}_n^\varepsilon| \ \le\ 2^{n(h(X,Y)+\varepsilon)}. \tag{6.31}
$$

In addition, if $\widetilde{X}_1^n$ and $\widetilde{Y}_1^n$ are independent with the same marginals as $P(X_1^n, Y_1^n)$, then

$$
(1-\varepsilon) 2^{-n(I(X,Y)+3\varepsilon)} \le \Pr\{(\widetilde{X}_1^n, \widetilde{Y}_1^n) \in \mathcal{G}_n^\varepsilon\} \le 2^{-n(I(X,Y)-3\varepsilon)} \tag{6.32}
$$

for sufficiently large n.

Proof. Since $\{X_k\}$ and $\{Y_k\}$ are mixing processes, by Shannon-McMillan-Breiman theorem we observe that for given $\varepsilon > 0$ there exists n_1 such that for $n > n_1$

$$\begin{aligned}\mathsf{Pr}\left\{\left|-\frac{1}{n}\log P(X_1^n) - h(X)\right| > \varepsilon\right\} &< \frac{\varepsilon}{3},\\ \mathsf{Pr}\left\{\left|-\frac{1}{n}\log P(Y_1^n) - h(Y)\right| > \varepsilon\right\} &< \frac{\varepsilon}{3}.\end{aligned}$$

Also, since the pair $\{X_k, Y_k\}$ is a mixing process, by the subadditive ergodic theorem (applied to $\log P(X_1^n, Y_1^n)$) we show that for given $\varepsilon > 0$ there exists n_2 such that for $n > n_2$

$$\mathsf{Pr}\left\{\left|-\frac{1}{n}\log P(X_1^n, Y_1^n) - h(X,Y)\right| > \varepsilon\right\} < \frac{\varepsilon}{3}.$$

In view of the above, for $n > \max\{n_1, n_2\}$ the probability of the union of the above sets is less than ε. Thus, the intersection of the complements of these sets form $\mathcal{G}_n^\varepsilon$, and therefore

$$\mathsf{Pr}\{\mathcal{G}_n^\varepsilon\} > 1 - \varepsilon.$$

To prove (6.31) we proceed exactly as in the proof of AEP of a single sequence, so we omit it here. Finally, for (6.32) we observe that

$$\begin{aligned}\mathsf{Pr}\{(\widetilde{X}_1^n, \widetilde{Y}_1^n) \in \mathcal{G}_n^\varepsilon\} &= \sum_{(x_1^n, y_1^n)\in\mathcal{G}_n^\varepsilon} P(x_1^n)P(y_1^n)\\ &\le 2^{n(h(X,Y)+\varepsilon)}2^{-n(h(X)-\varepsilon)}2^{-n(h(Y)-\varepsilon)}\\ &= 2^{-n(I(X,Y)-3\varepsilon)}.\end{aligned}$$

Using the left-hand side of (6.31), we establish in a similar manner the lower bound of (6.32). This completes the proof. ■

The above finding has profound implications for channel coding. We shall discuss it in depth in Section 6.4.2, however, a brief discussion is in order. Observe that there are about $2^{nh(X)}$ typical X sequences, about $2^{nh(Y)}$ typical Y sequences, but only $2^{nh(X,Y)}$ jointly typical sequences (recall, $h(X,Y) \le h(X) + h(Y)$). The probability that any randomly chosen pair (X, Y) is jointly typical is approximately

$$\frac{2^{nh(X,Y)}}{2^{nh(X)}2^{nh(Y)}} = 2^{-nI(X,Y)}.$$

In other words, for fixed Y_1^n, we may consider about $2^{nI(X,Y)}$ pairs before we find a jointly typical pair.

6.3.3 AEP for Biased Distributions

Consider the following *classification problem*: A nineteen century manuscript written in Polish was found in Paris. One suspects that it is a lost manuscript of a famous Polish poet Adam Mickiewicz. To prove this assertion a classification algorithm is proposed that will compare (cf. [360] for possible algorithms) the found text to samples of Mickiewicz's writings.

From information theory point of view, we may assume that the found text is generated by an unknown distribution $P = \Pr\{\cdot\}$ that is compared to the known distribution Q obtained from Mickiewicz's samples. In other words, we estimate the probability of a sample x_1^n generated by source P with respect to the known distribution Q. We call this problem: *AEP for a biased distribution.*

To start with, we must define the relative entropy or *divergence* $D(P \parallel Q)$ for random sequences. Let, as before, $X_1^n = (X_1, \ldots, X_n)$, be a random sequence generated by source P over a finite alphabet. We define the **divergence rate** as

$$D(P \parallel Q) := -\lim_{n\to\infty} \frac{\mathbf{E}_P[\log Q(X_1^n)/P(X_1^n)]}{n} \tag{6.33}$$

provided the limit exists. Unlike the entropy rate, the above limit may not exist for general stochastic processes, and in general one needs lim sup in the above definition. Below, we restrict our discussion to jointly mixing processes that will guarantee the existence of the limit. In such a case, we can also write

$$\begin{aligned} -\lim_{n\to\infty} \frac{\mathbf{E}_P[\log Q(X_1^n)]}{n} &= -\lim_{n\to\infty} \frac{\mathbf{E}_P[\log Q(X_1^n)/P(X_1^n)] + \mathbf{E}_P[\log P]}{n} \\ &= h(P) + D(P \parallel Q). \end{aligned}$$

Let us now extend AEP to biased distributions. First, however, we prove the existence of the limit in (6.33) for the so called *jointly* mixing processes.

(JMX) Let the measures P and Q be defined on the same probability space and let $\mathcal{F}_m^n$ be the σ-algebra generated by X_m^n. The measures P and Q are *weakly mixing* if there exit constants c_1, C_1, c_2, C_2 such that for any $m, g \geq 1$ and any two events $A \in \mathcal{F}_1^m$ and $B \in \mathcal{F}_{m+g}^\infty$ the following holds

$$c_1 P(A)P(B) \leq \quad P(AB) \quad \leq c_2 P(A)P(B), \tag{6.34}$$

$$C_1 Q(A)Q(B) \leq \quad Q(AB) \quad \leq C_2 Q(A)Q(B) \tag{6.35}$$

and $c_1, C_1 > 0$.

In passing, we should observe that the condition $c_1, C_1 > 0$ is quite restrictive. For example, not all Markov chains belong to (JMX), but Markov chains with *positive* transition probabilities are jointly mixing.

It is easy to see that if both Q and P are mixing, then $Q(X_1^n)/P(X_1^n)$ is a mixing sequence. Therefore

$$\log Q(X_1^n)/P(X_1^n) \leq \log C_2/c_1 + \log Q(X_1^m)/P(X_1^m) + \log Q(X_m^n)/P(X_m^n),$$

and $\log Q/P$ is subadditive. By Theorem 5.8 we have

$$-\lim_{n\to\infty} \frac{\log Q(X_1^n)/P(X_1^n)}{n} = D(P \parallel Q) \quad \text{(a.s)},$$

and hence

$$-\lim_{n\to\infty} \frac{\log Q(X_1^n)]}{n} = h(P) + D(P \parallel Q) \quad \text{(a.s.)}. \tag{6.36}$$

This leads to the following generalization of AEP (details of the proof are left as an exercise).

Theorem 6.10 (AEP for Biased Distributions) *Let $\{X_k\}$ be generated according to P, which together with another distribution Q, is jointly mixing as defined in (JMX). Then the state space $\mathcal{A}^n$ can be partitioned into two subsets $\mathcal{B}_n^\varepsilon$ ("bad set") and $\mathcal{G}_n^\varepsilon(Q)$ ("good set") such that for given $\varepsilon > 0$ there is N_ε so that for $n \geq N_\varepsilon$ we have*

$$2^{-n(h(P)+D(P\|Q))(1+\varepsilon)} \leq \quad Q(x_1^n) \quad \leq 2^{-n(h(P)+D(P\|Q))(1-\varepsilon)}, \qquad x_1^n \in \mathcal{G}_n^\varepsilon(Q), \tag{6.37}$$

$$Q(\mathcal{B}_n^\varepsilon) \quad \leq \varepsilon. \tag{6.38}$$

Moreover,

$$(1-\varepsilon)2^{n(h(P)+D(P\|Q))(1-\varepsilon)} \leq |\mathcal{G}_n^\varepsilon(Q)| \leq 2^{n(h(P)+D(P\|Q))(1+\varepsilon)} \tag{6.39}$$

for sufficiently large n.

We should point out that the above AEP for biased distributions cannot be extended to general stationary ergodic processes (cf. [146, 147]). In fact, as we mentioned above even the limit in the definition (6.33) may not exist in a general case. However, the following general result is known (the reader is asked to prove it in Exercise 9; cf. also Gray [59, 148]).

Lemma 6.11 *If P is a stationary process and P^k represents the measure of a k order stationary Markov process, then the limit in (6.33) exits and*

$$D(P \parallel P^k) = -h_P(X) - \mathbf{E}_P[\ln P^k(X_k|X_0^{k-1})] = -h_P(X) - h_{P^k}(X_k|X_0^{k-1}).$$

6.3.4 Lossy Generalizations of AEP

Let $w \in \mathcal{A}^k$ be a word of length k. Often, one asks for the probability that a prefix of a sequence generated by a stationary source is equal to w. Using AEP, we can asymptotically estimate this probability (e.g., by checking whether w is a typical sequence or not). But, in some situations (e.g., imagine a continuous signal) another question is more appropriate, namely, how well w approximates a k-prefix of a sequence. To answer such a question a lossy or approximate extension of AEP is needed. We discuss it next.

In order to define more precisely the notion of *approximate*, we need to introduce a *distortion measure* or a *distance*. In information theory, one defines the *distortion* function between symbols as

$$d: \ \mathcal{A} \times \hat{\mathcal{A}} \to \mathbb{R}^+.$$

Then, the distortion–per-symbol becomes

$$d_n(x_1^n, \hat{x}_1^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i). \tag{6.40}$$

For example, for the Hamming distance one sets $d(x_i, \hat{x}_i) = 0$ if $x_i = \hat{x}_i$ and $d(x_i, \hat{x}_i) = 1$ otherwise. Observe that in this case, the distortion $d_n(x_1^n, \hat{x}_1^n)$ represents the fraction of mismatches between x_1^n and $\hat{x}_1^n$. Another distortion measure is useful for image compression; e.g., the *square error* distortion defined as $d(x_i, \hat{x}_i) = (x_i - \hat{x}_i)^2$.

Let us now fix $D > 0$ and define a *ball* of radius D and center $c_1^k \in \mathcal{A}^k$ as

$$B_D(c_1^k) = \{x_1^k : \ d_k(x_1^k, c_1^k) \le D\}.$$

In particular, $B_0(x_1^k) = x_1^k$. If a random sequence X_1^k is generated by source P, we shall write $P(B_D(X_1^k)) = \sum_{\hat{X}_1^k \in \mathcal{A}^k} \mathsf{Pr}\{d_k(X_1^k, \hat{X}_1^k)\}$. With these definitions in mind, we can now generalize Rényi entropy.

Definition 6.12 (Generalized Rènyi's Entropy of order b) *For any $-\infty \le b \le \infty$*

$$\begin{aligned} r_b(D) &= -\lim_{k\to\infty} \frac{\log \mathbf{E}[P^b(B_D(X_1^k))]}{bk} \qquad (6.41)\\ &= -\lim_{k\to\infty} \frac{\log\left(\sum_{w_1^k \in \mathcal{A}^k} P^b(B_D(w_1^k))P(w_1^k)\right)}{bk}, \end{aligned}$$

where for $b = 0$ we understand $r_0(D) = \lim_{b\to 0} r_b(D)$, that is,

$$r_0(D) = -\lim_{k\to\infty} \frac{\mathbf{E}[\log P(B_D(X_1^k))]}{k}, \tag{6.42}$$

provided the above limits exist.

In theorem below, we first establish the existence of the generalized Rényi entropies for mixing processes. Then, we prove a generalization of the Shannon-McMillian-Breiman theorem which will imply AEP for approximate/lossy case.

Theorem 6.13 (AEP for Approximate/Lossy Case) (i) *Let $\{X_k\}$ be a weakly mixing process with $c_1 > 0$ (cf. (6.34)). Then the generalized b-th order entropy $r_b(D)$ is well defined (i.e., the limit in (6.41) exists) for any $-\infty \le b \le \infty$.*
(ii) *The generalized Shannon entropy $r_0(D)$ satisfies*

$$r_0(D) = -\lim_{k\to\infty} \frac{\log P(B_D(X_1^k))}{k} \quad \text{(a.s.)} \ . \tag{6.43}$$

(iii) *For a weakly mixing sequence $\{X_k\}$ and fixed $\varepsilon > 0$, the state space $\mathcal{A}^n$ can be partitioned into two subsets $\mathcal{B}_n^\varepsilon$ ("bad set") and $\mathcal{G}_n^\varepsilon$ ("good set") such that there exists N_ε so that for all $n > N_\varepsilon$ we have*

$$\begin{aligned} 2^{-n(r_0(D)+\varepsilon)} \le \quad & P(B_D(x_1^n)) && \le 2^{-n(r_0(D)-\varepsilon)}, && (6.44)\\ & P(\mathcal{B}_n^\varepsilon) && \le \varepsilon, && (6.45)\\ (1-\varepsilon)2^{n(r_0(D)-\varepsilon)} \le \quad & |\mathcal{G}_n^\varepsilon|. &&&& (6.46) \end{aligned}$$

Proof. We only consider $0 \le b < \infty$ while $b < 0$ case is left as an exercise. For (i) it suffices to show that for some constant $c > 0$

$$\mathbf{E}[P^b(B_D(w_1^{n+m}))] \ge c\mathbf{E}[P^b(B_D(w_1^n))]\mathbf{E}[P^b(B_D(w_1^m))]. \tag{6.47}$$

Provided (6.47) is true we simply use Fekete's Theorem 5.1 to establish our claim. In the course of proving (6.47) we shall see that $P^b(B_D(X_1^{n+m})) \ge cP^b(B_D(X_1^n))P^b(B_D(X_1^m))$ which by the *Subadditive Ergodic Theorem* 5.8 will imply (6.43) of part (ii).

We now wrestle with (6.47). Observe that for any string $w_1^{n+m} \in \mathcal{A}^{n+m}$ of length $n+m$ we have

$$\begin{aligned} P^b(B_D(w_1^{n+m})) &= \left(\sum_{z_1^{n+m}\in B_D(w_1^{n+m})} P(z_1^{n+m})\right)^b \ge c\left(\sum_{z_1^{n+m}\in B_D(w_1^{n+m})} P(z_1^n)P(z_1^m)\right)^b \\ &\ge c\left(\sum_{z_1^n\in B_D(w_1^n)} P(z_1^n)\right)^b \left(\sum_{z_1^m\in B_D(w_1^m)} P(z_1^m)\right)^b \\ &= cP^b(B_D(w_1^n))P^b(B_D(w_1^m)) \end{aligned}$$

where c is an universal constant whose value can change from line to line but always $c > 0$. The first inequality of the above follows from the weak mixing condition (6.34), and the second one is a simple consequence of the following easy to establish property of the distortion–per–symbol symbol

$$d_{n+m}(x_1^{n+m}, \hat{x}_1^{n+m}) = \frac{n}{n+m} d_n(x_1^n, \hat{x}_1^n) + \frac{m}{n+m} d_m(x_1^m, \hat{x}_1^m).$$

In fact, for the proof we need inequality in the above which is satisfied by any distortion measure having subadditivity property. To complete the proof we use again the mixing condition to get

$$\begin{aligned}
\mathbf{E}[P^b(B_D(w_1^{n+m}))] &= \sum_{w_1^{n+m} \in \mathcal{A}^{n+m}} P^b(B_D(w_1^{n+m}))P(w_1^{n+m}) \\
&\geq c \sum_{w_n \in \mathcal{A}^n} P^b(B_D(w_1^n))P(w_1^n) \sum_{w_1^m \in \mathcal{A}^m} P^b(B_D(w_1^m))P(w_1^m) \\
&= c\mathbf{E}[P^b(B_D(w_1^n))]\mathbf{E}[P^b(B_D(w_1^m))].
\end{aligned}$$

Part (i) and (ii) are proved by Fekete's theorem and the Superadditive Ergodic Theorem (cf. Chapter 5).

Part (iii) is a direct consequence of Part (ii). For example, to prove (6.46) we proceed as follows

$$1 - \varepsilon \leq \sum_{x_1^n \in \mathcal{G}_n^\varepsilon} P(x_1^n) \leq \sum_{x_1^n \in \mathcal{G}_n^\varepsilon} P(B_D(x_1^n)) \leq |\mathcal{G}_n^\varepsilon| 2^{-n(r_0(D)-\varepsilon)}.$$

This completes the proof. ■

The evaluation of the generalized R'enyi's entropies is harder than expected. Rényi's entropies for memoryless sources are computed in the lemma below. There are no simple formulas for the generalized Rényi's entropies for Markovian sources, however, some useful relationships may be found in Yang and Kieffer [347].

Lemma 6.14 (Łuczak and Szpankowski 1997) *Assume a binary alphabet* $\mathcal{A} = \{0, 1\}$ *with symbol occurrence* p *and* $q = 1-p$. *Define* $h(D, x) = (1-D)\log((1-D)/x) + D\log(D/(1-x))$ *for* $0 < D, x < 1$. *Then*
(i) *Let* $p_{\min} = \min\{p, q\}$ *and* $p_{\max} = \max\{p, q\}$, *then*

$$r_{-\infty}(D) = \begin{cases} h(D, p_{\min}) & \textit{for } D \leq p_{\max} \\ 0 & \textit{for } D > p_{\max} \end{cases}.$$

and

$$r_{\infty}(D) = \begin{cases} h(D, p_{\max}) & \textit{for } D \leq p_{\min} \\ 0 & \textit{for } D > p_{\min} \end{cases}.$$

In addition, $r_{-\infty}(D)$ *and* $r_{\infty}(D)$ *are convex functions of* D.

(ii) *If* $p = q = 1/2$ *then, for every* $-\infty \le b \le \infty$ *and* $D \le p_{\min}$*, we have* $r_b(D) = h(D, 1/2)$.

(iii) *Let* $p \neq q$ *and* $-\infty < b < \infty$. *Then,* $r_b(D) = 0$ *whenever* $D > 2pq$*, while for* $0 \le D \le 2pq$ *and* $b \neq 0$

$$\begin{aligned} r_b(D) &= (1/b) \min_{0 \le x \le 1} \Big\{ x \log(x/p) + (1-x) \log((1-x)/q) - b\Big(D \log(p/q) \\ &+ x \log(px) + (1-x) \log(q(1-x)) - x \log(x - F(x)) \\ &- D \log(D - F(x)) - (1 - x - D) \log(1 - x - D + F(x))\Big)\Big\} , \end{aligned} \tag{6.48}$$

where $F(x)$ *is defined as*

$$F(x) = \frac{x+D}{2} + \frac{\sqrt{(p^2 + (x+D)(q-p))^2 + 4xq^2 D(p-q)} - p^2}{2(p-q)} . \tag{6.49}$$

In particular, we have

$$r_1(D) = \begin{cases} h(D, P) & \text{for } D \le 1 - P = 2pq \\ 0 & \text{for } D > 1 - P = 2pq \end{cases}$$

where $P = p^2 + q^2$. *The function* $r_1(D)$ *is convex with respect to* D.

(iv) *If* $p \neq q$ *then* $r_0(D) = 0$ *for* $D > 2pq$*, and for* $0 \le D \le 2pq$

$$\begin{aligned} r_0(D) &= -\Big(D \log(p/q) + 2p \log p + 2q \log q - p \log(p - F(p)) \\ &\quad - D \log(D - F(p)) - (q - D) \log(q - D + F(p))\Big) , \end{aligned} \tag{6.50}$$

where F *is the function defined by (6.49). In addition,* $r_0(D)$ *is convex with respect to* D.

Proof. We only prove part (iii). The reader is referred to [248] for the whole proof, or even better, is encouraged to provide the missing parts.

Let $p \neq q$ and $-\infty < b < \infty$, $b \neq 0$. From the definition of the expectation, for $\mathbf{E}[P^b(B_D(X_1^k))]$ we have

$$\mathbf{E}[P^b(B_D(X_1^k))] = \sum_{i=0}^{k} \binom{k}{i} p^i q^{k-i} \left(\sum_{j=0}^{\lfloor Dk \rfloor} \sum_{\ell=\max\{0, i+j-k\}}^{\min\{i,j\}} \binom{i}{\ell} \binom{k-i}{j-\ell} p^{i+j-2\ell} q^{k-i-j+2\ell} \right)^b ,$$

where i counts the number of ones in X_1^k, j stays for the overall number of mismatches and ℓ is the number of disagreements among ones. Let us look first at the sum

$$\begin{aligned} s(k,i) &= \sum_{j=0}^{\lfloor Dk \rfloor} \sum_{\ell=\max\{0, i+j-k\}}^{\min\{i,j\}} \binom{i}{\ell} \binom{k-i}{j-\ell} p^{i+j-2\ell} q^{k-i-j+2\ell} \\ &= \sum_{j=0}^{\lfloor Dk \rfloor} \sum_{\ell=\max\{0, i+j-k\}}^{\min\{i,j\}} r(k,i,j,\ell) . \end{aligned}$$

Since there are at most k^2 terms in the sum, certainly we have

$$\max_{j,\ell} r(k,i,j,\ell) \le s(k,i) \le k^2 \max_{j,\ell} r(k,i,j,\ell) \ .$$

(Note that all ratios which grows polynomially with k will disappear if we divide the logarithm of $s(k,i)$ by k, thus they will not affect the value of $r_b(D)$.) Similarly,

$$\max_i \binom{k}{i} p^i q^{k-i} s^b(k,i) \le \mathbf{E}[P^b(B_D(X_1^k))] \le k \max_i \binom{k}{i} p^i q^{k-i} s^b(k,i) \ .$$

By Stirling's formula (cf. (8.31) for every $x \in (0,1)$ we have

$$\binom{k}{xk} \sim \left(\frac{1}{(1-x)^{1-x}x^x}\right)^k \ . \tag{6.51}$$

We use above to estimate the binomial coefficients and set $i = xk$, $j = yk$, $\ell = zk$ to arrive at the following asymptotic formula for $\mathbf{E}[P^b(B_D(X_1^k))]$

$$\max_{0\le x\le 1} \left\{ \left(\frac{p^x q^{1-x}}{x^x(1-x)^{(1-x)}} \left(\max_{A(x,D)} \left\{ \frac{p^{x+y-2z} q^{1-x-y+2z} x^x (1-x)^{1-x}}{z^z (x-z)^{x-z} (y-z)^{y-z} (1-x-y+z)^{1-x-y+z}} \right\} \right)^b \right)^k \right\},$$

where $A(x,D) \subset \mathbb{R}^2$ is defined as

$$A(x,D) = \{(y,z) \in \mathbb{R}^2 : 0 \le y \le D,\ \max\{0, x+y-1\} \le z \le \min\{x,y\}\} \ .$$

Consequently, we get the following formula

$$\begin{aligned} r_b(D) &= (1/b) \min_{0\le x\le 1} \Big\{ x\log(x/p) + (1-x)\log((1-x)/q) \\ &- b \max_{A(x,D)} \Big\{ (x+y-2z)\log(p/q) + \log q + x\log x \\ &+ (1-x)\log(1-x) - z\log z - (x-z)\log(x-z) \\ &- (y-z)\log(y-z) - (1-x-y+z)\log(1-x-y+z) \Big\}\Big\} \ . \end{aligned} \tag{6.52}$$

A simple algebra reveals that (6.48) follows from (6.52). Indeed, let us assume first that $D > 2pq$. Then, the value of the maximum in (6.52) is 0 and is achieved for $y - z = (1-x)p$ and $z = qx$. Furthermore, the first two terms vanish for $x = p$. Hence, for such a D, we have $r_b(D) = 0$ for every $-\infty < b < \infty$.

If $D \le 2pq$, then the function which appears under the maximum in (6.52) grows with y, so we must put $y = D$. Furthermore, easy calculations show that to choose an optimal value of z one must solve the equation

$$(p-q)z^2 + (p^2 + (q-p)(x+D))z - xDq^2 = 0 \ .$$

Thus, we should set $z = F(x)$, where F is defined by (6.49), and (6.48) follows.

Finally, let us notice that the case when $b = 0$ can be easily deduced from (6.48). Indeed, for $b \to 0$ the sum of the first two terms $x\log(x/p)$ and $(1-x)\log((1-x)/q)$ must vanish, which is possible only for $x = p$. Thus, (6.50) follows. ∎

We must point out, however, that one can envision another AEP for lossy case that is actually quite useful in the rate distortion theory discussed in Section 6.4.3. We briefly discuss it here.

We start with a definition of *distortion typical set* denoted as $\mathcal{G}^{\varepsilon}_{d,n}$.

Definition 6.15 *The distortion ε-typical set $\mathcal{G}^{\varepsilon}_{d,n}$ of sequences $(x_1^n, \hat{x}_1^n)$ with respect to the probability measure $P = \Pr\{\cdot,\cdot\}$ is the set of all n–sequences such that for given $\varepsilon > 0$*

$$\begin{aligned}
\mathcal{G}^{\varepsilon}_{d,n} = \Big\{(x_1^n, \hat{x}_1^n) :\ & 2^{-nh(X)(1+\varepsilon)} \le & P(x_1^n) & \le 2^{-nh(X)(1-\varepsilon)}, \\
& 2^{-nh(\hat{X})(1+\varepsilon)} \le & P(\hat{x}_1^n) & \le 2^{-nh(\hat{X})(1-\varepsilon)}, \\
& 2^{-n(h(X,Y)+\varepsilon)} \le & P(x_1^n, \hat{x}_1^n) & \le 2^{-n(h(X,Y)-\varepsilon)} \\
& \mathbf{E}_{X_1^n}[d(X_1^n, \hat{X}_1^n)] - \varepsilon \le & d(x_1^n, \hat{x}_1^n) & \le \mathbf{E}_{X_1^n}[d(X_1^n, \hat{X}_1^n)] + \varepsilon\Big\}.
\end{aligned}$$

The following extension of joint–AEP can be stated.

Theorem 6.16 (AEP with Distortion) *Let $\{X_k\}$, $\{\hat{X}_k\}$ and $\{X_k, \hat{X}_k\}$ be mixing processes. Then for given $\varepsilon > 0$ there is N_ε so that for $n \ge N_\varepsilon$ we have*

$$\Pr\{(X_1^n, \hat{X}_1^n) \in \mathcal{G}^{\varepsilon}_{d,n}\} \ge 1 - \varepsilon, \tag{6.53}$$

while for $(x_1^n, \hat{x}_1^n) \in \mathcal{G}^{\varepsilon}_{n,d}$ we obtain

$$P(\hat{x}_1^n | x_1^n) 2^{-n(I(X;\hat{X})-3\varepsilon)} \le P(x_1^n) \le P(\hat{x}_1^n | x_1^n) 2^{-n(I(X;\hat{X})+3\varepsilon)} \tag{6.54}$$

for sufficiently large n.

Proof. We only prove (6.54) since (6.53) follows from the subadditivity of $\log P(X_1^n)$, $\log P(\hat{X}_1^n)$, $\log P(X_1^n, \hat{X}_1^n)$ and $d(X_1^n, \hat{X}_1^n)$ for mixing processes. Observe that for $(x_1^n, \hat{x}_1^n) \in \mathcal{G}^{\varepsilon}_{d,n}$

$$\begin{aligned}
P(\hat{x}_1^n | x_1^n) &= P(\hat{x}_1^n) \frac{P(x_1^n, \hat{x}_1^n)}{P(x_1^n) P(\hat{x}_1^n)} \\
&\le P(\hat{x}_1^n) \frac{2^{-n(h(X,\hat{X})-\varepsilon)}}{2^{-n(h(X)+\varepsilon)} 2^{-n(h(\hat{X})+\varepsilon)}} = P(\hat{x}_1^n) 2^{n(I(X,\hat{X})-3\varepsilon)}.
\end{aligned}$$

The lower bound is proved in a similar fashion. ∎

6.4 Three Theorems of Shannon

In 1948 Claude Shannon published his seminal paper *A Mathematical Theory of Communication* where he formulated the three fundamental results of information theory. Namely, that of *universal source coding* theorem, *channel coding* theorem and *rate distortion* theorem. In this section, we discuss them and provide simple proofs. We shall follow the presentation of Cover and Thomas [59]. In Shannon's spirit, we shall deal mostly with memoryless sources to present succinctly the main ideas. Throughout this section, we use a new technique, yet not seen on pages of this book, called *random coding.* This method selects codes at random to establish the existence of a good code that achieves the best possible code rate.

6.4.1 Source Coding Theorem

A *source code* $\mathscr{C}$ for a random variable X is a mapping

$$\phi: \ \mathcal{A} \rightarrow \{0,1\}^*,$$

that we also write as $\mathscr{C} = \phi(X)$. Such a code is one-to-one if

$$x_i \neq x_j \in \mathcal{A} \quad \Rightarrow \quad \mathscr{C}(x_i) \neq \mathscr{C}(x_j).$$

The code $\mathscr{C}_n$ for a sequence $X_1, \ldots, X_n$ of random variables is an *extension* of $\mathscr{C}$ and it is defined by a mapping

$$\phi_n: \ \mathcal{A}^n \rightarrow \{0,1\}^*.$$

Such a code is:

- *uniquely decodable* if it is one-to-one,
- *prefix code* or *instantaneous code* if no codeword is a prefix of another codeword.

It turns out that lengths of prefix codewords cannot be completely arbitrary and must satisfy certain constraint known as *Kraft's inequality.* We prove it below for the simplest case. Its extensions can be found in [59].

Theorem 6.17 (Kraft's Inequality) *For any prefix code (over a binary alphabet), the codeword lengths $\ell_1, \ell_2, \ldots, \ell_m$ satisfy the inequality*

$$\sum_{i=1}^{m} 2^{-\ell_i} \leq 1. \tag{6.55}$$

Conversely, if codeword lengths satisfy this inequality, then one can build a prefix code.

Proof. This is an easy exercise on trees. Observe that one can associate a binary tree with a code by assigning, say, "0" to left branch and "1" to the right branch. A path from the root to a node creates a code. Prefix codes guarantee that *none* of the codeword is assigned to an *internal node*, that is, all codewords are terminal nodes (or *leaves* of the tree). Let $\ell_{\max}$ be the maximum codeword length. Observe that at level $\ell_{\max}$ some nodes are codewords, some are descendants of codewords, and some are neither. Since the number of descendants at level $\ell_{\max}$ of a codeword located at level ℓ_i is $2^{\ell_{\max}-\ell_i}$ (if root is at level 0 of the tree, then at level ℓ there are no more than 2^ℓ nodes), we obtain

$$\sum_{i=1}^{m} 2^{\ell_{\max}-\ell_i} \leq 2^{\ell_{\max}},$$

which is the desired inequality. The converse part is easy to prove, and we omit it here. ∎

The average length $\bar{\ell}_n(\mathscr{C}_n)$ of the source code $\mathscr{C}_n$ is defined as

$$\bar{\ell}_n(\mathscr{C}_n) = \sum_{x_1^n \in \mathcal{A}^n} \ell(x_1^n) P(x_1^n),$$

where P is the source distribution, that is, $P(x_1^n) = \mathsf{Pr}\{X_1^n = x_1^n\}$. We cannot make $\bar{\ell}(\mathscr{C}_n)$ arbitrarily small, as the theorem below indicates. In other words, there is a limit how much we can compress on the average.

Theorem 6.18 *For any source code $\mathscr{C}_n$ satisfying the Kraft inequality, the average code length $\bar{\ell}(\mathscr{C}_n)$ must be bigger equal than the entropy $h(X_1^n)$ of the source, that is, $\bar{\ell}(\mathscr{C}_n) \geq h_n(X_1^n)$.*

Proof. Let $K = \sum_{x_1^n} 2^{-\ell(x_1^n)} \leq 1$. Then

$$\begin{aligned}
\bar{\ell}(\mathscr{C}_n) - h_n(X_1^n) &= \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n)\ell(x_1^n) + \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \log P(x_1^n) \\
&= \sum_{x_1^n \in \mathcal{A}^n} P(x_1^n) \log \frac{P(x_1^n)}{2^{-\ell(x_1^n)}/K} - \log K \\
&\geq 0
\end{aligned}$$

since the first term is a divergence and cannot be negative by Theorem 6.3, while $K \leq 1$ by Kraft's inequality. ∎

The quantity

$$\bar{r}_n(\mathscr{C}_n) = \bar{\ell}(\mathscr{C}_n) - h_n(X_1^n)$$

is known as the average *code redundancy.* We just proved that for prefix codes it cannot be negative. We also observe that the average bits per symbol $\ell(C_n)/n$ is bounded from the below by the entropy rate of the source. In Theorem 6.19 below we show that this is true even if we allow errors on the decoding side. We shall return to the code redundancy in Section 8.7.2 and Section 9.4.2 where we use analytic tools to evaluate precisely the code redundancy for some classes of processes.

To formulate a *universal source coding problem*, we start by introducing the bit rate R of a code $\mathscr{C}_n$. Hereafter, we only deal with extended codes for $X_1, \ldots, X_n$ generated by a source with the underlying probability measure P. Let M be the number of codewords for $\mathscr{C}_n$. The *bit rate* of such a code is defined as

$$R = \frac{\log M}{n}.$$

We often write $M = 2^{nR}$ to mean $M = \lceil 2^{nR} \rceil$, and the corresponding code we denote as $(2^{nR}, n)$. Observe that R represents the number of bits per source symbol, and it can be viewed as the reciprocal of the compression ratio. We expect $R < 1$ for good compression codes.

The *fixed rate block code* $\mathscr{C}_n$ of fixed rate R for $X_1, \ldots, X_n$ is defined by two mappings:

- encoder
$$\phi_n : \mathcal{A}^n \to \{1, 2, \ldots, 2^{nR}\};$$
- decoder
$$\psi_n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{A}^n.$$
where we explicitly set $M = \lceil 2^{nR} \rceil$.

Let $\hat{X}_1^n = \psi_n(\phi_n(X_1^n)) = \psi_n(\mathscr{C}_n)$, and we allow that $\hat{X}_1^n \neq X_1^n$. In this case, we define the probability of error P_E with respect to the source probability P as

$$P_E := \mathsf{Pr}\{X_1^n \neq \hat{X}_1^n\}.$$

The problem of **universal source coding** is to find the smallest bit rate R (or biggest compression ratio) such that $P_E < \varepsilon$ for given $\varepsilon > 0$ irrespectively of the (usually unknown) distribution P.

The first Shannon theorem states that one can build a universal source code as long as the bit rate R is bigger than the entropy $h(P)$ of the source. In other words, on average the best bit rate is equal to the entropy of the source.

Theorem 6.19 (Shannon 1948) (i) *There exists a sequence of* $(2^{nR}, n)$ *universal source codes with rate* $R > h(P)$ *such that for a given* $\varepsilon > 0$ *the probability of error* $P_E < \varepsilon$.
(ii) *Conversely, for all universal codes with* $P_E \to 0$ *the bit rate* $R > h(P)$.

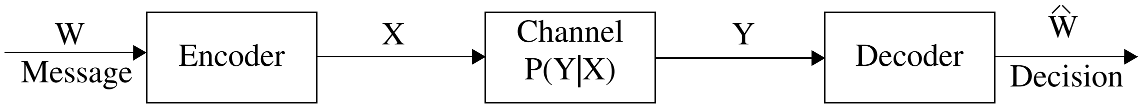

Figure 6.1: A communication system.

Proof. We prove the achievability part of the theorem by constructing a *a random code* that satisfies the conditions of the theorem. For given $\varepsilon > 0$, let $\mathcal{G}_n^\varepsilon$ be a set of $\mathcal{A}^n$ such that $P(\mathcal{G}_n^\varepsilon) \geq 1 - \varepsilon$. We assign codes that are reproduced correctly to elements of the set $\mathcal{G}_n^\varepsilon$. Clearly, then the error P_E is bounded by ε. By AEP, we know that $\mathcal{G}_n^\varepsilon$ is the set of P-typical sequences and its cardinality can be bounded as $(1-\varepsilon)2^{n(h(P)-\varepsilon)} \leq |\mathcal{G}_n^\varepsilon| \leq 2^{n(h(P)+\varepsilon)}$. Thus, $R = \log|\mathcal{G}_n^\varepsilon|/n$ is bounded between $h(P) - \varepsilon$ and $h(P) + \varepsilon$, for sufficiently large n. This proves the first part of the theorem.

To prove the converse, we establish that if $R < h(P)$, then $P_E > 1 - \varepsilon$ for a given $0 < \varepsilon < 1$. Let $\mathscr{C}'$ be the subset of all codewords $\mathscr{C}$ that are encoded without an error. Observe that $|\mathscr{C}'| = 2^{nR}$. Since we assume that $R < h(P)$ there exists $\delta > 0$ such that

$$|\mathscr{C}'| < 2^{n(h-2\delta)}.$$

We now use AEP with $2\delta = \varepsilon$ to obtain (we write below $\mathcal{B}_n^\varepsilon = \mathcal{A}^n - \mathcal{G}_n^\varepsilon$)

$$\begin{aligned} 1 - P_E &= \sum_{x_1^n \in \mathscr{C}'} P(x_1^n) \\ &= \sum_{x_1^n \in \mathscr{C}' \cap \mathcal{B}_n^{\varepsilon/2}} P(x_1^n) + \sum_{x_1^n \in \mathscr{C}' \cap \mathcal{G}_n^{\varepsilon/2}} P(x_1^n) \\ &\leq \varepsilon/2 + 2^{n(h-2\delta)} 2^{-n(h-\delta)} < \varepsilon, \end{aligned}$$

for sufficiently large n. This proves the theorem. ∎

6.4.2 Channel Coding Theorem

Let A relies a message W drawn from the index set of size M to B who receives message $\hat{W}$ and must guess W. Such a communication is performed by sending the signal $X(W)$ that is received as a signal Y depending on X, hence on the message W. This communication process is illustrated in Figure 6.1 in which the communication medium is called the *channel* and is characterized by the conditional probability $P(Y|X)$. As before, we shall write P for the probability measure characterizing the source and the channel. In fact, we should point out that to send M messages one needs a code of length $n \geq \log M$, so from now on we shall deal with X_1^n and Y_1^n for some $n \geq \log M$.

Throughout this section, we consider channel $P(Y_1^n|X_1^n)$ satisfying the following three properties:

$$
\begin{aligned}
P(y_1^n|x_1^n) &= \prod_{i=1}^{n} P(y_i|x_i), && (6.56)\\
P(y_k|x_1^k, y_1^{k-1}) &= P(y_k|x_k), \qquad k = 1, 2, \ldots, n, && (6.57)\\
P(x_k|x_1^{k-1}, y_1^{k-1}) &= P(x_k|x_1^{k-1}), \quad k = 1, 2, \ldots, n. && (6.58)
\end{aligned}
$$

Such a channel is called *discrete memoryless channel* (cf. (6.56)–6.57)) *without feedback* (cf. (6.58)).

For channel $P(Y_1^n|X_1^n)$, we define an (M, n) channel code with rate R such that $M = \lceil 2^{nR} \rceil$ as follows:

- encoding function
$$X_1^n : \{1, 2, \ldots, M\} \to \mathcal{A}^n$$
yielding *codewords* $X_1^n(1), \ldots, X_1^n(M)$ that constitutes the *codebook*;
- decoding function
$$\chi : \mathcal{B}^n \to \{1, 2, \ldots, M\},$$
where $\mathcal{B}$ is the alphabet of the output signal Y. Observe that the received message is $\hat{W} = \chi(Y_1^n)$.

The point to observe is that we consider a *noisy* channel that can alter the signal, so that the received signal is not necessary equal to the original one. To capture this situation, we introduce the (average) probability of error P_E defined as

$$
\begin{aligned}
P_E &= \mathsf{Pr}\{W \neq \hat{W}\} && (6.59)\\
&= \frac{1}{M} \sum_{i=1}^{M} \mathsf{Pr}\{\chi(Y_1^n) \neq i | W = i\} = \frac{1}{M} \sum_{i=1}^{M} \sum_{y_1^n : \chi(y_1^n) \neq i} P(y_1^n | x_1^n(i)).
\end{aligned}
$$

This definition implies that a message to be sent is selected uniformly from the index set $\{1, 2, \ldots, M\}$ where $M = \lceil 2^{nR} \rceil$.

The **channel coding problem** is to find the largest possible R (hence M) for which there exists a channel code $(2^{nR}, n)$ such that the probability of error $P_E \to 0$ as $n \to \infty$. In a sense, the channel coding is an opposing problem to the source coding. In the latter, we try to find the best compression, thus the smallest possible bit rate R, while in the former we want to send reliably as many messages as possible, hence making the code rate R as big as possible.

To formulate the main result of this section, we introduce the *channel capacity* defined as

$$C = \max_{P(X)}\{I(X;Y)\} \tag{6.60}$$

(In Exercises 12–13 we ask to compute the channel capacity for the binary symmetric channel and the binary erasure channel.) We shall prove that the channel capacity is the achievability rate of channel codes. To see that this intuitively, we observe that there are approximately $2^{nH(Y)}$ typical Y-sequences. We can divide this set into $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$ distinguishable sets of size $2^{nH(Y|X)}$ corresponding to different input sequences (cf. Theorem 6.9). Hence, we can send at most $2^{nI(X;Y)}$ distinguishable sequences of length n. We make below this statement more rigorous. In our presentation, we shall follow quite closely the excellent exposition of Cover and Thomas [59].

To follow this idea, we first derive the *converse* channel coding theorem by showing that if $R > C$ then $P_E > 1-\varepsilon$ for sufficiently large n. Then, we formulate and complete the proof of the Shannon coding theorem.

Our goal now is to show that

$$P_E \geq 1 - \frac{C}{R} - \frac{1}{nR} \tag{6.61}$$

which would imply that if $R > C$, then the probability of error is bounded away from zero for sufficiently large n. Let us now adopt a simplifying (but realistic) assumption that all the codewords are distinct, hence $nR = h(W) = h(X_1^n(W))$ (since messages are chosen uniformly). This implies

$$h(X_1^n|Y_1^n) = h(X_1^n) - I(X_1^n;Y_1^n) \geq n(R-C), \tag{6.62}$$

where we used the fact that $I(X_1^n;Y_1^n) \leq nC$. To see the latter is true, we write the following chain of relationships that follow directly from the definition of discrete memoryless channel (6.57) and (6.16)

$$\begin{aligned}
I(X_1^n;Y_1^n) &= h(Y_1^n) - h(Y_1^n|X_1^n) = h(Y_1^n) - \sum_{i=1}^{n} h(Y_i|Y_1,\ldots,Y_{i-1},X_1^n) \\
&= h(Y_1^n) - \sum_{i=1}^{n} h(Y_i|X_i) \leq \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Y_i|X_i) \\
&= nI(X_i,Y_i) \leq nC.
\end{aligned}$$

In order to complete the proof of (6.61), we must find an upper bound on $h(X_1^n|Y_1^n)$ in (6.62). We start with an observation that if $P_E = 0$, then $h(X_1^n|Y_1^n) = h(W|Y_1^n) = 0$, since in such a situation W can be completely determined from Y_1^n (cf. Exercise 4). In the case $P_E > 0$, the conditional entropy $h(W|Y_1^n)$, hence also $h(X_1^n|Y_1^n) \leq h(W|Y_1^n)$, are positive and should depend on the probability of the error. This fact is known as *Fano's inequality*, and we derive it next.

Let $E = I(\hat{W} \neq W)$, where, as always, I is the indicator function. Using $h(X, Y|Z) = h(X|Z) + h(X|Y, Z) = h(Y|Z) + h(Y|X, Z)$, and noting that $h(E|W, Y_1^n) = 0$, we have

$$h(E, W|Y_1^n) = h(W|Y_1^n) = h(E|Y_1^n) + h(W|E, Y_1^n) \leq 1 + h(W|E, Y_1^n)$$

where the inequality follows from $h(E|Y_1^n) \leq h(E) \leq 1$. We estimate $h(W|E, Y_1^n)$ as follows

$$\begin{aligned} h(W|E, Y_1^n) &= \mathsf{Pr}\{E = 0\} h(W|Y_1^n, E = 0) + \mathsf{Pr}\{E = 1\} h(W|Y_1^n, E = 1) \\ &\leq (1 - P_E)0 + P_E nR \leq nRP_E \end{aligned}$$

since $h(W|Y_1^n, E = 0) = 0$ due to the fact that in this case W is completely determined by Y. Combining everything we obtain the Fano inequality

$$h(X_1^n|Y_1^n) \leq h(W|Y_1^n) \leq 1 + nRP_E, \tag{6.63}$$

which, after substituting in (6.62), proves (6.61).

In view of the above, we know that the rate of a reliable transmission (i.e., $P_E < \varepsilon$) cannot exceed the capacity of the channel C. But, can we achieve this capacity. Shannon in 1948 was the first to use a number of new ideas (e.g., random coding technique) to prove this rather counterintuitive fact. We formulate it in the following channel coding theorem.

Theorem 6.20 (Shannon's Channel Coding Theorem) (i) *For every $\varepsilon > 0$ and rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $P_E < \varepsilon$ for sufficiently large n.*
(ii) *Conversely, any sequence of $(2^{nR}, n)$ codes with $P_E \to 0$ must have $R < C$.*

Proof. The converse part was already proved above, so we now concentrate on the achievability part. The proof of this part is a classic example of the *random coding technique.* Thus, let $R < C$ and we construct a random code that achieves P_E as small as desire for sufficiently large n. We independently generate a $(2^{nR}, n)$ code according to the distribution $P(x_1^n) = \prod_{i=1}^n P(x_i)$. The wth codeword is denoted as $x_1^n(w)$ for $1 \leq w \leq 2^{nR}$. Clearly, the probability of generating a particular code $\mathscr{C}$ is

$$P(\mathscr{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} P(x_i(w)).$$

We postulate that:

- The code $\mathscr{C}$ is revealed to both sender and receiver.

- A message W is chosen uniformly, that is,
$$\Pr\{W = w\} = 2^{-nR}$$
for $w = 1, 2, \dots, 2^{nR}$.
- The codeword of $x_1^n(w)$ sent over the channel is received as y_1^n where
$$P(y_1^n|x_1^n(w) = \prod_{i=1}^{n} P(y_i|x_i(w)).$$
- *Typical Decoding Procedure.* The receiver declares that $\hat{W}$ was sent if: (i) $(X_1^n(W), Y_1^n)$ is *jointly typical*; (ii) there is no other index $K \neq W$ such that $(X_1^n(K), Y_1^n)$is jointly typical, that is, $(X_1^n(K), Y_1^n) \notin \mathcal{G}_n^\varepsilon$ (cf. Joint AEP Theorem 6.9).

Under the above decoding rule, we now estimate the probability of error $P_E = \Pr\{W \neq \hat{W}\}$. Let $P_E^w(\mathscr{C}) = \Pr\{\hat{W} \neq w|W = w\}$ when $\mathscr{C}$ code was used. Clearly

$$P_E = \sum_{\mathscr{C}} P(\mathscr{C})\Pr\{\hat{W} \neq W|\mathscr{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathscr{C}} P(\mathscr{C}) P_E^w(\mathscr{C}).$$

But, by symmetry of the code construction the average probability of error averaged over all codes does not depend on the particular index w, hence we further assume $w = 1$. The above can be written as

$$P_E = \sum_{\mathscr{C}} P(\mathscr{C}) P_E^1(\mathscr{C}).$$

Let now

$$A_i = \{(X_1^n(i), Y_1^n) \in \mathcal{G}_n^\varepsilon, \quad i \in \{1, 2, \dots, 2^{nR}\}$$

where $\mathcal{G}_n^\varepsilon$ is the set of jointly (X,Y)–typical sequences. Clearly,

$$P_E = \Pr\{\bar{A}_1 \cup A_2 \cup \cdots \cup A_{2^{nR}}\} \leq \Pr\{\bar{A}_1\} + \sum_{i=2}^{2^{nR}} \Pr\{A_i\},$$

where $\bar{A}_1$ is the complement of A_i and represents the event that $X_1^n(1)$ and Y_1^n are not jointly typical. Observe that $X_1^n(1)$ and $X_1^n(i)$ are independent, as well as $X_1^n(i)$ and Y_1^n for $i \neq 1$ due to the way the code is constructed. Then, by Theorem 6.9

$$\begin{aligned}
P_E &\leq \Pr\{\bar{A}_1\} + \sum_{i=2}^{2^{nR}} \Pr\{A_i\} \\
&\leq \varepsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X,Y)-3\varepsilon)} \\
&\leq \varepsilon + 2^{3n\varepsilon} 2^{-n(I(X,Y)-R)} \\
&\leq 2\varepsilon
\end{aligned}$$

for sufficiently large n and $R < I(X;Y) - 3\varepsilon$. To complete the proof, we choose the source distribution P to be the distribution P that achieves the capacity C. Hence, the condition $R < I(X,Y)$ in the above can be replaced by $R < C$. ■

6.4.3 Rate Distortion Theorem

In some situations, during the decoding we cannot or do not want to recover exactly the message sent by the encoder. The former situation arises when dealing with continuous signals while the latter in a lossy data compression when one agrees to lose some information in order to get better compression. Rate distortion theory deals with such problems and can be succinctly described as follows: given a source distribution and a distortion measure, what is the minimum rate required to achieve a particular distortion.

To state our problem in precise mathematical terms, we introduce some definitions. Let $M = \lceil 2^{nR} \rceil$. A rate distortion code consists of an encoding mapping

$$\phi_n : \mathcal{A}^n \to \{1, 2, \ldots, 2^{nR}\},$$

and a decoding (reproduction) function

$$\psi_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{A}}^n,$$

where $\hat{\mathcal{A}}$ is the reproduction alphabet. We often write

$$\hat{X}_1^n = \psi_n(\phi_n(X_1^n)),$$

and refer to $\hat{X}_1^n$ as the *reproduction* vector, and $\hat{\mathcal{A}}^n$ as the reproduction space. The point to observe is that $X_1^n \neq \hat{X}_1^n$ and $\hat{X}_1^n$ approximates the source code X_1^n. We now introduce a distortion measure that was already mentioned in Section 6.3.4. Throughout this section, we only consider bounded distortion function, that is,

$$\max_{a \in \mathcal{A}, \hat{\in} \hat{\mathcal{A}}} \{d(a, \hat{a})\} = d_{\max} < \infty.$$

Roughly speaking, given $D > 0$, we call a rate distortion code $(2^{nR}, n, D)$ a D–semifaithful if either $d(X_1^n, \hat{X}_1^n) \leq D$ with high probability or $\mathbf{E}_{X_1^n}[d(X_1^n, \hat{X}_1^n)] \leq D$. We shall show below that these two definitions are asymptotically equivalent, at least for memoryless source which are assumed throughout this section. We shall follow here Csiszár and Körner [62], and Cover and Thomas [59].

Definition 6.21 (i) *A pair (R, D) is called ε–achievable rate at distortion level D for source $\{X_k\}$, if for every $\delta > 0$ and sufficiently large n there exists a sequence of rate distortion codes (ϕ_n, ψ_n) of rate less than $R + \delta$ such that*

$$\Pr\{d_n(X_1^n, \psi(\phi(X_1^n)) \leq D + \delta\} \geq 1 - \varepsilon. \tag{6.64}$$

(ii) *The real number R is an achievable rate at distortion D if it is ε–achievable for every* $0 < \varepsilon < 1$.
(iii) *The pair (R, D) (resp. (R_ε, D)) is called achievable rate-distortion pair (resp. ε–achievable rate-distortion) if it is achievable (resp. ε–achievable).*
(iv) *The infimum of achievable (resp. ε–achievable) rates at distortion D is called the rate distortion function and will be denoted $R(D)$ (resp. $R_\varepsilon(D)$). Clearly*

$$R(D) = \lim_{\varepsilon \to 0} R_\varepsilon(D). \tag{6.65}$$

Before, we go any further, we would like to observe that condition (6.64) can be replaced by a more global one, namely

$$\lim_{n\to\infty} \mathbf{E}_{X_1^n}[d(X_1^n, \psi_n(\phi_n(X_1^n)))] \le D. \tag{6.66}$$

without any harm for the final results. To see this, let us first assume (6.64) holds. We show that (6.66) follows. Indeed, let $\mathcal{S}_n$ the set of $\hat{x}_1^n$ for which (6.64) holds. Then

$$\begin{aligned}
\mathbf{E}[d(X_1^n, \hat{X}_1^n)] &= \sum_{x_1^n} P(x_1^n) d(x_1^n, \hat{x}_1^n) \\
&= \sum_{x_1^n \in \mathcal{S}_n} P(x_1^n) d(x_1^n, \hat{x}_1^n) + \sum_{x_1^n \notin \mathcal{S}_n} P(x_1^n) d(x_1^n, \hat{x}_1^n) \\
&\le D + \delta + \varepsilon d_{\max}.
\end{aligned}$$

Now assume (6.66) holds, and let $N(a, \hat{a}|x_1^n, \hat{x}_1^n)$ be the number of $a \in \mathcal{A}$ and $\hat{a} \in \hat{\mathcal{A}}$ occurring in x_1^n and $\hat{x}_1^n$, respectively. If $(X_k, \hat{X}_k)$ are i.i.d., then from the law of large numbers we infer for any $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\{|N(a, \hat{a}|X_1^n, \hat{X}_1^n)/n - P(a, \hat{a})| > \varepsilon\} = 0, \tag{6.67}$$

where, as always in this section, $P(\cdot, \cdot)$ is the probability measure of $(X_k, \hat{X}_k)$. (Observe that (6.67) for a much larger class of processes.) In view of this, we have:

$$\begin{aligned}
d(X_1^n, \psi_n(\phi_n(X_1^n)) &= \frac{1}{n} \sum_{a\in\mathcal{A}} \sum_{\hat{a}\in\hat{\mathcal{A}}} N(a, \hat{a}|X_1^n, \hat{X}_1^n) d(a, \hat{a}) \\
&\overset{(6.67)}{\le} \sum_{a\in\mathcal{A}} \sum_{\hat{a}\in\hat{\mathcal{A}}} P(a, \hat{a}) d(a, \hat{a}) + 2\varepsilon|\mathcal{A}||\hat{\mathcal{A}}| d_{\max} \quad \text{with prob. } 1 - \varepsilon \\
&\overset{(6.66)}{\le} D + \delta \qquad \text{with prob. } 1 - \varepsilon
\end{aligned}$$

where $\delta > 0$. This implies (6.64). Therefore, we shall use freely either (6.66) or (6.64) in the discussions below.

We now define the so called *information rate distortion function* $R_I(D)$, and later prove that for memoryless sources it is equal to the rate distortion function $R(D)$, that is, $R_I(D)$ represents the infimum of rates that achieve a particular distortion. This is the main result of this section which was proved by Shannon in 1959.

Definition 6.22 *The information rate distortion function $R_I(D)$ for source $\{X_k\}$ with distortion measure $d(x, \hat{x})$ is defined as*

$$R_I(D) = \min_{P(\hat{x}|x): \mathbf{E}[d(X,\hat{X})] \le D} I(X; \hat{X}) \tag{6.68}$$

where the minimization is over all conditional distributions $P(\hat{x}|x)$ for which the joint distribution $P(x, \hat{x}) = P(x)P(\hat{x}|x)$ satisfies

$$\mathbf{E}[d(X, \hat{X})] = \sum_{(x,\hat{x})} P(x)P(\hat{x}|x)d(x, \hat{x}) \le D.$$

In Exercise 14 we ask the reader to prove the following property of $R_I(D)$.

Lemma 6.23 *The information rate distortion function $R_I(D)$ is a non-increasing convex and continuous function of D.*

The main result of this section can be stated as follows.

Theorem 6.24 (Shannon Rate Distortion Theorem) *The rate distortion function $R(D)$ for memoryless sources X with distribution $P(\cdot)$ and bounded distortion function $d(x, \hat{x})$ is equal to the information rate distortion function $R_I(D)$, that is,*

$$R(D) = R_I(D) = \min_{P(\hat{x}|x): \mathbf{E}[d(X,\hat{X})] \le D} I(X; \hat{X}) \tag{6.69}$$

is the minimum achievable rate at distortion level D.

Proof. CONVERSE. We start with the converse part which is easier. We need to show that that for any $(2^{nR}, n, D)$ rate distortion code, the code rate R satisfies $R > R_I(D)$. We proceed now as in Cover and Thomas [59], taking into account the fact that $X_1, \ldots, X_n$ are independent:

$$\begin{aligned}
nR &\ge h(\hat{X}_1^n) \ge h(\hat{X}_1^n) - h(\hat{X}_1^n | X_1^n) = I(\hat{X}_1^n; X_1^n) \\
&\overset{\text{Theorem 6.3}}{=} h(X_1^n) - h(X_1^n|\hat{X}_1^n) = \sum_{i=1}^{n} h(X_i) - \sum_{i=1}^{n} h(X_i|\hat{X}_1^n, X_1, \ldots, X_i) \\
&\overset{(6.15)}{\ge} \sum_{i=1}^{n} h(X_i) - \sum_{i=1}^{n} h(X_i|\hat{X}_i) = \sum_{i=1}^{n} I(X_i; \hat{X}_i)
\end{aligned}$$

$$\begin{aligned}
&\overset{(6.68)}{\geq} \quad \sum_{i=1}^{n} R_I(\mathbf{E}[d(X_i,\hat{X}_i)]) \\
&\overset{\text{Lemma 6.23}}{\geq} \quad nR_I\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{E}[d(X_i,\hat{X}_i)]\right) = nR_I(\mathbf{E}[d(X_1^n,\hat{X}_1^n)]) \\
&\overset{\text{Lemma 6.23}}{\geq} \quad R_I(D).
\end{aligned}$$

ACHIEVABILITY. Let $P(\hat{x}|x)$ be the conditional probability that achieves equality in the definition of $R_I(D)$, that is, $I(X;\hat{X}) = R_I(D)$. We will prove that for any $\delta > 0$ there exists a rate distortion code $(2^{nR}, n, D)$ with rate $R > R_I(D)$ and distortion less equal than $D + \delta$. As in the channel coding theorem, we construct a random code as follows:

- *Generation.* Generate randomly a rate distortion codebook $\mathscr{C}$ consisting of 2^{nR} sequences $\hat{X}_1^n$ drawn i.i.d. according to $\prod_{i=1}^{n} P(\hat{x}_i)$. Index this codewords by $w \in \{1, 2, \ldots, 2^{nR}\}$.
- *Encoding.* Encode x_1^n by such w that $(X_1^n, \hat{X}_1^n(w) \in \mathcal{G}_{n,d}^{\varepsilon}$ where $\mathcal{G}_{n,d}^{\varepsilon}$ is the set of distortion typical sequences, as defined in Definition 6.15. If there is no such w, set $w = 1$. Observe that nR bits suffice to encode such an index w.
- *Decoding.* The reproduction sequence is $\hat{X}_1^n(w)$.

We need to estimate the expected distortion $\mathbf{E}_{X_1^n,\,\mathscr{C}}[d(X_1^n, \hat{X}_1^n)]$ over *the random choice of codebooks* $\mathscr{C}$. Observe that

$$\mathbf{E}_{X_1^n,\,\mathscr{C}}[d(X_1^n,\hat{X}_1^n)] \leq D + \varepsilon + P_E d_{\max}$$

where P_E is the total probability of all sequences x_1^n for which there is no codeword $\hat{X}_1^n(w)$ which is distortion typical with x_1^n. It suffices now to show that $P_E \to 0$ for sufficiently large n.

We estimate now P_E. Let $\mathcal{S}_n(\mathscr{C})$ be the set of source sequences x_1^n for which there is at least one codeword in $\mathscr{C}$ that is distortion typical with x_1^n. We obtain

$$\begin{aligned}
P_E &= \sum_{\mathscr{C}} P(\mathscr{C}) \sum_{x_1^n \notin \mathcal{S}_n(\mathscr{C})} P(x_1^n) = \sum_{x_1^n} P(x_1^n) \sum_{\mathscr{C}: x_1^n \notin \mathcal{S}_n(\mathscr{C})} P(\mathscr{C}) \\
&= \sum_{x_1^n} P(x_1^n)\left(1 - \Pr\{(x_1^n,\hat{X}_1^n) \in \mathcal{G}_{d,n}^{\varepsilon}\}\right)^{2^{nR}} \qquad (6.70)
\end{aligned}$$

where the last equation is a consequence of a random generation of 2^{nR} independent codewords. Using distortion AEP Theorem 6.16, we obtain

$$\Pr\{(x_1^n,\hat{X}_1^n) \in \mathcal{G}_{d,n}^{\varepsilon}\} = \sum_{\hat{x}_1^n : (x_1^n,\hat{x}_1^n) \in \mathcal{G}_{d,n}^{\varepsilon}} P(\hat{x}_1^n)$$

$$\geq \quad 2^{-n(I(X;\hat{X})+3\varepsilon)} \sum_{\hat{x}_1^n:(x_1^n,\hat{x}_1^n)\in\mathcal{G}_{d,n}^{\varepsilon}} P(\hat{x}_1^n|x_1^n). \tag{6.71}$$

Let $A(x_1^n) = \sum_{\hat{x}_1^n:(x_1^n,\hat{x}_1^n)\in\mathcal{G}_{d,n}^{\varepsilon}} P(\hat{x}_1^n|x_1^n)$. We continue now with (6.70), and use the following simple inequality

$$(1-xy)^n \leq 1 - x + e^{-ny}, \qquad 0 \leq x, y \leq 1 \tag{6.72}$$

to obtain

$$\begin{aligned}
P_E &\overset{(6.70)}{\leq} \sum_{x_1^n} P(x_1^n)\left(1 - 2^{-n(I(X;\hat{X})+3\varepsilon)}A(x_1^n)\right)^{2^{nR}} \\
&\overset{(6.71)}{\leq} \sum_{x_1^n} P(x_1^n)(1 - A(x_1^n)) + \exp\left(-2^{n(R-I(X;\hat{X})-3\varepsilon)}\right) \\
&\overset{(6.72)}{=} \sum_{(x_1^n,\hat{x}_1^n)\notin\mathcal{G}_{d,n}^{\varepsilon}} P(x_1^n,\hat{x}_1^n) + \exp\left(-2^{n(R-I(X;\hat{X})-3\varepsilon)}\right) \\
&= \Pr\{(X_1^n,\hat{X}_1^n) \notin \mathcal{G}_{d,n}^{\varepsilon}\} + \exp\left(-2^{n(R-I(X;\hat{X})-3\varepsilon)}\right) \\
&\to 0 \qquad \text{if } \; R > I(X;\hat{X}) + 3\varepsilon.
\end{aligned}$$

since $\Pr\{(X_1^n, \hat{X}_1^n \notin \mathcal{G}_{d,n}^{\varepsilon}\} < \varepsilon$ due to (6.53) of Theorem 6.16. In summary, if we choose $P(\hat{x}|x)$ to be the conditional distribution that achieves the minimum in the rate information function $R_I(D)$, then $R > R_I(D)$ implies $R > I(X;\hat{X})$, and we can choose ε small enough so that $\mathbf{E}[d(X_1^n, \hat{X}_1^n)] \leq D + \delta$. This completes the proof. ∎

Before we leave this section, we formulate yet another representation for the rate distortion function $R(D)$ that captures quite well the main thrust of $R(D)$ and is easy to understand intuitively.

Definition 6.25 (Operational Rate Distortion) *Let $\mathcal{A} = \hat{\mathcal{A}}$, and define a D-ball in $\mathcal{A}^n$ with center y_1^n as*

$$B_D(y_1^n) = \{x_1^n : \; d(x_1^n, y_1^n) \leq D\}$$

Let $N_\varepsilon(n, D)$ be the minimum number of D-balls covering $\mathcal{A}^n$ up to a set of probability ε, that is, the smallest N for which there exist $y_1^n(1), \ldots, y_1^n(N)$ such that that set $\mathcal{S}_n = \bigcup_{i=1}^N B_D(y_1^n(i))$ satisfies

$$\Pr\{\mathcal{S}_n\} \geq 1 - \varepsilon.$$

Then define

$$R_\varepsilon(n, D) = \min_{\mathcal{S}_n:\; P(\mathcal{S}_n)\geq 1-\varepsilon} \frac{\log N_\varepsilon(n, D)}{n},$$

and the following

$$R_O(D) = \lim_{\varepsilon \to 0} \lim_{n \to \infty} R_\varepsilon(n, D) \tag{6.73}$$

defines the so called **operational rate-distortion**.

We shall now argue that $R(D) = R_O(D)$. As for the converse, let us consider ε-achievable codes $(2^{nR}, n, D)$, and let $\mathcal{S}_n$ be the set of x_1^n for which (6.64) holds. We encode sequentially $x_1^n \in \mathcal{S}_n$ by assigning to it the index of the closest $\hat{x}_1^n \in \mathcal{S}_n$ that is within distance D. The other x_1^n are encoded arbitrary, e.g., we assign to them index $w = 1$. Clearly, the set $\mathcal{S}_n$ can be cover by 2^{nR} balls of radius $D + \delta$. But, $N_\varepsilon(n, D)$ is the smallest number of balls covering $1 - \varepsilon$ set, hence $N_\varepsilon(n, D) \leq 2^{nR_\varepsilon(D+\delta)}$, and therefore

$$\limsup_{n \to \infty} \frac{\log N_\varepsilon(n, D)}{n} \leq R_\varepsilon(D),$$

since $R(D)$ is a continuous function of D. To obtain the achievability part, we construct a code $(2^{nR}, n, D)$ with any $R > R_O(D)$ that achieves asymptotically the distortion D. Actually, it suffices to set $\hat{x}_1^n(i) = y_1^n(i)$ (center of D-balls) for $i = 1, \ldots, 2^{nR}$. The encoder assigns x_1^n to the closest $y_1^n(i) = \hat{x}_1^n(i)$ such that $d(x_1^n, \hat{x}_1^n) \leq D$. Clearly, $\Pr\{d(x_1^n, \hat{x}_1^n) \leq D\} \geq 1 - \varepsilon$.

We just sketched the derivation of the following important result.

Theorem 6.26 *For memoryless sources and bounded distortion function*

$$R(D) = R_O(D) = \lim_{\varepsilon \to 0} \lim_{n \to \infty} R_\varepsilon(n, D). \tag{6.74}$$

6.5 Applications

In this application section, we discuss three problems involving entropy and AEP. At first, we look at the typical depth in a suffix tree (cf. Section 6.5.1). This turns out to be equivalent to the computation of the phrase length in the Wyner-Ziv [357] compression scheme, as discussed in Section 1.2. We follow up this path and extend the analysis to lossy Lempel-Ziv scheme. The next problem we tackle is more of interest to DNA recombination than to data compression. We compare greedy and optimal algorithms that find the Shortest Common Superstring described Section 1.4. We prove that these algorithms are asymptotically optimal (cf. Section 6.5.2). Finally, we turn our attention to the *fixed database* model of a lossy data compression and estimate the compression ratio (cf. Section 6.5.3).

6.5.1 Phrase Length in the Lempel-Ziv Scheme and Depth in a Suffix Tree

We start with a more general case, namely, a lossy extension of the Lempel-Ziv scheme, and conclude with a lossless scheme. Let us assume that a mixing source MX (cf. Section 2.1) emits a stationary mixing sequence $\{X_k\}$. We also assume that the first n symbols, X_1^n, called the *database* or the *training sequence*, are revealed to the decoder and encoder. To make the analysis a little more interesting, we actually postulate that the first n symbols are coming from a distorted source, and we denote them as $\hat{X}_1^n$. Let $\hat{P}$ be the underlying probability measure of $\hat{X}_1^n$, and we shall postulate that $(\hat{X}_1^n, X_{n+1}^\infty)$ is ψ-strongly mixing process.

We consider now a string X_{n+1}^∞ generated by a ψ-mixing source with the underlying probability P. For simplicity, we only consider binary alphabets $\mathcal{A} = \hat{A} = \{0,1\}$ and the Hamming distance as the distortion measure. Let D be fixed. Define two important parameters:

> Let D_n be the largest K such that a prefix of X_{n+1}^∞ of length K is within distance D from X_i^{i-1+K} for some $1 \le i \le n-K+1$, that is, $d(X_i^{i-1+K}, X_{n+1}^{n+K}) \le D$;

and

> For *fixed* $M \le n$, let $D_n(M)$ be the length K of the longest prefix of X_M^∞ for which there exists $M+K \le i \le n+1$ such that $d(X_i^{i+K}, X_M^{M+K}) \le D$.

We observe that for $D=0$ the phrase lengths D_n and $D_n(M)$ become the *depth of insertion* and the M-th depth of a suffix tree built from $\hat{X}_1^n$ (cf. Definition 1.1 of Section 1.1).

To formulate the main result of this subsection, we recall the definition of the generalized Shannon entropy $\hat{r}_0(D)$ with respect to measure $\hat{P}$ (cf. Definition 6.12)

$$\hat{r}_0(D) = -\lim_{k\to\infty} \frac{\mathbf{E}_P[\log \hat{P}(B_D(X_1^k))]}{k} \tag{6.75}$$

where $\mathbf{E}_P$ is the average operator with respect to P.

Theorem 6.27 (Łuczak and Szpankowski 1997) *Let $(\hat{X}_1^n, X_{n+1}^\infty)$ be a ψ-strongly mixing process. Let also $\hat{r}_0(D) > 0$. Then*

$$\lim_{n\to\infty} \frac{D_n(M)}{\log n} = \lim_{n\to\infty} \frac{D_n}{\log n} = \frac{1}{\hat{r}_0(D)} \qquad (\text{pr.}) - P \times \hat{P} \tag{6.76}$$

provided

$$\sum_{\ell \ge k} \mathsf{Pr}\{\mathcal{B}_\ell^\varepsilon\} = \delta(k) \to 0 \tag{6.77}$$

where $\mathcal{B}_k^\varepsilon$ is the set of bad states in the lossy AEP with respect to $\hat{r}_0(D)$ (cf. Theorem 6.13). More precisely, for any $\varepsilon > 0$

$$\mathsf{Pr}\left\{\left|\frac{D_n}{\log n} - \frac{1}{\hat{r}_0(D)}\right| > \varepsilon\right\} = O\left(\max\{\psi(\log n), \delta(\log n), \frac{\log n}{n^{\varepsilon/4}}\}\right). \tag{6.78}$$

Proof. We establish separately an upper bound and a lower bound for D_n. The quantity $D_n(M)$ can be treated in a similar manner. Throughout, we use the lossy AEP formulated in Theorem 6.13 from which we know that the space $\mathcal{A}^n$ can be divided into sets $\mathcal{G}_n^\varepsilon$ and $\mathcal{B}_n^\varepsilon$ such that for $\hat{x}_1^n \in \mathcal{G}_n^\varepsilon$

$$2^{-n\hat{r}_0(D)(1+\varepsilon)} \leq \hat{P}(B_D(x_1^n)) \leq 2^{-n\hat{r}_0(D)(1-\varepsilon)}, \tag{6.79}$$

and $P(\mathcal{B}_n^\varepsilon) < \varepsilon$. We start with the upper bound and use the first moment method (cf. Chapter 3). Let Z_n be the number of positions $1 \leq i \leq n-k+1$ such that the prefix of X_{n+1}^{n+k} is within distance D from $\hat{X}_i^{i+k-1}$, i.e.

$$Z_n = |\{1 \leq i \leq n-k+1: \ d(\hat{X}_i^{i+k-1}, X_{n+1}^{n+k}) \leq D\}|.$$

To find an upper bound we need to estimate the probability $\Pr\{(D_n \geq k\}$ for $k = (1+\varepsilon)\hat{r}_0^{-1}(D)\log n$. Surprisingly, there is a certain detail we have to take care of to get it right. Surprisingly, $\{D_n \geq k\}$ does not imply that there exists a position i in the database such that $X_{n+1}^{n+k} \in B_D(\hat{X}_i^{i+k-1})$. To see it, let us assume that $d(\hat{X}_1^n, w_1^n) \leq D$ for a word $w_1^n \in \mathcal{A}^n$. When we increase n, we might have $d(\hat{X}_1^{n+l}, w_1^{n+l}) \leq D$ as well as $d(\hat{X}_1^{n+l}, w_1^{n+l}) \geq D$ for $l \geq 1$. Roughly speaking, the set $\{n: \ d(\hat{X}_1^n, w_1^n) \leq D\}$ does not consist of consecutive integers, that is, it has *gaps*. The correct implication is as follows:

$$\{D_n \geq k\} \quad \Longrightarrow \quad \exists_{\ell \geq k} \exists_{1 \leq i \leq n-\ell+1} \ d(X_{i+1}^{i+\ell}, X_{n+1}^{n+\ell}) \leq D \ . \tag{6.80}$$

Then

$$\begin{aligned}
\Pr\{D_n \geq k\} &\leq \sum_{\ell \geq k} \sum_{i=1}^{n-\ell} \Pr\{\hat{X}_i^{i+\ell-1} \in B_D(X_{n+1}^{n+\ell}), \ X_{n+1}^{n+\ell} \in \mathcal{G}_n^{\varepsilon/2}\} + \sum_{\ell \geq k} \Pr\{\mathcal{B}_\ell^{\varepsilon/2}\} \\
&= \sum_{\ell \geq k} \sum_{i=1}^{n-\ell} \sum_{w_1^\ell \in \mathcal{G}_\ell^{\varepsilon/2}} \Pr\{\hat{X}_i^{i+\ell-1} \in B_D(X_{n+1}^{n+\ell}), \ w_1^\ell = X_{n+1}^{n+\ell}\} + \sum_{\ell \geq k} \Pr\{\mathcal{B}_\ell^{\varepsilon/2}\} \\
&\leq \sum_{\ell \geq k} \sum_{i=1}^{n-\ell} (1 + \psi(n+2-\ell-i))\, 2^{-\ell\hat{r}_0(D)(1-\varepsilon/2)} + \sum_{\ell \geq k} \Pr\{\mathcal{B}_\ell^{\varepsilon/2}\} \\
&\leq nC2^{-k\hat{r}_0(D)(1-\varepsilon/2)} + \sum_{\ell \geq k} \Pr\{\mathcal{B}_\ell^{\varepsilon/2}\}.
\end{aligned}$$

where $C > 0$ is a constant. Set $k = \lfloor (1+\varepsilon)\hat{r}_0^{-1}(D)\log n \rfloor$. But $\sum_{\ell \geq k} \Pr\{\mathcal{B}_\ell^{\varepsilon/2}\} = \delta(k) \to 0$ (e.g., $\Pr\{\mathcal{B}_k^\varepsilon\} = O(1/k^{1+\delta})$ for some $\delta > 0$ suffices). Finally, we arrive at

$$\Pr\{D_n \geq \lfloor (1+\varepsilon)\hat{r}_0^{-1}(D)\log n \rfloor\} \leq \frac{c}{n^{\varepsilon/2(1-\varepsilon)}} + \delta(\log n)$$

for some constant $c > 0$. This completes the proof of the upper bound.

For the lower bound, we use the second moment method (cf. Section 3). Let $k = \lfloor (1-\varepsilon)\hat{r}_0^{-1}(D)\log n \rfloor$, and define

$$Z_n' = |\{1 \le i \le n/(k+g) :\ d(\hat{X}_{i(k+g)+1}^{(i+1)k+ig}, X_{n+1}^{n+k}) \le D\}|$$

and $g = \Theta(\log n)$ is a *gap* between $\lfloor n/(k+g) \rfloor = \Theta(n/\log n)$ *non-overlapping* substrings of length k. In words, instead of looking at all strings of length k we consider only $m = \lfloor n/(k+g) \rfloor$ strings with gaps of length g among them. These gaps are used to "weaken" dependency between the substrings of length k. Observe now that

$$\Pr\{D_n < k\} \le \Pr\{Z_n' = 0\},$$

as one can expect. Indeed, if $Z_n' > 0$ then by the definition $D_n \ge k$. Note also that

$$\begin{aligned}\Pr\{Z_n' = 0\} &= \Pr\{Z_n' = 0, X_{n+1}^{n+k} \in \mathcal{G}_k^{\varepsilon/2}\} + \Pr\{Z_n' = 0, X_{n+1}^{n+k} \in \mathcal{B}_k^{\varepsilon/2}) \\ &\le \sum_{w_1^k \in \mathcal{G}_k^{\varepsilon/2}} \Pr\{Z_n' = 0 | X_{n+1}^{n+k} = w_1^k\}\Pr\{w_1^k \in \mathcal{G}_k^{\varepsilon/2}\} + \Pr\{\mathcal{B}_k^{\varepsilon/2}\} \\ &\le \sum_{w_1^k \in \mathcal{G}_k^{\varepsilon/2}} \Pr\{Z_n'(w_1^k) = 0\}\Pr\{w_1^k \in \mathcal{G}_k^{\varepsilon/2}\} + \Pr\{\mathcal{B}_k^{\varepsilon/2}\}\ ,\end{aligned}$$

where

$$Z_n'(w_1^k) = |\{1 \le i \le n/(k+g) :\ d(\hat{X}_{i(k+g)+1}^{(i+1)k+ig}, w_1^k) \le D\}|\ ,$$

and $\Pr\{Z_n'(w_1^k) = 0\} = \Pr\{Z_n' = 0 | X_{n+1}^{n+k} = w_1^k\}$. Thus, it is suffices to show that $\Pr\{Z_n'(w_1^k) = 0\} \to 0$ *uniformly* for all $w_1^k \in \mathcal{G}_k^{\varepsilon/2}$. Hereafter, we assume that $w_1^k \in \mathcal{G}_k^{\varepsilon/2}$.

Let now $m = n/k = \Theta(n/\log n)$. From the definition of the set $\mathcal{G}_k^{\varepsilon/2}$ for every $w_1^k \in \mathcal{G}_k^{\varepsilon/2}$ we have

$$\mathbf{E}[Z_n'(w_1^k)] = m\Pr\{B_D(w_1^k)\} \ge m2^{-k\hat{r}_0(D)(1+\varepsilon/2)} \ge c\frac{n^{\varepsilon/2(1+\varepsilon)}}{\log n}$$

for a constant c. We now compute the variance $\mathbf{Var}[Z_n'(w_1^n)]$ for $w_1^k \in \mathcal{G}_k^{\varepsilon/2}$. Let $Z_n^i = 1$ if w_1^k occurs approximately at position $i(k+g)$, otherwise $Z_n^i = 0$. Certainly, $Z_n'(w_1^n) = \sum_{i=1}^m Z_n^i$, and $\mathbf{Var}[Z_n'(w_1^n)] = \sum_{i=1}^m \mathbf{Var}[Z_n^i] + \sum_{i=1}^m \sum_{i,j=1}^m \mathbf{Cov}[Z_n^i, Z_n^j]$. Simple algebra reveals that

$$\sum_{i=1}^m \mathbf{Var}[Z_n^i] \le m\mathbf{E}[Z_n^i] = m\Pr\{B_D(w_1^k)\} = \mathbf{E}[Z_n'(w_1^k)].$$

To compute the second term in the sum above, we split it as $\sum_{i,j=1}^n \mathbf{Cov}[Z_n^i, Z_n^j] = S_1 + S_2$ where

$$S_1 = \sum_{i=1}^m \sum_{|i-j| \le n^{\varepsilon/4}} \mathbf{Cov}[Z_n^i, Z_n^j],$$

$$S_2 = \sum_{i=1}^{m} \sum_{|i-j| \geq n^{\varepsilon/4}} \mathbf{Cov}[Z_n^i, Z_n^j] .$$

Observe that

$$\mathbf{Cov}[Z_n^i, Z_n^j] = \Pr\{Z_n^i Z_n^j = 1\} - \Pr\{Z_n^i = 1\}\Pr\{Z_n^j = 1\} \leq \Pr\{Z_n^i = 1\} = \mathbf{E}[Z_n^i] .$$

Hence $S_1 \leq 2n^{\varepsilon/4}\mathbf{E}[Z_n'(w_1^k)]$.

On the other hand, proceeding as in the above and using the mixing condition we also have $\mathbf{Cov}[Z_n^i, Z_n^j] \leq \psi(g)\Pr\{Z_n^i = 1\}\Pr\{Z_n^j = 1\}$ where $g \geq n^{\varepsilon/4}$. Thus, $S_2 \leq 2\psi(g)(\mathbf{E}^2[Z_n'(w_1^k)]$. Consequently, for every $w_1^k \in \mathcal{G}_k^{\varepsilon/2}$ we have ($\varepsilon < 1$)

$$\Pr\{Z_n'(w_1^k) = 0) \leq \frac{\mathbf{Var}[Z_n'(w_1^k)]}{(\mathbf{E}^2[Z_n'(w_1^k)]} \leq 2\psi(g) + O\left(\frac{n^{\varepsilon/4}}{\mathbf{E}[Z_n'(w_1^k)]}\right) \leq 2\psi(g) + O\left(\frac{\log n}{n^{\varepsilon/4}}\right) ,$$

and finally we obtain

$$\Pr\{D_n < \lfloor (1-\varepsilon)\hat{r}_0^{-1}(D) \log n \rfloor\} \to 0$$

as $n \to \infty$ which completes the lower bound. ∎

Here is another interesting question: Can we extend convergence in probability of $D_n/\log n$ in Theorem 6.27 to almost sure convergence? Observe that the rate of convergence in Theorem 6.27 does not justify such an extension. We recall, however, that in Section 4.2.3 we obtained such a generalization for the height H_n in a trie even when the rate of convergence was not good enough to apply Borel-Cantelli Lemma. But, the height H_n is a nondecreasing function of n, and this allows to extend convergence in probability to almost sure convergence. Clearly, D_n does not possess this property, and surprisingly enough we shall prove that $D_n/\log n$ does *not* convergence almost surely. Actually, this was an open problem posed by Wyner and Ziv in [357] for the lossless case. It was answered in the negative by the author in [324] (cf. also Ornestain and Weiss [265] for another formulation of this problem).

We start with a definition of the generalized height H_n and the generalized fill-up level F_n.

> The height H_n is equal to the largest K for which there exist $1 \leq i < j \leq n+1-K$ such that $d(\hat{X}_i^{i-1+K}, X_j^{j-1+K}) \leq D$;

and

> The fill-up level F_n is the largest k such that for *every* $w_1^k \in \mathcal{A}^k$ there exists $1 \leq i \leq n+1-k$ such that $d(\hat{X}_i^{i-1+k}, w_1^k) \leq D$.

Observe now that $F_n \leq D_n \leq H_n$. If we show that $F_n/\log n$ and $H_n/\log n$ converge (a.s.) to *different*(!) constants, then we should expect $D_n/\log n$ oscillates between these two constants. We shall prove this fact below.

Regarding F_n, we establish the following generalization of the result presented in Exercise 10 of Chapter 4.

Theorem 6.28 (Łuczak and Szpankowski 1997) *Let $\{X_k\}$ be a mixing process with ψ-mixing coefficients such that for every $\kappa \geq 0$*

$$\lim_{g\to\infty} g^{\kappa}\alpha(g) = 0. \tag{6.81}$$

Then

$$\lim_{n\to\infty} \frac{F_n}{\log n} = \frac{1}{r_{-\infty}(D)} \qquad \text{(a.s.)} \tag{6.82}$$

for $D \geq 0$.

Proof. We start with an upper bound which is quite simple in this case. Let us define $p_{\min}(k) = \min_{w_1^k\in\mathcal{A}^k}\{P(B_D(w_1^k))\}$ where we assume that the underlying probability measure of the source is P. By Theorem 6.13, $p_{\min}(k) \leq 2^{-kr_{-\infty}(D)(1-\varepsilon)}$. Observe that – unlike the lossless case – by definition of F_n we have

$$\{F_n > \ell\} \qquad \Longrightarrow \qquad \exists_{k>\ell}\forall_{w_1^k\in\mathcal{A}^k}\exists_{1\leq i\leq n+1-k}\ \ d(X_i^{i-1+k}, w_1^k) \leq D\ .$$

Thus, in particular, $\Pr\{F_n > \ell\} \leq (n+1)\sum_{k>\ell} P(B_D(w_k^{\min}))$, where $w_k^{\min}$ is a word from $\mathcal{A}^k$ for which $\log(P(B_D(w_k^{\min})) \sim -kr_{-\infty}(D)$. Hence, for $\ell = \lfloor(1+\varepsilon)r_{-\infty}^{-1}(D)\log n\rfloor$ we have

$$\Pr\{s > \ell\} \leq (n+1)\sum_{k>\ell} p_{\min}(k) = O(1/n^{\varepsilon})\ .$$

The lower bound requires a bit more work. Let us set $k = \lfloor(1-\varepsilon)r_{-\infty}^{-1}(D)\log n\rfloor$ and consider a set of *non-overlapping* substrings of X_1^n of length $k = O(\log n)$ between which one inserts *gaps* of length $g = O(\log n)$. Thus, there are $m = \lfloor(n+1)/(k+g)\rfloor = O(n/\log n)$ substrings $\{X_{i(k+g)+1}^{(i+1)k+ig}\}_{i=1}^m$. We show that with probability tending to 1 as $n\to\infty$ for every $w_1^k \in \mathcal{A}^k$ one can find among these m substrings at least one which are within distance D from w_1^k and consequently $F_n \geq k$. Indeed, using mixing conditions we have

$$\begin{aligned}
\Pr\{F_n < k\} &\leq \Pr\{\bigcup_{w_1^k\in\mathcal{A}^k}\bigcap_{i=1}^{m}\left(X_{i(k+g)+1}^{(i+1)k+ig} \neq w_1^k\right)\} \\
&\leq \sum_{w_1^k\in\mathcal{A}^k}(1+\psi(g))^m(1-P(B_D(w_1^k)))^m \leq 2^k(1+\psi(g))^m(1-p_{\min}(k))^m\ .
\end{aligned}$$

Taking into account (6.81) we immediately prove that

$$\Pr\{F_n < \lfloor(1-\varepsilon)r_{-\infty}^{-1}(D)\log n\rfloor\} \leq O(\exp(-n^{\varepsilon/2}/\log n))$$

which completes the proof of the convergence in probability of F_n.

The rate of convergence for the upper bound does not yet justify to apply Borel-Cantelli lemma. But, F_n is a nondecreasing sequence of n, hence taking exponentially increasing skeleton such as $n_l = 2^l$, as we explained in Section 4.2.6, we obtain almost sure convergence for the fill-up level. ■

The height H_n is harder to handle. First of all, we formulate it for the lossless case. Its proof follows the idea already presented in Section 4.2.6 so we ask the reader to prove it in Exercise 19.

Theorem 6.29 (Szpankowski 1993) *Let $\{X_k\}$ be a mixing process with ψ- mixing coefficients such that*

$$\sum_{g\geq 1} \psi(g) < \infty.$$

Then the height H_n in the lossless case ($D = 0$) satisfies

$$\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{2}{h_1} \qquad \text{(a.s.)} \tag{6.83}$$

where h_1 is the first Rényi entropy.

A generalization of Theorem 6.29 to the lossy situation is not easy. One expects that (6.83) becomes in lossy case

$$\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{2}{r_1(D)} \qquad \text{(a.s.)}$$

where $r_1(D)$ is the first generalized R'enyi entropy. This fact was only proved for memoryless and Markovian sources by Arratia and Waterman [22]. The difficulty arises when analyzing overlapping substrings. The reader is asked to prove it in Exercise 20 for memoryless sources.

Now, we are able to answer the question posed just after the proof of Theorem 6.27. Does $D_n/\log n$ converge (a.s.) to a limit? We only consider the lossless case to simplify the presentation. We shall show that

$$\liminf_{n\to\infty} \frac{D_n}{\log n} = \frac{1}{h_{-\infty}} \qquad \text{(a.s.)} \qquad\qquad \limsup_{n\to\infty} \frac{D_n}{\log n} = \frac{2}{h_1} \tag{6.84}$$

provided (6.81) holds. Indeed, $D_n \leq H_n$, hence $D_n/\log n \leq H_n/\log n$, and obviously

$$\limsup_{n\to\infty} \frac{D_n}{\log n} \leq \lim_{n\to\infty} \frac{H_n}{\log n} \qquad \text{(a.s.)}.$$

We now prove that the above hold with $\leq$ replaced by $\geq$ which will complete the proof. Note that almost surely $D_n = H_n$ whenever $H_{n+1} > H_n$, which happens infinitely often (i.o.) since

$H_n \to \infty$ (a.s.), and $\{X_k\}$ is an ergodic sequence. Therefore, $\Pr\{D_n = H_n \text{ i.o.}\} = 1$ implies that almost surely there exists a subsequence, say $n_k \to \infty$, such that $D_{n_k} = H_{n_k}$. Thus

$$\lim_{n_k\to\infty} D_{n_k}/\log n_k = \lim_{n_k\to\infty} H_{n_k}/\log n_k \quad \text{(a.s.)}.$$

This implies that

$$\limsup_{n\to\infty} \frac{D_n}{\log n} \geq \lim_{n\to\infty} \frac{H_n}{\log n} \qquad \text{(a.s.)},$$

that is, $\limsup_{n\to\infty} D_n/\log n = \lim_{n\to\infty} H_n/\log n$ (a.s.), and by (6.83) this proves the lim sup part of (6.84). In a similar manner we can prove the lim inf part by using F_n and Theorem 6.28. Can (6.84) be extended to the lossy case?

6.5.2 Shortest Common Superstring Problem

We now turn our attention to the *Shortest Common Superstring* problem that we described in Section 1.4. To recall and to re-establish notation, there are n strings, say $X_1^\ell(1), \ldots, X_1^\ell(n)$ of equal length $\ell \to \infty$. These strings are generated by n independent memoryless sources. Our goal is to construct the shortest superstring, where by a superstring we mean a string that contains given strings $X_1^\ell(1), \ldots, X_1^\ell(n)$ as substrings. In fact, instead of minimizing the superstring length we will maximize the *overlap* O_n^{opt} defined as the difference between $n\ell$ (the total length of all strings) and the shortest superstring (cf. Section 1.4). We compare O_n^{opt} to an overlap O_n^{gr} that is obtained by an application of a greedy algorithm to construct a superstring. A class of greedy algorithms was also discussed in Section 1.4. Here we concentrate on **RGREEDY** that builds a superstring Z by finding the maximum overlap between (suffix of) Z and (prefixes of) those original strings that yet have not been used in Z. We prove a surprising result showing that both O_n^{opt} and O_n^{gr} asymptotically grow like $\frac{1}{h} n \log n$ where h is the entropy rate of the source.

Before we formulate our findings, let us get some feeling about the problem. As explained in Section 1.4, to construct a superstring one takes a string, say $X_1^\ell(i_1)$, and attach to its end another string, say $X_1^\ell(i_2)$, trying to overlap the longest suffix of $X_1^\ell(i_1)$ with the longest common prefix of $X_1^\ell(i_2)$. We continue this process until we exhaust the set of available strings. This, of course, does not construct an optimal superstring which would require to examine all $n!$ possible permutations of $X_1^\ell(1), \ldots, X_1^\ell(n)$. In fact, the problem is NP-hard (see Section 1.4 for a reduction to a Hamiltonian path problem). In the **RGREEDY** algorithm we choose the first string, say $X_1^\ell(i_1)$ arbitrary, but our next choice, say $X_1^\ell(i_2)$, is such that its overlap with $X_1^\ell(i_1)$ is the largest among all strings excluding $X_1^\ell(i_1)$. And so on, that is, we choose in every step a remaining string that has the largest overlap with the superstring built so far.

Let us define C'_{ij} to be the length of the longest suffix of $X_1^\ell(i)$ that is equal to the prefix of $X_1^\ell(j)$ for $1 \le i \neq j \le n$. We introduce

$$\begin{aligned} D_n'(i) &= \max_{1\le j\le n, j\neq i}\{C_{ij}'\} , \\ H_n &= \max_{1\le i\le n}\{D_n'(i)\} . \end{aligned}$$

We should observe that C_{ij}' is distributed exactly as the alignment C_{ij} introduced in Definition 1.2 (i.e., the longest common prefix of $X(i)$ and $X(j)$). But then, $D_n'(i)$ is distributed as the ith depth $D_n(i)$ in a trie built from $X_1^\ell(1), \ldots X_1^\ell(n)$. Therefore, we write $D_n(i)$ for $D_n'(i)$ for $1 \le i \le n$. By Theorem 6.27 and Theorem 4.7 of Section 4.2.3, we know that for any $\varepsilon > 0$

$$\begin{aligned} \lim_{n\to\infty} \Pr\left\{(1-\varepsilon)\frac{1}{h}\ln n \le D_n(i) \le (1+\varepsilon)\frac{1}{h}\ln n\right\} &= 1 - O(1/n^\varepsilon) \quad 1 \le i \le n, \quad (6.85) \\ \lim_{n\to\infty} \Pr\left\{(1-\varepsilon)\frac{2}{h_1}\ln n \le H_n \le (1+\varepsilon)\frac{2}{h_1}\ln n\right\} &= 1 - O(1/n^\varepsilon). \quad (6.86) \end{aligned}$$

To get an upper bound on $O_n^{\rm opt}$, we observe that

$$O_n^{\rm opt} \le \sum_{i=1}^n D_n(i).$$

But with high probability (i.e., $1 - O(n^{-\varepsilon})$) all but $n\varepsilon$ depths $D_n(i)$ are bounded by $(1+\varepsilon)\frac{1}{h}\ln n$. Those $n\varepsilon$ longer strings cannot have depths bigger than H_n, hence they total contribution to the sum above is bounded by $O(n\varepsilon \ln n)$. In summary, with high probability (or in short: **whp**) for any $\varepsilon > 0$

$$\Pr\{O_n^{\rm opt} \le (1+\varepsilon)\frac{1}{h} n \ln n\} = 1 - O(1/n^\varepsilon) \tag{6.87}$$

Can we establish a matching lower bound for $O_n^{\rm opt}$? Definitely *not*, unless the length ℓ of strings are unbounded (indeed, take $\ell = 1$ to see that $O_n^{\rm opt} \le n$). If the original strings are "long" enough, as specified below, then we prove the following main result.

Theorem 6.30 (Alexander 1996, Frieze and Szpankowski 1998) *Consider the Shortest Common Superstring problem in which n memoryless sources generate independent strings $X_1^\ell(1), \ldots, X_1^\ell(n)$ over the alphabet $\mathcal{A} = \{\omega_1, \ldots, \omega_V\}$ of size $V = |\mathcal{A}|$. Then*

$$\lim_{n\to\infty} \frac{O_n^{\rm opt}}{n \log n} = \frac{1}{h}, \qquad (\text{pr.}) \qquad \lim_{n\to\infty} \frac{O_n^{\rm gr}}{n \log n} = \frac{1}{h} \tag{6.88}$$

provided

$$\ell > \frac{4}{h_1} \ln n \tag{6.89}$$

where $h_1 = -\ln(p_1^2 + \cdots + p_V^2)$ is the first order Rényi's entropy and $p_t = \Pr\{X_k(i) = \omega_t\}$ for $1 \le i \le n$.

Proof. Since $O_n^{\rm gr} \le O_n^{\rm opt}$, we need only to establish a lower bound for the greedy algorithm. We first justify (6.89). Let $\mathcal{E}$ denote the event that there is no such a pair, say i,j, that the overlap between $X_1^\ell(i)$ and $X_1^\ell(j)$ is bigger than $\ell/2$. Since $\mathcal{E}$ is equivalent to postulate that $H_n \le \ell/2$, we conclude from (6.83) that **whp** (i.e., $1-O(n^{-\varepsilon})$ all overlaps are smaller than $\ell/2$ provided (6.89) holds. In view of this, we split every string $X_1^\ell(i)$ into a prefix $Y_1^{\ell/2}(i)$ and suffix $W_{\ell/2+1}^\ell(i)$ such that $X_1^\ell(i) = Y_1^{\ell/2}(i)W_{\ell/2+1}^\ell(i)$.

From the above, we conclude that a suffix of the superstring Z being constructed is random and independent of the previous history of the algorithm. Having this in mind, let us define for $1 \le k \le \ell/2$

$$N_t(y_1^k) = |\{1 \le i \le k:\ y_i = \omega_t \in \mathcal{A},\ 1 \le t \le V\}|,$$

where $\mathcal{A} = \{\omega_1, \ldots, \omega_V\}$. In words, $N_t(y_1^k)$ denotes the number of positions in y_1^k that are equal to the tth symbol of the alphabet. When k is large, we expect that $N_t(y_1^k) \sim kp_t$ since $N_t(y_1^k)$ is distributed as the *binomial*$(k,p_t) := B(k,p_t)$. Indeed, let us define the set of typical sequences as

$$\mathcal{G}_k^\varepsilon = \{y_1^k:\ N_t(y_1^k) \le (1+\varepsilon)kp_t,\ 1 \le t \le V\}.$$

From AEP we know that $\mathsf{Pr}\{\mathcal{G}_k^\varepsilon\} \ge 1-\varepsilon$, but we need a stronger version of it. In particular, we need to control the probability of $\mathsf{Pr}\{\mathcal{B}_k^\varepsilon\}$ where $\mathcal{B}_k^\varepsilon = \mathcal{A}^k - \mathcal{G}_k^\varepsilon$.

Recall that in Example 6 of Chapter 5 we proved the following bound on the tails of the binomial distribution $B = B(n,p)$ (cf. (5.38))

$$\begin{aligned}
\mathsf{Pr}\{B \le (1-\varepsilon)np\} &\le e^{-\varepsilon^2 np/3}, \\
\mathsf{Pr}\{B \ge (1+\varepsilon)np\} &\le e^{-\varepsilon^2 np/3}.
\end{aligned}$$

Thus

$$\mathsf{Pr}\{\mathcal{B}_k^\varepsilon\} = \mathsf{Pr}\{y_1^k(i) \notin \mathcal{G}_k^\varepsilon\} \le \sum_{t=1}^V e^{-\varepsilon^2 kp_t/3} = \theta, \qquad (6.90)$$

and θ decays to zero for the following choice of ε and k

$$\varepsilon = (\log n)^{-1/3} \quad \text{and} \quad k = \left\lfloor (1-2\varepsilon)\frac{1}{h}\log n \right\rfloor$$

that we assume from now on. In other words, $\varepsilon^2 k \to \infty$ with n and

$$e^{-k(1-\varepsilon)h} \le \mathsf{Pr}\{Y_1^k(i) = y_1^k\} = P(y_1^k) \le e^{-k(1+\varepsilon)h}, \quad y_1^k \in \mathcal{G}_k^\varepsilon. \qquad (6.91)$$

As a side effect, we observe that **whp**

$$|\{i:\ y_1^k(i) \notin \mathcal{G}_k^\varepsilon\}| = o(\varepsilon n). \qquad (6.92)$$

Finally, for $y_1^k \in \mathcal{G}_k^\varepsilon$ define

$$R(y_1^k) = |\{i : \ Y_1^k(i) = y_1^k\}|.$$

Then

$$\begin{aligned}\mathsf{Pr}\{R(y_1^k) \le (1-\varepsilon)nP(y_1^k)\} &\le |\mathcal{G}_k^\varepsilon| e^{-\varepsilon^2 nP(y_1^k)/3} \\ &\le V^k e^{-\varepsilon^2 n^\varepsilon/3} = o(1) \qquad (6.93)\end{aligned}$$

since $R(y_1^k)$ is distributed as $B(n, P(y_1^k))$ and $nP(y_1^k) \ge n^\varepsilon$ due to (6.91).

To model the progress of **RGREEDY**, we build a trie $\mathcal{T}_n$ from $Y_1^{\ell/2}(1), \ldots, Y_1^{\ell/2}(n)$ and label every internal node v by $\nu(v)$ that denotes the size of the subtree rooted at v. In other words, if node v is at depth d and path to v can be labeled as $v = y_1 y_2 \ldots y_d$, then $\nu(v)$ is the number of i such that $y_1^d(i) = y_1 y_2 \ldots y_d$. We sometimes write $\nu(y_1^d)$ instead of $\nu(v)$.

We now make the trie $\mathcal{T}_n$ a dynamic tree, by allowing random deletions leading to tries $\mathcal{T}_{n-1}, \ldots \mathcal{T}_1$. Let $W_1^{\ell/2}$ be an independent string. We delete from the trie $\mathcal{T}_n$ a string $Y_1^{\ell/2}(i)$ that has the largest overlap with the a suffix of $W_1^{\ell/2}$. After the deletion, we modify labels $\nu(v)$ along the path of $W_1^{\ell/2}$. The reader should observe that $W_1^{\ell/2}(i)$ is a suffix of Z (i.e., it represents the second half of the string just added to the superstring Z at step i of **RGREEDY**).

Let κ_i be the length of the longest path along $W_1^{\ell/2}(i)$ in the trie $\mathcal{T}_{n-i+1}$ when deleting string $Y_1^{\ell/2}(i)$ for $1 \le i \le n$. We will show that **whp**

$$\kappa_1 + \kappa_2 + \cdots + \kappa_n \ge (1-5\varepsilon)\frac{1}{h} n \ln n. \qquad (6.94)$$

The final argument goes as follows. We want to show that **whp** we will have $\kappa_t \ge k$ for $1 \le t \le n_0 = \lceil (1-3\varepsilon)n \rceil$, where $k = (1-2\varepsilon)\frac{1}{h}\log n$. Now, most of the time the k-suffix of Z lies in $\mathcal{G}_k^\varepsilon$. Indeed, the probability it doesn't is at most θ which we prove in (6.90) to be small. If the k-suffix of Z belongs to $\mathcal{G}_k^\varepsilon$ and

$$\nu(y_1^k) \ne 0 \quad \text{for all} \quad y_1^k \in \mathcal{G}_k^\varepsilon, \qquad (6.95)$$

then $\kappa \ge k$, where $\nu(y_1^k)$ is the label of the node reached by y_1^k. We argue next that **whp** (6.95) holds up to $n_0 = \lceil (1-3\varepsilon)n \rceil$. If we consider a fixed $y_1^k \in \mathcal{G}_k^\varepsilon$, then at this point the number of decrements $D(y_1^k)$ in $\nu(y_1^k)$ (when deleting n_0 strings) is distributed as $B(n_0, P(y_1^k))$. Hence, using $n_0 P(y_1^k) \ge (1-3\varepsilon)n^\varepsilon$, for $y_1^k \in \mathcal{G}_k^\varepsilon$ we have

$$\begin{aligned}\mathsf{Pr}\{D(y_1^k) \ge (1+\varepsilon)n_0 P(y_1^k)\} &\le 2|\mathcal{G}_k^\varepsilon| e^{-(1-3\varepsilon)\varepsilon^2 n^\varepsilon/3} \\ &= o(1).\end{aligned}$$

So **whp** at this point $\nu(y_1^k) \ge n(1-\varepsilon)P(y_1^k) - n_0(1+\varepsilon)P(y_1^k) > 0$ for every $y_1^k \in \mathcal{G}_k^\varepsilon$. Thus, (6.94) follows and this completes the proof. ∎

In Exercise 21 we ask the reader to extend the above result to mixing models while in Exercise 22 to other greedy algorithms discussed in Section 1.4.

In passing we should mention that the Shortest Common Superstring is not a good compression algorithm. We start with a description of the compression code based on the Shortest Common Superstring. Observe that instead of storing all n strings of total length $n\ell$ we can store the Shortest Common Superstring and n pointers indicating the beginning of an original string (plus lengths of all strings). The compression ratio c_n (understood as the ratio of the number of bits needed to transmit the compression code to the length of the original set of strings) is

$$c_n = \frac{n\ell - \frac{1}{H} n \log n + n \log_2(n\ell - \frac{1}{H} n \log n)}{n\ell}$$

where the first term of the numerator represents the length of the shortest superstring and the second term corresponds to the number of bits needed to encode the pointers. When the length of a string ℓ grows faster than $\log n$, then $c_n \to 1$ (i.e., no compression). When $\ell = O(\log n)$ some compression might take place. The fact that SCS does not compress well is hardly surprising: In the construction of SCS we do not use all available redundancy of all strings (as in the Lempel-Ziv schemes) but only that contained in suffixes/prefixes of the original strings.

6.5.3 Fixed-Database Lossy Lempel-Ziv Algorithm

We wrap-up this long chapter with a computation of the compression ratio for a lossy extension of the Lempel-Ziv algorithm in the *fixed database model.* The reader is referred to Section 1.2.1 for a more detailed description. Below we only sketch the model.

Let us assume that the encoder and decoder have access to a common *fixed* database $\hat{X}_1^n$ generated according to a Markovian source with distribution $\hat{P}$. The source sequence X_1^M is emitted also by a Markovian source with distribution P. We assume that the database sequence and the source sequence are independent, but one can extend this analysis to mixing dependent sequences.

For fixed $D > 0$, the source sequence X_1^M is partitioned into $\Pi_n = \Pi_n(X_1^M)$ variable length phrases $Z^1, Z^2, \ldots, Z^{\Pi_n}$ of respective lengths $D_n^1, \ldots, D_n^{\Pi_n}$. The first phrase Z^1 and its length D_n^1, are computed as

$$\begin{aligned} D_n^1 &= \max\{k : \ d(\hat{X}_i^{i+k-1}, X_1^k) \leq D, \quad 1 \leq i \leq n-k+1\} \\ Z^1 &= X_1^{D_n^1}. \end{aligned}$$

where $d(\cdot,\cdot)$ is an additive distortion measure. This implies that the first phrase comprises of the longest prefix of the input X_1^M that matches a substring in the database $\hat{X}_1^n$ within distortion D or less. The string $\hat{Z}^1$ recovered by the decoder is therefore given by

$$\hat{Z}^1 = \hat{X}_i^{i+D_n^1-1}, \tag{6.96}$$

Let now $K_n(m) = \sum_{i=1}^{m} D_n^i$ for some $m \geq 1$. Subsequent phrases are computed in a similar manner, that is, the $(m+1)$-st phrase is defined as

$$\begin{aligned}
D_n^{m+1} &= \max\{k : \ d(\hat{X}_i^{i+k-1}, X_{K_n(m)+1}^{K_n(m)+k}) \leq D, \quad 1 \leq i \leq n-k+1\}, \\
Z^{m+1} &= X_{K_n(m)}^{K_n(m)+D_n^m-1}, \\
\hat{Z}^{m+1} &= \hat{X}_i^{i+D_n^m-1}.
\end{aligned}$$

Observe that the source sequence is partitioned into Π_n as $X_1^M = Z^1 Z^2 \dots Z^{\Pi_n}$ while the decoder recovers the string $\hat{Z}^1 \hat{Z}^2 \dots \hat{Z}^{\Pi_n}$ that is within distortion D from X_1^M. We represent each $\hat{Z}^i$ by a pointer `ptr` to the database and its length, hence its description cost is $\log n + \Theta(\log L_n^i)$ bits. The total data compression code length is

$$\ell_n(X_1^n) = \sum_{i=1}^{\Pi_n} \log n + \Theta(\log L_n^i).$$

Thus, the bit rate (or the compression ratio in bits per symbol) is defined as

$$r_n(X_1^n) = \frac{1}{M} \sum_{i=1}^{\Pi_n} \left(\log n + \Theta(\log L_n^i) \right). \tag{6.97}$$

Our goal is to evaluate asymptotically $r_n(X_1^M)$ as $n, M \to \infty$. We shall prove that the asymptotic average bit rate is equal to the generalized Shannon entropy $\hat{r}_0(D)$ defined in (6.75). For reader's convenience we repeat the definition of $\hat{r}_0(D)$

$$\hat{r}_0(D) = \lim_{n\to\infty} \frac{\mathbf{E}_P[-\log \hat{P}(B_D(X_1^n))]}{n},$$

where $B_D(\hat{x}_1^n) = \{y_1^n : \ d(\hat{x}_1^n, y_1^n) \leq D\}$ is the ball of radius D and center $\hat{x}_1^n$. To see why $\hat{r}_0(D)$ appears as the limit of the bit rate, we recall that by Theorem 6.27 (in fact, by exactly the same arguments that led to Theorem 6.27) for any ε

$$\Pr\left\{(1-\varepsilon)\frac{1}{\hat{r}_0(D)} \log n \leq D_n^1 \leq (1+\varepsilon)\frac{1}{\hat{r}_0(D)} \log n\right\} = 1 - O\left(\frac{\log n}{n^\varepsilon}\right) \tag{6.98}$$

for Markovian sources. (Indeed, for Markovian sources the $\psi(n)$ coefficients and $\delta(n)$ defined in (6.77) decay exponentially with n.)

Theorem 6.31 *Let us consider the fixed database model with the database $\hat{X}_1^n$ generated by a Markovian source $\hat{P}$ and the source sequence X_1^M emitted by a Markovian source P where all transition probabilities of both Markovian sources are positive. The average bit rate attains asymptotically*

$$\lim_{n\to\infty} \lim_{M\to\infty} \mathbf{E}[r_n(X_1^M)] = \hat{r}_0(D), \tag{6.99}$$

where $\mathbf{E}$ denotes the expectation with respect to $P \times \hat{P}$.

Proof. We start with an upper bound. We adopt the approach of Wyner and Ziv [359] and its simplification proposed by Kontoyiannis [227]. Let us partition all phrases into "long" phrases and "short" phrases. A phrase k is long if $D_n^i \geq (1-\varepsilon)\log n$; otherwise it is a short phrase. Let $\mathcal{L}_M$ and $\mathcal{S}_M$ be the sets of long and short phrases, respectively. Observe that

$$N = |\mathcal{L}_M| \leq \frac{M(\hat{r}_0(D)}{(1-\varepsilon)\log n}. \tag{6.100}$$

Now, we proceed as follows (below c is a constant and $\varepsilon > 0$ is arbitrary small that can change from line to line):

$$\begin{aligned}
r_n(X_1^N) &\leq \frac{1}{M}\sum_{i\in\mathcal{L}_M}(\log n + c\log D_n^i) + \frac{1}{M}\sum_{i\in\mathcal{S}_M}(\log n + c\log D_n^i) \\
&\overset{(6.100)}{\leq} (1+\varepsilon)\hat{r}_0(D) + c\frac{N}{M}\sum_{i\in\mathcal{L}_M}\frac{1}{N}\log D_n^i + \frac{1}{M}\sum_{i\in\mathcal{S}_M}(\log n + c\log D_n^i) \\
&\overset{\text{Jensen}}{\leq} (1+\varepsilon)\hat{r}_0(D) + c\frac{N}{M}\log\left(\frac{1}{N}\sum_{i\in\mathcal{L}_M}D_n^i\right) + \frac{1}{M}\sum_{i\in\mathcal{S}_M}(\log n + c\log D_n^i) \\
&\leq (1+\varepsilon)\hat{r}_0(D) + c\frac{N}{M}\log\left(\frac{M}{N}\right) + \frac{1}{M}\sum_{i\in\mathcal{S}_M}(\log n + c\log D_n^i) \\
&\leq (1+\varepsilon)\hat{r}_0(D) + \frac{c\log\log n}{\log n} + \frac{1}{M}\sum_{i\in\mathcal{S}_M}(\log n + c\log D_n^i).
\end{aligned}$$

We now evaluate the expected value of the third term above. We shall write below $\widetilde{D}_n^1$ for a match (phrase) that can occur in any position.

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{M}\sum_{i\in\mathcal{S}_M}(\log n + c\log D_n^i)\right] &\leq \frac{c}{M}\log n\ \mathbf{E}\left[\sum_{i=1}^{|\Pi_n|} I(D_n^i \text{ is short})\right] \\
&\leq \frac{c}{M}\log n\ \mathbf{E}\left[\sum_{i=1}^{M} I(\widetilde{D}_n^1 \text{ is short})\right] \\
&\leq c\log n\ \Pr\left\{\widetilde{D}_n^1 \leq (1-\varepsilon)\frac{1}{\hat{r}_0(D)}\log n\right\} \\
&\overset{(6.98)}{\leq} c\frac{\log^2 n}{n^\varepsilon}
\end{aligned}$$

where the third inequality above follows by considering not just all i's but all possible positions on X_1^M where a short match (i.e., $\widetilde{D}_n^1$) can occur. Thus

$$\limsup_{n\to\infty}\limsup_{M\to\infty}\mathbf{E}[r_n(X_1^M)] \leq \hat{r}_0(D),$$

and this establishes the desired upper bound.

We now deal with the lower bound. Let $\mathcal{H}$ be the set of those phrases whose length is not bigger than $(1+\varepsilon)\frac{1}{\hat{r}_0(D)}\log n$. Clearly

$$\mathbf{E}[r_n(X_1^M)] \geq \frac{1}{M}\mathbf{E}\left[\sum_{\mathcal{H}} \log n\right] = \frac{1}{M}\mathbf{E}[|\mathcal{H}|]\log n.$$

Thus it suffices to prove that

$$\mathbf{E}[|\mathcal{H}|] \geq \hat{r}_0(D)\frac{M(1-o(1))}{(1+\varepsilon)\log n} \tag{6.101}$$

to establish the desired lower bound. We start with the following fact: For all $i \geq 1$, any $\delta > 0$ and a positive constant we have for n

$$\Pr\{D_n^i > n^\delta\} \leq n\exp(-An^\delta) \tag{6.102}$$

The above is a simple consequence of our previous findings (cf. Section 4.2.6, Section 6.5.1). Nevertheless, the reader is asked to prove (6.102) independently in Exercise 27. Now, let $\bar{\mathcal{H}}$ be the complementary set to $\mathcal{H}$. We call phrases in $\bar{\mathcal{H}}$ as "very long". Observe that

$$M \leq |\mathcal{H}|\frac{1}{\hat{r}_0(D)}(1+\varepsilon)\log n + \sum_{i\in\bar{\mathcal{H}}} D_n^i$$

Taking the expectation of both side of the above yields (the constant c below can vary from line to line)

$$\begin{aligned}
M &\leq \mathbf{E}[|\mathcal{H}|]\frac{1}{\hat{r}_0(D)}(1+\varepsilon)\log n + \mathbf{E}\left[\sum_{\{i:\ i\in\bar{\mathcal{H}}\&D_n^i\leq n^{\varepsilon/2}\}} D_n^i + \sum_{\{i:\ D_n^i > n^{\varepsilon/2}\}} D_n^i\right] \\
&\overset{(6.102),\delta=\varepsilon/2}{\leq} \mathbf{E}[|\mathcal{H}|]\frac{1}{\hat{r}_0(D)}(1+\varepsilon)\log n + n^{\varepsilon/2}\cdot\mathbf{E}\left[\sum_{i=1}^M I(\widetilde{D}_n^1 \text{ is very long})\right] + e^{-An^{\varepsilon/2}}n^2M \\
&\leq \mathbf{E}[|\mathcal{H}|]\frac{1}{\hat{r}_0(D)}(1+\varepsilon)\log n \\
&+ n^{\varepsilon/2}M\Pr\left\{\widetilde{D}_n^1 > (1+\varepsilon)\frac{1}{\hat{r}_0(D)}\log n\right\} + e^{-An^{\varepsilon/2}}nM \\
&\overset{(6.98)}{\leq} \mathbf{E}[|\mathcal{H}|]\frac{1}{\hat{r}_0(D)}(1+\varepsilon)\log n + cM\frac{\log n}{n^{\varepsilon/2}}
\end{aligned}$$

where $c > 0$ is a constant. Hence (6.101) follows and Theorem 6.31 is proved. ■

6.6 Extensions and Exercises

6.1 Consider a discrete random variable, X, defined over a finite set $\mathcal{A}$. Prove that the average number of bits, L, required to represent such a random variable satisfies the following inequality
$$h(X) \leq L \leq h(X) + 1.$$

6.2 Prove or disprove that $h(X|Y) = h(Y|X)$. What about $h(X) - h(X|Y) = h(Y) - h(Y|X)$?

6.3 Prove Theorem 6.2.

6.4 Let $Y = g(X)$ where g is a measurable function. Prove

- $h(g(X)) \leq h(X)$;
- $h(Y|X) = 0$.

6.5 Random variables X, Y, Z form a Markov chain in that order (denoted $X \to Y \to Z$) if the conditional distribution of Z depends on Y and is independent of X, that is,
$$\mathsf{Pr}\{X = x, Y = y, Z = z\} = \mathsf{Pr}\{X = x\}\mathsf{Pr}\{Y = y|X = x\}\mathsf{Pr}\{Z = z|Y = y\}$$
for all possible x, y, z. Prove the *data processing inequality* that states
$$I(X;Y) \geq I(X;Z)$$
if $X \to Y \to Z$.

6.6 Consider a probability vector $\mathbf{p} = (p_1, \ldots, p_n)$ such that $\sum_{i=1}^{n} p_i = 1$. What distribution $\mathbf{p}$ minimizes the entropy $h(\mathbf{p})$?

6.7 This is the problem considered by Alon and Orlitsky [10]. Let X be a discrete random variable over a finite set $\mathcal{A}$. A binary encoding $\phi : \mathcal{A} \to \{0,1\}^*$ is an injection to the set of finite binary strings. Let $\ell(X)$ be the minimum expected length of the encoding ϕ of X (the encoding ϕ is *not* necessary prefix–free!). Prove that
$$\begin{aligned} \ell(X) &\leq h(X) && \text{(Wyner)} \\ h(X) - \log\log(|\mathcal{A}| + 1) &\leq \ell(X) && \text{(Dunham).} \end{aligned}$$

6.8 Provide details of the proof of Theorem 6.10.

6.9 Prove Lemma 6.11. Furthermore, prove the following representation for the divergence rate $D(P \parallel P^k)$ (cf. [147]).

Lemma 6.32 *Let $\{X_k\}$ is a source described by process distributions P and P^k, where P is stationary and P^k is kth order stationary Markov process. Then for $n \geq k$*

$$\begin{aligned} D(P \parallel P^k) &= \lim_{n\to\infty} D_{P\|P^k}(X_0|X_{-1},\ldots,X_{-n}) \\ &= -h_P(X) - \mathbf{E}_P[\log P^k(X_k|X_0^{k-1})]. \end{aligned}$$

6.10 ⚠ Using Lemma 6.32 established in Exercise 9 above, prove the following Markov approximation of a stationary distribution (cf. [59, 147] and Section 4.2.1).

Theorem 6.33 (Markov Approximation) *Let $\{X_k\}$ be a stationary process with distribution P. Define the kth order Markov approximation as*

$$P^k(x_1^n) := P(x_1^k) \prod_{i=k}^{n} P(x_i|X_{i-k}^{i-1} = x_{i-k}^{i-1}).$$

Then

$$\lim_{k\to\infty} D(P \parallel P^k) = 0.$$

6.11 Prove Part (i) of Theorem 6.13 for $b < 0$.

6.12 Consider a *binary symmetric channel* in which the binary input $\{0,1\}$ is transformed into the binary output $\{0,1\}$ such that $P(Y=0|X=0) = P(Y=1|X=1) = 1-p$ and $P(Y=0|X=1) = P(Y=1|X=0) = p$. Prove that the capacity C of such a channel is

$$C = 1 - h(p) = 1 + p\log p + (1-p)\log(1-p).$$

6.13 Consider a *binary erasure channel* which transmits the binary input $\{0,1\}$ into a ternary alphabet $\{0,1,e\}$ where e means that a symbol is erased. Let $P(Y=0|X=0) = P(Y=1|X=1) = 1-\alpha$ and $P(Y=e|X=1) = P(Y=e|X=0) = \alpha$. Prove that

$$C = 1 - \alpha.$$

6.14 Prove Theorem 6.23.

6.15 Consider a Bernoulli(p) source. For the Hamming distance, prove that the rate distortion functions is

$$R(D) = \begin{cases} h(p) - h(D) & 0 \le D \le \min\{p, 1-p\} \\ 0 & \text{otherwise} \end{cases}$$

where $h(x) = -x \log x - (1-x)\log(1-x)$ for $0 \le x \le 1$.

6.16 ⚠ (Strassen 1962) Let X_1^n be generated by a memoryless source P. Define $M_\varepsilon(n)$ as the minimum cardinality of sets $A \in \mathcal{A}^n$ with $P(A) \ge 1 - \varepsilon$. Prove that (cf. Strassen [313])

$$\log M_\varepsilon(n) = nh(P) + \sqrt{n\mathbf{Var}[X_1]}\lambda - \frac{1}{2}\log n + O(1)$$

where λ is a solution of $\Phi(\lambda) = 1 - \varepsilon$ with Φ being the distribution of the standard normal distribution.

6.17 ⚠ (Wolfowitz 1961) (i) Consider the channel coding problem, and let $N_\varepsilon(n)$ be the maximum number of messages that can be sent reliably (i.e., the probability of error $P_E < \varepsilon$). Prove that (cf. [313, 354])

$$\exp(nC - K\sqrt{n}) \le N_\varepsilon(n) \le \exp(nC + K\sqrt{n})$$

for a constant $K > 0$ and c is the channel capacity.

(ii) Find an asymptotic expansion of $N_\varepsilon(n)$.

6.18 ⍒ Prove that $R_I(D) = R_O(D)$ with*out* using the random coding technique.

6.19 ⚠ (Szpankowski 1993) Using the analysis from Section 4.2.6 prove Theorem 6.29 (cf. [324, 325]).

6.20 ⚠ (Arratia and Waterman 1989) Consider the generalized height H_n in a lossy situation as defined in Section 6.5.1. For a memoryless source prove that

$$\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{2}{r_1(D)} \qquad \text{(a.s.)}$$

where $r_1(D)$ is the first generalized R'enyi entropy. This is a difficult problem.

6.21 ⚠ Consider the Shortest Common Superstring problem as in Section 6.5.2 but with mixing sources, that is, the strings $X(1), \ldots, X(n)$ are generated by n independent mixing source. Extend Theorem 6.30 to this situation, that is, prove the following.

Theorem 6.34 (Frieze and Szpankowski 1998) *Consider the Shortest Common Superstring problem under the mixing model. Then (**whp**)*

$$\lim_{n\to\infty} \frac{O_n^{\text{opt}}}{n\log n} = \frac{1}{h} \qquad \text{(pr.)} \qquad \lim_{n\to\infty} \frac{O_n^{\text{gr}}}{n\log n} = \frac{1}{h}$$

provided

$$|\mathbf{x}^i| > \frac{4}{h_1}\log n$$

for all $1 \le i \le n$ where h_1 is the Rényi entropy of order one.

6.22 ⚠ (Frieze and Szpankowski 1998) Extend Theorem 6.30 for the Shortest Common Supetsring problem to other greedy algorithms discussed in Section 1.4, namely, **GREEDY** and **MGREEDY**.

6.23 ⚠ (Kontoyiannis 1999) Let P^* be the probability mass function that achieves infimum in Definition 6.22 (cf. (6.68)) of the information rate distortion function $R(D)$. As before, we write $B_D(X_1^n)$ for the distortion-ball of all strings of length n that are within distortion D from the center X_1^n. Using the Gärtner-Ellis Theorem 5.20 prove that

$$\lim_{n\to\infty} \frac{-\log P^*(B_D(X_1^n))}{n} = R(D) \qquad \text{(a.s.)}.$$

6.24 ⚠ Consider the *lossless* fixed database model where Y_1^n is database and X_1^M the source string. Both sequences are generated independently by a mixing source P. Let, as before, $Z^1 = X_1^{D_n^1}$ be the first phrase. Define $\mathcal{C}(X_1^n)$ to be the set of all possible Z^1. It is also called the *complete prefix set.* Observe that it can be easily generated from the suffix tree of Y_1^n. Indeed, $\mathcal{C}(X_1^n)$ is the set of all depths of insertion. The following lemma of Yang and Kieffer [346] is very useful in the analysis of lossless data compression. The reader is asked to prove it.

Lemma 6.35 (Yang and Kieffer 1997) *Assume that the source sequence $\{X_k\}$ is strongly mixing with summable mixing coefficients. Then for any string $x_1^k \in \mathcal{A}^k$ with $k \le n$*

$$\Pr\{\exists_{i\le k} :\ x_1^i \in \mathcal{C}(X_1^n)\} \le \frac{c}{(n-k+1)P(x)}$$

where c is a constant.

6.25 ⚠ Consider the *waiting time* N_ℓ defined as follows:

The waiting time N_ℓ is the smallest $N \geq 2\ell$ such that $d(X_1^\ell, X_{N-\ell+1}^N) \leq D$.

Alternatively, in the fixed database model, let $\{X_k\}$ be a source sequence and $(\ldots, Y_{-2}, Y_{-1})$ be an independent database sequence. Then, N_ℓ is defined as the smallest $i \geq \ell$ such that $d(Y_{-i}^{-i+\ell-1}, X_0^{\ell-1}) \leq D$. Prove the following results.

Theorem 6.36 (Łuczak & Szpankowski 1997, Yang & Kieffer 1998) *Let $\{X_k\}$ and $\{Y_k\}$ be strongly mixing sequence with summable mixing coefficients. The following holds*

$$\lim_{\ell\to\infty} \frac{\log N_\ell}{\ell} = r_0(D) \qquad \text{(a.s.)}$$

where $r_0(D)$ is the generalized Shannon entropy.

6.26 Consider the kth phrase length D_n^k in the fixed database model of Section 6.5.3. Prove that for all $k \leq n$

$$\lim_{n\to\infty} \frac{D_n^k}{\log n} = \frac{1}{\hat{r}_0(D)} \quad \text{(pr.)},$$

but $D_n^k / \log n$ does *not* converge almost surely when $k \to \infty$.

6.27 Prove (6.102) from Section 6.5.3, that is, for any $\delta > 0$ and for all $i \geq 1$

$$\Pr\{D_n^i > n^\delta\} \leq n \exp(-An^\delta)$$

where D_n^i is the length of the ith phrase and A is a positive constant.

6.28 Consider the longest phrase $\max_{1\leq i\leq M} D_n^i$ in the fixed database lossy Lempel-Ziv model. It appears that **whp**

$$\max_{1\leq i\leq M} D_n^i \to \frac{2}{\hat{r}(D)} \log n$$

where $\hat{r}(D)$ is related to the generalized Rényi entropy of order one? Is it true? How to define $\hat{r}(D)$? (For a related problem see Ziv and Merhav [367].)

6.29 Extend Theorem 6.31 to almost sure convergence, that is,

$$\lim_{n\to\infty} \lim_{M\to\infty} r_n(X_1^M) = \hat{r}_0(D) \quad \text{(a.s.)}$$

6.30 Consider the lossy extension of the Lempel-Ziv'78 algorithm and the lossy growing database model. What is the asymptotic bit rate $r_n(X_1^n)$ in these cases? Observe that the comparison (to obtain the next phrase) is made to a distorted database whose distribution is unknown.