We employ the random coding argument as follows. Write the error probability in (6.3) as a function of the codewords,

\[ e_n(x_1, \ldots, x_N) := \frac{1}{N} \sum_{i=1}^{N} W_n(\{y : \varphi(y) \neq i\} | x_i). \]

If we can find a random codebook \((X_1, \ldots, X_N)\) such that \(E[e_n(X_1, \ldots, X_N)] < \lambda\), then there must be at least one codebook \((x_1, \ldots, x_N) \in (X^n)^N\) with \(e_n(x_1, \ldots, x_N) < \lambda\).

For later use, note that

\[
E[e_n(X_1, \ldots, X_N)] = E\left[ \frac{1}{N} \sum_{i=1}^{N} W_n(\{y : \varphi(y) \neq i\} | X_i) \right] = \frac{1}{N} \sum_{i=1}^{N} E\left[ W_n(\{y : \varphi(y) \neq i\} | X_i) \right].
\]

Our approach is to obtain a bound on \(E[W_n(\{y : \varphi(y) \neq i\} | X_i)]\) that does not depend on \(i\). To this end, we first compute the conditional expectation

\[
E[W_n(\{y : \varphi(y) \neq i\} | X_i)] = \sum_{x} E[W_n(\{y : \varphi(y) \neq i\} | X_i, X_i = x)] P(X_i = x).
\]

6.4.3. Rates of Reduced Codebooks

**An Underlying Observation**

Suppose \(\theta_1, \ldots, \theta_N\) are nonnegative numbers whose numerical average is less than \(\lambda\). Then more than half of the \(\theta_i\) are individually less than \(2\lambda\). To establish this claim, we first express it in symbolic terms. Suppose that

\[
\frac{1}{N} \sum_{i=1}^{N} \theta_i < \lambda,
\]

\(^b\) In general, the decoder \(\varphi\) is defined in terms of the codewords. However, to keep the notation from getting out of hand, we just write \(\varphi(y)\) instead of the more explicit \(\varphi_{x_1, \ldots, x_N}(y)\).
where $0 < \lambda < \infty$. Put $G := \{ i : \theta_i < 2\lambda \}$. We show that $|G| > N/2$. Write

$$
\lambda > \frac{1}{N} \sum_{i=1}^{N} \theta_i = \frac{1}{N} \left( \sum_{i \in G^c} \theta_i + \sum_{i \in G} \theta_i \right) \geq \frac{1}{N} \sum_{i \in G^c} \theta_i \geq \frac{1}{N} \sum_{i \in G^c} 2\lambda = \frac{2\lambda}{N} |G^c|.
$$

It follows that $|G^c| < N/2$, and then $|G| = N - |G^c| > N - N/2 = N/2$.

**Putting It All Together**

Suppose we can find a random codebook $(X_1, \ldots, X_N)$ with $\mathbb{E}[e_n(X_1, \ldots, X_N)] < \lambda$. Then by the random coding argument, there is a codebook $(x_1, \ldots, x_N)$ such that

$$
e_n(x_1, \ldots, x_N) = \frac{1}{N} \sum_{i=1}^{N} W_n(\{y : \varphi(y) \neq i\} | x_i) < \lambda.
$$

By the observation above, there is a subset $G$ of $\{1, \ldots, N\}$ such that $|G| > N/2$ and

$$W_n(\{y : \varphi(y) \neq i\} | x_i) < 2\lambda, \quad i \in G.$$

By the throwing away the codewords $x_i$ for $i \notin G$, and using the construction in Section 6.4.1 we have an encoder $f_G$ and decoder $\varphi_G$ such that

$$W_n(\{y : \varphi_G(y) \neq i\} | x_i) < 2\lambda, \quad i \in G. \quad (6.9)$$

Note that the rate of this modified code satisfies

$$\frac{\log |G|}{n} > \frac{\log N/2}{n} = \frac{\log N}{n} - \frac{\log 2}{n}.$$

Hence, even though we discard half of the original codewords, the rate of the modified code is nearly $(\log N)/n$ for large $n$.

**CHAPTER 7**

**Typicality Decoding**

The success of the asymptotic equipartition property (AEP) in summarizing the key concepts used to prove Shannon’s Source Coding Theorem suggests the following extension.

**7.1. Joint Typicality**

When $P_{XY}(x,y)$ is any joint probability mass function on $X \times Y$, we use the notation

$$P_X(x) := \sum_y P_{XY}(x,y) \quad \text{and} \quad P_Y(y) := \sum_x P_{XY}(x,y)$$

50
for the marginal probability mass functions on $X$ and $Y$, respectively. Next, for $x = (x_1, \ldots, x_n) \in X^n$ and $y = (y_1, \ldots, y_n) \in Y^n$, we put

$$P^n_{XY}(x, y) := \prod_{k=1}^{n} P_{XY}(x_k, y_k).$$

The reader should verify that the marginals on $X^n$ and $Y^n$ satisfy

$$\sum_y P^n_{XY}(x, y) = \prod_{k=1}^{n} P_X(x_k) =: P^n_X(x) \quad \text{and} \quad \sum_x P^n_{XY}(x, y) = \prod_{k=1}^{n} P_Y(y_k) =: P^n_Y(y).$$

Given $\varepsilon > 0$, we say that $x$ and $y$ are jointly typical if pair $(x, y)$ belongs to the set

$$A_n := \left\{(x, y) \in X^n \times Y^n : \left| \frac{1}{n} \log \frac{1}{P^n_X(x)} - H(P_X) \right| \leq \varepsilon, \left| \frac{1}{n} \log \frac{1}{P^n_Y(y)} - H(P_Y) \right| \leq \varepsilon, \text{and} \left| \frac{1}{n} \log \frac{1}{P^n_{XY}(x, y)} - H(P_{XY}) \right| \leq \varepsilon \right\}.$$ 

By writing $|t| \leq \varepsilon$ as $-\varepsilon \leq t \leq \varepsilon$, it is easy to show that

$$A_n := \left\{(x, y) \in X^n \times Y^n : \exp(-n[H(P_X) + \varepsilon]) \leq P^n_X(x) \leq \exp(-n[H(P_X) - \varepsilon]), \exp(-n[H(P_Y) + \varepsilon]) \leq P^n_Y(y) \leq \exp(-n[H(P_Y) - \varepsilon]),\text{ and } \exp(-n[H(P_{XY}) + \varepsilon]) \leq P^n_{XY}(x, y) \leq \exp(-n[H(P_{XY}) - \varepsilon]) \right\}.$$ 

We thus have upper and lower bounds on the marginal probabilities $P^n_X(x)$ and $P^n_Y(y)$ and on the joint probabilities $P^n_{XY}(x, y)$.

### 7.1.1. Mutual Information

In proving the joint AEP, the expression $H(P_X) + H(P_Y) - H(P_{XY})$ arises several times. Before analyzing this expression, it is convenient to introduce the following notation. If $p$ and $q$ are pmfs on $X$ and $Y$, respectively, then $p \times q$ is the pmf on $X \times Y$ defined by

$$(p \times q)(x, y) := p(x)q(y).$$

Now observe that

$$H(P_X) + H(P_Y) - H(P_{XY}) = \mathbb{E}[-\log P_X(X)] + \mathbb{E}[-\log P_Y(Y)] - \mathbb{E}[-\log P_{XY}(X, Y)].$$
\[ E \left[ \log \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)} \right] = D(P_{XY} \parallel P_X \times P_Y), \]

where \( D \) is the Kullback–Leibler informational divergence or relative entropy. We call \( I(X \land Y) := I(P_{XY}) := E \left[ \log \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)} \right] = \sum_x \sum_y P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \)

the (average) mutual information between \( X \) and \( Y \).

Several properties of mutual information (and entropy) are readily apparent. First, because mutual information can be expressed as a divergence, we see that \( I(X \land Y) \geq 0 \), with equality if and only if \( X \) and \( Y \) are independent. Second,

\[ H(XY) \leq H(X) + H(Y), \]

with equality if and only if \( X \) and \( Y \) are independent. Third,

\[ I(X \land Y) = H(Y) - H(Y|X) = H(X) - H(X|Y), \]

from which it follows that

\[ H(X) \geq H(X|Y). \]

Since entropy is a measure of uncertainty, this last property is summarized by saying, “conditioning reduces uncertainty.”

7.1.2. The Joint Asymptotic Equipartition Property

**Theorem 7.1** (Joint Asymptotic Equipartition Property (Joint AEP)). The set of jointly typical pairs has small cardinality in the sense that

\[ |A_n| \leq \exp(n[H(P_{XY}) + \varepsilon]). \]

For large \( n \), the set has high \( P_{XY}^n \) probability in the sense that

\[ \lim_{n \to \infty} P_{XY}^n(A_n) = 1. \]

It then follows that for large \( n \),

\[ |A_n| \geq (1 - \varepsilon) \exp(n[H(P_{XY}) - \varepsilon]). \]

\(^a\) Many authors write \( I(X;Y) \) instead of \( I(X \land Y) \).
If $Q_n(x, y) := P^n_X(x)P^n_Y(y)$, then

$$Q_n(A_n) \leq \exp(-n[I(P_{XY}) - 3\varepsilon]),$$

and for sufficiently large $n$,

$$Q_n(A_n) \geq (1 - \varepsilon) \exp(-n[I(P_{XY}) + 3\varepsilon]).$$