Foreword

This chapter is based on lecture notes from coding theory courses taught by Venkatesan Guruswami at University at Washington and CMU; by Atri Rudra at University at Buffalo, SUNY and by Madhu Sudan at MIT.

This version is dated **July 21, 2015**. For the latest version, please go to

http://www.cse.buffalo.edu/ atri/courses/coding-theory/book/

The material in this chapter is supported in part by the National Science Foundation under CAREER grant CCF-0844796. Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

Chapter 6

What Happens When the Noise is Stochastic: Shannon's Theorem

Shannon was the first to present a rigorous mathematical framework for communication, which (as we have already seen) is the problem of reproducing at one point (typically called the "receiver" of the channel) a message selected at another point (called the "sender" to the channel). Unlike Hamming, Shannon modeled the noise stochastically, i.e. as a well defined random process. He proved a result that pin-pointed the best possible rate of transmission of information over a very wide range of stochastic channels. In fact, Shannon looked at the communication problem at a higher level, where he allowed for compressing the data first (before applying any error-correcting code), so as to minimize the amount of symbols transmitted over the channel.

In this chapter, we will study some stochastic noise models (most of) which were proposed by Shannon. We then prove an optimal tradeoff between the rate and fraction of errors that are correctable for a specific stochastic noise model called the Binary Symmetric Channel.

6.1 Overview of Shannon's Result

Shannon introduced the notion of reliable communication[1] over noisy channels. Broadly, there are two types of channels that were studied by Shannon:

- (Noisy Channel) This type of channel introduces errors during transmission, which result in an incorrect reception of the transmitted signal by the receiver. Redundancy is added at the transmitter to increase reliability of the transmitted data. The redundancy is taken off at the receiver. This process is termed as *Channel Coding*.

- (Noise-free Channel) As the name suggests, this channel does not introduce any type of error in transmission. Redundancy in source data is used to compress the source data at the transmitter. The data is decompressed at the receiver. The process is popularly known as *Source Coding*.

[1] That is, the ability to successfully send the required information over a channel that can lose or corrupt data.

Figure 6.1 presents a generic model of a communication system, which combines the two concepts we discussed above.

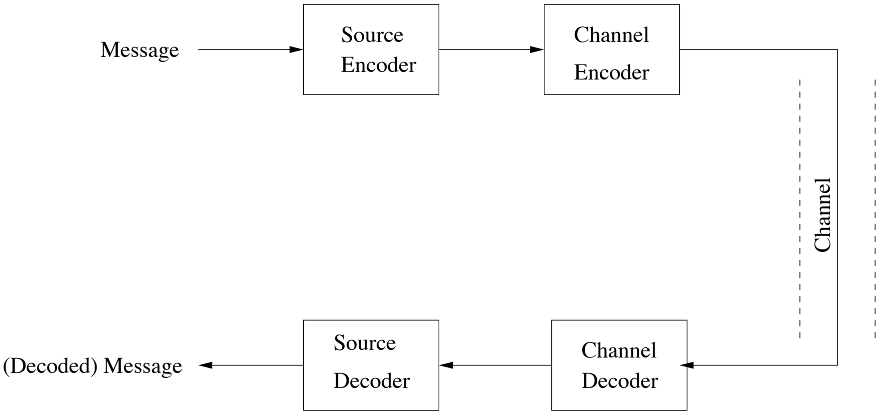

Figure 6.1: The communication process

In Figure 6.1, source coding and channel coding are coupled. In general, to get the optimal performance, it makes sense to design both the source and channel coding schemes simultaneously. However, Shannon's source coding theorem allows us to decouple both these parts of the communication setup and study each of these parts separately. Intuitively, this makes sense: if one can have reliable communication over the channel using channel coding, then for the source coding the channel effectively has no noise.

For source coding, Shannon proved a theorem that precisely identifies the amount by which the message can be compressed: this amount is related to the *entropy* of the message. We will however, not talk much more about source coding in in this book. (However, see Exercises 6.10, 6.11 and 6.12.) From now on, we will exclusively focus on the channel coding part of the communication setup. Note that one aspect of channel coding is how we model the channel noise. So far we have seen Hamming's worst case noise model in some detail. Next, we will study some specific stochastic channels.

6.2 Shannon's Noise Model

Shannon proposed a stochastic way of modeling noise. The input symbols to the channel are assumed to belong to some *input alphabet* $\mathcal{X}$, while the channel selects symbols from its *output alphabet* $\mathcal{Y}$. The following diagram shows this relationship:

$$\mathcal{X} \ni x \to \boxed{\text{channel}} \to y \in \mathcal{Y}$$

The channels considered by Shannon are also *memoryless*, that is, noise acts independently on each transmitted symbol. In this book, we will only study *discrete* channels where both the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite. For the sake of variety, we will define one channel that is continuous, though we will not study it in any detail later on.

The final piece in specification of a channel is the *transition matrix* **M** that governs the process of how the channel introduces error. In particular, the channel is described in form of a matrix with entries as cross over probability over all combination of the input and output alphabets. For any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, let $\Pr(y|x)$ denote the probability that y is output by the channel when x is input to the channel. Then the transition matrix is given by $\mathbf{M}(x, y) = \Pr(y|x)$. Specific structure of the matrix is shown below.

$$\mathbf{M} = \begin{pmatrix} & \vdots & \\ \cdots & \Pr(y|x) & \cdots \\ & \vdots & \end{pmatrix}$$

Next, we look at some specific instances of channels.

Binary Symmetric Channel (BSC)**.** Let $0 \le p \le 1$. The Binary Symmetric Channel with *crossover probability* p or BSC_p is defined as follows. $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. The 2×2 transition matrix can naturally be represented as a bipartite graph where the left vertices correspond to the rows and the right vertices correspond to the columns of the matrix, where $\mathbf{M}(x, y)$ is represented as the weight of the corresponding (x, y) edge. For BSC_p, the graph is illustrated in Figure 6.2.

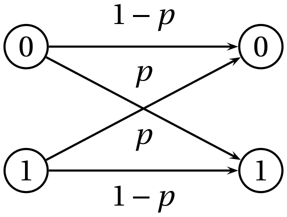

Figure 6.2: Binary Symmetric Channel BSC_p

In other words, every bit is flipped with probability p. We claim that we need to only consider the case when $p \le \frac{1}{2}$, i.e. if we know how to ensure reliable communication over BSC_p for $p \le \frac{1}{2}$, then we can also handle the case of $p > \frac{1}{2}$. (See Exercise 6.1.)

q**-ary Symmetric Channel (**qSC**).** We now look at the generalization of BSC_p to alphabets of size $q \ge 2$. Let $0 \le p \le 1 - \frac{1}{q}$. (As with the case of BSC_p, we can assume that $p \le 1 - 1/q$– see Exercise 6.2.) The q-ary Symmetric Channel with crossover probability p, or $q\text{SC}_p$ is defined as follows. $\mathcal{X} = \mathcal{Y} = [q]$. The transition matrix **M** for $q\text{SC}_p$ is defined as follows.

$$M(x, y) = \begin{cases} 1 - p & \text{if } y = x \\ \frac{p}{q-1} & \text{if } y \neq x \end{cases}$$

In other words, every symbol is retained as it at the output with probability $1 - p$ and is distorted to each of the $q - 1$ possible different symbols with equal probability of $\frac{p}{q-1}$.

Binary Erasure Channel (BEC) In the previous two examples that we saw, $\mathcal{X} = \mathcal{Y}$. However this might not always be the case.

Let $0 \le \alpha \le 1$. The Binary Erasure Channel with *erasure probability* α (denoted by BEC_α) is defined as follows. $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1,?\}$, where ? denotes an *erasure*. The transition matrix is as follows:

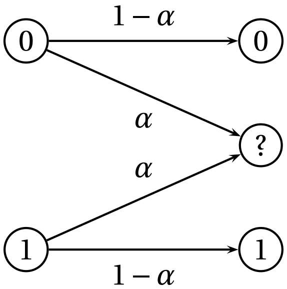

Figure 6.3: Binary Erasure Channel BEC_α

In Figure 6.3 any missing edge represents a transition that occurs with 0 probability. In other words, every bit in BEC_α is erased with probability α (and is left unchanged with probability $1-\alpha$).

Binary Input Additive Gaussian White Noise Channel (BIAGWN). We now look at a channel that is continuous. Let $\sigma \ge 0$. The Binary Input Additive Gaussian White Noise Channel with standard deviation σ or BIAGWN_σ is defined as follows. $\mathcal{X} = \{-1,1\}$ and $\mathcal{Y} = \mathbb{R}$. The noise is modeled by continuous Gaussian probability distribution function. The Gaussian distribution has lots of nice properties and is a popular choice for modeling noise continuous in nature. Given $(x,y) \in \{-1,1\} \times \mathbb{R}$, the noise $y-x$ is distributed according to the Gaussian distribution of mean of zero and standard deviation of σ. In other words,

$$\Pr\left(y \mid x\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\left(\frac{(y-x)^2}{2\sigma^2}\right)\right)$$

6.2.1 Error Correction in Stochastic Noise Models

We now need to revisit the notion of error correction from Section 1.3. Note that unlike Hamming's noise model, we cannot hope to *always* recover the transmitted codeword. As an example, in BSC_p there is always some positive probability that a codeword can be distorted into another codeword during transmission. In such a scenario no decoding algorithm can hope to recover the transmitted codeword. Thus, in some stochastic channels there is always some *decoding error probability* (where the randomness is from the channel noise): see Exercise 6.14 for example channels where one can have zero decoding error probability. However, we would like this error probability to be small for every possible transmitted codeword. More precisely, for every message, we would like the decoding algorithm to recover the transmitted message with probability $1 - f(n)$, where $\lim_{n\to\infty} f(n) \to 0$, that is $f(n)$ is $o(1)$. Ideally, we would like to have $f(n) = 2^{-\Omega(n)}$. We will refer to $f(n)$ as the decoding error probability.

6.2.2 Shannon's General Theorem

Recall that the big question that we are interested in this book is the tradeoff between the rate of the code and the fraction of errors that can be corrected. For stochastic noise models that we have seen, it is natural to think of the fraction of errors to be the parameter that governs the amount of error that is introduced by the channel. For example, for BSC_p, we will think of p as the fraction of errors.

Shannon's remarkable theorem on channel coding was to *precisely* identify when reliable transmission is possible over the stochastic noise models that he considered. In particular, for the general framework of noise models, Shannon defined the notion of *capacity*, which is a real number such that reliable communication is possible if and only if the rate is less than the capacity of the channel. In other words, given a noisy channel with capacity C, if information is transmitted at rate R for any $R < C$, then there exists a coding scheme that guarantees negligible probability of miscommunication. On the other hand if $R > C$, then regardless of the chosen coding scheme there will be some message for which the decoding error probability is bounded from below by some constant.

In this chapter, we are going to state (and prove) Shannon's general result for the special case of BSC_p.

6.3 Shannon's Result for BSC_p

We begin with a notation. For the rest of the chapter, we will use the notation $\mathbf{e} \sim \mathrm{BSC}_p$ to denote an error pattern $\mathbf{e}$ that is drawn according to the error distribution induced by BSC_p. We are now ready to state the theorem.

Theorem 6.3.1 (Shannon's Capacity Theorem for BSC)**.** *For real numbers p, ε such that $0 \le p < \frac{1}{2}$ and $0 \le \varepsilon \le \frac{1}{2} - p$, the following statements are true for large enough n:*

1. *There exists a real $\delta > 0$, an encoding function $E : \{0,1\}^k \to \{0,1\}^n$ and a decoding function $D : \{0,1\}^n \to \{0,1\}^k$ where $k \le \lfloor (1 - H(p+\varepsilon))\, n \rfloor$, such that the following holds for every $\mathbf{m} \in \{0,1\}^k$:*

$$\Pr_{\mathbf{e} \sim \mathrm{BSC}_p} [D(E(\mathbf{m}) + \mathbf{e})) \ne \mathbf{m}] \le 2^{-\delta n}.$$

2. *If $k \ge \lceil (1 - H(p) + \varepsilon) n \rceil$ then for every pair of encoding and decoding functions, $E : \{0,1\}^k \to \{0,1\}^n$ and $D : \{0,1\}^n \to \{0,1\}^k$, there exists $\mathbf{m} \in \{0,1\}^k$ such that*

$$\Pr_{\mathbf{e} \sim \mathrm{BSC}_p} [D(E(\mathbf{m}) + \mathbf{e})) \ne \mathbf{m}] \ge \frac{1}{2}.$$

Note that Theorem 6.3.1 implies that the capacity of BSC_p is $1 - H(p)$. It can also be shown that the capacity of $q\mathrm{SC}_p$ and BEC_α are $1 - H_q(p)$ and $1 - \alpha$ respectively. (See Exercises 6.6 and 6.7.)

Entropy function appears in Theorem 6.3.1 due to the same technical reason that it appears in the GV bound: the entropy function allows us to use sufficiently tight bounds on the volume of a Hamming ball (Proposition 3.3.1).

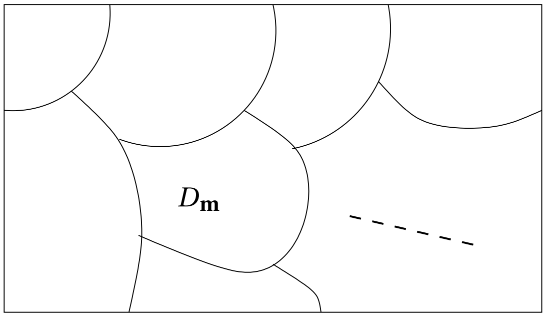

Figure 6.4: The sets $D_{\mathbf{m}}$ partition the ambient space $\{0,1\}^n$.

6.3.1 Proof of Converse of Shannon's Capacity Theorem for BSC

We start with the proof of part (2) of Theorem 6.3.1. (Proof of part (1) follows in the next section.)

For the proof we will assume that $p > 0$ (since when $p = 0$, $1 - H(p) + \varepsilon > 1$ and so we have nothing to prove). For the sake of contradiction, assume that the following holds for every $\mathbf{m} \in \{0,1\}^k$:

$$\Pr_{\mathbf{e} \sim \mathrm{BSC}_p}[D(E(\mathbf{m}) + \mathbf{e}) \neq \mathbf{m}] \le 1/2.$$

Define $D_{\mathbf{m}}$ to be the set of received words $\mathbf{y}$ that are decoded to $\mathbf{m}$ by D, that is,

$$D_{\mathbf{m}} = \left\{\mathbf{y} | D(\mathbf{y}) = \mathbf{m}\right\}.$$

The main idea behind the proof is the following: first note that the sets $D_{\mathbf{m}}$ partition the entire space of received words $\{0,1\}^n$ (see Figure 6.3.1 for an illustration). (This is because D is a function.) Next we will argue that since the decoding error probability is at most a $1/2$, then $D_{\mathbf{m}}$ for every $\mathbf{m} \in \{0,1\}^k$ is "large." Then by a simple packing argument, it follows that we cannot have too many distinct $\mathbf{m}$, which we will show implies that $k < (1 - H(p) + \varepsilon)n$: a contradiction. Before we present the details, we outline how we will argue that $D_{\mathbf{m}}$ is large. Let $S_{\mathbf{m}}$ be the shell of radius $[(1-\gamma)pn, (1+\gamma)pn]$ around $E(\mathbf{m})$, that is,

$$S_{\mathbf{m}} = B\left(E(\mathbf{m}), (1+\gamma)pn\right) \setminus B\left(E(\mathbf{m}), (1-\gamma)pn\right).$$

(We will set $\gamma > 0$ in terms of ε and p at the end of the proof.) See Figure 6.3.1 for an illustration. Then we argue that because the decoding error probability is bounded by $1/2$, most of the received words in the shell $S_{\mathbf{m}}$ are decoded correctly, i.e. they fall in $D_{\mathbf{m}}$. To complete the argument, we show that number of such received words is indeed large enough.

Fix an arbitrary message $\mathbf{m} \in \{0,1\}^k$. Note that by our assumption, the following is true (where from now on we omit the explicit dependence of the probability on the BSC_p noise for clarity):

$$\Pr[E(\mathbf{m}) + \mathbf{e} \notin D_{\mathbf{m}}] \le 1/2. \tag{6.1}$$

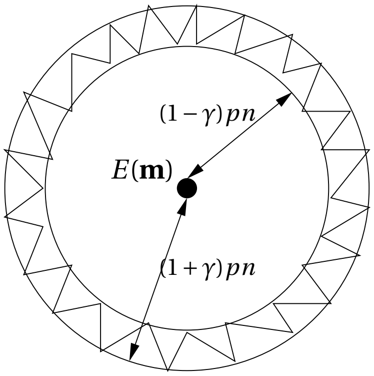

Figure 6.5: The shell $S_{\mathbf{m}}$ of inner radius $(1-\gamma)pn$ and outer radius $(1+\gamma)pn$.

Further, by the (multiplicative) Chernoff bound (Theorem 3.1.6),

$$\Pr[E(\mathbf{m})+\mathbf{e}\not\in S_{\mathbf{m}}] \le 2^{-\Omega(\gamma^2 n)}. \tag{6.2}$$

(6.1) and (6.2) along with the union bound (Proposition 3.1.3) imply the following:

$$\Pr[E(\mathbf{m})+\mathbf{e}\not\in D_{\mathbf{m}}\cap S_{\mathbf{m}}] \le \frac{1}{2}+2^{-\Omega(\gamma^2 n)}.$$

The above in turn implies that

$$\Pr[E(\mathbf{m})+\mathbf{e}\in D_{\mathbf{m}}\cap S_{\mathbf{m}}] \ge \frac{1}{2}-2^{-\Omega(\gamma^2 n)} \ge \frac{1}{4}, \tag{6.3}$$

where the last inequality holds for large enough n. Next we upper bound the probability above to obtain a lower bound on $|D_{\mathbf{m}}\cap S_{\mathbf{m}}|$.

It is easy to see that

$$\Pr[E(\mathbf{m})+\mathbf{e}\in D_{\mathbf{m}}\cap S_{\mathbf{m}}] \le |D_{\mathbf{m}}\cap S_{\mathbf{m}}|\cdot p_{\max},$$

where

$$p_{\max} = \max_{\mathbf{y}\in S_{\mathbf{m}}}\Pr\left[E(\mathbf{m})+\mathbf{e}=\mathbf{y}\right] = \max_{d\in[(1-\gamma)pn,(1+\gamma)pn]} p^d(1-p)^{n-d}.$$

In the above, the second equality follows from the fact that all error patterns with the same Hamming weight appear with the same probability when chosen according to BSC$_p$. Next, note that $p^d(1-p)^{n-d}$ is decreasing in d for $p\le 1/2$.[2] Thus, we have

$$p_{\max} = p^{(1-\gamma)pn}(1-p)^{n-(1-\gamma)pn} = \left(\frac{1-p}{p}\right)^{\gamma pn}\cdot p^{pn}(1-p)^{(1-p)n} = \left(\frac{1-p}{p}\right)^{\gamma pn} 2^{-nH(p)}.$$

[2] Indeed $p^d(1-p)^{n-d} = (p/(1-p))^d(1-p)^n$ and the bound $p\le 1/2$ implies that the first exponent is at most 1, which implies that the expression is decreasing in d.

Thus, we have shown that

$$\Pr\left[E(\mathbf{m})+\mathbf{e}\in D_{\mathbf{m}}\cap S_{\mathbf{m}}\right]\le |D_{\mathbf{m}}\cap S_{\mathbf{m}}|\cdot\left(\frac{1-p}{p}\right)^{\gamma pn}2^{-nH(p)},$$

which, by (6.3), implies that

$$|D_{\mathbf{m}}\cap S|\ge\frac{1}{4}\cdot\left(\frac{1-p}{p}\right)^{-\gamma pn}2^{nH(p)}. \tag{6.4}$$

Next, we consider the following sequence of relations:

$$\begin{aligned}
2^n &= \sum_{\mathbf{m}\in\{0,1\}^k}|D_{\mathbf{m}}| && (6.5)\\
&\ge \sum_{\mathbf{m}\in\{0,1\}^k}|D_{\mathbf{m}}\cap S_{\mathbf{m}}|\\
&\ge \frac{1}{4}\left(\frac{1}{p}-1\right)^{-\gamma pn}\sum_{\mathbf{m}\in\{0,1\}^k}2^{H(p)n} && (6.6)\\
&= 2^{k-2}\cdot 2^{H(p)n-\gamma p\log(1/p-1)n}\\
&> 2^{k+H(p)n-\varepsilon n}. && (6.7)
\end{aligned}$$

In the above, (6.5) follows from the fact that for $\mathbf{m}_1\neq\mathbf{m}_2$, $D_{\mathbf{m}_1}$ and $D_{\mathbf{m}_2}$ are disjoint. (6.6) follows from (6.4). (6.7) follows for large enough n and if we pick $\gamma=\frac{\varepsilon}{2p\log\left(\frac{1}{p}-1\right)}$. (Note that as $0<p<1/2$, $\gamma=\Theta(\varepsilon)$.)

(6.7) implies that $k<(1-H(p)+\varepsilon)n$, which is a contradiction. The proof of part (2) of Theorem 6.3.1 is complete.

Remark 6.3.1. It can be verified that the proof above can also work if the decoding error probability is bounded by $1-2^{-\beta n}$ (instead of the $1/2$ in part (2) of Theorem 6.3.1) for small enough $\beta=\beta(\varepsilon)>0$.

Next, we will prove part (1) of Theorem 6.3.1

6.3.2 Proof of Positive Part of Shannon's Theorem

Proof Overview. The proof of part (1) of Theorem 6.3.1 will be done by the probabilistic method (Section 3.2). In particular, we randomly select an encoding function $E:\{0,1\}^k\to\{0,1\}^n$. That is, for every $\mathbf{m}\in\{0,1\}^k$ pick $E(\mathbf{m})$ uniformly and independently at random from $\{0,1\}^n$. D will be the maximum likelihood decoding (MLD) function. The proof will have the following two steps:

- (Step 1) For any arbitrary $\mathbf{m}\in\{0,1\}^k$, we will show that for a random choice of E, the probability of failure, over BSC_p noise, is small. This implies the existence of a good encoding function for any arbitrary message.

- (Step 2) We will show a similar result for *all* **m**. This involves dropping half of the code words.

Note that there are two sources of randomness in the proof:

1. Randomness in the choice of encoding function E and

2. Randomness in the noise.

We stress that the first kind of randomness is for the probabilistic method while the second kind of randomness will contribute to the decoding error probability.

"Proof by picture" of Step 1**.** Before proving part (1) of Theorem 6.3.1, we will provide a pictorial proof of Step 1. We begin by fixing $\mathbf{m} \in \{0,1\}^k$. In Step 1, we need to estimate the following quantity:

$$\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right) \neq \mathbf{m}\right]\right].$$

By the additive Chernoff bound (Theorem 3.1.6), with all but an exponentially small probability, the received word will be contained in a Hamming ball of radius $\left(p+\varepsilon'\right)n$ (for some $\varepsilon' > 0$ that we will choose appropriately). So one can assume that the received word $\mathbf{y}$ with high probability satisfies $\Delta(E(\mathbf{m}),\mathbf{y}) \le (p+\varepsilon')n$. Given this, pretty much the only thing to do is to estimate the decoding error probability for such a $\mathbf{y}$. Note that by the fact that D is MLD, an error can happen only if there exists another message $\mathbf{m}'$ such that $\Delta(E(\mathbf{m}'),\mathbf{y}) \le \Delta(E(\mathbf{m}),\mathbf{y})$. The latter event implies that $\Delta(E(\mathbf{m}'),\mathbf{y}) \le (p+\varepsilon')n$ (see Figure 6.6). Thus, the decoding error probability is upper bounded by

$$\Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[E\left(\mathbf{m}'\right) \in B\left(\mathbf{y},\left(p+\varepsilon'\right)n\right)\right] = \frac{Vol_2\left(\left(p+\varepsilon'\right)n,n\right)}{2^n} \approx \frac{2^{H(p)n}}{2^n},$$

where the last step follows from Proposition 3.3.1. Finally, by the union bound (Proposition 3.1.3), the existence of such a "bad" $\mathbf{m}'$ is upper bounded by $\approx \frac{2^k 2^{nH(p)}}{2^n}$, which by our choice of k is $2^{-\Omega(n)}$, as desired.

The Details. For notational convenience, we will use $\mathbf{y}$ and $E\left(\mathbf{m}\right)+\mathbf{e}$ interchangeably:

$$\mathbf{y} = E\left(\mathbf{m}\right)+\mathbf{e}.$$

That is, $\mathbf{y}$ is the received word when $E\left(\mathbf{m}\right)$ is transmitted and $\mathbf{e}$ is the error pattern.

We start the proof by restating the decoding error probability in part (1) of Shannon's capacity theorem for BSC_p (Theorem 6.3.1) by breaking up the quantity into two sums:

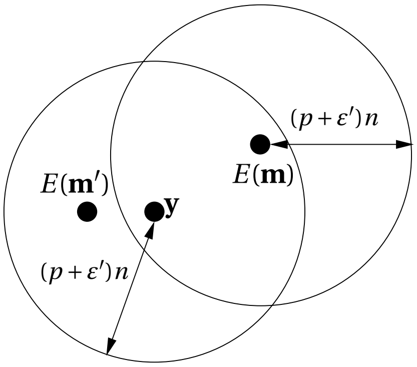

Figure 6.6: Hamming balls of radius $\left(p+\varepsilon'\right)n$ and centers $E(\mathbf{m})$ and $E(\mathbf{m}')$ illustrates Step 1 in the proof of part (1) of Shannon's capacity theorem for the BSC.

$$\Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq \mathbf{m}\right] = \sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right]\cdot \mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}} + \sum_{\mathbf{y}\notin B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right]\cdot \mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}},$$

where $\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}$ is the indicator function for the event that $D(\mathbf{y})\neq\mathbf{m}$ *given* that $E(\mathbf{m})$ was the transmitted codeword and we use $\mathbf{y}|E(\mathbf{m})$ as a shorthand for "$\mathbf{y}$ is the received word given that $E(\mathbf{m})$ was the transmitted codeword." As $\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\le 1$ (since it takes a value in $\{0,1\}$) and by the (additive) Chernoff bound (Theorem 3.1.6) we have

$$\Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq \mathbf{m}\right] \le \sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)} \Pr\left[\mathbf{y}|E(\mathbf{m})\right]\cdot \mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}} + e^{-(\varepsilon')^2 n/2}.$$

In order to apply the probabilistic method (Section 3.2), we will analyze the expectation (over the random choice of E) of the decoding error probability, which by the upper bound above satisfies

$$\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq \mathbf{m}\right]\right] \le e^{-\varepsilon'^2 n/2} + \sum_{\mathbf{y}\in B\left(E(\mathbf{m}),(p+\varepsilon')n\right)} \Pr_{\mathbf{e}\sim \mathrm{BSC}_p}\left[\mathbf{y}|E(\mathbf{m})\right]\cdot \mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right]. \quad (6.8)$$

In the above we used linearity of expectation (Proposition 3.1.2) and the fact that the distributions on $\mathbf{e}$ and E are independent.

Next, for a fixed received word $\mathbf{y}$ and the transmitted codeword $E(\mathbf{m})$ such that $\Delta(\mathbf{y},E(\mathbf{m}))\le(p+\varepsilon')n$ we estimate $\mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right]$. Since D is MLD, we have

$$\mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right] = \Pr_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}|E(\mathbf{m})\right] \le \sum_{\mathbf{m}'\neq\mathbf{m}} \Pr\left[\Delta\left(E\left(\mathbf{m}'\right),\mathbf{y}\right)\le\Delta\left(E\left(\mathbf{m}\right),\mathbf{y}\right)|E(\mathbf{m})\right], \quad (6.9)$$

where in the above "$|E(\mathbf{m})$" is short for "being conditioned on $E(\mathbf{m})$ being transmitted" and the inequality follows from the union bound (Proposition 3.1.3) and the fact that D is MLD.

Noting that $\Delta(E(\mathbf{m}'),\mathbf{y}) \le \Delta(E(\mathbf{m}),\mathbf{y}) \le (p+\varepsilon')n$ (see Figure 6.6), by (6.9) we have

$$\begin{aligned}
\mathbb{E}_E\left[\mathbb{1}_{D(\mathbf{y})\neq\mathbf{m}}\right] &\le \sum_{\mathbf{m}'\neq\mathbf{m}} \Pr\left[E\left(\mathbf{m}'\right)\in B\left(\mathbf{y},\left(p+\varepsilon'\right)n\right)|E(\mathbf{m})\right] \\
&= \sum_{\mathbf{m}'\neq\mathbf{m}} \frac{\left|B\left(\mathbf{y},\left(p+\varepsilon'\right)n\right)\right|}{2^n} \quad (6.10)\\
&\le \sum_{\mathbf{m}'\neq\mathbf{m}} \frac{2^{H(p+\varepsilon')n}}{2^n} \quad (6.11)\\
&< 2^k\cdot 2^{-n(1-H(p+\varepsilon'))} \\
&\le 2^{n(1-H(p+\varepsilon))-n(1-H(p+\varepsilon'))} \quad (6.12)\\
&= 2^{-n(H(p+\varepsilon)-H(p+\varepsilon'))}. \quad (6.13)
\end{aligned}$$

In the above (6.10) follows from the fact that the choice for $E(\mathbf{m}')$ is independent of $E(\mathbf{m})$. (6.11) follows from the upper bound on the volume of a Hamming ball (Proposition 3.3.1) while (6.12) follows from our choice of k.

Using (6.13) in (6.8), we get

$$\begin{aligned}
\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim\text{BSC}_p}\left[D(E(\mathbf{m})+\mathbf{e})\neq\mathbf{m}\right]\right] &\le e^{-\varepsilon'^2 n/2} + 2^{-n(H(p+\varepsilon)-H(p+\varepsilon'))}\sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)}\Pr\left[\mathbf{y}|E(\mathbf{m})\right] \\
&\le e^{-\varepsilon'^2 n/2} + 2^{-n(H(p+\varepsilon)-H(p+\varepsilon'))} \le 2^{-\delta' n}, \quad (6.14)
\end{aligned}$$

where the second inequality follows from the fact that

$$\sum_{\mathbf{y}\in B(E(\mathbf{m}),(p+\varepsilon')n)}\Pr\left[\mathbf{y}|E(\mathbf{m})\right] \le \sum_{\mathbf{y}\in\{0,1\}^n}\Pr\left[\mathbf{y}|E(\mathbf{m})\right] = 1$$

and the last inequality follows for large enough n, say $\varepsilon' = \varepsilon/2$ and by picking $\delta' > 0$ to be small enough. (See Exercise 6.3.)

Thus, we have shown that for any arbitrary $\mathbf{m}$ the average (over the choices of E) decoding error probability is small. However, we still need to show that the decoding error probability is exponentially small for *all* messages *simultaneously*. Towards this end, as the bound holds for each $\mathbf{m}$ we have

$$\mathbb{E}_{\mathbf{m}}\left[\mathbb{E}_E\left[\Pr_{\mathbf{e}\sim\text{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq\mathbf{m}\right]\right]\right] \le 2^{-\delta' n}.$$

The order of the summation in the expectation with respect to $\mathbf{m}$ and the summation in the expectation with respect to the choice of E can be switched (as the probability distributions are defined over different domains), resulting in the following expression:

$$\mathbb{E}_E\left[\mathbb{E}_{\mathbf{m}}\left[\Pr_{\mathbf{e}\sim\text{BSC}_p}\left[D\left(E\left(\mathbf{m}\right)+\mathbf{e}\right)\neq\mathbf{m}\right]\right]\right] \le 2^{-\delta' n}.$$

By the probabilistic method, there exists an encoding function E^* (and a corresponding decoding function D^*) such that

$$\mathbb{E}_{\mathbf{m}}\left[\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D^*\left(E^*(\mathbf{m})+\mathbf{e}\right)\neq\mathbf{m}\right]\right]\leq 2^{-\delta' n}. \tag{6.15}$$

(6.15) implies that the *average* decoding error probability is exponentially small. However, recall we need to show that the *maximum* decoding error probability is small. To achieve such a result, we will throw away half of the messages, i.e. *expurgate* the code. In particular, we will order the messages in decreasing order of their decoding error probability and then drop the top half. We claim that the maximum decoding error probability for the remaining messages is $2\cdot 2^{-\delta' n}$. Next, we present the details.

From Average to Worst-Case Decoding Error Probability. We begin with the following "averaging" argument.

Claim 6.3.2. *Let the messages be ordered as* $\mathbf{m}_1,\mathbf{m}_2,\ldots,\mathbf{m}_{2^k}$ *and define*

$$P_i=\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D(E(\mathbf{m}_i)+\mathbf{e})\neq\mathbf{m}_i\right].$$

Assume that $P_1\leq P_2\leq\ldots\leq P_{2^k}$ *and (6.15) holds, then* $P_{2^{k-1}}\leq 2\cdot 2^{-\delta' n}$

Proof. By the definition of P_i,

$$\begin{aligned}\frac{1}{2^k}\sum_{i=1}^{2^k}P_i&=\mathbb{E}_{\mathbf{m}}\Pr_{\mathbf{e}\sim\mathrm{BSC}_p}\left[D(E(\mathbf{m})+\mathbf{e})\neq\mathbf{m}\right]\\&\leq 2^{-\delta' n},\end{aligned}\tag{6.16}$$

where (6.16) follows from (6.15). For the sake of contradiction assume that

$$P_{2^{k-1}}>2\cdot 2^{-\delta' n}. \tag{6.17}$$

So,

$$\frac{1}{2^k}\sum_{i=1}^{2^k}P_i\geq\frac{1}{2^k}\sum_{i=2^{k-1}+1}^{2^k}P_i \tag{6.18}$$

$$>\frac{2\cdot 2^{-\delta' n}\cdot 2^{k-1}}{2^k} \tag{6.19}$$

$$>2^{-\delta' n}, \tag{6.20}$$

where (6.18) follows by dropping half the summands from the sum. (6.19) follows from (6.17) and the assumption on the sortedness of P_i. The proof is now complete by noting that (6.20) contradicts (6.16). □

Thus, our final code will have $\mathbf{m}_1, \dots, \mathbf{m}_{2^{k-1}}$ as its messages and hence, has dimension $k' = k-1$. Define $\delta = \delta' + \frac{1}{n}$. In the new code, maximum error probability is at most $2^{-\delta n}$. Also if we picked $k \le \lfloor (1 - H(p+\varepsilon))\, n \rfloor + 1$, then $k' \le \lfloor (1 - H(p+\varepsilon))\, n \rfloor$, as required. This completes the proof of Theorem 6.3.1.

We have shown that a random code can achieve capacity. However, we do not know of even a succinct representation of general codes. A natural question to ask is if random linear codes can achieve the capacity of BSC_p. The answer is yes: see Exercise 6.4.

For linear code, representation and encoding are efficient. But the proof does not give an explicit construction. Intuitively, it is clear that since Shannon's proof uses a random code it does not present an "explicit" construction. However, in this book, we will formally define what we mean by an explicit construction.

Definition 6.3.1. A code C of block length n is called *explicit* if there exists a poly(n)-time algorithm that computes a succinct description of C given n. For linear codes, such a succinct description could be a generator matrix or a parity check matrix.

We will also need the following stronger notion of an explicitness:

Definition 6.3.2. A linear $[n,k]$ code C is called *strongly explicit*, if given any index pair $(i,j) \in [k] \times [n]$, there is a poly($\log n$) time algorithm that outputs $G_{i,j}$, where G is a generator matrix of C.

Further, Shannon's proof uses MLD for which only exponential time implementations are known. Thus, the biggest question left unsolved by Shannon's work is the following.

Question 6.3.1. *Can we come up with an explicit construction of a code of rate* $1 - H(p+\varepsilon)$ *with efficient decoding and encoding algorithms that achieves reliable communication over* BSC_p*?*

As a baby step towards the resolution of the above question, one can ask the following question:

Question 6.3.2. *Can we come up with an explicit construction with* $R > 0$ *and* $p > 0$*?*

Note that the question above is similar to Question 2.7.1 in Hamming's world. See Exercise 6.13 for an affirmative answer.

6.4 Hamming vs. Shannon

As a brief interlude, let us compare the salient features of the works of Hamming and Shannon that we have seen so far:

HAMMING	SHANNON
Focus on codewords itself	Directly deals with encoding and decoding functions
Looked at explicit codes	Not explicit at all
Fundamental trade off: rate vs. distance (easier to get a handle on this)	Fundamental trade off: rate vs. error
Worst case errors	Stochastic errors

Intuitively achieving positive results in the Hamming world is harder than achieving positive results in Shannon's world. The reason is that the adversary in Shannon's world (e.g. BSC_p) is much weaker than the worst-case adversary in Hamming's world (say for bits). We make this intuition (somewhat) precise as follows:

Proposition 6.4.1. *Let $0 \le p < \frac{1}{2}$ and $0 < \varepsilon \le \frac{1}{2} - p$. If an algorithm A can handle $p + \varepsilon$ fraction of worst case errors, then it can be used for reliable communication over* BSC_p

Proof. By the additive Chernoff bound (Theorem 3.1.6), with probability $\ge 1 - e^{\frac{-\varepsilon^2 n}{2}}$, the fraction of errors in BSC_p is $\le p + \varepsilon$. Then by assumption on A, it can be used to recover the transmitted message. □

Note that the above result implies that one can have reliable transmission over BSC_p with any code of relative distance $2p + \varepsilon$ (for any $\varepsilon > 0$).

A much weaker converse of Proposition 6.4.1 is also true. More precisely, if the decoding error probability is exponentially small for the BSC, then the corresponding code must have constant relative distance (though this distance does not come even close to achieving say the Gilbert-Varshamov bound). For more see Exercise 6.5.

6.5 Exercises

Exercise 6.1. Let (E, D) be a pair of encoder and decoder that allows for successful transmission over BSC_p for every $p \le \frac{1}{2}$. Then there exists a pair (E', D') that allows for successful transmission over $\mathrm{BSC}_{p'}$ for any $p' > 1/2$. If D is (deterministic) polynomial time algorithm, then D' also has to be a (deterministic) polynomial time algorithm.

Exercise 6.2. Let (E, D) be a pair of encoder and decoder that allows for successful transmission over $q\mathrm{SC}_p$ for every $p \le 1 - \frac{1}{q}$. Then there exists a pair (E', D') that allows for successful transmission over $q\mathrm{SC}_{p'}$ for any $p' > 1 - \frac{1}{2}$. If D is polynomial time algorithm, then D' also has to be a polynomial time algorithm though D' can be a randomized algorithm even if D is deterministic.[3]

Exercise 6.3. Argue that in the positive part of Theorem 6.3.1, one can pick $\delta = \Theta(\varepsilon^2)$. That is, for $0 \le p < 1/2$ and small enough ε, there exist codes of rate $1 - H(p) - \varepsilon$ and block length n that can be decoded with error probability at most $2^{-\Theta(\varepsilon^2)n}$ over BSC_p.

[3] A randomized D' means that given a received word $\mathbf{y}$ the algorithm can use random coins and the decoding error probability is over both the randomness from its internal coin tosses as well as the randomness from the channel.

Exercise 6.4. Prove that there exists linear codes that achieve the BSC_p capacity. (Note that in Section 6.3 we argued that there exists not necessarily a linear code that achieves the capacity.)

Hint: Modify the argument in Section 6.3: in some sense the proof is easier.

Exercise 6.5. Prove that for communication on BSC_p, if an encoding function E achieves a maximum decoding error probability (taken over all messages) that is exponentially small, i.e., at most $2^{-\gamma n}$ for some $\gamma > 0$, then there exists a $\delta = \delta(\gamma, p) > 0$ such that the code defined by E has relative distance at least δ. In other words, good distance is *necessary* for exponentially small maximum decoding error probability.

Exercise 6.6. Prove that the capacity of the $q\mathrm{SC}_p$ is $1 - H_q(p)$.

Exercise 6.7. The binary erasure channel with erasure probability α has capacity $1-\alpha$. In this problem, you will prove this result (and its generalization to larger alphabets) via a sequence of smaller results.

1. For positive integers $k \le n$, show that less than a fraction q^{k-n} of the $k \times n$ matrices G over $\mathbb{F}_q$ fail to generate a linear code of block length n and dimension k. (Or equivalently, except with probability less than q^{k-n}, the rank of a random $k \times n$ matrix G over $\mathbb{F}_q$ is k.)

   Hint: Try out the obvious greedy algorithm to construct a $k \times n$ matrix of rank k. You will see that you will have many choices every step: from this compute (a lower bound on) the number of full rank matrices that can be generated by this algorithm.

2. Consider the q-ary erasure channel with erasure probability α ($q\mathrm{EC}_\alpha$, for some α, $0 \le \alpha \le 1$): the input to this channel is a field element $x \in \mathbb{F}_q$, and the output is x with probability $1-\alpha$, and an erasure '?' with probability α. For a linear code C generated by an $k \times n$ matrix G over $\mathbb{F}_q$, let $D : (\mathbb{F}_q \cup \{?\})^n \to C \cup \{\mathsf{fail}\}$ be the following decoder:

   $$D(\mathbf{y}) = \begin{cases} \mathbf{c} & \text{if } \mathbf{y} \text{ agrees with exactly one } \mathbf{c} \in C \text{ on the unerased entries in } \mathbb{F}_q \\ \mathsf{fail} & \text{otherwise} \end{cases}$$

   For a set $J \subseteq \{1, 2, \ldots, n\}$, let $P_{\mathrm{err}}(G|J)$ be the probability (over the channel noise and choice of a random message) that D outputs fail conditioned on the erasures being indexed by J. Prove that the average value of $P_{\mathrm{err}}(G|J)$ taken over all $G \in \mathbb{F}_q^{k\times n}$ is less than $q^{k-n+|J|}$.

3. Let $P_{\mathrm{err}}(G)$ be the decoding error probability of the decoder D for communication using the code generated by G on the $q\mathrm{EC}_\alpha$. Show that when $k = Rn$ for $R < 1 - \alpha$, the average value of $P_{\mathrm{err}}(G)$ over all $k \times n$ matrices G over $\mathbb{F}_q$ is exponentially small in n.

4. Conclude that one can reliably communicate on the $q\mathrm{EC}_\alpha$ at any rate less than $1-\alpha$ using a linear code.

Exercise 6.8. Consider a binary channel whose input/output alphabet is $\{0, 1\}$, where a 0 is transmitted faithfully as a 0 (with probability 1), but a 1 is transmitted as a 0 with probability $\frac{1}{2}$ and a 1 with probability $1/2$. Compute the capacity of this channel.

Hint: This can be proved from scratch using only simple probabilistic facts already stated/used in the book.

Exercise 6.9. Argue that Reed-Solomon codes from Chapter 5 are strongly explicit codes (as in Definition 6.3.2).

Exercise 6.10. In this problem we will prove a special case of the source coding theorem. For any $0 \le p \le 1/2$, let $\mathscr{D}(p)$ be the distribution on $\{0,1\}^n$, where each of the n bits are picked independently to be 1 with probability p and 0 otherwise. Argue that for every $\varepsilon > 0$, strings from $\mathscr{D}(p)$ can be compressed with $H(p+\varepsilon)\cdot n$ bits for large enough n.

More precisely show that for any constant $0 \le p \le 1/2$ and every $\varepsilon > 0$, for large enough n there exists an encoding (or compression) function $E : \{0,1\}^n \to \{0,1\}^*$ and a decoding (or decompression) function $D : \{0,1\}^* \to \{0,1\}^n$ such that[4]

1. For every $\mathbf{x} \in \{0,1\}^n$, $D(E(\mathbf{x})) = \mathbf{x}$, and
2. $\mathbb{E}_{\mathbf{x}\leftarrow\mathscr{D}(p)}[|E(\mathbf{x})|] \le H(p+\varepsilon)\cdot n$, where we use $|E(\mathbf{x})|$ to denote the length of the string $E(\mathbf{x})$. In other words, the *compression rate* is $H(p+\varepsilon)$.

Hint: Handle the "typical" strings from $\mathscr{D}$ and non-typical strings separately.

Exercise 6.11. Show that if there is a constructive solution to Shannon's channel coding theorem with E being a linear map, then there is a constructive solution to Shannon's source coding theorem in the case where the source produces a sequence of independent bits of bias p.

More precisely, let (E,D) be an encoding and decoding pairs that allows for reliable communication over BSC_p with exponentially small decoding error and E is a linear map with rate $1-H(p)-\varepsilon$. Then there exists a compressing and decompressing pair (E',D') that allows for compression rate $H(p)+\varepsilon$ (where compression rate is as defined in part 2 in Exercise 6.10). The decompression algorithm D' can be randomized and is allowed exponentially small error probability (where the probability can be taken over both the internal randomness of D' and $\mathscr{D}(p)$). Finally if (E,D) are both polynomial time algorithms, then (E',D') have to be polynomial time algorithms too.

Exercise 6.12. Consider a Markovian source of bits, where the source consists of a 6-cycle with three successive vertices outputting 0, and three successive vertices outputting 1, with the probability of either going left (or right) from any vertex is exactly 1/2. More precisely, consider a graph with six vertices $v_0, v_1, \ldots, v_5$ such that there exists an edge $(v_i, v_{(i+1) \bmod 6})$ for every $0 \le i \le 5$. Further the vertices v_i for $0 \le i < 3$ are labeled $\ell(v_i) = 0$ and vertices v_j for $3 \le j < 6$ are labeled $\ell(v_j) = 1$. Strings are generated from this source as follows: one starts with some *start* vertex u_0 (which is one of the v_i's): i.e. the start *state* is u_0. Any any point of time if the current state if u, then the source outputs $\ell(u)$. Then with probability 1/2 the states moves to each of the two neighbors of u.

Compute the optimal compression rate of this source.

[4] We use $\{0,1\}^*$ to denote the set of all binary strings.

Hint: Compress "state diagram" to a minimum and then make some basic observations to compress the source information.

Exercise 6.13. Given codes C_1 and C_2 with encoding functions $E_1 : \{0,1\}^{k_1} \to \{0,1\}^{n_1}$ and $E_2 : \{0,1\}^{k_2} \to \{0,1\}^{n_2}$ let $E_1 \otimes E_2 : \{0,1\}^{k_1 \times k_2} \to \{0,1\}^{n_1 \times n_2}$ be the encoding function obtained as follows: view a message $\mathbf{m}$ as a $k_1 \times k_2$ matrix. Encode the columns of $\mathbf{m}$ individually using the function E_1 to get an $n_1 \times k_2$ matrix $\mathbf{m}'$. Now encode the rows of $\mathbf{m}'$ individually using E_2 to get an $n_1 \times n_2$ matrix that is the final encoding under $E_1 \otimes E_2$ of $\mathbf{m}$. Let $C_1 \otimes C_2$ be the code associated with $E_1 \otimes E_2$ (recall Exercise 2.18).

For $i \geq 3$, let H_i denote the $[2^i - 1, 2^i - i - 1, 3]_2$-Hamming code. Let $C_i = H_i \otimes C_{i-1}$ with $C_3 = H_3$ be a new family of codes.

1. Give a lower bound on the relative minimum distance of C_i. Does it go to zero as $i \to \infty$?

2. Give a lower bound on the rate of C_i. Does it go to zero as $i \to \infty$?

3. Consider the following simple decoding algorithm for C_i: Decode the rows of the rec'd vector recursively using the decoding algorithm for C_{i-1}. Then decode each column according to the Hamming decoding algorithm (e.g. Algorithm 4). Let δ_i denote the probability of decoding error of this algorithm on the BSC_p. Show that there exists a $p > 0$ such that $\delta_i \to 0$ as $i \to \infty$.

   Hint: First show that $\delta_i \leq 4^i \delta_{i-1}^2$.

Exercise 6.14. We consider the problem of determining the best possible rate of transmission on a stochastic memoryless channel with *zero decoding error probability.* Recall that a memoryless stochastic channel is specified by a transition matrix $\mathbf{M}$ s.t. $\mathbf{M}(x,y)$ denotes the probability of y being received if x was transmitted over the channel. Further, the noise acts independently on each transmitted symbol. Let $\mathbb{D}$ denote the input alphabet. Let $R(\mathbf{M})$ denote the best possible rate for a code C such that there exists a decoder D such that for every $\mathbf{c} \in C$, $\Pr[D(\mathbf{y}) \neq \mathbf{c}] = 0$, where $\mathbf{y}$ is picked according to the distribution induced by $\mathbf{M}$ when $\mathbf{c}$ is transmitted over the channel (i.e. the probability that $\mathbf{y}$ is a received word is exactly $\prod_{i=1}^{n} \mathbf{M}(c_i, y_i)$ where C has block length n). In this exercise we will derive an alternate characterization of $R(\mathbf{M})$.

We begin with some definitions related to graphs $\mathcal{G} = (V,E)$. An *independent set* S of $\mathcal{G}$ is a subset $S \subseteq V$ such that there is no edge contained in S, i.e. for every $u \neq v \in S$, $(u,v) \notin E$. For a given graph $\mathcal{G}$, we use $\alpha(\mathcal{G})$ to denote the size of largest independent set in $\mathcal{G}$. Further, given an integer $n \geq 1$, the *n-fold product* of $\mathcal{G}$, which we will denote by $\mathcal{G}^n$, is defined as follows: $\mathcal{G}^n = (V^n, E')$, where $((u_1,\ldots,u_n),(v_1,\ldots,v_n)) \in E'$ if and only if for every $i \in [n]$ either $u_i = v_i$ or $(u_i, v_i) \in E$.

Finally, define a *confusion graph* $\mathcal{G}_{\mathbf{M}} = (V,E)$ as follows. The set of vertices $V = \mathbb{D}$ and for every $x_1 \neq x_2 \in \mathbb{D}$, $(x,y) \in E$ if and only if there exists a y such that $\mathbf{M}(x_1,y) \neq 0$ and $\mathbf{M}(x_2,y) \neq 0$.

1. Prove that

$$R(\mathbf{M}) = \lim_{n\to\infty} \frac{1}{n}\cdot\log_{|\mathbb{D}|}\left(\alpha\left(\mathscr{G}^n_{\mathbf{M}}\right)\right).^5 \tag{6.21}$$

2. A *clique cover* for a graph $\mathscr{G} = (V, E)$ is a partition of the vertices $V = \{V_1, \ldots, V_c\}$ (i.e. V_i and V_j are disjoint for every $i \neq j \in [c]$ and $\cup_i V_i = V$) such that the graph induced on V_i is a *complete graph* (i.e. for every $i \in [c]$ and $x \neq y \in V_i$, we have $(x, y) \in E$). We call c to be the *size* of the clique cover $V_1, \ldots, V_c$. Finally, define $\nu(\mathscr{G})$ to be the size of the smallest clique cover for $\mathscr{G}$. Argue that

$$\alpha(\mathscr{G})^n \le \alpha(\mathscr{G}^n) \le \nu(\mathscr{G})^n.$$

Conclude that

$$\log_{|\mathbb{D}|}\alpha(\mathscr{G}) \le R(\mathbf{M}) \le \log_{|\mathbb{D}|}\nu(\mathscr{G}). \tag{6.22}$$

3. Consider any transition matrix $\mathbf{M}$ such that the corresponding graph $\mathscr{C}_4 = \mathscr{G}_{\mathbf{M}}$ is a 4-cycle (i.e. the graph $(\{0,1,2,3\}, E)$ where $(i, i+1 \mod 4) \in E$ for every $0 \le i \le 3$). Using part 2 or otherwise, argue that $R(\mathbf{M}) = \frac{1}{2}$.

4. Consider any transition matrix $\mathbf{M}$ such that the corresponding graph $\mathscr{C}_5 = \mathscr{G}_{\mathbf{M}}$ is a 5-cycle (i.e. the graph $(\{0,1,2,4\}, E)$ where $(i, i+1 \mod 5) \in E$ for every $0 \le i \le 4$). Using part 2 or otherwise, argue that $R(\mathbf{M}) \ge \frac{1}{2}\cdot\log_5 5$. (This lower bound is known to be tight: see Section 6.6 for more.)

6.6 Bibliographic Notes

Shannon's results that were discussed in this chapter appeared in his seminal 1948 paper [63]. All the channels mentioned in this chapter were considered by Shannon except for the *BEC* channel, which was introduced by Elias.

The proof method used to prove Shannon's result for BSC_p has its own name– "random coding with expurgation."

Elias [17] answered Question 6.3.2 (the argument in Exercise 6.13 is due to him).

[5] In literature, $R(\mathbf{M})$ is defined with $\log_{|\mathbb{D}|}$ replaced by $\log_2$. We used the definition in (6.21) to be consistent with our definition of capacity of a noisy channel. See Section 6.6 for more.