Introduction to Coding Theory **CMU: Spring 2010**

Notes 3: Stochastic channels and noisy coding theorem bound

January 2010

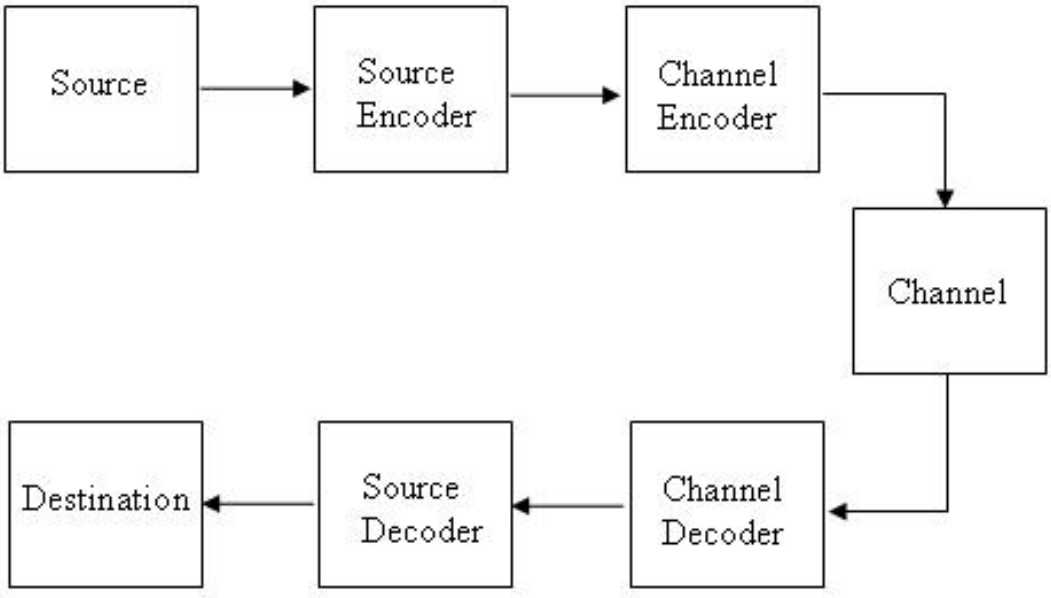
Lecturer: Venkatesan Guruswami *Scribe: Venkatesan Guruswami*

We now turn to the basic elements of Shannon's theory of communication over an intervening noisy channel.

1 Model of information communication and noisy channel

To quote Shannon from his paper *A Mathematical theory of communication*: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." The basic setup of the communication problem consists of a source that generates digital information which is to reliably communicated to a destination through a channel, preferably in the most efficient manner possible. This "destination" could be spatially separated (eg., a distant satellite is sending images of Jupiter back to the space station on Earth), or could be temporally separated (eg., we want to retrieve date stored on our hard disk at a later point of time).

The following is a schematic of the communication model:



The first step in the communication model is to exploit the redundancy in the output of the source and compress the information to economize the amount of "raw, non-redundant" data that must be transmitted across the channel. This data compression step in called *source coding.* If at each time step the source outputs an i.i.d copy of a random variable Z supported on a finite set $\mathcal{Z}$, then

Shannon's source coding theorem states that one can compress its output to $H(Z)$ bits per time step (on average, over n i.i.d samples from the source Z, as $n \to \infty$). In other words n samples from the source can be coded as one of $M \approx 2^{H(Z)n}$ possible outputs. Here $H(Z)$ is the fundamental Shannon entropy defined as

$$H(Z) = \sum_{z \in \mathcal{Z}} \mathbb{P}[Z = z] \log \frac{1}{\mathbb{P}[Z = z]} \ . \tag{1}$$

where log is to the base 2. Thus the entropy of a fair coin toss is 1, and that of a p-biased coin toss is $h(p) = -p \log p - (1-p) \log(1-p)$. The *source decoder* at the other end of the communication channel then decompresses the received information into (hopefully) the original output of the source.

The output of the source coder, say m, must then be communicated over a noisy channel. The channel's noisy behavior causes errors in the received symbols at the destination. To recover from the errors incurred due to the channel, one should *encode* the information output by the source coder by adding systematic redundancy to it. This is done through channel coding which maps m to a codeword c of some suitable error-correcting code (the study of channel coding will be our focus in this course).

1.1 Modeling the noisy channel

The basic channel model consists of an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. We will focus on *memoryless channels* — for each $x \in \mathcal{X}$ there is a distribution D_x on $\mathcal{Y}$ such that when input $x \in \mathcal{X}$ is fed at one end of the channel, the channel distorts it to $y \in \mathcal{Y}$ according to an independent sample drawn according to D_x. (In particular, the channel has no "state," and its behavior is independent of the history of previously transmitted symbols.) The collection of the distributions D_x comprise the "channel law" for the behavior of the channel. In a discrete memoryless channel (DMC), given by a triple $\Lambda = (\mathcal{X}, \mathcal{Y}, \Pi)$, the input and output alphabets $\mathcal{X}, \mathcal{Y}$ are finite, and therefore the channel law can be specified by a $|\mathcal{X}| \times |\mathcal{Y}|$ conditional probability matrix Π which is a stochastic matrix where each row sums to 1:

$$|\mathcal{X}| \left\{ \overbrace{\left(\quad \Pi(y|x) \quad \right)}^{|\mathcal{Y}|} \right.$$

The (x, y)'th entry $\Pi(y|x)$ is the conditional probability $\mathbb{P}(Y = y | X = x)$ of the receiving y when x was transmitted on the channel.

1.2 Noisy coding and joint source-channel coding theorems

Suppose at the output of the source coder, we have a message m from one of $M \approx 2^{H(Z)n}$ possible messages (that encode n samples from the source Z), which is to be communicated across a channel $(\mathcal{X}, \mathcal{Y}, \Pi)$. Then the channel encoder it into a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in C \subseteq \mathcal{X}^n$ for some error-correcting code C and the information is sent via n uses of the channel. At the other end, a sequence $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \mathcal{Y}^n$ is received with conditional probability

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} \Pi(y_i|x_i) \tag{2}$$

(due to the memoryless nature of the channel). The decoder must then map this sequence $\mathbf{y}$ into a legal codeword $c \in C$ (or equivalently into a message $m \in \mathcal{M}$).

A piece of notation: For a DMC $(\mathcal{X}, \mathcal{Y}, \Pi)$, a positive integer n, and $\mathbf{x} \in \mathcal{X}^n$, let us denote by $\Pi(\mathbf{x})$ the above distribution (2) on $\mathcal{Y}^n$ induced by the Π on input sequence $\mathbf{x}$.

Theorem 1 (Shannon's noisy coding theorem) *For every discrete memoryless channel $\Lambda = (\mathcal{X}, \mathcal{Y}, \Pi)$, there exists a real number $C_0 = C_0(\Lambda)$ called its channel capacity, such that the following holds for every $R < C_0$. For all large enough n, there exists an integer $M \geq 2^{Rn}$ and*

1. *an encoding map* $\mathrm{Enc} : \{1, 2, \ldots, M\} \to \mathcal{X}^n$ *(of some error-correcting code over alphabet $\mathcal{X}$ of rate $R/\log|\mathcal{X}|$), and*
2. *a decoding map* $\mathrm{Dec} : \mathcal{Y}^n \to \{1, 2, \ldots, M\} \cup \{\text{fail}\}$

such that for every $m \in \{1, 2, \ldots, M\}$

$$\mathbb{P}_{\Pi}[\mathrm{Dec}(\Pi(\mathrm{Enc}(m))) = m] \geq 1 - 2^{-\Omega_{R,C_0}(n)}$$

where the probability is over the behavior of the channel Π (on input $\mathrm{Enc}(m)$*).*

Further, the capacity C_0 is given by the expression

$$\max_{p \in \mathrm{Dist}_{\mathcal{X}}} H(Y) - H(Y|X)$$

where the maximum is taken over all probability distributions p on $\mathcal{X}$. In the above, $H(Y)$ is the entropy of the $\mathcal{Y}$-valued random variable Y with distribution function

$$\mathbb{P}[Y = y] = \sum_{x \in \mathcal{X}} \mathbb{P}(Y = y|X = x)p(x) = \sum_{x \in \mathcal{X}} \Pi(y|x)p(x)$$

and $H(Y|X)$ is the conditional entropy

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x)H(Y|X = x) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} \Pi(y|x) \log \frac{1}{\Pi(y|x)} \ .$$

Remark 2 *The quantity $H(Y) - H(Y|X)$ is called the mutual information between X and Y, and denoted $I(X,Y)$. It represents the decrease in uncertainty of a random variable Y given the knowledge of random variable X, which intuitively captures how much information X reveals about Y. If Y is independent of X, then $H(Y|X) = H(Y)$, and $I(X,Y) = 0$. On the other hand if $Y = f(X)$ for some function f (i.e., Y is determined by X), then $H(Y|X) = 0$ and $I(X,Y) = H(Y)$.*

Combining Shannon's source coding and noisy coding theorems, and the two-stage communication process comprising a separate source coding stage followed by channel coding stage, one can conclude that reliable communication of the output of a source Z on a noisy channel Λ is possible as long as $H(Z) < C_0(\Lambda)$, i.e., the source outputs data at a rate that is less than the capacity of the channel. This result has a converse (called the converse to the joint source-channel coding theorem) that says that if $H(Z) > C_0(\Lambda)$ then reliable communication is not possible.

Together, these imply a "separation theorem," namely that it is information-theoretically optimal to do source and channel coding separately, and thus one can gain modularity in communication system design without incurring any loss in rate of data transfer. While this converse to the joint source-channel coding theorem is rather intuitive in the setting of point-to-point communication between a sender and a receiver, it is worth remarking that the separation theorem breaks down in some scenarios with multiple users and correlated sources.

We will not prove Shannon's theorem in the above generality here, but content ourselves with establishing a special case (for the binary symmetric channel). The proof for the general case follows the same general structure once some basic information theory tools are set up, and we will remark briefly about this at the end. But first we will see some important examples of noisy channels.

2 Examples of channels

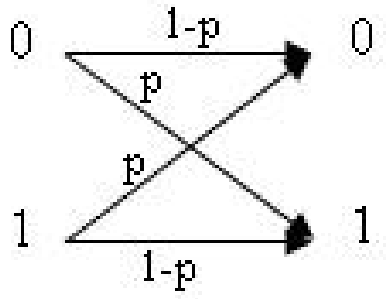
A discrete channel with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively, specified by the conditional probability matrix $\Pi(y|x)$, can also be represented pictorially by an input-output diagram, which is a bipartite graph with nodes on left identified with $\mathcal{X}$ and nodes on right identified with $\mathcal{Y}$ and a directed edge between $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ with weight $\Pi(y|x)$.

2.1 Binary symmetric channel

The *Binary Symmetric Channel* (BSC) has input alphabet $\mathcal{X} = \{0,1\}$ and output alphabet $\mathcal{Y} = \{0,1\}$. The BSC is parameterized by a real number p, $0 \leq p \leq 1/2$ called the *crossover probability*, and often denoted BSC_p. The channel flips its input with probability p, in other words,

$$\Pi(y|x) = \begin{cases} p & \text{if } y = x \\ 1-p & \text{if } y = 1-x \end{cases}$$

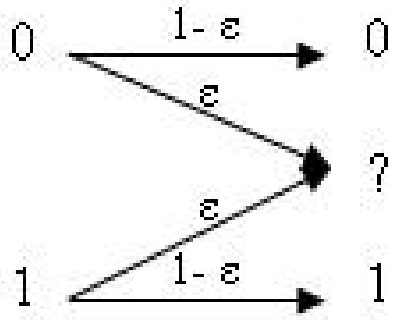
Pictorially, BSC_p can be represented as



If a uniform input $X \in \{0,1\}$ is fed as input to BSC_p, then the output Y is also uniformly distributed. Given $X = x$, Y is distributed as a p-biased coin, and $H(Y|X = x) = h(p)$. Thus $H(Y|X) = h(p)$, and therefore $I(X,Y) = H(Y) = H(Y|X) = 1 - h(p)$. It can be checked that the uniformly distributed X maximizes $I(X,Y)$, and so Shannon's theorem implies that $1 - h(p)$ is the capacity of BSC_p. We will shortly prove this special case of Shannon's theorem.

2.2 Binary erasure channel

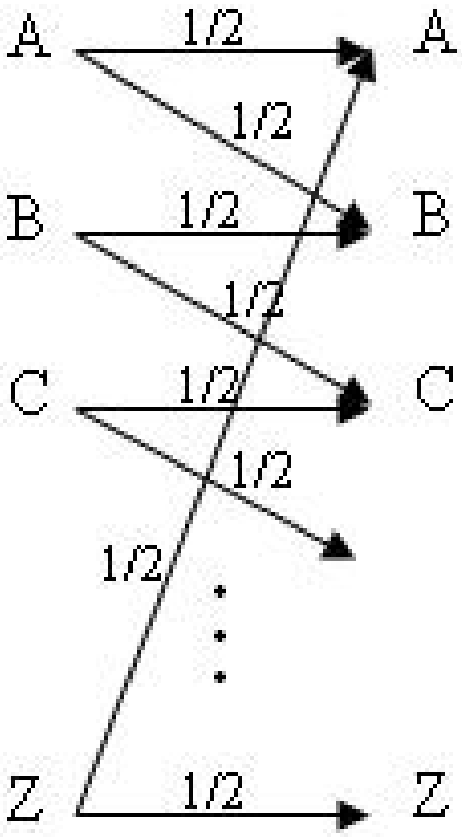
The Binary Erasure Channel (BEC) is parameterized by a real ϵ, $0 \leq \epsilon \leq 1$, which is called the *erasure probability*, and is denoted BEC_ϵ. Its input alphabet is $\mathcal{X} = \{0,1\}$ and output alphabet is $\mathcal{Y} = \{0,1,?\}$. Upon input $x \in \mathcal{X}$, the channel outputs x with probability $1 - \epsilon$, and outputs ? (corresponding to erasing the symbol) with probability ϵ. (It never flips the value of a bit.) Pictorially:



When a bit string of length n for large n is transmitted across BEC_ϵ, with high probability only $\approx (1-\epsilon)n$ bits are received unerased at the other end. This suggests that the maximum rate at which reliable communication is possible is at most $1 - \epsilon$. It turns out that a rate approaching $1 - \epsilon$ can be achieved, and the capacity of BEC_ϵ equals $1 - \epsilon$.

2.3 Noisy Typewriter Channel

The noisy typewriter channel is given by the following diagram:



If we restrict the code to send only one of the symbols $\{A, C, E, \ldots, Y\}$ in each channel use, we can communicate one of 13 possible messages with **zero** error. Therefore the capacity of the channel is at least $\log_2 13$. One can prove that this rate is the maximum possible and the capacity of the channel is exactly $\log 13$. (Indeed, this follows from Shannon's capacity formula: Since $|\mathcal{Y}| = 26$, $H(Y)$ is at most $\log 26$. Also $H(Y|X) = 1$ for every distribution of the channel input X. Hence $H(Y) - H(Y|X) \leq \log 13$.)

Note that we can achieve a rate equal to capacity with *zero* probability of miscommunication. For the BSC_p with $p > 0$ on the other hand, zero error communication is not possible at *any* positive rate, since for every pair of strings $x, x' \in \{0,1\}^n$, there is a positive probability that x will get distorted to x' by the noise caused by the BSC_p.

The study of zero error capacity of channels was introduced in another classic work of Shannon. Estimating the zero error capacity of even simple channels (such as the 5-cycle) has led to some beautiful results in combinatorics, including Lovász's celebrated work on the Theta function.

2.4 Continuous Output Channel

We now see an example of a continuous output channel that is widely used to model noise and compare the performance (typically via simulation) of different coding schemes. The binary input

additive white Gaussian noise (BIAWGN) channel has input alphabet $\mathcal{X} = \{1, -1\}$ (it is more convenient to encode binary symbols by ± 1 instead of $\{0, 1\}$) and output alphabet $\mathcal{Y} = \mathbb{R}$. The input $x \in \{1, -1\}$ is "modulated" into the real number βx and the channel adds additive noise distributed according to $N(0, \sigma^2)$ to βx. Thus the output distribution is a Gaussian with mean βx and variance σ^2. Formally

$$\mathbb{P}[Y \leq y | X = x] = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z-\beta x)^2/(2\sigma^2)} \, dz \ .$$

The quantity $(\beta/\sigma)^2$ is commonly referred to as the *signal-to-noise ratio* (SNR for short), with β^2 corresponding to the energy per input bit and σ^2 corresponding to the amount of noise. The SNR is usually measured in decibel units (dB), and expressed as the value $10 \log_{10}(\beta/\sigma)^2$. As one might expect, the capacity of the AWGN channel increases as its SNR increases.

3 Shannon's capacity theorem for the binary symmetric channel

We now turn to establishing the capacity of BSC_p to be $1 - h(p)$.

3.1 Connection to minimum distance

First, let us connect this question to the Hamming world. If we have a family of binary codes of relative distance more than $(2p + \epsilon)$, then we claim that this enables communicating on the BSC_p with exponentially small probability of miscommunication. The reason is that by the Chernoff bound for independent Bernoulli random variables (stated below), the probability that at least $(p + \epsilon/2)n$ are corrupted out of n bits transmitted on a BSC_p is exponentially small. When the number of errors is less than $(p+\epsilon/2)n$, the received word has a unique closest codeword in Hamming distance, which is also the original transmitted codeword.

Lemma 3 (Chernoff bound for i.i.d. Bernoulli random variables) *If $X_1, X_2, \ldots X_n$ are i.i.d. $\{0, 1\}$-valued random variables with $\mathbb{P}[X_i = 1] = p$, then for every $\epsilon > 0$, for large enough n the following tail estimates hold:*

$$\mathbb{P}[\sum_{i=1}^{n} X_i \geq (p + \epsilon)n] \leq 2^{\frac{-\epsilon^2 n}{3}}$$

$$\mathbb{P}[\sum_{i=1}^{n} X_i \leq (p - \epsilon)n] \leq 2^{\frac{-\epsilon^2 n}{3}}$$

Together with the Gilbert-Varshamov bound, we conclude the existence of codes of rate at least $1 - h(2p)$ for reliable communication on BSC_p. This rate is positive only for $p < 1/4$, and falls short of the bound $1 - h(p)$ which we "know" to be the capacity of BSC_p from Shannon's general theorem.

The Hamming upper bound on rate for codes of relative distance $2p$ was also equal to $1 - h(p)$. So if the Hamming bound could be attained, we could achieve the capacity of BSC_p simply by using

codes of relative distance $2p$. However, we will soon see that the Hamming upper bound can be improved, and there are no codes of positive rate for relative distance $2p$ for $p \geq 1/4$ or of rate $1 - h(p)$ for $p < 1/4$.

3.2 Intuition: mostly disjoint packings

The key to Shannon's theorem is that we do not need every pair of codewords to differ in a fraction $2p$ of locations, but only that for *most* (as opposed to for all) points obtained by flipping about a p fraction of bits of a codeword c have no other codeword closer than c. In other words, it suffices to be able to pack $\approx 2^{(1-h(p))n}$ "mostly-disjoint" Hamming balls of radius pn so that most points in $\{0,1\}^n$ belong to at most one such Hamming ball. Indeed, we will show below (Theorem 4) that such a packing exists, and therefore one can reliably communicate on BSC_p with rate approaching $1 - h(p)$.

The intuition for the case of general discrete memoryless channels as stated in Theorem 1 is similar. For a typical sequence $x \in \mathcal{X}^n$ (chosen according to the product distribution $p^{\otimes n}$), when x is transmitted, there are $\approx 2^{H(Y|X)n}$ possible received sequences in $\mathcal{Y}^n$ (call this the "neighborhood" of x), out of a total volume of $2^{H(Y)n}$. It turns out it is possible to pick a collection of $\approx 2^{(H(Y)-H(Y|X))n}$ sequences in $\mathcal{X}^n$ whose neighborhoods are mostly disjoint. This enables reliable communication at rate approaching $H(Y) - H(Y|X)$.

3.3 Converse to capacity theorem for BSC

We now give an explanation for why $1 - h(p)$ ought to be an *upper bound* on capacity of the BSC_p. Suppose a code $C \subset \{0,1\}^n$ achieves negligible error probability for communication on BSC_p with some decoding rule $D : \{0,1\}^n \to C \cup \{\text{fail}\}$. When c is transmitted, with overwhelming probability the received word belongs to a set Typical_c of $\approx 2^{h(p)n}$ possible strings whose Hamming distance to c is close to pn (say in the range $[(p - o(1))n, (p + o(1))n]$, and these possibilities are all roughly equally likely. Therefore, in order to ensure that c is recovered with high probability from its noisy version, the decoding rule D must map most of the strings in Typicakl_c to c. Thus we must have $|D^{-1}(c)| \approx 2^{h(p)n}$ for each $c \in C$, leading to the upper bound $|C| \leq 2^{(1-h(p)+o(1))n}$.

A different way to argue about the $1 - h(p)$ upper bound is related to a discussion in our very first lecture. It is based on the observation that when communication is successful, the decoder not only recovers the transmitted codeword but also the locations of the (typically around pn) errors. The former carries $\log |C|$ bits of information, whereas the latter typically conveys $\approx h(p)n$ bits of information. Since the total amount of non-redundant information that can be reliably conveyed by n bits cannot exceed n, we again get the upper bound $|C| \leq 2^{(1-h(p)+o(1))n}$.

Exercise 1 *Develop the above arguments into a formal proof that communication at a rate of $1 - h(p) + \epsilon$ on BSC_p incurs a probability of error bounded below by an absolute constant, and in fact by $1 - 2^{-\Omega_{\epsilon,p}(n)}$ where n is the block length of the code.*

3.4 The theorem

We conclude these notes with the formal statement and proof of the capacity theorem for BSC_p.

Theorem 4 *For every $p \in (0, 1/2)$ such that $0 \leq p < \frac{1}{2}$, and every $0 < \gamma < 1/2 - p$ and all large enough integers n, there exists a $\delta = \delta(\gamma, p)$ and a code with encoding map $\mathrm{Enc} : \{0,1\}^k \to \{0,1\}^n$ for $k = (1 - h(p+\gamma))n$ and a decoding rule $D : \{0,1\}^n \to \{0,1\}^k \cup \{\mathrm{fail}\}$ such that*

$$\mathbb{P}_z[D(E(m) + z) = m] \geq 1 - 2^{-\delta n}$$

where the probability is over the noise z caused by BSC_p.

PROOF: The construction is by the probabilistic method. Let $\ell = k + 1$. The encoding function $\mathrm{Enc} : \{0,1\}^\ell \to \{0,1\}^n$ is chosen uniformly at random from all possible functions. In other words, for every message $m \in \{0,1\}^\ell$, the corresponding codeword, $\mathrm{Enc}(m)$ is chosen uniformly at random from $\{0,1\}^n$. (Note that this might assign the same codeword to two different messages but this (tiny) probability will be absorbed into the decoding error probability.)

Pick $\epsilon = \epsilon(\gamma) > 0$ to be a small enough constant. The decoding function D is defined as follows: $D(y) = m$ if $\mathrm{Enc}(m)$ is the unique codeword such that $\Delta(y, \mathrm{Enc}(m)) \leq (p + \epsilon)n$ and $D(y) = \mathrm{fail}$ otherwise.

For $z \in \{0,1\}^n$, let $\mathrm{prob}(z)$ denote the probability that the noise caused by BSC_p on input the all 0's vector equals z (note that $\mathrm{prob}(z) = p^{\mathrm{wt}(z)}(1-p)^{n-\mathrm{wt}(z)}$).

Fix a message m. For each possible Enc, the probability that $D(\mathrm{Enc}(m) + z) \neq m$ taken over the noise z caused by BSC_p is at most

$$\begin{aligned}
\mathbb{P}_z[D(\mathrm{Enc}(m) + z) \neq m)] &\leq \mathbb{P}_z[\mathrm{wt}(z) > (p+\epsilon)n] + \sum_{z \in B(0,(p+\epsilon)n)} \mathrm{prob}(z)\mathbf{1}(D(\mathrm{Enc}(m) + z) \neq m) \\
&\leq 2^{-\Omega(\epsilon^2 n)} + \sum_{z \in B(0,(p+\epsilon)n)} \mathrm{prob}(z) \sum_{m' \neq m} \mathbf{1}(\Delta(\mathrm{Enc}(m) + z, \mathrm{Enc}(m')) \leq (p+\epsilon)n)
\end{aligned}$$

where the notation $\mathbf{1}(E)$ stands for the indicator random variable of the event E, the first estimate follows from the Chernoff bound, and the second estimate because when the decoding is unsuccessful when at most $(p + \epsilon)n$ errors occur, there must be some other codeword besides $\mathrm{Enc}(m)$ that is close to the received word $\mathrm{Enc}(m) + z$.

Now let us bound the expected value of this probability of miscommunication over the random choice of Enc. For each fixed z, and $m \neq m'$,

$$\begin{aligned}
\mathbb{E}_{\mathrm{Enc}}[\mathbf{1}(\Delta(\mathrm{Enc}(m) + z, \mathrm{Enc}(m')))] &\leq \mathbb{P}_{\mathrm{Enc}}[\Delta(\mathrm{Enc}(m) + z, \mathrm{Enc}(m')) \leq (p+\epsilon)n] \\
&= \frac{\mathrm{Vol}(n, (p+\epsilon)n)}{2^n} \\
&\leq 2^{-(1-h(p+\epsilon))n}
\end{aligned}$$

Therefore, by linearity of expectation

$$\begin{aligned}
\mathbb{E}_{\mathrm{Enc}}\mathbb{P}_z[D(\mathrm{Enc}(m) + z) \neq m)] &\leq 2^{-\Omega(\epsilon^2 n)} + \sum_{z \in B(0,(p+\epsilon)n)} \mathrm{prob}(z) 2^\ell 2^{-(1-h(p+\epsilon))n} \\
&\leq 2^{-\Omega(\epsilon^2 n)} + 2 \cdot 2^{-(h(p+\gamma) - h(p+\epsilon))n} < \frac{1}{2} \cdot 2^{-\delta n}
\end{aligned}$$

for some $\delta = \delta(\gamma, p) > 0$ when ϵ is chosen small enough.

We can conclude from the above for each fixed m that the probability over Enc that the error probability in communicating m (over the channel noise) exceeds $2^{-\delta n/2}$ is $2^{-\delta n/2}$. We would like to find an encoding Enc for which the error probability is low for every m simultaneously. The bound is too weak to do a union bound over all 2^ℓ messages. So we proceed as follows.

Since $\mathbb{E}_{\text{Enc}}\mathbb{P}_z[D(\text{Enc}(m) + z) \neq m)] < 2^{-1-\delta n}$ for each fixed m, this also holds on average over all choices of m. That is

$$\mathbb{E}_m\mathbb{E}_{\text{Enc}}\mathbb{P}_z[D(\text{Enc}(m) + z) \neq m)] < 2^{-1-\delta n} .$$

Changing the order of expectations

$$\mathbb{E}_{\text{Enc}}\mathbb{E}_m\mathbb{P}_z[D(\text{Enc}(m) + z) \neq m)] < 2^{-1-\delta n} .$$

Therefore there must exist an encoding Enc^* for which

$$\mathbb{E}_m\mathbb{P}_z[D(\text{Enc}^*(m) + z) \neq m)] < 2^{-1-\delta n} .$$

By an averaging argument, for at most $1/2$ the messages $m \in \{0,1\}^\ell$ one can have $\mathbb{P}_z[D(\text{Enc}^*(m) + z) \neq m)] \geq 2^{-\delta n}$. Expurgating these messages, we get an encoding $\text{Enc}' : \{0,1\}^{\ell-1} \to \{0,1\}^n$ and a decoding function D' such that for every $m' \in \{0,1\}^k$, $\mathbb{P}_z[D'(\text{Enc}'(m') + z) \neq m'] < 2^{-\delta n}$. This finishes the proof of the theorem. □

We remark that neither the encoding function nor the decoding function in the above proof are efficiently computable. The challenge put forth by Shannon's work is to "constructivize" his result and find explicit codes with polynomial time encoding and decoding that achieve capacity.

Exercise 2 *Prove that Theorem 4 also holds with a linear code, and that a random linear code achieves capacity of the BSC with high probability. (In fact, the proof becomes easier in this case, as no expurgation is needed at the end.)*

We end these notes by noting another connection between the Shannon and Hamming worlds. Though minimum distance is not the governing factor for achieving capacity on the BSC, a large minimum distance is necessary to to have a positive error exponent (i.e., achieve exponentially small error probability). We leave it as an exercise to justify this claim.