CS 229r Essential Coding Theory	January 29, 2020
Lecture 2	
Instructor: Madhu Sudan	*Scribe: Alec Sun*

1 Introduction

The zero-th problem set will be due on Friday. It is worth zero credit, but you are expected to work on the problems. Feedback on solution write-ups will likely be provided.

Today we will talk about the following topics:

- We will talk about Hamming's contributions, in particular the notion of distance in a code, as well as thinking of such a code as a set of balls. We will also talk about bounds on rate for these codes. All of these things came out of a single paper by Hamming.
- Then we will talk about Shannon's contributions. We will introduce the concept of Shannon capacity and also prove a converse to Shannon's Theorem.

2 Distances in Hamming Codes

Last class we represented a coding function as

$$E : \Sigma^k \to \Sigma^n,$$

where Σ is the finite set, also known as the **alphabet**, in question. Denote the size $|\Sigma|$ of the alphabet by q. We also considered a decoding function

$$\Sigma^n \to \Sigma^k.$$

We will not stress whether or not D serves to detect errors or correct errors today. Recall the following definitions.

Definition 1. *A **code** is defined as a set*

$$C = \left\{ E(m) \mid m \in \Sigma^k \right\}.$$

Definition 2. *An element of Σ^k is called a **message**. The message length k is always relative to the size of the alphabet Σ.*

Definition 3. *We denote by n the **block length**, or simply **length**.*

Definition 4. *Define*

$$\Delta(x, y) = \#\{\text{coordinates where } x \text{ and } y \text{ differ}\}.$$

This is a valid metric, in particular, it satisfies the triangle inequality.

Definition 5. *Define the* ***distance*** *of a code C as*

$$d = \Delta(C) = \min_{\substack{x,y \in C \\ x \neq y}} \{\Delta(x,y)\}.$$

The four parameters n, k, d, q in the definition of a code above define a $(n,k,d)_q$-code. If q is suppressed, it is assumed that the code is binary.

Note 6. *Here is a special case of a $(n,k,d)_q$-code. If Σ is a field and the code C is linear, namely the encoding function E is linear, then we use square brackets and call C a $[n,k,d]_q$-code.*

One combinatorial question we will ask is the following: Which $(n,k,d)_q$ codes are achievable?

Remark Here are some easy observations:

1. A $(n,k,d)_q$-code can be extended to a $(n+1,k,d)_q$-code.
2. A $(n,k,d)_q$-code can be modified to a $(n,k-1,d)_q$-code.
3. A $(n,k,d)_q$-code can be modified to a $(n,k,d-1)_q$-code.

There is no general monotonicity in the achievability of q. In general, one might expect that a larger q is easier to achieve. Hence, one might desire to construct codes with q small. As for the other parameters, the general goal will be to minimize n, maximize k, maximize d, and minimize q.

The Hamming code we constructed last class produced a family of codes depending on a parameter ℓ with $n = 2^\ell - 1, k = 2^\ell - \ell - 1$. In particular, we showed the existence of a $[n,k,3]_2$-code. Put another way, for infinitely many n, there exists a $[n, n - \log(n+1), 3]_2$ code. The size of the code is

$$|C| = 2^{n - \log(n+1)} = \frac{2^n}{n+1}.$$

Remark Why is the constant 3 here? We stated that a t-error correcting code is equivalent to a $2t$-error detecting code last time. This in turn implies a distance of at least $2t+1$ in order to detect for $2t$ errors. Hence for the 1-error correcting code we presented last time, the distance is at least 3.

3 Balls in Hamming Codes

We will define the concept of a **ball** in order to prove an upper bound on the size of a code $C \subseteq \mathbb{F}_2^n$.

Definition 7. *Define a ball of radius r around x to be*

$$Ball(x,r) = \{y \in \Sigma^n \mid \Delta(x,y) \leq r\}.$$

Define the ***volume*** *of the ball as the number of points in the ball, namely*

$$Vol_q(n,r) = |Ball(x,r)|.$$

In this definition we have implicitly used the fact that all balls with the same radius have the same volume, hence we can remove the dependence of Vol *on the center of the ball x.*

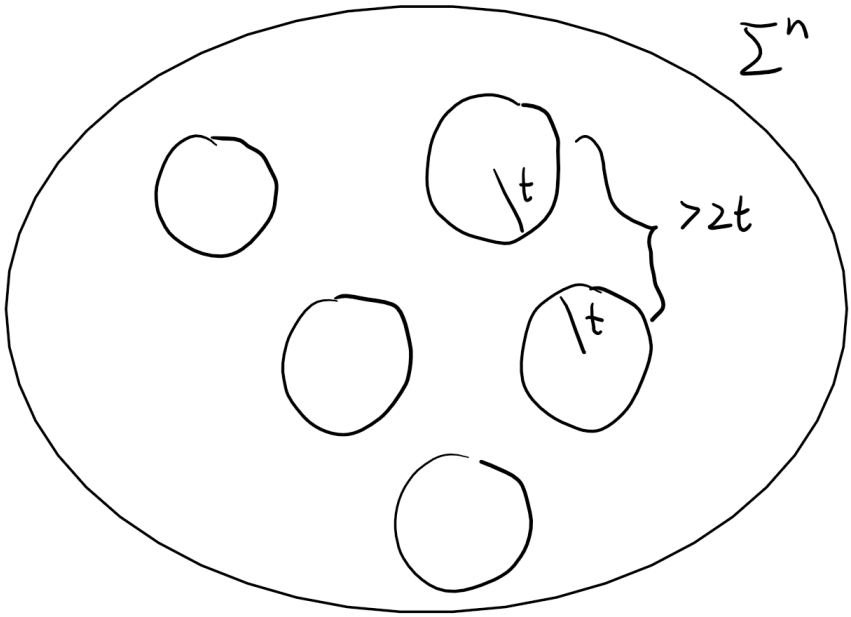


Figure 1: In the Hamming code context, the general picture to keep in mind is that of a disjoint set of balls in Σ^n. A set of balls with radii t have to be disjoint in order to correct t errors.

Lemma 8 (Hamming Bound)**.** *Let C be a code a distance d in Σ^n. Then*

$$|C| \leq \frac{|\Sigma|^n q^n}{Vol_q\left(n, \lfloor \frac{d-1}{2} \rfloor\right)}.$$

Example 9. *Consider the specific case of distance $d \geq 3$. Note that $Vol_2(n,1) = n+1$ because there are n bits that could be switched, and we also have to count the center of the ball. Hence the lemma tells us that if $C \subseteq \mathbb{F}_2^n$ is a code of distance at least 3, then*

$$|C| \leq \frac{2^n}{n+1}.$$

In other words, the Hamming code we constructed last class is optimal with respect to size of the code.

So far we have constructed the following explicit examples of $[n,k,d]_q$-codes:

- A $[n,n,1]_2$-code exists by simply considering the identity function $E : \{0,1\}^n \to \{0,1\}^n$, and this is the best possible.
- A $[n,n-1,2]_2$-code exists by defining

$$E : \{0,1\}^{n-1} \to \{0,1\}^n$$
$$E(m_1 \circ \cdots \circ m_{n-1}) = m_1 \circ \cdots \circ m_{n-1} \circ \left(\oplus_{i=1}^{n-1} m_i\right)$$

  and this is the best possible.

- A $[n, n - \log(n+1), 3]_2$-code exists, and this is the best possible.

One initially surprising fact is that a $[n, n - \log(n+1) - 1, 4]_2$-code exists by simply adding a parity check bit to the distance 3 code. The paradigm is that the appearance of the expression

$$\left\lfloor \frac{d-1}{2} \right\rfloor$$

in the Hamming bound

$$|C| \leq \frac{|\Sigma|^n q^n}{\mathrm{Vol}_q\left(n, \left\lfloor \frac{d-1}{2} \right\rfloor\right)}.$$

in the Hamming Bound is ingrained in the sense that k decreases by a significant amount in terms of n only after every pair of distances d.

In general, there is the following lemma.

Lemma 10. *If d is odd, then a $(n, k, d)_2$-code can be modified to create a $(n+1, k, d+1)_2$-code.*

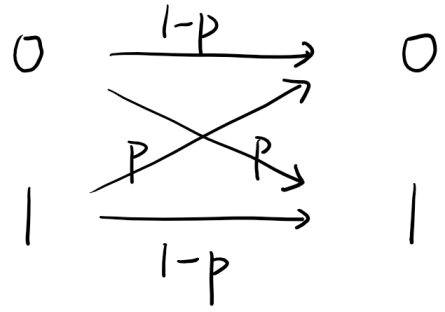
Proof of Lemma 10: See the exercises. □

Note 11. *Later in the course, we will prove that $\left[n, n - \frac{d}{2}\log n, d\right]_2$-codes exist. We will also spend a lot of time later in the course talking about efficiency in the encoding and the decoding functions.*

4 Binary Symmetric Channel

We now turn to a very different foundational paper by Shannon. Shannon's theory differs from Hamming's theory in the sense that Hamming explored the adversarial side of errors in codes while Shannon explored the probabilistic errors. In the Hamming model, it is necessary that every two encodings differ by greater than $2t$ to correct t errors because if not, the worst case in which these two encodings degenerate into the same string after t errors are applied could happen. However, if two encoding differ, for example, by exactly $2t$, in the Shannon model this degeneration event happens with so little probability that it is negligible. When we use the word "negligible," it will mean "exponentially small."

Shannon considers a much more benign model of errors. Shannon introduces the concept of an **error channel**. See below for a beautiful artist's rendition of the binary symmetric channel.



Definition 12 (Binary Symmetric Channel)**.** *To define the binary symmetric channel BSC(p), we consider someone who first transmits either the bit 0 or the bit 1. With probability p the bit gets flipped, and with probability $1-p$ the bit remains the same.*

Note 13. *It is natural to consider $p \in (0, 1/2)$, otherwise the code is erred more often than not, although technically the results should extend to all $p \in [0, 1]$. Furthermore, the binary symmetric channel will act independently on bits.*

We informally defined the **rate** of a code in the Hamming context last class. We define it here for the probabilistic Shannon context.

Definition 14. *For all n, consider the encoding function $E_n : \{0,1\}^{k_n} \to \{0,1\}^n$ and the decoding function $D_n : \{0,1\}^n \to \{0,1\}^{k_n}$. The goal in Shannon coding theory is for us to decode the message with high probability. In particular, for a binary symmetric channel acting on a message space M, we desire*

$$\Pr_{m \sim \text{Uniform}(M)} [D(BSC(E(m))) \neq m] = o_n(1).$$

*Given this condition, we then want to maximize the **rate** of a code, also known as the **capacity** of the Shannon code, which we define as*

$$\textit{Capacity} = \lim_{n \to \infty} \frac{k_n}{n}.$$

Shannon proved the following theorem in his paper.

Theorem 15 (Shannon)**.** *The capacity of a binary symmetric channel code BSC(p) is at most $1 - H(p)$, where*

$$H(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$

is the entropy of the binary symmetric channel.

We will explain some elements of the proof. One can derive by using Stirling's Approximation that the volume of a radius pn ball in $\mathbb{F}_2^n$ can be approximated as

$$\text{Vol}_2(n, pn) \approx 2^{(H(p)+o(1))n}.$$

In other words, this volume is related to the entropy $H(p)$ of the binary symmetric channel BSC(p).

How do we define an error correcting code in this model? One of the key ideas is the **probabilistic method**. It is debated who invented this method even though Erdős published a paper on it. In particular, one of the pieces of evidence in the debate is Shannon's paper, which is an old and excellent application of this method. We now highlight the idea in Shannon's proof of Shannon's Theorem. We note that the following proof of Theorem 15 is non-constructive, as is common with proofs using the probabilistic method.

Proof of Theorem 15: First we pick a random encoding function. $E : \{0,1\}^k \to \{0,1\}^n$. That is, for every message $m \in \{0,1\}^k$, we map m uniformly and independently into $\{0,1\}^n$. Recall that the code associated to this function is

$$C = \{E(m) \mid m \in \{0,1\}^n\}.$$

We define the decoding function $D : \{0,1\}^n \to \{0,1\}^k$, which takes in an input $y = \text{BSC}(E(m))$, as follows. On input $y \in \{0,1\}^n$:

- If $\text{Ball}\,(y, (p+\varepsilon)(n))) \cap C = \emptyset$, output "error."
- If $\text{Ball}\,(y, (p+\varepsilon)(n))) \cap C = \{E(m)\}$, output m.
- Otherwise, if $|\text{Ball}\,(y, (p+\varepsilon)(n))) \cap C| > 1$, output "error."

In other words, we output if a small ball of radius slightly greater than pn around the encoding is associated to a unique message, otherwise we output an error.

Now we do the analysis of the construction. For a fixed message m, what is the probability that such a randomly chosen encoding function E has an exponentially small error probability? Shannon proved the following.

Theorem 16. *We have*

$$\Pr_{m,E,BSC}[D\,(BSC(E(m))) \neq m] \leq \exp(-n).$$

By the probabilistic method paradigm, the above probability is an average over all E, so in particular, proving the above theorem shows that there is at at least one E such that

$$\Pr_{m,\text{BSC}}[D\,(\text{BSC}\,(E(m))) \neq m] \leq \exp(-n).$$

To prove the theorem, We will prove that the probability of an error, namely that the first or third cases happen, is small:

- If $\text{Ball}\,(y, (p+\varepsilon)(n))) \cap C = \emptyset$, output "error."
- If $\text{Ball}\,(y, (p+\varepsilon)(n))) \cap C = \{E(m)\}$, output m.
- Otherwise, if $|\text{Ball}\,(y, (p+\varepsilon)(n))) \cap C| > 1$, output "error."

We analyze the first case. By the Chernoff bound, the event $\text{Ball}\,(y, (p+\varepsilon)(n))) \cap C = \emptyset$, which happens only if there are at least $(p+\varepsilon)(n)$ errors, occurs with exponentially small probability.

Now we analyze the third case. Recall that we have fixed a message m, which also fixes $E(m)$. We can view the Shannon error model acting on $E(m)$, as well as choosing our random encoding function E on the other messages $m' \neq m$, respectively as follows:

- Consider the errors being generated in the process $E(m) \mapsto \text{BSC}\,(E(m))$. Note that $\text{BSC}\,(E(m))$ is within a ball $\text{Ball}(E(m), (p+\varepsilon'')(n))$ with overwhelming possibility for every $\varepsilon'' > 0$ by the Chernoff bound, hence can we can condition on this event. More formally, if the error probability is exponentially small conditioned on the event

  $$\text{BSC}\,(E(m)) \in \text{Ball}(E(m), (p+\varepsilon)(n)),$$

  then it is also exponentially small not conditioned on this event.
- Finally, randomly choose all $\{E(m'), m' \neq m\}$ uniformly and independently.

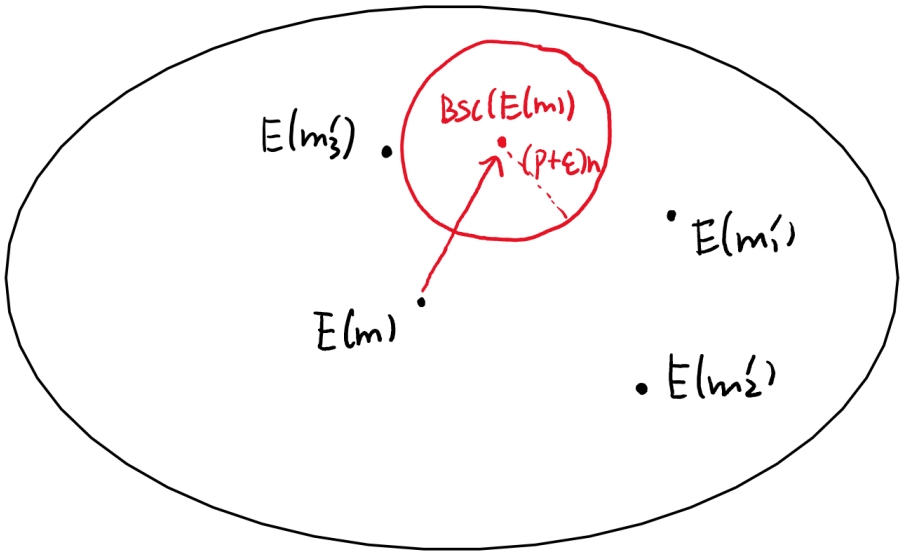


Figure 2: Error in the first case.

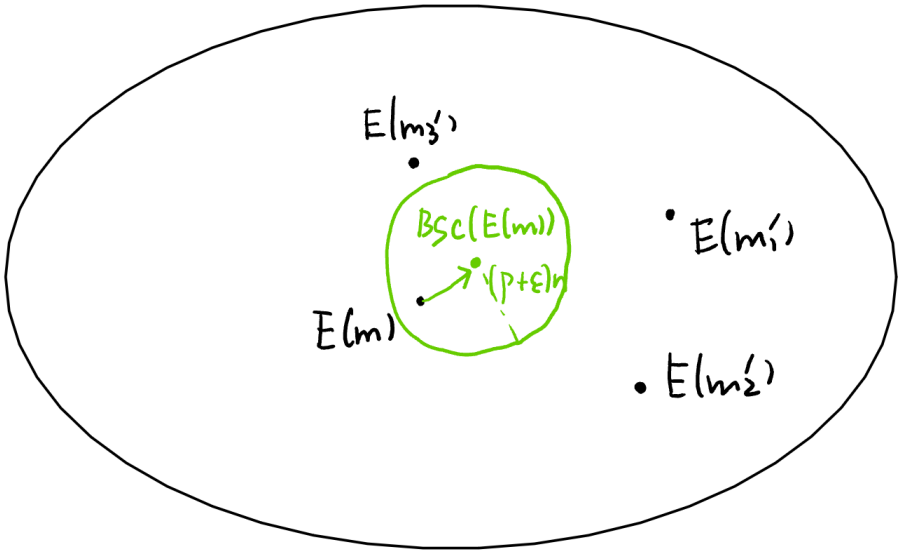


Figure 3: No error in the second case.

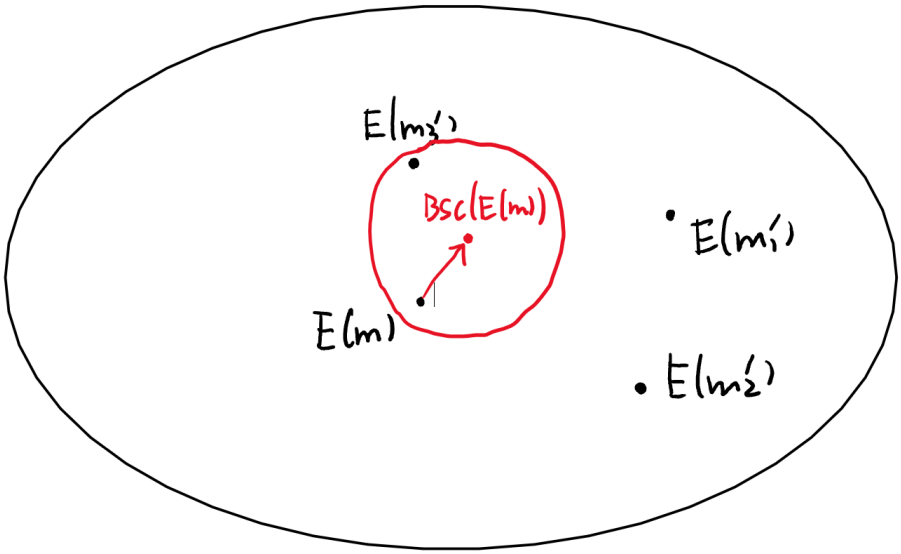


Figure 4: Error in the third case.

For any particular m', the probability that $E(m') = \text{BSC}\,(E(m))$ is exactly 2^{-n}. Hence by a union bound, the probability that no $E(m')$ for $m' \neq m$ equals $\text{BSC}\,(E(m))$ is bounded above by a union bound over all m' by

$$2^k \cdot 2^{-n} \cdot \text{Vol}(n, (p+\varepsilon)(n))$$

since there are at most 2^k messages m'. The last step to show that the probability of the third case is exponentially small is to show that

$$\begin{aligned} 2^k \cdot 2^{-n} \cdot \text{Vol}(n, (p+\varepsilon)(n)) &\leq 2^k \cdot 2^{-n} \cdot 2^{(H(p)+\varepsilon'')(n)} \\ &\leq 2^{-\varepsilon' n} \end{aligned}$$

for some $\varepsilon' > 0$. In other words, we need

$$\begin{aligned} k - n + (H(p) + \varepsilon'')n &\leq -\varepsilon' n \\ k &\leq (1 - H(p) - (\varepsilon' + \varepsilon''))n. \end{aligned}$$

Letting $\varepsilon', \varepsilon'' \to 0$, this shows that the maximum Shannon capacity of $\text{BSC}(p)$, which recall is defined by k/n, is at least $1 - H(p) - \varepsilon$ for all $\varepsilon > 0$.

□

Next class, we will prove the other direction of Shannon capacity of a binary symmetric channel, namely that

$$\text{Capacity}(\text{BSC}(p)) \leq 1 - H(p) + \varepsilon$$

for all $\varepsilon > 0$. In particular, we show that if

$$\text{Capacity} > 1 - H(p) + \varepsilon$$

for some $\varepsilon > 0$, then failure in decoding in fact happens with overwhelming probability.

Then we will return to studying general $[n, k, d]_q$-codes. The two parameters we will study are the rate $0 \leq R \leq 1$, namely the limit k/n as $n \to \infty$, as well as the **normalized distance** $0 \leq \delta \leq 1$, defined by the limit d/n as $n \to \infty$. In particular, there happen to be trade-offs between these two parameters which make the question of determining which ordered pairs (R, δ) in the region $[0, 1] \times [0, 1]$ are possible.

5 Exercises

Below are two exercises related to the class material that should be straightforward with the hints provided. We have also included a tricky third exercise in the form of a puzzle that relates to creating an error correcting code using Hamming distance. The full details of the solutions will be posted in the later version of the scribe notes.

1. Prove Lemma 10: If d is odd, then a $(n, k, d)_2$-code can be modified to create a $(n+1, k, d+1)_2$-code.

   As a hint, consider adding a parity-type bit to the end of the encoding similar to the $d = 1$ to $d = 2$ case discussed last class. After you have proven that this works, explain why the proof fails when d is even.

2. In our goal for the binary symmetric channel, the instructor mentioned that he desired

   $$\Pr_{m \sim \text{Uniform}(M)} [D(\text{BSC}(E(m))) \neq m] = o_n(1).$$

   However, we can also desire the stronger condition that

   $$\Pr [D(\text{BSC}(E(m))) \neq m] = o_n(1)$$

   for every message $m \in M$. This would correspond with the idea that every message should be decodable with high probability, as opposed to the average decoding probability across messages being high. The construction we gave using the probabilistic method can be modified slightly to achieve this goal as well while maintaining the same Shannon capacity of $1-H(p)-\varepsilon$ for $\varepsilon > 0$. In particular, consider the following idea.

   Fix a constant $0 < \kappa < 1$. Consider dropping a fraction κ of the messages with the greatest error probabilities from the message space $M = \{0,1\}^k$. For example, if $\kappa = 1/2$ then we would purge the 2^{k-1} messages for which

   $$\Pr [D(\text{BSC}(E(m))) \neq m]$$

   is the highest. Prove that in the new message space M' with $|M'| = (1-\kappa)|M|$,

   $$\Pr [D(\text{BSC}(E(m))) \neq m] = o_n(1)$$

   for every message $m \in M'$.

3. Fix an integer $k \geq 2$, and suppose that $2^k - 1$ wombats are sitting at a table. For each of them, Madhu randomly draws either 0 or 1 on that wombat's head uniformly and independently. This is done so that each wombat can see all other wombats' numbers but not their own number. No communication between wombats is allowed.

   Now, each wombat must either privately write down what they think their number is, or abstain from guessing. Each wombat either guesses or abstains simultaneously with every other wombat, so the decision of any wombat should only be a function of the numbers of the other wombats. The wombats will win if at least one wombat guesses their number correctly, and no wombat guesses their number incorrectly. The goal is for the wombats can devise a strategy that has a winning probability of at least

   $$\frac{2^k - 1}{2^k}.$$

(a) Prove that the wombats can achieve the goal when $k = 2$.

As a hint, prove that there exists a strategy that is guaranteed to succeed long as the wombats are not assigned either 000 or 111.

(b) Prove that the wombats can achieve the goal for all $k \geq 2$ by creating a strategy modeled off of an error correcting code in which there is some guaranteed success set $S \subseteq \{0,1\}^{2^k-1}$ of Madhu's assignments where the wombats win, as well as some guaranteed failure set $F = S^c$ of Madhu's assignments where the wombats lose. By definition of S and F, given an element $s \in S$, what can you say about the minimum Hamming distance between s and an element of F?