6.897 Algorithmic Introduction to Coding Theory	September 5, 2001
Lecture 1	
Lecturer: Madhu Sudan	*Scribe: Ryan O'Donnell*

1 Error correcting codes and this class

This course teaches the mathematics underlying objects called error correcting codes (ECCs). We won't define what they are today - for reasons to be clarified later. For now it will suffice to know that they are combinatorial objects that possess certain extremal properties which are very useful for the communication and storage of information.

Modulo the definition of ECCs, we can still discuss the big outline of this course. The contents of this course can be divided into four roughly equal parts. Our first part will be explain some of the basic constructions of error-correcting codes. Next we will show "negative results" showing limitations, or bounds, on the performance of codes. The two parts above are essentially combinatorial in nature, with few computational themes. We will get into the computational aspects in the third and fourth parts of this course. The third part of the course is where we focus on algorithmic tasks associated with ECCs and some solutions to these tasks. Finally, in the fourth part we will discuss some consequences one can derive in computational complexity theory as a result of the existence of ECCs and from the algorithmic capabilities surrounding them.

1.1 Standard references

Since the subject is quite old, there are plenty of texts. Yet we won't follow any one of them. Here are comments on some of them.

1. The text by MacWilliams and Sloane [3].
   (a) This is possibly the most referenced text in CS?
   (b) Unfortunately, it is getting outdated pretty fast.
   (c) Has detailed coverage of (too) many families of codes.
   (d) Coverage of algorithms is not so good.
2. The text by van Lint [7].
   (a) The book is much more concise and handy than MacWilliams and Sloane. Easier to get to some of the results here, if they are available.
   (b) Still, the emphasis is more combinatorial than algorithmic.
3. A text by Blahut [1].
   (a) Book was targetted at engineers rather than mathematics, or so the author claims.

(b) Yet the coverage is excellent, especially for insight, motivations, definitions, even algorithms

(c) The main drawback is that it is out of print and not easily available (not in MIT library, e.g.).

4. The Handbook of Coding Theory [5].

   (a) Offers extensive coverage of many recent themes.

   (b) Sometimes excessive. (E.g. 130 pages of table of best known codes.)

   (c) But contains many interesting chapters. E.g., chapter on algebraic-geometry codes, and the chapter on deep-space applications.

2 Information Theory

Even though this course is on coding theory, we start with a brief coverage of information theory. Part of the reason is historic. Coding theory was initiated by two seminal papers:

1. In 1948, Shannon wrote a detailed treatise on the mathematics behind communication [6].

2. In 1950, Hamming, motivated by the task of correcting small number of errors on magnetic storage media, wrote the first paper introducing error-correcting codes [2].

In this lecture we will discuss the main results from the Shannon paper, which founded the theory of information, while also co-founding (with Hamming) the theory of error-correcting codes. The theory of information provides the motivation for many questions in coding theory and so we will study this first.

Shannon considered the problem of two parties, say Alice and Bob, who wish to communicate digitally. In particular, Alice wishes to send a message to Bob over a certain digital channel. Shannon considered both "noiseless" and "noisy" channels. We start by considering the "noiseless" case.

2.1 Noiseless coding

In this case, the channels perfectly transmits the bits Alice wants to send to Bob. However, we might imagine that the channel is slow, so our goal is to compress the information we want to send down to as few bits as possible. Then we send these bits across the channel for Bob to receive and decompress. We start with an example which shows how to effect such a compression scheme.

2.2 An example

Consider transmitting the contents of a piece of paper, which contains some handwritten material on it, by fax. If the paper has just a small amount of black text on a white background, when it is digitized it might consist of 99% 0's and 1% 1's. Let's say for the sake of argument that we want to transmit a message in which each bit is independently 0 with probability .99 and 1 with probability .01.

Consider the following encoding scheme: We split the message up into blocks of 10 bits. If the block is 0000000000, we send the bit 0. If the block is anything else, say x, we send the string $1x$. Now the expected length of the encoding of a block is:

$$1 \cdot \Pr[\text{block is } 0000000000] + 11 \cdot \Pr[\text{block is not } 0000000000] \quad = \quad 11 - 10q$$

where $q := \Pr[\text{block is } 0000000000] = .99^{10} \geq .9$. Hence the expected length of the encoding of one 10 bit block is at most 2 bits. Thus the original message has been compressed to within 20% of its length! (Food for thought: Do we really need the probabilistic model to acheive this compression?)

Can we do better? This is exactly the question Shannon raised and answered with his theorem on noiseless coding.

2.2.1 Entropy

To analyze how much a distribution could be compressed, Shannon introduced some mathematical definitions associated with information. The source of information is modeled as a probability distribution, and a message is a random variable drawn from this distribution. The goal of compression is to encode the support of the probability space (using, say binary strings) so as to minimize the expected length of the message. To any source (or more correctly, to any distribution), Shannon assigned a non-negative real number, that he termed "entropy", that measures the information content of the distribution. He then showed that this quantity (to within an additive term of *one bit*) measures the compressibility of a bit.

The simplest possible distribution is a coin flip, in which heads has probability p and tails has probability $1-p$. The entropy assigned to this is:

$$H(p) \stackrel{\text{def}}{=} p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}.$$

We can generalize this notion to arbitrary distributions.

Definition 1 *Let U be any finite set, and let D be a probability distribution on U, i.e., $D : U \to [0,1]$ with $\sum_{x \in U} D(x) = 1$. Let X be a random variable distributed according to D. Then the entropy of D is*[1]*:*

$$H(D) \stackrel{\text{def}}{=} \sum_{x \in U} D(x) \log_2 \frac{1}{D(x)}.$$

The noiseless coding theorem of Shannon essentially says the following:

Theorem 2 *For every finite set U and for every distribution $D : U \to [0,1]$, There exists an encoding function $\mathsf{Enc} : U \to \{0,1\}^*$ and a decoding function $\mathsf{Dec} : \{0,1\}^* \to U$ such that for every $x \in U$, $\mathsf{Dec}(\mathsf{Enc}(x)) = x$, and*

$$\mathrm{Exp}_{x \leftarrow D}[|\mathsf{Enc}(x)|] \in [H(D), H(D)+1],$$

where $|s|$ denotes the length of a string s. Conversely, no error-free encoding can do better.

[1] In lecture, we used a random variable X drawn according to D and defined this to be the entropy of X, rather than D. In these notes we will switch to the more appropriate notation, which defines it to be a function of the distribution and not the variable.

Shannon's actual theorem is often stated differently, but we will state the theorem in the above since it is clearer. We don't prove the theorem here, but the main idea for the existence result is the following: (1) Round all probabilities down so that they become powers of 2 (i.e., construct D' such that for every x, $D'(x) = 2^{-i}$ for some integer i, and $D(x)/2 < D'(x) \leq D(x)$. (2) Show that there exists an encoding Enc that encodes an element x, whose new probability $D'(x)$ equals 2^{-i}, with i bits (so that no strings have the same encoding). (3) Conclude that this encoding has an expected length in the desired range! The converse is harder to prove - we won't get into it.

2.2.2 An elegant algebraical identity

The entropy function exhibits some very nice mathematical properties. For example if a distribution D decomposes into two independent distributions D_1 and D_2 (i.e., $U = U_1 \times U_2$, $D_1 : U_1 \to [0,1]$ and $D_2 : U_2 \to [0,1]$ and $D(x,y) = D_1(x)D_2(y)$), then $H(D) = H(D_1) + H(D_2)$. This fact can be use to prove an interesting algebraic identity, which is otherwise quite tricky to prove.

From the above property and the entropy of a bit, we find that the entropy in n independent p-biased coin flips is $nH(p)$. On the other hand, this should be equal to the entropy of the distribution D given by $D(x) := p^{\#1's \text{ in } x}(1-p)^{\#0's \text{ in } x}$. Using the second formula, we calculate this as:

$$\sum_{t=0}^{n} \binom{n}{t} D_t \log_2 \frac{1}{D_t}$$

where $D_t := p^t(1-p)^{n-t}$. Thus we get

$$\sum_{t=0}^{n} \binom{n}{t} D_t \log_2 \frac{1}{D_t} = nH(p).$$

Entropy and its variants play an important role in combinatorics and probability. Some very useful notions to keep track of in this area are those of mutual information, conditional entropy, and relative entropy!

To turn back to our example, what is the optimal compression factor for the case of our message to be faxed, whose bits were 1 with probability 1%? The answer is $H(.01) \approx 8\%$ — message can be compressed to within $H(.01)$ of their original length, if we know ones occur with probability about 1%.

2.3 Noisy channels

For the purposes of this course, the more interesting notion of a channel is the class of "noisy" channels considered by Shannon — i.e., channels that flip some of the bits sent across them. The problem here is for Alice to encode the message she wishes to send in such a way that even if the channel corrupts some of the bits in the encoding, Bob will be able to decode the result into Alice's original message, with high probability. Shannon considered a large class of probabilistic models for noisy channels, and for each proved the surprising theorem that you could always overcome a constant error rate by sending an encoded message that was longer by a constant factor.

The general model Shannon gave for channels is as follows. There is an input alphabet Σ and an output alphabet Γ (both usually finite). Then we have a bipartite graph with Σ on the left and Γ on the right; each edge (σ, γ) is labeled with a probability of the channel converting a σ to a γ. (The channel operates independently on each character.) Of course, for each σ on the left, the sum of the labels on the edges touching it must be 1.

We will illustrate his results for such channels only in cases where $\Sigma \subseteq \Gamma$. (In such cases, it is clear what an error is.)

Two commonly considered channels:

1. Binary Symmetric Channel — the channel considered in the theorem. $\Sigma = \Gamma = \{0,1\}$, $(0,0)$ and $(1,1)$ get probability $1-p$, and $(0,1)$ and $(1,0)$ get probability p.
2. Binary Erasure Channel — in this channel, bits don't get flipped — rather they get erased. Specfically, $\Sigma = \{0,1\}$, $\Gamma = \{0,1,?\}$, $(0,0)$ and $(1,1)$ get probability $1-p$, and $(0,?)$ and $(1,?)$ get probability p.

2.3.1 Noisy coding theorem

The noisy coding theorem of Shannon is a powerful and general one. When specialized to the case of a binary symmetric channel with error probability p, we get the following result.

Theorem 3 *For every $p < 1/2$, there exists a constant $c < \infty$ and a pair of functions $E : \{0,1\}^k \to \{0,1\}^{ck}$, and $D : \{0,1\}^{ck} \to \{0,1\}^k$ with the following property: If we pick a message uniformly at random from $\{0,1\}^k$, encode with E and then send the result across the noisy channel, and decode the result, then we recover the original message with probability $1 - o(1)$.*[2]

Theorem 3 applies — with different constants — to the binary erasure channel as well.

2.3.2 Hamming notations

We need just a few definitions before proceeding with the proof of the theorem.

Definition 4 *If x and y are in Σ^n, then the* Hamming distance between them is $\Delta(x,y) :=$ # of coordinates on which x and y differ.

Definition 5 *The* Hamming ball of radius r centered at y is $B(y,r) := \{x \in \Sigma^n : \Delta(x,y) \leq r\}$.

Definition 6 $\mathrm{Vol}(r,n)$ *denotes the volume of (any) radius-r ball in $\{0,1\}^n$; i.e., $|B(y,r)|$. Exactly, this quantity is $\sum_{i=0}^{r} \binom{n}{i}$. If we fix some $p > 0$ and let $n \to \infty$ then we get* $\mathrm{Vol}(pn,n) = 2^{(H(p)+o(1))n}$.

2.3.3 Proof of Theorem 3

Proof The proof is highly non-constructive; it uses the probabilistic method.

Let $n > k$ be decided upon later. Pick the encoding function $E : \{0,1\}^k \to \{0,1\}^n$ at random, i.e., for every $m \in \{0,1\}^k$, $E(m)$ is chosen uniformly at random from $\{0,1\}^n$, independently of all other choices.

[2]The $o(1)$ term depends only on k.

The decoding function $D : \{0,1\}^n \rightarrow \{0,1\}^k$ works as follows. Given a string $y \in \{0,1\}^n$, we find (non-constructively) the $m \in \{0,1\}^k$ such that $\Delta(y, E(m))$ is minimized. This m is the value of $D(y)$.

We now prove that D and E have the required property, with high probability, over the random choice of E.

Fix $m \in \{0,1\}^k$ as also $E(m)$. (The rest of the proof will use the fact that $E(m')$ for $m' \neq m$ is still random.) the message being sent. Denote by y the corrupted version of $E(m)$ that is received by Bob. Let η denote the error vector $y - E(m)$. Note that the η is a random variable with each of its coordinates being 1 w.p. p and 0 w.p. $1-p$, independent of other coordinates.

Fix $\epsilon > 0$. Let $r = (p+\epsilon)n$. In order for $D(y) \neq m$ at least one of the following two events must occur:

1. $y \notin B(E(m), r)$ (i.e., too many errors occurred in transmission.)
2. There exists some $m' \neq m$ such that $E(m') \in B(y.r)$. (The errors take the received word too close to the encoding of some other message.)

In particular, if neither event occurs, then m is the unique message such that $E(m)$ is within a distance of r from y and so $D(y) = m$.

We show that for an appropriate choice of n, the events above happen with low probability. For the first event to happen, it must be that η has more than $p+\epsilon$ fraction of 1s. We can apply Chernoff bounds (see appendix at the end of this lecture) to see that:

$$\Pr_{\eta}[y \notin B(E(m), r))] \leq 2^{-(\epsilon^2/2)n}.$$

For any $\epsilon > 0$, we can pick n to be large enough that the above quantity is as small as we want.

We now move onto the second event. First fix y and an $m' \neq m$ and consider the event that $E(m') \notin B(y, r)$. The probability of this event, taken over the random variable $E(m')$, is exactly $\mathrm{Vol}(B(y,r))/2^n$. Using the approximation $\mathrm{Vol}(B(y, pn) \approx 2^{H(p)n}$ and ignoring ϵ (which is arbitrarily small), we find that for every $m' \neq m$.

$$\Pr_{E}[E(m') \in B(y,r)] \approx 2^{H(p)n-n}.$$

Using the union bound, we get that

$$\Pr_{E}[\exists m' \neq m \text{ s.t. } E(m') \in B(y,r)] \approx 2^{k-n+H(p)n}.$$

Thus we find that if $c > \frac{1}{1-H(p)}$, then $k/n < 1 - H(p)$ and thus the quantity above is less than one. This gives the choice of c that we need.

Our proof is not yet complete! Why? Well, we have argued that a fixed m is likely to be decoded correctly, but what about other messages? To complete the argument, we will actually need to lose a little bit.

Let $\delta = 2^{-(\epsilon^2/2)n} + 2^{k-n+H(p)n}$, denote the total probability of error in either of the two steps above. We have shown that for any fixed m, for a random choice of E and the associated D, the expected decoding error probability is at most δ. Thus we can now conclude that this expectation continues to hold if we make m a random variable chosen uniformly from $\{0,1\}^k$. We get

$$\mathrm{Exp}_{m,E,\eta}[D(E(m)+\eta) \neq m] \leq \delta.$$

In particular this implies that there exists an E such that for the corresponding D, we have

$$\mathrm{Exp}_{m,\eta}[D(E(m)+\eta) \neq m] \leq \delta.$$

This concludes the proof of Shannon's theorem. ■

2.3.4 Capacity of the channel

Shannon's theorem above shows that if we are willing to slow down the effective transmission rate of the channel to $1/c < 1 - H(p)$, then we can acheive error-free communication with vanishingly small probability. Furthermore this quantity $1 - H(p)$ is only a function of the "noisy channel" and not of the number of bits we wish to transmit over it. I.e., the effective rate of transmission can be an absolute constant independent of n. Shannon called this quantity (the limiting rate of k/n, as $n \to \infty$) the *capacity* of the noisy channel.

For the Binary Symmetric Channel, how large can its capacity be? The proof above shows that the capacity, denoted $C_{\mathrm{BSC}}(p)$, is at least $1 - H(p)$. It is natural to ask if this quantity is an artifact of the proof, or is it the correct capacity for the channel. Shannon proved that the latter was the case. We will prove this in the next lecture, but first let us see why this is the case intuitively.

Suppose that we are actually is a setup there is a noiseless channel between Alice's location and Bob's, but this channel has been "hijacked" by Eve and Fred. Say Alice and Eve are in the same physical location while Fred and Bob are at the other. To send a message over, Alice must hand it over to Eve who then sends it through the channel (after some potential corruption). Similarly at the receiving end, Fred receives the message and hands it over to Bob. Suppose Eve and Fred want to use this channel to exchange some messages of their own (at Alice & Bob's expense). They do so by informing Alice and Bob that the channel is noisy with bits being flipped with probability p. They advise Alice and Bob to use some encoding/decoding schemes. Alice and Bob agree on an encoding scheme E and a decoding scheme D and in their naivette share these functions with Eve and Fred as well.

In truth it may be that Eve wishes to use the "noise" to send some messages of her own to Fred. Say she has a message η which is a plain paper image, where each pixel of the page is 1 independently with probability p. The way she sends η to Fred (at Alice's expense) is that when Alice gives her a encoded message $E(m)$ to transmit, Eve sends over $E(m) + \eta$. Fred receives $y = E(m) + \eta$ (the channel does not introduce any noise) at the receiving end and passes it on, untampered, to Bob, but also retains a copy. As far as Alice and Bob are concerned, nothing mailicious is occuring - with high probability $D(y) = m$ and so they are exchanging messages at capacity of the "noisy channel". But note the situation w.r.t. Eve and Fred. Fred also knows E and D and can compute $\eta = y - E(D(y)) = E(m) + \eta - E(m)$, with high probability. So Eve and Fred are also exchanging messages among themselves (with some small probability of exchanging incorrect messages). But now if we consider Alice & Eve together at one end of the noiseless channel, and Bob & Fred together at the other end, the parties are exchanging bits at a rate of at least $C(p) + H(p)$ across the noiseless channel (assuming we believe the tightness of the noiseless coding theorem). Since we are normalizing so that the rate the noiseless channel is 1, we get $C(p) + H(p) < 1$! This is the link between the noisy case and the noiseless case of the Shannon theorems.

Appendix

Notation used in the lecture. We mention some notation that we used earlier or may use in later lectures. $\mathbb{Z}$ denotes the set of integers, $\mathbb{Z}^{\geq 0}$ denotes the set of non-negative integers, and $\mathbb{Z}^+$ the set of positive integers. $\mathbb{R}$ denote the set of reals, and $\mathcal{Q}$ the rationals. For real numbers a and b, the notation $[a, b]$ stands for the closed interval from a to b, i.e., the set $\{x \in \mathbb{R} | a \leq x \leq b\}$, while (a, b) is the open interval between a and b. For an integer k, we will use $[k]$ to denote the set $\{1, \ldots, k\}$. If D is a distribution on the universe U, then $X \leftarrow D$ denotes a random variable X drawn from U according to the distribution D. For an event $\mathcal{E} \subseteq U$, the quantity $\Pr_{X \leftarrow D}[X \in \mathcal{E}]$ denotes the probability of the event $\mathcal{E}$ when X is chosen according to D. For a real-valued function $f : U \to \mathbb{R}$, the quantity $\mathrm{Exp}_{X \leftarrow D}[f(X)]$ denotes the expected value of $f(X)$ when X is chosen according to D. When the distribution D is clear from context, we may abbreviate these quantities to $\Pr[\mathcal{E}]$ and $\mathrm{Exp}_X[f(X)]$. Similarly $\mathrm{Var}_{X \leftarrow D}[f(X)]$ denotes the variance of $f(X)$ (i.e., $\mathrm{Var}(f(X)) = \mathrm{Exp}[(f(X))^2] - \mathrm{Exp}[f(X)]^2$).

Some basic probability facts Here is a quick recap of basic facts on probability and expectations.

Probability One of the most used facts on probability is the union bound: $\Pr[\mathcal{E}_1 \cup \mathcal{E}_2] \leq \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2]$. Note that the bound makes no "independence" assumptions.

Expectations Analogous to the above we have: $\mathrm{Exp}[X_1 + X_2] = \mathrm{Exp}[X_1] + \mathrm{Exp}[X_2]$. Note that this is an equality! If random variables are independent, then we get a product relationship $\mathrm{Exp}[X_1 X_2] = \mathrm{Exp}[X_1]\mathrm{Exp}[X_2]$.

Converting probabilities to expectations: Since expectations are more amenable to algebraic manipulations, it is often useful to convert statements on probability to statements of events. The standard way to do this is to use "indicator variables". For an event $\mathcal{E}$, let $I_{\mathcal{E}}$ be the 0/1-valued variable given by $I_{\mathcal{E}}(X) = 1$ if $X \in \mathcal{E}$ and $I_{\mathcal{E}}(X) = 0$ otherwise. Then we have $\mathrm{Exp}_X[I_{\mathcal{E}}(X)] = \Pr[\mathcal{E}]$.

Converting expectations to probabilities: Since probabilities are the quantities that have more intuitive meaning, these are the more standard targets of our investigation. Since expectations figure in proofs, we would like to find ways to convert statements on expectations back into probability statements. This conversion is not standard. Several "tail inequalities" are used to acheive this conversion, e.g., Markov's, Chebychev's, and Chernoff's: We state them below:

Markov's inequality For a non-negative random variable X and positive real α, then $\Pr[X \geq \alpha] \leq \frac{E[X]}{\alpha}$.

Chebychev's inequality In its general form, this inequality is just an application of Markov's inequality to the random variable Y^2, for arbitrary Y. In the general form, it is quite hard to see its strength, so we give a special form.

Let $Y = \sum_{i=1}^n Y_i$, where the Y_i's identically distributed random variables that are pairwise (but not fully) independent. Let $\mathrm{Exp}[Y_i] = \mu$ and $\mathrm{Var}[Y_i] = \sigma^2$. Then for $\lambda > 0$, we have $\Pr[Y \geq (\mu + \lambda)n] \leq \frac{\sigma^2}{n\lambda^2}$.

Chernoff bounds If $Y_1, \ldots, Y_n$ are completely independent it is possible to get stronger bounds on the probability that their sum deviates much from their expectation. We consider the special case of variables taking values in the interval $[0, 1]$

Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables taking values in the interval $[0, 1]$ with mean μ. Let $Y = \sum_i Y_i$. Then $\Pr[Y \geq (\mu + \lambda)n] \leq e^{-(\lambda^2/2)n}$, where e is the base of the natural logarithm.

For further elaboration on probabilities and expectations one may consult the text on Randomized Algorithms [4].

References

[1] Richard E. Blahut. *Theory and Practice of Error Control Codes.* Addison-Wesley, Reading, Massachusetts, 1983.

[2] Richard W. Hamming. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29:147–160, April 1950.

[3] F. J. MacWilliams and Neil J. A. Sloane. *The Theory of Error-Correcting Codes.* Elsevier/North-Holland, Amsterdam, 1981.

[4] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[5] Vera S. Pless and W. Cary Huffman (Eds.). *Handbook of Coding Theory (2 Volumes).* Elsevier, 1998.

[6] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[7] J. H. van Lint. *Introduction to Coding Theory.* Graduate Texts in Mathematics **86**, (Third Edition) Springer-Verlag, Berlin, 1999.