Capacity of the Binary Symmetric Channel

Martin Leslie

Department of Mathematics
University of Arizona

February 14, 2012

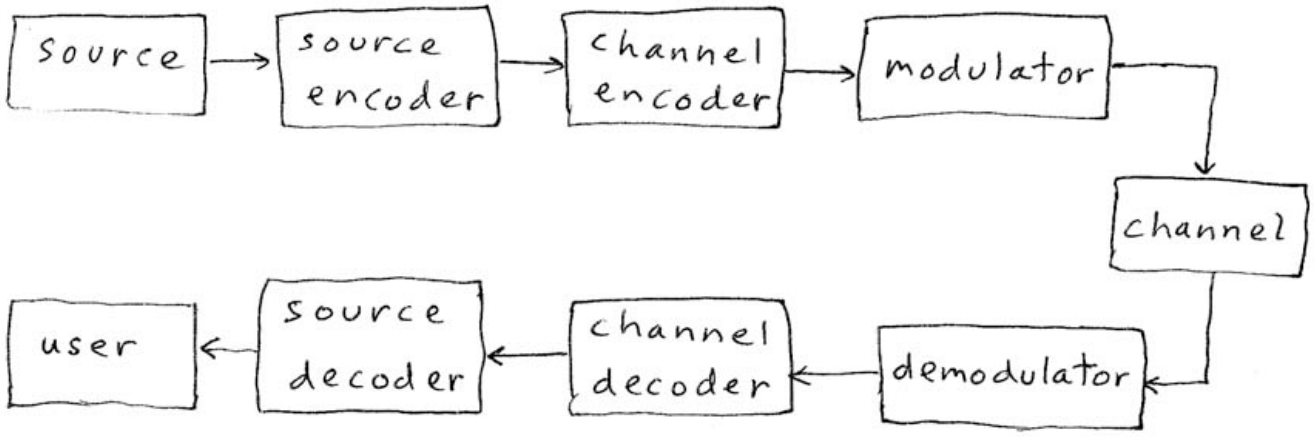
Some resources

- Elements of Information Theory, Thomas and Cover
- Information Theory, Inference and Learning Algorithms, David Mackay, available at `http://www.inference.phy.cam.ac.uk/mackay/itila/`
- Lecture Notes: Introduction to Coding Theory, Venkatesan Guruswami, available at `http://www.cs.cmu.edu/~venkatg/teaching/codingtheory/`

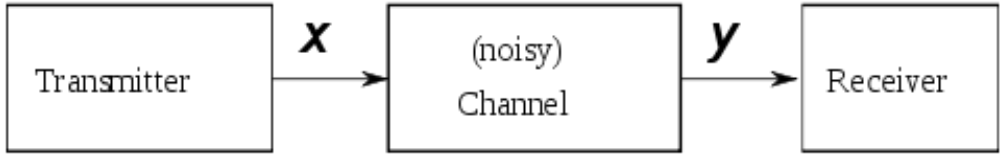
Outline

Communication

- Communication systems look something like

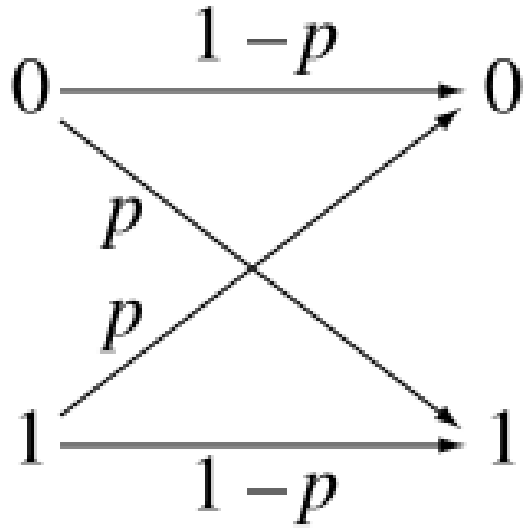


- Today we're just talking about channel coding so our picture is



The Binary Symmetric Channel

- The channel model we will use is the binary symmetric channel (BSC) which takes a binary input and with probability $p < 1/2$ switches it.



- This is a good model for deep space communications but not so good for hard drives or for terrestrial communications where errors often come in bursts.
- Can we find a good communication strategy for this channel?

Outline

Linear codes

- An $[n, k]_2$ linear code C is a k-dimensional linear subspace of $\mathbb{F}_2^n$.
- The basis vectors of C are the rows of the *generating matrix* G.
- With this matrix we can carry out encoding by the function from $\mathbb{F}_2^k \to \mathbb{F}_2^n$ that sends $u \mapsto uG$.
- Then the information rate is $R = k/n$.

The Hamming [7,4] code

- Let $G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$
- Then the codewords of C are

0000 000	1000 110	0010 111
1111 111	0100 011	1001 011
	1010 001	1100 101
	1101 000	1110 010
	0110 100	0111 001
	0011 010	1011 100
	0001 101	0101 110

The Parity Check Matrix

- If a code C has generating matrix $G = (I\ P)$ then define its *parity check matrix* to be $H^T = \begin{pmatrix} P \\ I \end{pmatrix}$.
- Notice that $GH^T = P + P = 0$.
- So if $c \in C$ we know $c = uG$ and thus

$$cH^T = uGH^T = 0.$$

- For the Hamming [7,4] code we have $H^T = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Hamming distance

- The Hamming distance $d_H(u, v)$ for $u, v \in \mathbb{F}_2^n$ is the number of places in which u and v differ.
- This satisfies all the axioms of a metric on $\mathbb{F}_2^n$.
- The Hamming weight is the number of 1's in a vector,

$$\mathrm{wt}(u) = d_H(u, 0).$$

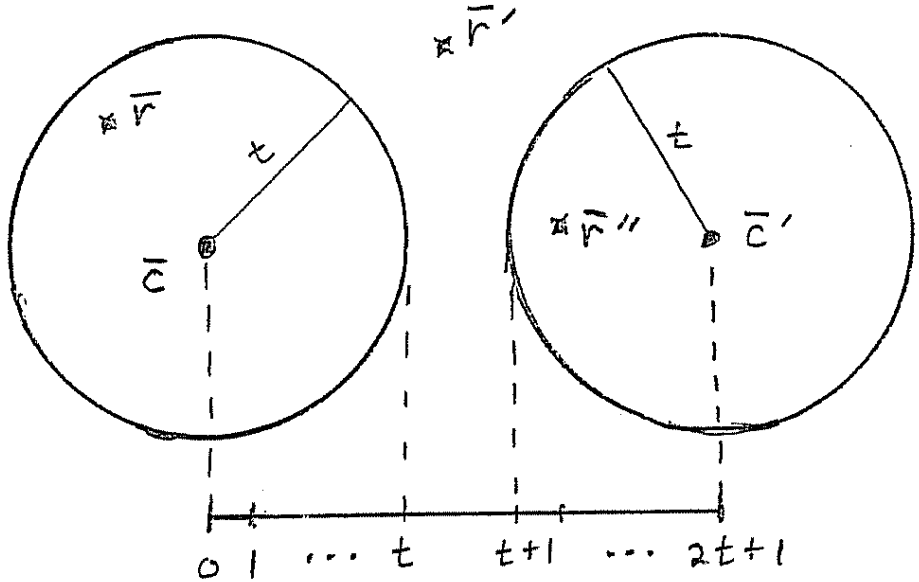
- Define d to be the minimum Hamming distance between any two vectors in C. Then

$$d = \min_{u \neq v \in C} d_H(u, v) = \min_{u \neq v \in C} \mathrm{wt}(u + v) = \min_{x \in C \setminus \{0\}} \mathrm{wt}(x).$$

- Then we talk about an $[n, k, d]_2$ code.

Nearest neighbour decoding

- If we receive $u \in \mathbb{F}_2^n$ we decode it to an element v of C for which $d_H(u, v)$ is minimum.
- It's possible that this is incorrect decoding but it is certainly the best choice on average.
- If d is the minimum distance of C then we can correct at least $t = \left\lfloor \frac{d-1}{2} \right\rfloor$ errors.



Syndrome decoding

- If we send codeword c but the channel adds error e, we receive $r = c + e$ and then can find the *syndrome* $rH^T = (c + e)H^T = eH^T$.
- We can find syndromes for all the possible errors (elements of $\mathbb{F}_2^n$) added by the channel and for each syndrome find a most likely error that leads to it - one with minimum weight.
- This gives us a *syndrome table* which allows us to decode more easily. For example for the Hamming [7,4] code we have

syndrome	000	110	011	111	101	100	010	001
likely error	0	e_1	e_2	e_3	e_4	e_5	e_6	e_7

Outline

Entropy

- Let X be a discrete random variable taking values in a finite alphabet $\mathcal{X}$ with probability mass function $p(x)$.
- The *self-information* (or *surprisal*) of $x \in \mathcal{X}$ is $\log \frac{1}{p(x)}$ where log means $\log_2$.
- The *entropy* of X is

$$\begin{aligned} H(X) &= \text{average self-information} \\ &= E_p\left[\log \frac{1}{p(X)}\right] \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \end{aligned}$$

How to think about entropy

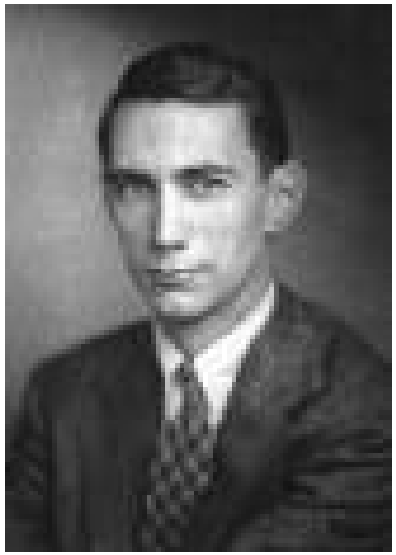
- Entropy is the average number of bits of information you gain about the value of X.
- Equivalently, it is the average uncertainty you have about each value of X before you receive it.
- The entropy of a fair coin flip is one bit. The entropy of a biased coin flip is less. For example $H(0.1, 0.9) = 0.47$.

Outline

Noisy–channel coding theorem

Theorem (Shannon, 1948)

A discrete memoryless channel has a capacity C. Fix an $R < C$. Then for all $\epsilon > 0$ there exists a sufficiently large n and an $[n, k]_2$ code with information rate $k/n \geq R$ and probability of decoding a block incorrectly less than ϵ.



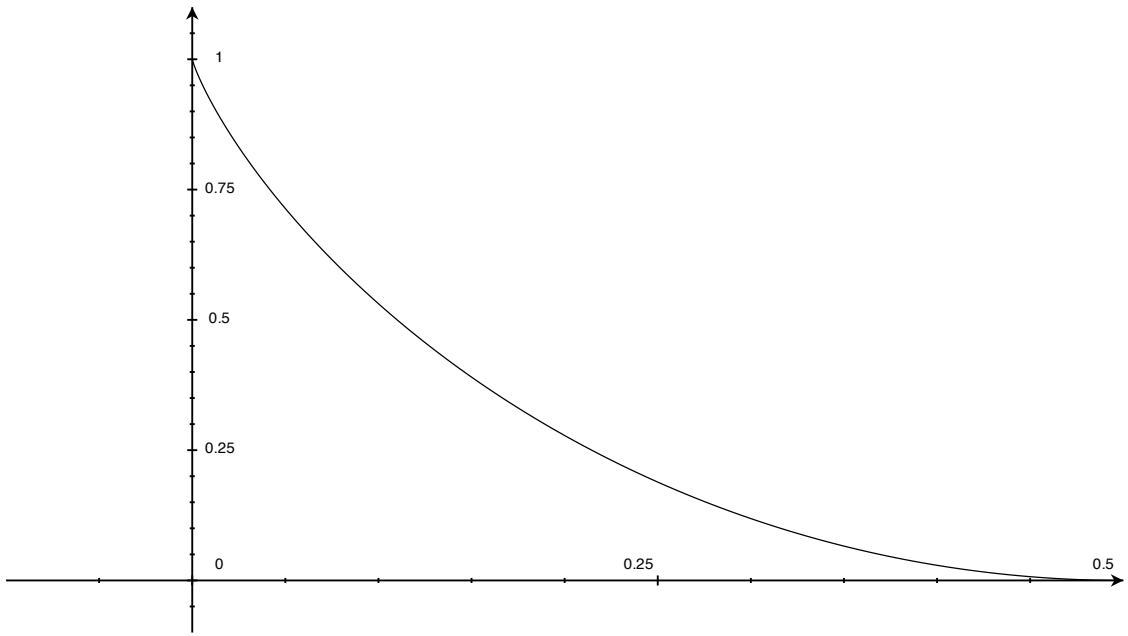
Capacity

- The capacity of the binary symmetric channel with crossover probability p is

$$C = 1 - h(p)$$

where $h(p) = H(X)$ for the random variable X which is 0 with probability $1 - p$ and 1 with probability p

$$\text{i.e. } h(p) = H(X) = -p \log p - (1 - p) \log(1 - p).$$



Typical Set

- If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution the same as X then

$$\begin{aligned} -\frac{1}{n}\log p(X_1,\ldots,X_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \\ &\to E[-\log p(X)] \\ &= H(X) \end{aligned}$$

  where the convergence is in probability and follows by the weak law of large numbers.

- Define the *typical set* with respect to X by

$$A_\epsilon^{(n)} = \left\{ x \in \mathcal{X}^n : \left| -\frac{1}{n}\log p(x_1,\ldots,x_n) - H(X) \right| < \epsilon \right\}.$$

Properties of the Typical Set

- For sufficiently large n,

$$P\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$$

and

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}.$$

- We always have $H(X) \leq \log|\mathcal{X}|$ so $2^{nH(X)} \leq |\mathcal{X}|^n$.
- So the typical set has almost all the probability but may be much smaller in size than the set of all sequences.

Setup of the proof

- We consider a linear code specified by a random $m \times n$ parity check matrix H of full rank (this code has rate $1 - m/n$).

Setup of the proof

- We consider a linear code specified by a random $m \times n$ parity check matrix H of full rank (this code has rate $1 - m/n$).
- We use syndrome decoding, so receive Hx where x is noise.

Setup of the proof

- We consider a linear code specified by a random $m \times n$ parity check matrix H of full rank (this code has rate $1 - m/n$).
- We use syndrome decoding, so receive Hx where x is noise.
- Then we use a *typical set decoder*: if there is a unique $x' \in A_\epsilon^{(n)}$ (the typical set for the BSC) such that $Hx' = Hx$ then we decode to x', otherwise decode to error.

Setup of the proof

- We consider a linear code specified by a random $m \times n$ parity check matrix H of full rank (this code has rate $1 - m/n$).
- We use syndrome decoding, so receive Hx where x is noise.
- Then we use a *typical set decoder*: if there is a unique $x' \in A_\epsilon^{(n)}$ (the typical set for the BSC) such that $Hx' = Hx$ then we decode to x', otherwise decode to error.
- We will show that the expected error probability can be made arbitrarily small, by taking n sufficiently large and ϵ sufficiently small, precisely when $1 - m/n < 1 - H(X)$.

Setup of the proof

- We consider a linear code specified by a random $m \times n$ parity check matrix H of full rank (this code has rate $1 - m/n$).
- We use syndrome decoding, so receive Hx where x is noise.
- Then we use a *typical set decoder*: if there is a unique $x' \in A_\epsilon^{(n)}$ (the typical set for the BSC) such that $Hx' = Hx$ then we decode to x', otherwise decode to error.
- We will show that the expected error probability can be made arbitrarily small, by taking n sufficiently large and ϵ sufficiently small, precisely when $1 - m/n < 1 - H(X)$.
- Since averaging over all H gives this performance we conclude that at least one H achieves at least this performance.

Estimating probability

- First we calculate the probability of decoding error given H

$$
\begin{aligned}
&P(\text{decoder error} \mid H) \\
&= P(x \notin A_\epsilon^{(n)}) + P(x \in A_\epsilon^{(n)}, \exists x' \in A_\epsilon^{(n)} \setminus \{x\} \text{ s.t. } Hx = Hx') \\
&< \epsilon + \sum_{x \in A_\epsilon^{(n)}} P(x) \sum_{x' \in A_\epsilon^{(n)} \setminus \{x\}} \mathbf{1}[H(x - x') = 0].
\end{aligned}
$$

Estimating probability

- First we calculate the probability of decoding error given H

$$
\begin{aligned}
&P(\text{decoder error} \mid H) \\
&= P(x \notin A_\epsilon^{(n)}) + P(x \in A_\epsilon^{(n)}, \exists x' \in A_\epsilon^{(n)} \setminus \{x\} \text{ s.t. } Hx = Hx') \\
&< \epsilon + \sum_{x \in A_\epsilon^{(n)}} P(x) \sum_{x' \in A_\epsilon^{(n)} \setminus \{x\}} \mathbf{1}[H(x - x') = 0].
\end{aligned}
$$

- So

$$
\begin{aligned}
&P(\text{decoder error}) \\
&< \epsilon + \sum_H P(H) \sum_{x \in A_\epsilon^{(n)}} P(x) \sum_{x' \in A_\epsilon^{(n)} \setminus \{x\}} \mathbf{1}[H(x - x') = 0] \\
&= \epsilon + \sum_{x \in A_\epsilon^{(n)}} P(x) \sum_{x' \in A_\epsilon^{(n)} \setminus \{x\}} \sum_H P(H) \mathbf{1}[H(x - x') = 0]
\end{aligned}
$$

Estimating probability, II

- But

$$\sum_H P(H)\mathbf{1}[H(x-x')=0] = (1/2)^m.$$

Estimating probability, II

- But

$$\sum_H P(H)\mathbf{1}[H(x-x')=0] = (1/2)^m.$$

- So

$$\begin{aligned} P(\text{decoder error}) &< \epsilon + \sum_{x\in A_\epsilon^{(n)}} P(x) \sum_{x'\in A_\epsilon^{(n)}\setminus\{x\}} 2^{-m} \\ &= \epsilon + P(A_\epsilon^{(n)})(|A_\epsilon^{(n)}|-1)2^{-m} \\ &< \epsilon + 1\cdot 2^{n(H(X)+\epsilon)}2^{-m} \\ &= \epsilon + 2^{n(H(X)-m/n+\epsilon)} \end{aligned}$$

Estimating probability, II

- But
$$\sum_H P(H)\mathbf{1}[H(x-x')=0] = (1/2)^m.$$
- So
$$\begin{aligned} P(\text{decoder error}) &< \epsilon + \sum_{x\in A_\epsilon^{(n)}} P(x) \sum_{x'\in A_\epsilon^{(n)}\setminus\{x\}} 2^{-m} \\ &= \epsilon + P(A_\epsilon^{(n)})(|A_\epsilon^{(n)}|-1)2^{-m} \\ &< \epsilon + 1\cdot 2^{n(H(X)+\epsilon)}2^{-m} \\ &= \epsilon + 2^{n(H(X)-m/n+\epsilon)} \end{aligned}$$
- Fix m/n such that $H(X) - m/n < 0$. Then choose ϵ less than half the arbitrarily small goal and less than the magnitude of $H(X) - m/n$. Finally, take n sufficiently large for our typical set bounds to hold and for the second term in the sum above to be less than half the goal.

A heuristic proof of the converse

- We show that for a (possibly nonlinear) code $C \subset \{0,1\}^n$ and a decoder $D\colon \{0,1\}^n \to C$ you can't expect to have rate
$$\frac{\log |C|}{n} > 1 - H(X)$$
and vanishing probability of decoding error.
- When $c \in C$ is transmitted, with high probability the received word is in $A_\epsilon^{(n)} + c$.
- This 'shell' of likely words has cardinality approximately $2^{nH(X)}$.
- Need to decode almost all the words in this shell to c so must have
$$|C| \cdot 2^{nH(X)} \leq 2^n$$
$$\text{i.e. } |C| \leq 2^{n(1-H(X))}.$$

Outline

Performance of codes