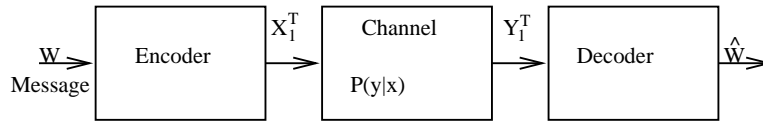


SySc/ECE INFORMATION THEORY¹: Notes on Channel Capacity (Chapter 8)



- Successful communication: $\hat{w} = w$.
- $N(T)$: # of distinguishable signals for T uses of the channel (X_1^T).
- If $N(T) \approx 2^{TC}$, then C is called the **Operational Channel Capacity**.
- For a channel described by $P_{Y|X}$, the **Information Channel Capacity** is:

$$C = \max_{P_X} I(X; Y)$$

- The *Big Theorem* (8.7.1 on page 198) says:

operational $C =$ information C .

Example: Memoryless Gaussian Channel (Chap 10)

Suppose $x_1^n \in \mathbb{R}^n$ and

$$P(y(t)|x(t)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y(t)-x(t))^2}{2\sigma^2}}$$

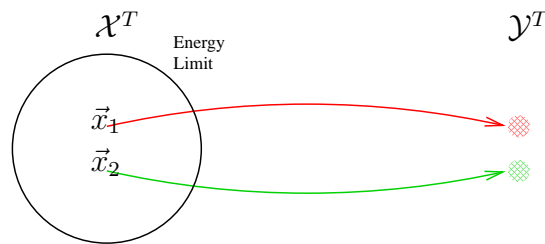


Figure 1: Gaussian channel with an energy limit. Code design = Sphere packing in $\mathbb{R}^T \leftrightarrow \mathcal{Y}^T$.

¹This document might be at http://www.ee.pdx.edu/~andy/info1/chan_notes.ps.

Example: Binary Channel

- Distinguishability = Distance.
- Hamming Distance = # of changed bits.
- Minimum distance decoding (in $B^3 =$ voting).
- Spheres in B^T .

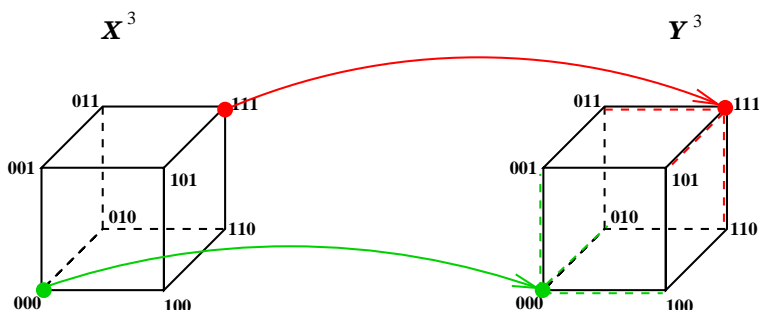


Figure 2: Geometry of codeword bits. The three bits sent or received specify the coordinates of a corner of the cube.

Example: Noiseless Binary Channel

		Y	
	$P_{Y X}$	0	1
X	0	1	0
	1	0	1

Q: What P_X maximizes $I(X; Y)$?

A: $P_X = (\frac{1}{2}, \frac{1}{2}) \rightarrow I(X; Y) = H(X) + H(Y) - H(X, Y) =$ Information C . Note: $N(T) = 2^T$ (# of distinguishable messages). So operational $C = \frac{1}{T} \log 2^T = 1$ bit.

Example: (Fig 8.3 on page 185 of text)

		Y			
	$P_{Y X}$	1	2	3	4
X	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0
	1	0	0	$\frac{1}{3}$	$\frac{2}{3}$

Maximize $I(X; Y)$ with $P_X = (\frac{1}{2}, \frac{1}{2}) \rightarrow 1$ bit. Note: $H(X; Y) = H(Y)$ or

$H(X|Y) = 0$. So $I(X; Y) = H(X)$.

Example: Noisy Typewriter

$P_{Y X}$	Y						
	1	2	3	4	...	25	26
1	$\frac{1}{2}$	$\frac{1}{2}$	0	0	...	0	0
2	0	$\frac{1}{2}$	$\frac{1}{2}$	0	...	0	0
3	0	0	$\frac{1}{2}$	$\frac{1}{2}$...	0	0
X	⋮	⋮		⋱	⋱		
25	0	0				$\frac{1}{2}$	$\frac{1}{2}$
26	$\frac{1}{2}$	0				0	$\frac{1}{2}$

Operational $C = \log 13$ (Use only odd x).

Claim $P(x) = \frac{1}{26} \forall x$ achieves information capacity $I(X; Y) = \log 13$

$P_{Y,X}$	Y						
	1	2	3	4	...	25	26
1	$\frac{1}{52}$	$\frac{1}{52}$	0	0	...	0	0
2	0	$\frac{1}{52}$	$\frac{1}{52}$	0	...	0	0
3	0	0	$\frac{1}{52}$	$\frac{1}{52}$...	0	0
X	⋮	⋮		⋱	⋱		
25	0	0				$\frac{1}{52}$	$\frac{1}{52}$
26	$\frac{1}{52}$	0				0	$\frac{1}{52}$

$H(X) = H(Y) = \log 26$ $H(X, Y) = \log 52$ $H(X) + H(Y) - H(X, Y) = \log \frac{26 \cdot 26}{52} = \log 13$. The distribution

$$P(x) = \begin{cases} \frac{1}{13} & x \text{ odd} \\ 0 & x \text{ even} \end{cases}$$

also achieves $I(X; Y) = \log 13$

$P_{Y,X}$	Y						
	1	2	3	4	...	25	26
1	$\frac{1}{26}$	$\frac{1}{26}$	0	0	...	0	0
2	0	0	0	0	...	0	0
3	0	0	$\frac{1}{26}$	$\frac{1}{26}$...	0	0
X	⋮	⋮		⋱	⋱		
25	0	0				$\frac{1}{26}$	$\frac{1}{26}$
26	0	0				0	0

$H(X) = \log 13$ $H(Y) = \log 26$ $H(X, Y) = \log 26$

The two distributions, P_1 and P_2 , above both achieve $I(X; Y) = \log 13$, so Theorem 2.7.4 on page 31 which says that $I(X; Y)$ is concave in P_X implies

that if

$$P_\alpha = \alpha P_1 + (1 - \alpha)P_2$$

then

$$I_\alpha(X; Y) \geq \log 13.$$

In fact $I_\alpha(X; Y) = \log 13 \forall \alpha : 0 \leq \alpha \leq 1$.

Symmetric Channels (8.2 in C&T)

Weakly Symmetric Channel: The two conditions are:

1. The rows of the transition matrix are permutations of each other. i.e., the vectors $P_{Y|X=x}$ are permutations of each other.
2. The column sums are the same, ie,

$$\sum_x \Pr(Y = i|X = x) = \sum_x \Pr(Y = j|X = x) \forall i, j$$

Symmetric Channel: The rows and columns of the transition matrix are permutations of each other.

Notice that for a weakly symmetric channel:

1. Because of property 1

$$H(Y|X) = \sum_x p(x)H(Y|X = x)$$

is independent of P_X .

- 2.

$$I(X; Y) = H(Y) - H(Y|X)$$

- 3.

$$C = -H(Y|X) + \max_{P_X} H(Y)$$

4. and because of property 2

$$C = \log |\mathcal{Y}| - H(Y|X)$$

is achieved by uniform P_X .

Definitions (8.5 in C&T)

Discrete Channel: $(\mathcal{X}, p(y|x), \mathcal{Y})$.

n th extension of DMC: (Discrete Memoryless Channel) $(\mathcal{X}_1^n, p(y_1^n|x_1^n), \mathcal{Y}_1^n)$.

Usually one also means without feedback which implies

$$p(y_1^n|x_1^n) = \prod_{k=1}^n p(y(k)|x(k)).$$

(M, n) code: M codewords, n^{th} extension.

- Index set $\mathcal{W} = (1, 2, \dots, M)$.
- Encoding function $X_1^n : \mathcal{W} \rightarrow \mathcal{X}_1^n$.
- Codebook $= \{X_1^n(w) : w \in \mathcal{W}\} \subset \mathcal{X}_1^n$.
- Decoding function $g : \mathcal{Y}_1^n \rightarrow \mathcal{W}$.

λ_i : Probability of error given $w = i$ was sent.

$\lambda^{(n)}$: Maximal probability of error for (M, n) code.

Arithmetic average error probability: $P_e^{(n)} = \frac{1}{M} \sum_{i=1}^M \lambda_i$.

Rate of an (M, n) code: $R = \frac{\log_2 M}{n}$ bits per use of the channel.

Achievable rate: \exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes with $\lim_{n \rightarrow \infty} \lambda^{(n)} = 0$.

Operational capacity: supremum of achievable rates.

Information capacity: $C \equiv \max_{P_X} I(X; Y)$.

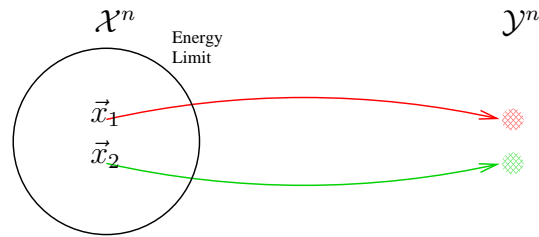
Paraphrase of The Channel Coding Theorem: Operational capacity equals information capacity, and any attempt to use a channel at a rate above its information capacity will result in a nonzero error rate.

Properties of *Information* Channel Capacity (8.3 in C&T)

1. $C \geq 0$ $I(X; Y) \geq 0$.
2. $C \leq \log |\mathcal{X}|$ $I(X; Y) = H(X) - H(X|Y)$.
3. $C \leq \log |\mathcal{Y}|$.
4. $I(X; Y)$ is continuous in P_X .
5. $I(X; Y)$ is concave in P_X .

Preview of Channel Coding Theorem (8.4 in C&T)

Consider $\vec{x}_1 \in \mathcal{X}_1^n$ and $\vec{x}_2 \in \mathcal{X}_1^n$.



In the sense of typical sets:

Q: How big are the spheres in \mathcal{Y}_1^n ?

A: $2^{nH(Y|X)}$.

Q: How big is \mathcal{Y}_1^n ?

A: $2^{nH(Y)}$.

Q: If no space is wasted, how many spheres fit in \mathcal{Y}_1^n ?

A: $\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X;Y)}$.

Q: If the number of spheres is reduced to $M = 2^{(1-\epsilon)nI(X;Y)}$, what fraction of \mathcal{Y}_1^n is filled?

A: $\text{frac}(\epsilon, n) = \frac{2^{(1-\epsilon)nI(X;Y)}}{2^{nI(X;Y)}} = 2^{-\epsilon nI(X;Y)}$.

Note: $\lim_{n \rightarrow \infty} \text{frac}(\epsilon, n) = 0$.

The point is that as n increases you can increase the number of different messages you design the channel to handle with any exponent less than C_{info} and expect to use the channel without error. There are two caveats:

- No procedure for sphere packing is offered.
- If sphere packing gets less efficient exponentially as dimension increases, then the argument might fall apart.

Rather than designing a code by carefully packing spheres, the proof of the channel coding theorem considers an ensemble of codes each member of which is created by placing the spheres randomly. One calculates the performance averaged over the ensemble, finds it adequate, and concludes that some members of the ensemble must perform adequately.

The Joint A.E.P. (Section 8.6 of C&T)

For a discrete memoryless channel (DMC) and iid X_1^n , i.e.,

$$p(x_1^n, y_1^n) = \prod_{k=1}^n p(x(k), y(k))$$

Definition: (x_1^n, y_1^n) are **Jointly Typical**, i.e., $(x_1^n, y_1^n) \in A_\epsilon^{(n)}$ if

$$\begin{aligned} & \left| \frac{-1}{n} \log p(x_1^n) - H(X) \right| < \epsilon \\ & \& \left| \frac{-1}{n} \log p(y_1^n) - H(Y) \right| < \epsilon \\ & \& \left| \frac{-1}{n} \log p(x_1^n, y_1^n) - H(X, Y) \right| < \epsilon \end{aligned}$$

The Joint AEP Theorem (Page 195 C&T): For (x_1^n, y_1^n) drawn according to $p(x_1^n, y_1^n)$

1.

$$\lim_{n \rightarrow \infty} \Pr \{A_\epsilon^{(n)}\} = 1$$

2.

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$$

3. If $(\tilde{X}_1^n, \tilde{Y}_1^n) \sim p(x_1^n)p(y_1^n)$ then for n large enough

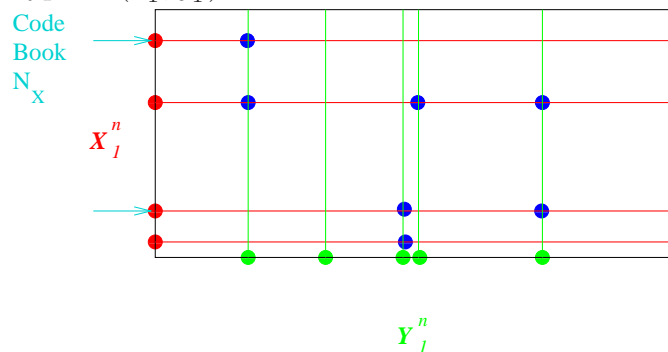
$$\Pr \left\{ (\tilde{X}_1^n, \tilde{Y}_1^n) \in A_\epsilon^{(n)} \right\} \leq 2^{-n(I(X;Y)-3\epsilon)}$$

and

$$\Pr \left\{ (\tilde{X}_1^n, \tilde{Y}_1^n) \in A_\epsilon^{(n)} \right\} \geq 2^{-n(I(X;Y)+3\epsilon)}$$

A color version of Fig. 8.9 on page 197 of C&T

# of typical x_1^n	$2^{nH(X)}$	red dots
# of typical y_1^n	$2^{nH(Y)}$	green dots
# of jointly typical (x_1^n, y_1^n)	$\leq 2^{nH(X,Y)}$	blue dots



If $H(X) + H(Y) > H(X, Y)$, i.e., $I(X; Y) > 0$, then some intersections don't get blue dots.

Q: Pick a red dot and a green dot at random. What is the probability that there is a corresponding blue dot?

A: $2^{-nI(X; Y)}$.

Q: Suppose a codebook with N_x entries is chosen at random and that a particular y_1^n is received. How many blue dots on the green column have corresponding bad codebook entries?

A: $N_x 2^{-nI(X; Y)}$

Note: if $N_x < 2^{nI(X; Y)}$ then the expected number of codebook entries that “interfere” with a received y_1^n is less than one.

Note: Pick x_1^n at random from $P_{X_1^n}$ and then pick y_1^n at random from $P_{Y_1^n | X_1^n = x_1^n}$, i.e., use the channel. This is the same as picking (x_1^n, y_1^n) at random from $P_{Y_1^n, X_1^n}$. Hence it will be in $A_c^{(n)}$ and there will be a blue dot.

Hamming Codes (Section 8.11 of C&T)

Parity Check Code: Choose $b(n)$ so that

$$\left(\sum_{i=1}^n b(i) \right) \bmod 2 = 0,$$

i.e.,

$$\begin{array}{r} b(1) \\ b(2) \\ \vdots \\ b(n-1) \\ \hline b(n) \end{array}$$

For a binary symmetric channel (BSC) with parameter q , the probability of an undetected error, i.e., even # of errors, is:

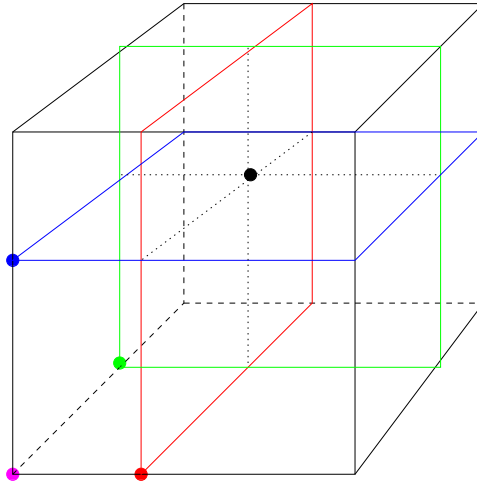
$$\Pr \{ \text{undetected} \} = q^2(1-q)^{n-2} \binom{n}{2} + q^4(1-q)^{n-4} \binom{n}{4} + \dots$$

Rectangular Code: Single error correction & Double error detection² or

Triple error detection.

0	1	0	1	0
0	1	1	1	1
0	0	0	1	1
1	0	0	0	1
1	0	1	1	1

Cube Code Single error correction. Coordinates of data bit (i, j, k) . Bit $b(0, 0, k)$ checks the $z = k$ plane.



$$b(0, 0, k) = \sum_{(i,j) \neq (0,0)} B(i, j, k) \pmod 2$$

So

$$\sum_{i,j} B(i, j, k) \pmod 2 = 0$$

Set $b(0, 0, 0)$ so that

$$\sum_{i,j,k} B(i, j, k) \pmod 2 = 0$$

to get double error detection³.

²The corner bit enables detection of errors in two other check bits.

³Without bit $(0, 0, 0)$, errors in check bits $b(0, 0, k)$ and $b(0, j, 0)$ will be interpreted as an error in data bit $(0, j, k)$.

Higher dimensions seem better

	Square	Cube	Hypercube (n=2)
Total bits	n^2	n^3	2^d
Check bits	$2n - 1$	$3n - 2$	$d + 1$
Data bits	$(n - 1)^2$	$n^3 - 3n + 2$	$2^d - d - 1$

(Hamming code doesn't use corner bit)

3-d Hamming Code. 7 bits. Coordinates of check bits have single bit set.

bit # (decimal)	0	1	2	3	4	5	6	7	
bit # (binary)	000	001	010	011	100	101	110	111	Check:
bit value		1	1	0	0	1	1	0	

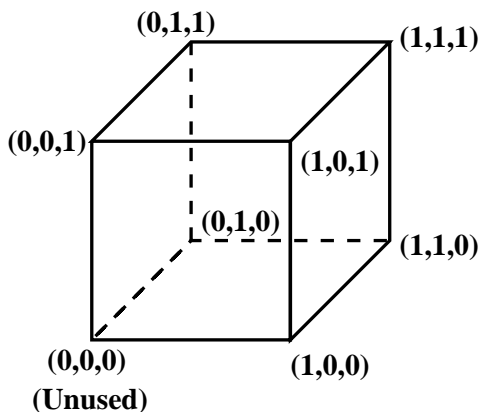


Figure 3: Geometry of bit addresses for a Hamming code with 4 message bits and 7 codeword bits. Three bits specify the address of a bit in the codeword. Note the difference to Fig. 2 where the three bits were the codeword itself.

$(0,0,1) = 1$	$(0,1,0) = 1$	$(1,0,0) = 0$
$(0,1,1) = 0$	$(0,1,1) = 0$	$(1,0,1) = 1$
$(1,0,1) = 1$	$(1,1,0) = 1$	$(1,1,0) = 1$
$(1,1,1) = 0$	$(1,1,1) = 0$	$(1,1,1) = 0$
0	0	0

2-d Hamming Code. 3 bits.

bit # (decimal)	0	1	2	3
bit # (binary)	00	01	10	11
bit value		0	0	0
or		1	1	1

Other values are corrected by voting.

Hamming codes: Minimum distance between message points = 3.

Use overall parity check bit in location 0 to get minimum distance = 4.

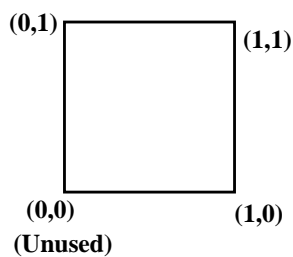


Figure 4: Geometry of bit addresses for a Hamming code with 1 message bit and 3 codeword bits.

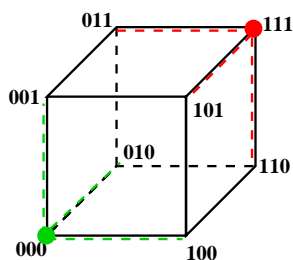


Figure 5: Geometry of codewords for a Hamming code with 1 message bit and 3 codeword bits.



Single error correction and double error detection, or triple error detection.

The Channel Coding Theorem

Paraphrase of The Channel Coding Theorem Operational capacity equals information capacity, and any attempt to use a channel at a rate above its information capacity will result in a nonzero error rate.

This is theorem 8.7.1 on page 198 of the text.

Typical set decoding If y_1^n is received and there is one and only one $w \in \mathcal{W}$ such that

$$(X_1^n(w), y_1^n) \in A_c^{(n)}$$

then $\hat{w} \leftarrow w$

else $\hat{w} \leftarrow 0$, i.e., error.

With typical set decoding selecting a codebook of M legal strings in \mathcal{X}_1^n specifies an (M, n) code.

Outline of the proof of the Channel Coding Theorem The goal is to show that for any $R < C$ there is a sequence of $(2^{nR-1}, n)$ codes with the maximal probability of error going to zero as $n \rightarrow \infty$. This result combined with a proof that rates $R > C$ are not achievable, establishes that C is the supremum of achievable rates.

1. Set P_X so that $I(X;Y) = C$.
2. Assign to each $(2^{nR}, n)$ code \mathcal{C} a probability⁴

$$\Pr(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{k=1}^n P_X((X_1^n(w))_k),$$

i.e., the probability of selecting the codebook at random.

3. Using a uniform distribution for the input w , calculate the probability of error averaged over all codebooks \mathcal{C} and inputs.

$$P(\mathcal{E}) = \sum_{i, \tilde{\mathcal{C}}} \Pr \left\{ \text{error} | w = i, \mathcal{C} = \tilde{\mathcal{C}} \right\} P(i, \tilde{\mathcal{C}})$$

Since the input i and the codebook $\tilde{\mathcal{C}}$ are independent, $P(i, \tilde{\mathcal{C}}) = \frac{P(\tilde{\mathcal{C}})}{2^{nR}}$. Later we will see that if $R < C$ then $\forall \epsilon > 0 \exists n_0$ such that $\forall n > n_0$

$$\begin{aligned} P(\mathcal{E}) &\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &< 2\epsilon. \end{aligned}$$

Note that this is averaged over w and \mathcal{C} but that we want a small probability of error for a good \mathcal{C} and the worst w .

4. Pick a good codebook \mathcal{C}^* with

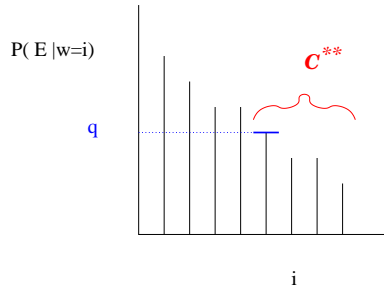
$$P(\mathcal{E}|\mathcal{C}^*) \leq 2\epsilon.$$

(This is still averaged over inputs w .)

5. Let \mathcal{C}^{**} be the half of the codewords in \mathcal{C}^* that are best. Note that if q is the probability of error for the worst w in \mathcal{C}^{**} , then

$$q < 2P(\mathcal{E})_{\text{avg}}.$$

⁴There are $|\mathcal{X}|^{n2^{nR}}$ different \mathcal{C} 's.



So

$$\forall i \in \mathcal{C}^{**}, P(\mathcal{E}|i) < 4\epsilon$$

$$\text{and Rate of } \mathcal{C}^{**} = \frac{1}{n} \log_2 \left(\frac{1}{2} 2^{nR} \right) = \frac{nR-1}{n}$$

Done: $\forall R < C, \exists$ a sequence of $(2^{nR-1}, n)$ codes such that the maximal probability of error $\rightarrow 0$.

Probability of Error: (Point 3. above)

$$\begin{aligned}
 P(\mathcal{E}) &= \sum_{w, \mathcal{C}} P(\mathcal{E}|w, \mathcal{C}) P(w, \mathcal{C}) \\
 P(w, \mathcal{C}) &= P(w) P(\mathcal{C}) \quad (\text{independent}) \\
 P(\mathcal{C}) &= \prod_{w=1}^{2^{nR}} \prod_{i=1}^n P_X((X_1^n(w))_i) \\
 P(w) &= \frac{1}{2^{nR}} \quad (\text{uniform}) \\
 P(\mathcal{E}) &= \sum_{w, \mathcal{C}} P(\mathcal{E}, \mathcal{C}|w) P(w) \\
 &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} P(\mathcal{E}, \mathcal{C}|w) \\
 &= \sum_{\mathcal{C}} P(\mathcal{E}, \mathcal{C}|w=1) \quad (\text{symmetry of codebooks}) \\
 &\equiv P(\mathcal{E}|w=1)
 \end{aligned}$$

For every codebook \mathcal{C} there is another codebook $\tilde{\mathcal{C}}$, with the same probability, i.e., $P(\mathcal{C}) = P(\tilde{\mathcal{C}})$, that is the same except that the codes for $w=1$ and $w=2$ are interchanged. Thus $P(\mathcal{E}, \mathcal{C}|w=1) = P(\mathcal{E}, \tilde{\mathcal{C}}|w=2)$ and

$$\sum_{\mathcal{C}} P(\mathcal{E}, \mathcal{C}|w=1) = \sum_{\mathcal{C}} P(\mathcal{E}, \mathcal{C}|w=2).$$

This is the symmetry argument.

Typical set decoding is:

$$\hat{w}(y_1^n) = \begin{cases} w' & \text{if } (X_1^n(w'), y_1^n) \in A_\epsilon^{(n)} \text{ and } w' \text{ is unique.} \\ 0 & \text{otherwise} \end{cases}$$

Note: $X_1^n(1) \sim P_{X_1^n}$ and $P(y_1^n | w = 1) = P(y_1^n)$.

Define Events or sets $\forall i \in \mathcal{W}$

$$\begin{aligned} E_i &= \{(X_1^n(i), y_1^n) \in A_\epsilon^{(n)}\} \\ \text{Error} &= \bar{E}_1 \cup E_2 \cup E_3 \cup \dots \cup E_{2^{nR}} \\ P(\mathcal{E} | w = 1) &\leq P(\bar{E}_1) + \sum_{i=2}^{2^{nR}} P(E_i) \end{aligned}$$

Joint AEP implies $P(\bar{E}_1) \rightarrow 0$, i.e., $\exists n_0$ such that $P(\bar{E}_1) < \epsilon \forall n > n_0$.

For $i \neq 1$ $(X_1^n(i), y_1^n) \sim P_{X_1^n} P_{Y_1^n}$ so

$$\Pr \{(X_1^n(i), y_1^n) \in A_\epsilon^{(n)}\} \equiv P(E_i) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

Since $I(X; Y) = C$

$$P(\mathcal{E}) \leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(C-3\epsilon)} < \epsilon + 2^{n(R-C+3\epsilon)}$$

If $R < C - 3\epsilon$, then for n large enough

$$P(\mathcal{E}) < 2\epsilon$$

Done!

Implementation: Suppose the channel is binary and the block size is $n = 1024$ and $R = 0.75$.

Q: How many codebooks?

A: ?

Skip section 8.8: Zero error codes and go to section 8.9:

Converse to the Channel Theorem

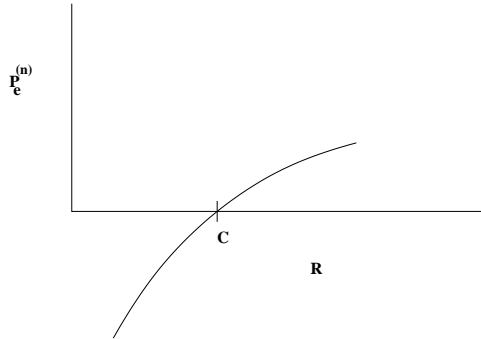
The book uses Fano's inequality to prove

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}, \quad (1)$$

which means that the probability of no error is

$$(1 - P_e) \leq \frac{C}{R} + \frac{1}{nR}$$

For large n (ignoring the annoying $\frac{1}{nR}$ term)



What about that annoying $\frac{1}{nR}$ term? Suppose that a block consists of m sub-blocks of size n . Let $1 - P_{e,nm}$ be the probability that there is no error. Equation 1 says

$$1 - P_{e,nm} \leq \frac{C}{R} + \frac{1}{mnR}.$$

One could use a code of length n to separately transmit each of the m sub-blocks. Let $1 - P_{e,n}$ be the probability that any particular sub-block is error free. Since coding the sub-blocks independently is a restriction on the codes that could be used,

$$(1 - P_{e,n})^m \leq (1 - P_{e,nm}) \leq \frac{C}{R} + \frac{1}{mnR}.$$

For any sub-block size n , the probability of error must obey

$$P_{e,n} \geq 1 - \left(\frac{C}{R} + \frac{1}{mnR} \right)^{\frac{1}{m}}$$

and the right hand side will be greater than zero for large values of m if $\frac{C}{R} < 1$.

Strong Converse

The material in this section is drawn from chapter 5 of Gallager's 1968 text. Gallager says the weak converse concerns the source and channel, but the strong converse concerns the channel alone. All the texts I've looked at emphasize that the strong converse does not imply the weak converse.

The Strong Converse⁵ For any discrete memoryless channel and any $(2^{nR}, n)$ code with $R > C$, there exists an $A > 0$ which depends on the

⁵Theorem 5.8.5 on page 173 of Gallager

channel but not on R or n such that

$$P_e^{(n)} \geq 1 - \frac{4A}{n(R-C)^2} - 2^{-\frac{n(R-C)}{2}} \quad (2)$$

Thus for large block size n , the function $P_e(R)$ switches abruptly from 0 to 1 at $R = C$.

Proof: Let $P_{Y|X}$ be the channel transition probability function and P_X be the input distribution that achieves capacity. Note

$$P_Y(y) = \sum_x P_{Y|X}(y|x)P_X(x).$$

For⁶ a particular input x

$$I(x; Y) \equiv E_{Y|x} \log \left(\frac{P(Y|x)}{P(Y)} \right) \leq C \quad (3)$$

Consider an $(2^{nR}, n)$ code with an encoding function $X_1^n : \mathcal{W} \mapsto \mathcal{X}_1^n$ and a decoding function $g : \mathcal{Y}_1^n \mapsto \mathcal{W}$. The average⁷ probability of correct decoding is

$$P_c = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \sum_{y_1^n : g(y_1^n) = w} P_{Y_1^n | X_1^n}(y_1^n | X_1^n(w)).$$

For $\epsilon > 0$ and $Q(y_1^n) \equiv \prod_{t=1}^n P(y(t))$ define the set

$$B_w = \left\{ y_1^n : \log \left(\frac{P(y_1^n | X_1^n(w))}{Q(y_1^n)} \right) > n(C + \epsilon) \right\}$$

and \bar{B}_w its complement. The elements of B_w are strings that are more strongly associated with $X_1^n(w)$ than average, i.e. they are atypical. Now

$$\begin{aligned} P_c &= \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \sum_{y_1^n : g(y_1^n) = w \& y_1^n \in B_w} P_{Y_1^n | X_1^n}(y_1^n | X_1^n(w)) \\ &+ \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \sum_{y_1^n : g(y_1^n) = w \& y_1^n \in \bar{B}_w} P_{Y_1^n | X_1^n}(y_1^n | X_1^n(w)) \\ &\equiv \text{Sum}_1 + \text{Sum}_2 \end{aligned}$$

⁶This is Theorem 4.5.1 on page 91 of Gallager

⁷Here the average is a uniformly weighted average over the input alphabet \mathcal{W} . The worst w will have a probability of error that is at least as bad as this average.

By the definition of \bar{B}_w , $P_{Y_1^n|X_1^n}(y_1^n|X_1^n(w)) \leq 2^{n(C+\epsilon)}Q(y_1^n)$ for each term in the second sum. Consequently the entire second sum satisfies

$$\begin{aligned}
\text{Sum}_2 &\leq \frac{2^{n(C+\epsilon)}}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \sum_{y_1^n: g(y_1^n)=w \& y_1^n \in \bar{B}_w} Q(y_1^n) \\
&\leq \frac{2^{n(C+\epsilon)}}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \sum_{y_1^n: g(y_1^n)=w} Q(y_1^n) \\
&\leq \frac{2^{n(C+\epsilon)}}{|\mathcal{W}|} \sum_{y_1^n} Q(y_1^n) \\
&= \frac{2^{n(C+\epsilon)}}{|\mathcal{W}|} \\
&= 2^{-n(R-C-\epsilon)}
\end{aligned}$$

The first sum is bounded by

$$\text{Sum}_1 \leq \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \sum_{y_1^n \in B_w} P_{Y_1^n|X_1^n}(y_1^n|X_1^n(w))$$

For any particular w

$$\sum_{y_1^n \in B_w} P(y_1^n|X_1^n(w)) = \Pr \left\{ \log \left(\frac{P(y_1^n|X_1^n(w))}{Q(y_1^n)} \right) > n(C + \epsilon) \right\},$$

and by Eqn. (3)

$$E_{Y_1^n} \log \left(\frac{P(y_1^n|X_1^n(w))}{Q(y_1^n)} \right) \leq nC.$$

Applying the Chebyshev inequality, $\Pr \{|Z - EZ| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}$, yields

$$\sum_{y_1^n: y_1^n \in B_w} P(y_1^n|X_1^n(w)) < \frac{A_{\text{total}}}{(n\epsilon)^2}$$

where A_{total} is the variance of $\log \left(\frac{P(Y_1^n|X_1^n(w))}{Q(Y_1^n)} \right)$. Note that:

$$\log \left(\frac{P(y_1^n|X_1^n(w))}{Q(y_1^n)} \right) = \sum_{t=1}^n \log \left(\frac{P(y(t)|(X_1^n(w))_t)}{P(y(t))} \right), \quad (4)$$

and since the channel is memoryless, the terms of the sum are independent and

$$A_{\text{total}} = \sum_{t=1}^n A_{x(t)}, \quad (5)$$

where $A_{x(t)}$ is the variance associated with $x(t) = (X_1^n(w))_t$. Denote

$$A = \max_{x \in \mathcal{X}} A_x \tag{6}$$

and observe that $\text{Sum}_1 \leq \frac{A}{n\epsilon^2}$. Thus

$$P_c \leq \frac{A}{n\epsilon^2} + 2^{-n(R-C-\epsilon)}$$

If $R > C$, choose $\epsilon = \frac{R-C}{2}$ to get

$$P_c \leq \frac{4A}{n(R-C)^2} + 2^{-n\frac{R-C}{2}}$$

Skip section 8.10 *Equality in the converse of the channel theorem*. We have already done Section 8.11, Hamming codes.

The key point of section 8.12 is that feedback doesn't increase the capacity of a discrete *memoryless* channel. Feedback can increase the capacity of stationary channels with memory, e.g. meteor burst communication.

The Joint Source Channel Coding Theorem

If V_1^n is a finite alphabet stochastic process that satisfies the AEP (an ergodic source) and $P_{Y|X}$ characterizes a DMC with capacity C , then

1. $H(\mathcal{V}) < C \implies \exists$ a source-channel code with $P_e^{(n)} \rightarrow 0$.
2. $H(\mathcal{V}) > C \implies \exists \delta > 0$ & n_0 such that $P_e^{(n)} > \delta \ \forall n > n_0$

Proof:

1. Combine source coding and channel coding.
2. To show

$$H(\mathcal{V}) > C \implies \exists \delta > 0 \ \& \ n_0 \ : \ P_e^{(n)} > \delta \ \forall n > n_0$$

Fano's inequality says (See page 39 of the text.)

$$\begin{aligned} H(V_1^n | \hat{V}_1^n) &\leq 1 + P_e \log |\mathcal{V}_1^n| \\ &= 1 + P_e^{(n)} n \log |\mathcal{V}| \end{aligned}$$

Now,

$$\begin{aligned} H(\mathcal{V}) &\leq \frac{1}{n} H(V_1^n) \\ &= \frac{1}{n} H(V_1^n | \hat{V}_1^n) + \frac{1}{n} I(V_1^n; \hat{V}_1^n) \\ &\leq \frac{1}{n} (1 + P_e^n n \log |\mathcal{V}|) + \frac{1}{n} I(X_1^n; \hat{Y}_1^n) \\ &\leq \frac{1}{n} + P_e^{(n)} \log |\mathcal{V}| + C \end{aligned}$$

So

$$P_e^{(n)} \geq \frac{1}{\log |\mathcal{V}|} \left(H(\mathcal{V}) - C - \frac{1}{n} \right)$$

and asymptotically⁸, you can't beat separate source and channel coding.

In *info1* module at downstairs:/home/andy/cvsroot. CVS version:

\$Id: chan_notes.tex,v 1.12 2003/11/21 20:02:06 andy Exp \$

⁸As $n \rightarrow \infty$