ECS452 2018/2 Part II.1 Dr.Prapun

4 Mutual Information and Channel Capacity

In Chapter 2, we have seen that entropy is used to measure the amount of randomness in a random variable. In this chapter, we introduce several more information-theoretic quantities. These quantities are important in the study of Shannon's results such as the calculation of channel capacity.

4.1 Information-Theoretic Quantities

Definition 4.1. Recall that, the **entropy** of a discrete random variable X is defined in Definition 2.41 to be

$$H(X) = -\sum_{x \in S_X} p_X(x) \log_2 p_X(x) = -\mathbb{E}\left[\log_2 p_X(X)\right]. \tag{19}$$

In this chapter, as in the previous chapter, X denotes the channel input. Recall that, in Section 3.1, S_X and $p_X(x)$ is denoted by $\mathcal{X}$ and $p(x)$, respectively. Under such notations, (19) becomes

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = -\mathbb{E}\left[\log_2 p(X)\right] = H(\underline{\mathbf{p}}) \tag{20}$$

and, similarly, for the channel output Y, we have

$$H(Y) = -\sum_{y \in \mathcal{Y}} q(y) \log_2 q(y) = -\mathbb{E}\left[\log_2 q(Y)\right] = H(\underline{\mathbf{q}}). \tag{21}$$

Definition 4.2. The **joint entropy** for two random variables X and Y is given by

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 p(x,y) = -\mathbb{E}\left[\log_2 p(X,Y)\right].$$

$\sum_{(x,y)}$

Example 4.3. Random variables X and Y have the following joint pmf matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \begin{bmatrix} \frac{1}{8} & \frac{1}{16} & \frac{1}{16} & \frac{1}{4} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} & 0 \\ \frac{1}{32} & \frac{1}{32} & \frac{1}{16} & 0 \\ \frac{1}{32} & \frac{1}{32} & \frac{1}{16} & 0 \end{bmatrix} \\ \begin{array}{cccc} 1/4 & 1/4 & 1/4 & 1/4 \end{array} \end{array} \begin{array}{c} p(x) \\ \to 1/2 \\ \to 1/4 \\ \to 1/8 \\ \to 1/8 \end{array}$$

Find $H(X)$, $H(Y)$ and $H(X,Y)$.

$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4}$ bits (per symbol)

$H(Y) = \left(-\frac{1}{4}\log_2\frac{1}{4}\right)\times 4 = \log_2 4 = 2$ bits (per symbol)

$H(X,Y) = \left(-\frac{1}{8}\log_2\frac{1}{8}\right)2 + \left(-\frac{1}{16}\log_2\frac{1}{16}\right)6 + \left(-\frac{1}{4}\log_2\frac{1}{4}\right)1 + \left(-\frac{1}{32}\log_2\frac{1}{32}\right)4$

$= \frac{27}{8}$ bits (per (X,Y)-pair)

Definition 4.4. The (conditional) entropy of Y when we know $X = x$ is denoted by $H(Y|X=x)$ or simply $H(Y|x)$. It can be calculated from

$$H(Y|x) = -\sum_{y\in\mathcal{Y}} Q(y|x)\log_2 Q(y|x)$$

$P[Y=y|X=x]$

- Note that the above formula is what we should expect it to be. When we want to find the entropy of Y, we use (21):

$$H(Y) = -\sum_{y\in\mathcal{Y}} q(y)\log_2 q(y).$$

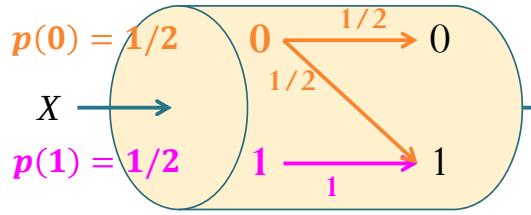
$P[Y=y]$

When we have an extra piece of information that $X = x$, we should update the probability about Y from the unconditional probability $q(y)$ to the conditional probability $Q(y|x)$.

- Note that when we consider $Q(y|x)$ with the value of x fixed and the value of y varied, we simply get the whole x-row from $\mathbf{Q}$ matrix. So, to

find $H(Y|x)$, we simply find the "usual" entropy from the probability values in the row corresponding to x in the $\mathbf{Q}$ matrix.

Example 4.5. Consider the following DMC (actually BAC)



$H(X,Y) = \left(-\frac{1}{4}\log_2\frac{1}{4}\right)2 + \left(-\frac{1}{2}\log_2\frac{1}{2}\right) = \frac{3}{2}$

$H(X) = \log_2 2 = 1$

$$Q = \begin{array}{c|cc} x\backslash y & 0 & 1 \\ \hline 0 & 1/2 & 1/2 \\ 1 & 0 & 1 \end{array} \xrightarrow{\times 1/2,\ \times 1/2} \begin{bmatrix} 1/4 & 1/4 \\ 0 & 1/2 \end{bmatrix} = P$$

$1/4 \quad 3/4$

$H(Y) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$

Originally $P[Y=y] = q(y) = \begin{cases} 1/4, & y=0, \\ 3/4, & y=1, \\ 0, & \text{otherwise.} \end{cases}$

(a) Suppose we know that $X = 0$.
The "x = 0" row in the $\mathbf{Q}$ matrix gives $Q(y|0) = \begin{cases} 1/2, & y=0,1, \\ 0, & \text{otherwise;} \end{cases}$
that is, given $x = 0$, the RV Y will be uniform.

$H(Y|X=0) = \log_2 2 = 1$

(b) Suppose we know that $X = 1$. The "x = 1" row in the $\mathbf{Q}$ matrix gives $Q(y|1) = \begin{cases} 1, & y=1, \\ 0, & \text{otherwise;} \end{cases}$ that is, given $x = 1$, the RV Y is degenerated (deterministic).

$H(Y|X=1) = 0$

Definition 4.6. **Conditional entropy**: The (average) conditional entropy of Y when we know X is denoted by $H(Y|X)$. It can be calculated from

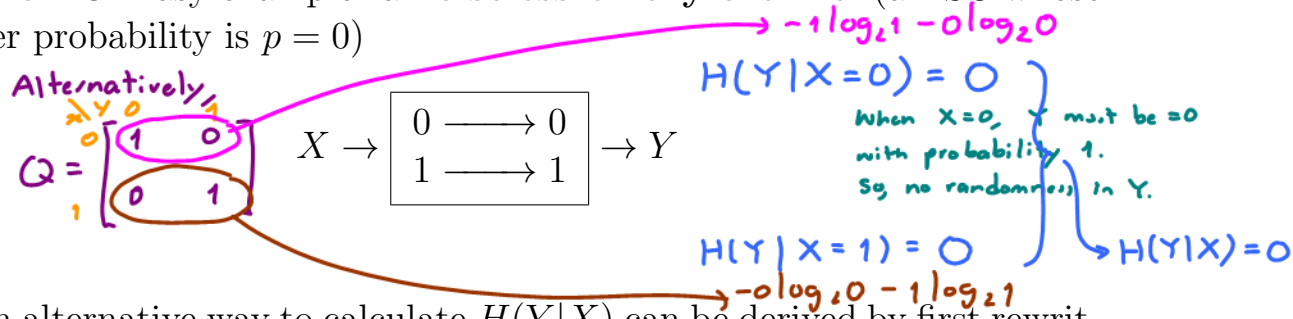
$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)\, H(Y|x).$$

Example 4.7. In Example 4.5,

$H(Y|X) = p(0)\,H(Y|X=0) + p(1)\,H(Y|X=1)$

$= \frac{1}{2}\times 1 + \frac{1}{2}\times 0 = \frac{1}{2}$

Example 4.8. Easy example: a **noiseless binary channel** (a BSC whose crossover probability is $p = 0$)



4.9. An alternative way to calculate $H(Y|X)$ can be derived by first rewriting it as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)\, H(Y|x) = -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} Q(y|x) \log_2 Q(y|x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log_2 Q(y|x) = -\mathbb{E}\left[\log_2 Q(Y|X)\right] \end{aligned}$$

Note that $Q(y|x) = \frac{p(x,y)}{p(x)}$. Therefore,

$$\begin{aligned} H(Y|X) &= -\mathbb{E}\left[\log_2 Q(Y|X)\right] = -\mathbb{E}\left[\log_2 \frac{p(X,Y)}{p(X)}\right] \\ &= \left(-\mathbb{E}\left[\log_2 p(X,Y)\right]\right) - \left(-\mathbb{E}\left[\log_2 p(X)\right]\right) \\ &= H(X,Y) - H(X) \end{aligned}$$

Example 4.10. In Example 4.5,

$H(Y|X) = H(X,Y) - H(X) = \frac{3}{2} - 1 = \frac{1}{2}$

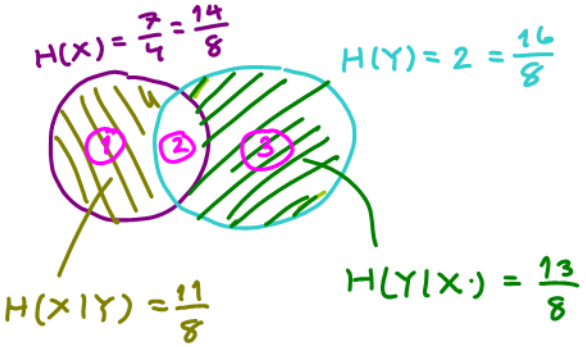
Example 4.11. Continue from Example 4.3. Recall that we got

$$H(X) = \frac{7}{4}, \quad H(Y) = 2, \quad H(X,Y) = \frac{27}{8}.$$ (1) + (2) + (3)

Find $H(Y|X)$ and $H(X|Y)$.

$H(Y|X) = H(X,Y) - H(X) = \frac{27}{8} - \frac{7}{4} = \frac{13}{8}$

$H(X|Y) = H(Y,X) - H(Y) = \frac{27}{8} - 2 = \frac{11}{8}$



$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cap B) = P(A) + P(B) - P(A \cup B)$

$= H(X) + H(Y) - H(X,Y)$

$= H(X) - H(X|Y) = \frac{14}{8} - \frac{11}{8} = \frac{3}{8}$

$= H(Y) - H(Y|X) = \frac{16}{8} - \frac{13}{8} = \frac{3}{8}$

$= H(X,Y) - H(X|Y) - H(Y|X)$

Definition 4.12. The **mutual information**[18] $I(X;Y)$ between two random variables X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y) \tag{22}$$

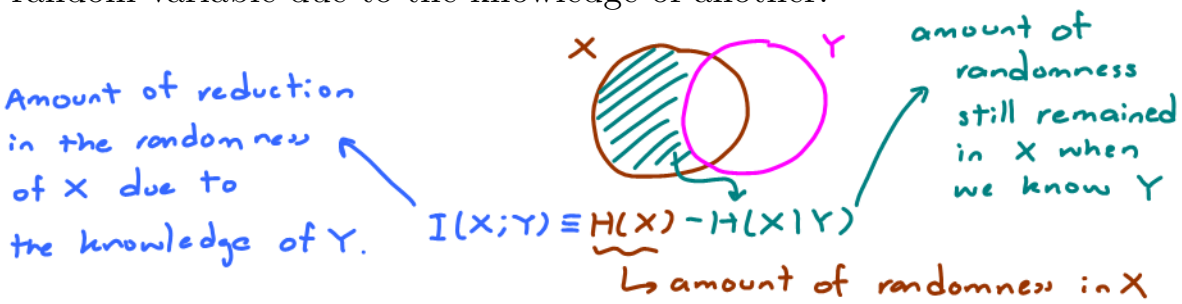
$$\equiv H(Y) - H(Y|X) \tag{23}$$

$$= H(X) + H(Y) - H(X,Y) \tag{24}$$

$$= \mathbb{E}\left[\log_2 \frac{p(X,Y)}{p(X)\,q(Y)}\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)\,q(y)} \tag{25}$$

$$= \mathbb{E}\left[\log_2 \frac{P_{X|Y}(X|Y)}{p(X)}\right] = \mathbb{E}\left[\log_2 \frac{Q(Y|X)}{q(Y)}\right]. \tag{26}$$

- Mutual information quantifies the reduction in the uncertainty of one random variable due to the knowledge of another.



- Mutual information is a measure of the amount of information one random variable contains about another [5, p 13].
- It is also natural to think of $I(X;Y)$ as a measure of how far X and Y are from being independent.
    - Technically, it is the (Kullback-Leibler) divergence between the joint and product-of-marginal distributions.

4.13. Some important properties

(a) $H(X,Y) = H(Y,X)$ and $I(X;Y) = I(Y;X)$.
However, in general, $H(X|Y) \neq H(Y|X)$.

(b) I and H are always ≥ 0.

(c) There is a one-to-one correspondence between Shannon's information measures and set theory. We may use an **information diagram**, which

[18] The name mutual information and the notation $I(X;Y)$ was introduced by [Fano, 1961, Ch 2].

is a variation of a Venn diagram, to represent relationship between Shannon's information measures. This is similar to the use of the Venn diagram to represent relationship between probability measures. These diagrams are shown in Figure 16.

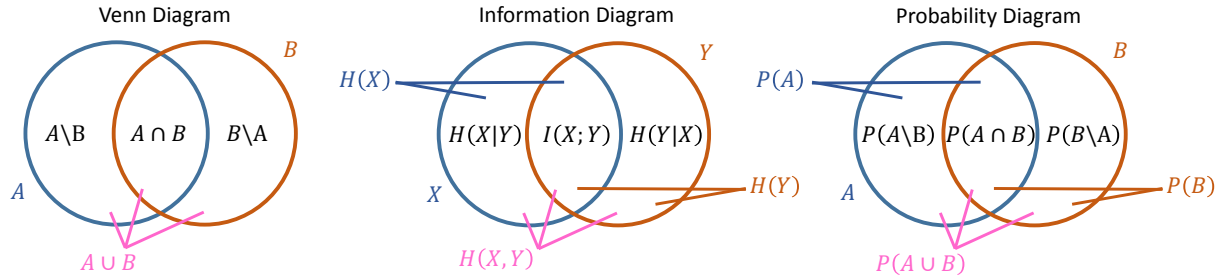


Figure 16: Venn diagram and its use to represent relationship between information measures and relationship between probabilities.

- Many information-theoretic properties can be easily "read" from the information diagram.
- Chain rule for information measures:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

- Caution: In probability theory, comma (",") is associated with "and" (intersection); that is, $P(A,B)$ is the same as $P(A \cap B)$.) However, for entropy, the notation is different. The use of "comma" in $H(X,Y)$ turns out to represent "union" of randomness. The "intersection" of randomness is denoted by semicolon (";") in $I(X;Y)$.

(d) $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.

- When this property is applied to the information diagram (or definitions (22), (23), and (24) for $I(X,Y)$), we have

  (i) $H(X|Y) \leq H(X)$,

  (ii) $H(Y|X) \leq H(Y)$,

  (iii) $H(X,Y) \leq H(X) + H(Y)$

  Moreover, each of the inequalities above becomes equality if and only if $X \perp\!\!\!\perp Y$.

(e) We have seen in Section 2.4 that

$$\underset{\text{deterministic (degenerated)}}{0} \leq H(X) \leq \underset{\text{uniform}}{\log_2 |\mathcal{X}|}. \tag{27}$$

Similarly,

$$\underset{\text{deterministic (degenerated)}}{0} \leq H(Y) \leq \underset{\text{uniform}}{\log_2 |\mathcal{Y}|}. \tag{28}$$

For conditional entropy, we have

$$\underset{\exists g\, Y=g(X)}{0} \leq H(Y|X) \leq \underset{X \perp\!\!\!\perp Y}{H(Y)} \tag{29}$$

and

$$\underset{\exists g\, X=g(Y)}{0} \leq H(X|Y) \leq \underset{X \perp\!\!\!\perp Y}{H(X)}. \tag{30}$$

For mutual information, we have

$$\underset{X \perp\!\!\!\perp Y}{0} \leq I(X;Y) \leq \underset{\exists g\, X=g(Y)}{H(X)} \tag{31}$$

and

$$\underset{X \perp\!\!\!\perp Y}{0} \leq I(X;Y) \leq \underset{\exists g\, Y=g(X)}{H(Y)}. \tag{32}$$

Combining 27, 28, 31, and 32, we have

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\} \leq \min\{\log_2 |\mathcal{X}|, \log_2 |\mathcal{Y}|\} \tag{33}$$

(f) $H(X|X) = 0$ and $I(X;X) = H(X)$.

Example 4.14. Find the mutual information $I(X;Y)$ between the two random variables X and Y whose joint pmf matrix is given by $\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 \end{bmatrix}$.

$I(X;Y) = H(X) + H(Y) - H(X,Y)$

$\approx 0.8113 + 0.8113 - 1.5 = 0.1226$

$$P = \begin{bmatrix} 1/2 & 1/4 \\ 1/4 & 0 \end{bmatrix} \begin{matrix} \Sigma \to 3/4 \\ \Sigma \to 1/4 \end{matrix}$$

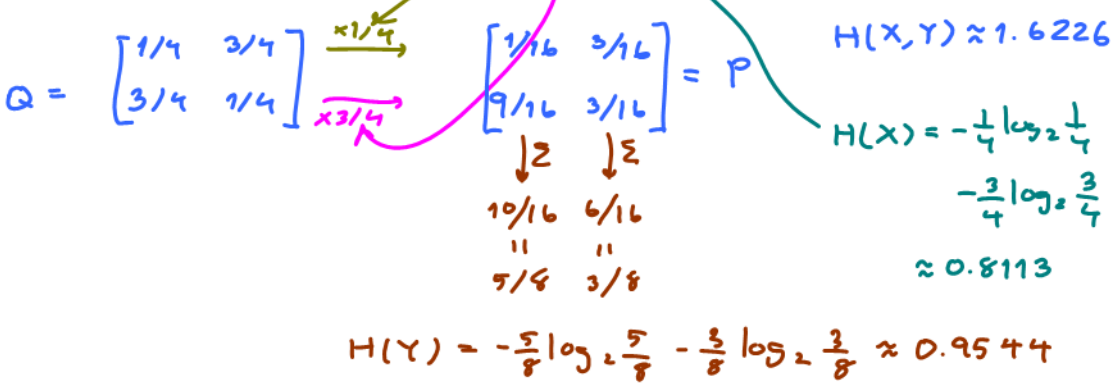
$$\begin{matrix} \Sigma \downarrow & \downarrow \Sigma \\ 3/4 & 1/4 \end{matrix}$$

$H(X,Y) = -\frac{1}{2}\log_2\frac{1}{2} - \left(\frac{1}{4}\log_2\frac{1}{4}\right)2 = 1.5$

$H(X) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \approx 0.8113 = H(Y)$

Example 4.15. Find the mutual information $I(X;Y)$ between the two random variables X and Y whose $\underline{\mathbf{p}} = \left[\frac{1}{4}\ \frac{3}{4}\right]$ and $\mathbf{Q} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$.

Method 1: First, convert the given information into the joint pmf matrix.



Then, $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

≈ 0.1432

Method 2: Use $I(X;Y) = H(Y) - H(Y|X)$.

(a) To find $H(Y)$, we need $q(y)$:

$$\underline{\mathbf{q}} = \underline{\mathbf{p}}\mathbf{Q} = \left[\frac{1}{4}, \frac{3}{4}\right] \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix} = \left[\frac{10}{16}, \frac{6}{16}\right] = \left[\frac{5}{8}, \frac{3}{8}\right].$$

This gives $H(Y) \approx 0.9544$.

(b) $H(Y|X) = \sum_x p(x) H(Y|x)$. So, we need $H(Y|x)$. Observe that each row of $\mathbf{Q}$ is $\left[\ \frac{1}{4}\ \ \frac{3}{4}\ \right]$. Therefore,

$$H(Y|x) = H\left(\left[\ \frac{1}{4}\ \ \frac{3}{4}\ \right]\right) \approx 0.8113$$

for any x (for any row of $\mathbf{Q}$). This gives

$$H(Y|X) = \sum_x p(x) H(Y|x) \approx \sum_x p(x) \times 0.8113$$

$$= 0.8113 \left(\sum_x p(x)\right) = 0.8113.$$

Finally,

$$I(X;Y) = H(Y) - H(Y|X) \approx 0.1432.$$

4.2 Operational Channel Capacity

4.16. In Chapter 3, we have studied how to compute the error probability $P(\mathcal{E})$ for digital communication systems over DMC. At the end of that chapter, we studied block encoding where the channel is used n times to transmit a k-bit info-block.

In this section, our consideration is "reverse".

4.17. In this and the next sections, we introduce a quantity called channel capacity which is crucial in benchmarking communication system. Recall that, in Chapter 2 where source coding was discussed, we were interested in the minimum rate (in bits per source symbol) to represent a source. Here, we are interested in the maximum rate (in bits per channel use) that can be sent through a given channel *reliably*.

4.18. Here, **reliable communication** means *arbitrarily* small error probability *can be achieved*.

- This seems to be an impossible goal.
  - If the channel introduces errors, how can one correct them all?
    - Any correction process is also subject to error, ad infinitum.

Definition 4.19. Given a DMC, its **"operational" channel capacity** is the maximum rate at which *reliable* communication over the channel *is possible*.

$\frac{k}{n}$

- The channel capacity is the maximum rate in bits per channel use at which information *can be* sent with **arbitrarily low** error probability.

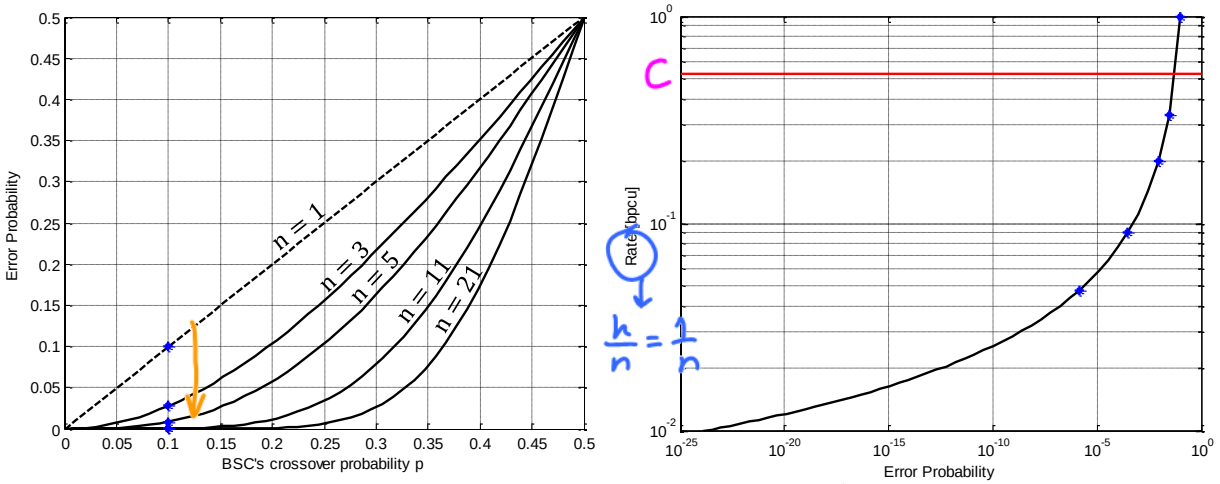
4.20. Claude Shannon showed, in his 1948 landmark paper, that this operational channel capacity is the same as the information channel capacity which we will discuss in the next section. From this, we can omit the words

"operational" and "information" and simply refer to both quantities as the *channel capacity*.

Example 4.21. In Example 4.35, we will find that the capacity of a BSC with crossover probability $p = 0.1$ is approximately 0.531 bits per channel use. This means that for any rate $R < 0.531$ and any error probability $P(\mathcal{E})$ that we desire, as long as it is greater than 0, we can find a suitable n, a rate R encoder, and a corresponding decoder which will yield an error probability that is at least as low as our set value.

- Usually, for very low desired value of $P(\mathcal{E})$, we may need large value of n.

Example 4.22. Repetition code is not good enough.



(a) (b) $p = 0.1$

Figure 17: Performance of repetition coding with majority voting at the decoder

- Continue from Example 4.21.
- In Figure 17b, with repetition code, trying to reduce the error probability to be less than the original p even a little bit already causes the rate to drop far below the capacity level indicated by the red horizontal line.
- In fact, for any rate > 0, we can see from Figure 17b that communication system based on repetition coding is not "reliable" according

to Definition 4.18. For example, for rate = 0.02 bits per channel use, repetition code can't satisfy the requirement that the error probability must be less than 10^{-15}. In fact, Figure 17b shows that as we reduce the error probability to 0, the rate also goes to 0 as well. Therefore, there is no positive rate that works for all error probability.

- However, because the channel capacity is 0.531 [bpcu], there must exist other encoding techniques which give better error probability than repetition code.
    - Although Shannon's result gives us the channel capacity, it does not give us any explicit instruction on how to construct codes which can achieve that value.

4.3 Information Channel Capacity

4.23. In Section 4.1, we have studied how to compute the value of mutual information $I(X;Y)$ between two random variables X and Y. Recall that, here, X and Y are the channel input and output, respectively. We have also seen, in Example 4.14, how to compute $I(X;Y)$ when the joint pmf matrix $\mathbf{P}$ is given. Furthermore, we have also worked on Example 4.15 in which the value of mutual information is computed from the prior probability vector $\underline{\mathbf{p}}$ and the channel transition probability matrix $\mathbf{Q}$. This second type of calculation is crucial in the computation of channel capacity. This kind of calculation is so important that we may write the mutual information $I(X;Y)$ as $I(\underline{\mathbf{p}}, \mathbf{Q})$.

Definition 4.24. Given a DMC channel, we define its "information" channel capacity as

$$C = \max_{\underline{\mathbf{p}}} I(X;Y) = \max_{\underline{\mathbf{p}}} I\left(\underline{\mathbf{p}}, \mathbf{Q}\right), \tag{34}$$

where the maximum is taken over all possible input pmfs $\underline{\mathbf{p}}$.

- Again, as mentioned in 4.20, Shannon showed that the "information" channel capacity defined here is equal to the "operational" channel capacity defined in Definition 4.19.
    - Thus, we may drop the word "information" in most discussions of channel capacity.

Calculating the channel capacity

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

1. Use (multi-variable) calculus
2. Blahut-Arimoto (MATLAB)
3. Check whether Q matches with any known special cases
   - (i) NO^2 $\Rightarrow C = \log_2 |\mathcal{X}|$ is obtained by uniform X
   - (ii) weakly symmetric $\Rightarrow C = \log_2 |\mathcal{Y}| - H(\underline{r})$ is obtained by uniform X
   - (iii) repeated rows $\Rightarrow C = 0$ is obtained by any $\underline{p}$
   - (iv) ...

Note: Do not assume that the input probabilities will have to be uniform to obtain C. See BAC in Ex. 4.25.

Example 4.25. The capacity of a BAC whose $Q(1|0) = 0.9$ and $Q(0|1) = 0.4$ can be found by first realizing that $I(X;Y)$ here is a function of a single variable: p_0. The plot of $I(X;Y)$ as a function of p_0 gives some rough estimates of the answers. One can directly solve for the optimal p_0 by simply taking derivative of $I(X;Y)$ and set it equal to 0. This gives the capacity value of 0.0918 bpcu which is achieved by $\underline{\mathbf{p}} = [0.5376,\ 0.4624]$.

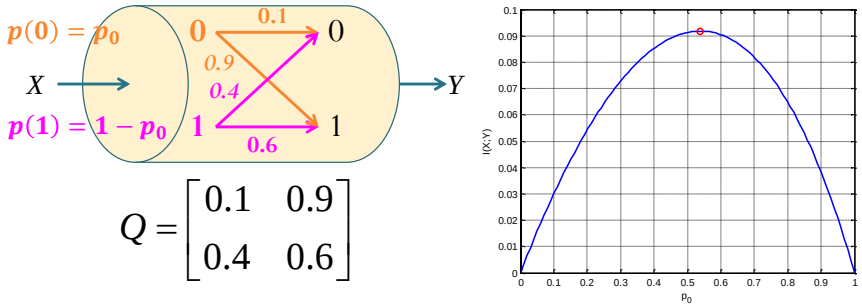


Figure 18: Maximization of mutual information to find capacity of a BAC channel. Capacity of 0.0918 bits is achieved by $\underline{\mathbf{p}} = [0.5376,\ 0.4624]$

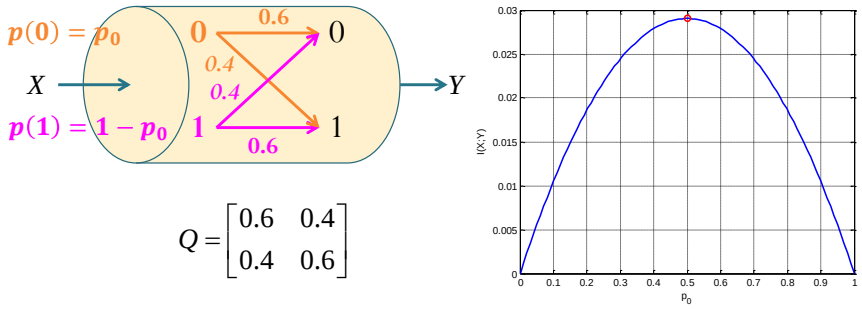


Figure 19: Maximization of mutual information to find capacity of a BSC channel. Capacity of 0.029 bits is achieved by $\underline{\mathbf{p}} = [0.5,\ 0.5]$

4.26. Blahut-Arimoto Algorithm [5, Section 10.8]: Alternatively, in 1972, Arimoto [1] and Blahut [2] independently developed an iterative algorithm to help us approximate the pmf $\underline{\mathbf{p}}^*$ which achieves capacity C. To do this, start with any (guess) input pmf $p_0(x)$, define a sequence of pmfs $p_r(x)$, $r = 0, 1, \ldots$ according to the following iterative prescription:

(a) $q_r(y) = \sum_x p_r(x)\, Q(y|x)$ for all $y \in \mathcal{Y}$.

(b) $c_r(x) = 2^{\left(\sum_y Q(y|x) \log_2 \frac{Q(y|x)}{q_r(y)}\right)}$ for all $x \in \mathcal{X}$.

(c) It can be shown that

$$\log_2\left(\sum_x p_r(x)\,c_r(x)\right) \le C \le \log_2\left(\max_x c_r(x)\right).$$

- If the lower-bound and upper-bound above are close enough. We take $p_r(x)$ as our answer and the corresponding capacity is simply the average of the two bounds.
- Otherwise, we compute the pmf

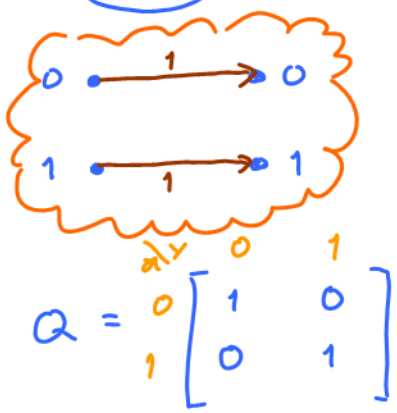
$$p_{r+1}(x) = \frac{p_r(x)\,c_r(x)}{\sum_x p_r(x)\,c_r(x)} \quad \text{for all } x \in \mathcal{X}$$

and repeat the steps above with index r replaced by $r+1$.

4.4 Special Cases for Calculation of Channel Capacity

In this section, we study special cases of DMC whose capacity values can be found (relatively) easy.

Example 4.27. Continue from Example 4.8 where we considered a noiseless binary channel. Find the corresponding channel capacity.



$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

$H(X|Y) = 0$ ← Intuitively, knowing Y, the value of X is completely determined.

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

$= H(X)$

$C = \max_P I(X;Y) = \max_P H(X)$

So, we maximize $I(X;Y)$ by $P = [\frac{1}{2} \quad \frac{1}{2}]$.

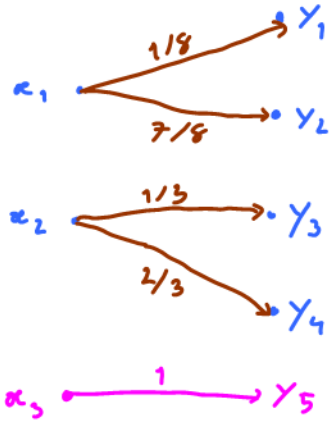
$C = \log_2|\mathcal{X}| = \log_2 2 = 1$ [bpcu]

bit per channel use.

Example 4.28. Noisy Channel with Nonoverlapping Outputs: Find the channel capacity of a DMC whose

Example 4.28b

$$\mathbf{Q} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} y_1 & y_2 & y_3 & y_4 & y_5 \end{array} \\ \begin{bmatrix} 1/8 & 7/8 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 2/3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$



$H(X|Y) = 0$

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

$= H(X)$

$C = \max_P I(X;Y) = \max_P H(X) = \log_2|\mathcal{X}| = 1$ [bpcu]

which happens when $P = [\frac{1}{2} \quad \frac{1}{2}]$

$C = \log_2|\mathcal{X}| = \log_2 3 \approx 1.585$ [bpcu] which happens when $P = [\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}]$

In this example, the channel appears to be noisy, but really is not. Even though the output of the channel is a random consequence of the input, the input can be determined from the output, and hence every transmitted bit can be recovered without error.

4.29. Reminder:

(a) Some definitions involving entropy

(i) Binary entropy function: $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$

(ii) $H(X) = -\sum_x p(x)\log_2 p(x)$

(iii) $H(\underline{\mathbf{p}}) = -\sum_i p_i \log_2(p_i)$

(b) A key entropy property that will be used frequently in this section is that for any random variable X,

$H(X) \le \log_2 |\mathcal{X}|$ with equality iff X is uniform.

4.30. A DMC is a **noisy channel with nonoverlapping outputs** (NO^2) when there is only one non-zero element in each column of its $\mathbf{Q}$ matrix. For such channel,

$C = \log_2 |\mathcal{X}|$ is achieved by uniform $p(x)$.

Definition 4.31. A DMC is called **symmetric** if (1) all the rows of its probability transition matrix $\mathbf{Q}$ are permutations of each other **and** (2) so are the columns.

Example 4.32. For each of the following $\mathbf{Q}$, is the corresponding DMC symmetric?

$$\begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}, \quad \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.2 & 0.3 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}, \quad \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}, \quad \begin{bmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \end{bmatrix}$$

(1) ✓ (2) ✓ ⇒ symmetric ⇓ (1) ✓ (2) ✓ Yes

(1) ✓ (2) × ⇒ not symmetric ⇑ (1) ✓ (2) × No

not symmetric (1) ✓ (2) ✓ Yes

Not symmetric (1) × (2) × No

weakly symmetric

4.33. Q: Does symmetric DMC always have square $\mathbf{Q}$? A: No

$$\begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \qquad \begin{bmatrix} 0.3 & 0.2 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.3 & 0.2 \end{bmatrix}$$

Example 4.34. Find the channel capacity of a DMC whose

$$\mathbf{Q} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix} \end{array}.$$

$I(X;Y) = H(X) - H(X|Y)$, $H(X|Y) \neq 0$

Solution: First, recall that the capacity C of a given DMC can be found by (34):

$$C = \max_{\underline{\mathbf{p}}} I(X;Y) = \max_{\underline{\mathbf{p}}} I(\underline{\mathbf{p}}, \mathbf{Q}).$$

$H(Y|x_1) = H([0.3 \ 0.2 \ 0.5]) = H(\underline{r})$

$H(Y|x_2) = H([0.5 \ 0.3 \ 0.2]) = H(\underline{r}) = H(Y|x_3)$

$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\underline{r})$

We see that, to maximize $I(X;Y)$, we need to maximize $H(Y)$. Of course, we know that the maximum value of $H(Y)$ is $\log_2 |\mathcal{Y}|$ which happens when Y is uniform. Therefore, if we can find $\underline{\mathbf{p}}$ which makes Y uniform, then this same $\underline{\mathbf{p}}$ will give the channel capacity.

$$\underline{q} = \underline{p}\mathbf{Q} = \begin{bmatrix} \alpha & \alpha & \alpha \end{bmatrix} \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix} = \begin{bmatrix} \alpha 0.3 + \alpha 0.5 + \alpha 0.2 & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} \alpha(0.3+0.5+0.2) & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Try uniform X $\rightarrow$ get uniform Y

$$C = \log_2 |\mathcal{Y}| - H(\underline{r}) = \log_2 3 - H([0.3 \ 0.2 \ 0.5]) \approx 0.0995 \text{ [bpcu]}$$

Remark: If we can't find $\underline{\mathbf{p}}$ that makes Y uniform, then $C < \log_2 |\mathcal{Y}| - H(\underline{\mathbf{r}})$ and we have to find a different technique to calculate C.

Example 4.35. Find the channel capacity of a BSC whose crossover probability is $p = 0.1$.

$$\mathbf{Q} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \end{array}$$

$C = \log_2 |\mathcal{Y}| - H(\underline{r}) = \log_2 2 - H([0.9 \ 0.1])$

$= 1 - (-0.9 \log_2 0.9 - 0.1 \log_2 0.1) \approx 0.531 \text{ [bpcu]}$

Definition 4.36. A DMC is called **weakly symmetric** if (1) all the rows of its probability transition matrix $\mathbf{Q}$ are permutations of each other and (2) all the column sums are equal.

- It should be clear from the definition that a symmetric channel is automatically weakly symmetric.

Condition (2) guarantees that uniform X gives uniform Y.

Condition (1) makes all $H(Y|x)$ the same; so, $H(Y|X)$ does not depend on $\underline{p}$. (It is a constant.)

4.37. For a weakly symmetric channel,

$$C = \log_2 |\mathcal{Y}| - H(\underline{\mathbf{r}}),$$

where $\underline{\mathbf{r}}$ is any row from the $\mathbf{Q}$ matrix. The capacity is achieved by a uniform pmf on the channel input.

- Important special case: For BSC, $C = 1 - H(p)$.

4.38. Properties of channel capacity

(a) $C \geq 0$

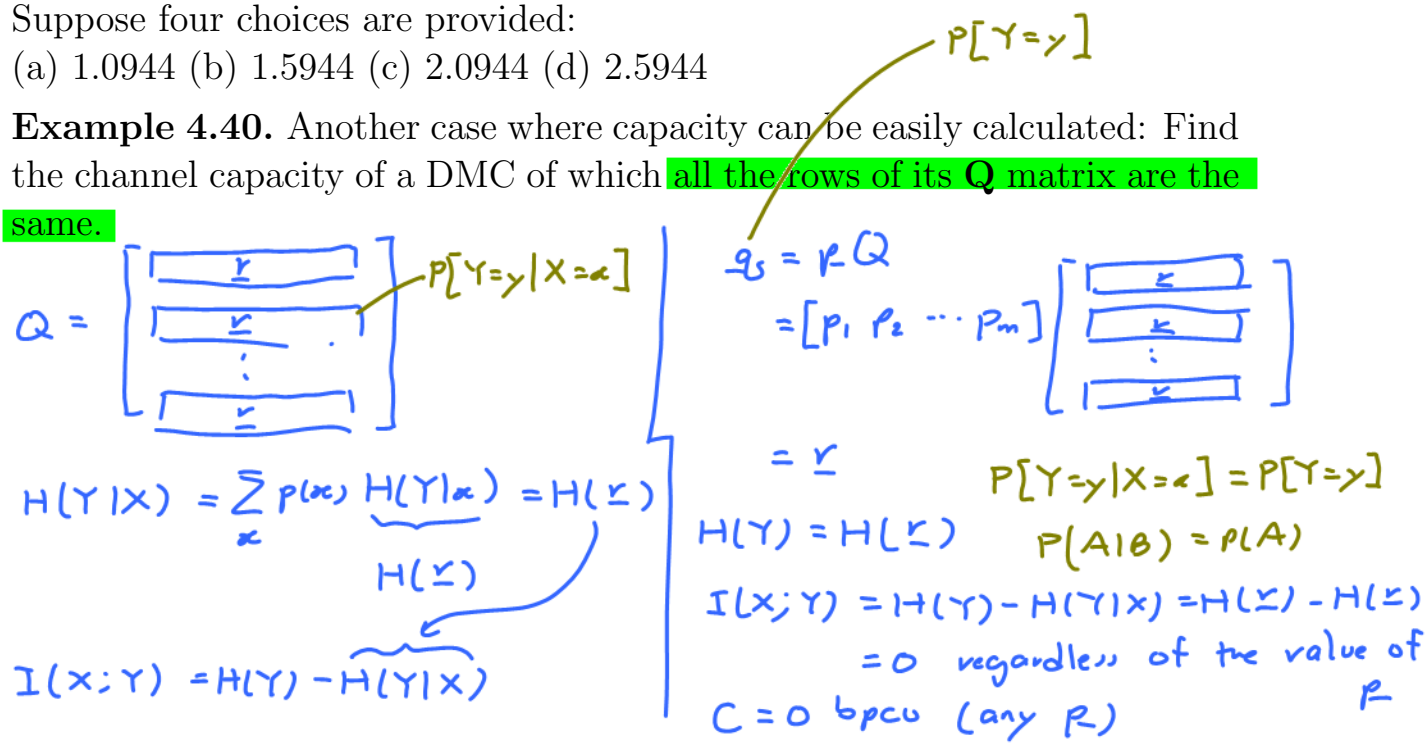
(b) $C \leq \min\{\log_2 |\mathcal{X}|, \log_2 |\mathcal{Y}|\}$

Example 4.39. Find the channel capacity of a DMC whose

$$\mathbf{Q} = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.025 & 0.025 & 0.95 \end{bmatrix}$$

Suppose four choices are provided:
(a) 1.0944 (b) 1.5944 (c) 2.0944 (d) 2.5944

Example 4.40. Another case where capacity can be easily calculated: Find the channel capacity of a DMC of which all the rows of its $\mathbf{Q}$ matrix are the same.



4.41. In this section, we worked with "toy" examples in which finding capacity is relatively easy. In general, there is no closed-form solution for computing capacity. When we have to deal with cases that do not fit in any special family of $\mathbf{Q}$ described in the examples above, the maximum may be found by standard nonlinear optimization techniques or the Blahut-Arimoto Algorithm discussed in 4.26.

4.5 Shannon's Coding theorem

4.42. Shannon's (Noisy Channel) Coding theorem [Shannon, 1948]

(a) Reliable communication over a (discrete memoryless) channel is possible if the communication rate R satisfies $R < C$, where C is the channel capacity.

In particular, for any $R < C$, there exist codes (encoders and decoders) with sufficiently large n such that

$$P(\mathcal{E}) \leq 2^{-n \times E(R)},$$

where $E(R)$ is

- a positive function of R for $R < C$ and
- completely determined by the channel characteristics

(b) At rates higher than capacity, reliable communication is impossible.

4.43. Significance of Shannon's (noisy channel) coding theorem:

(a) Express the limit to reliable communication

(b) Provides a yardstick to measure the performance of communication systems.

- A system performing near capacity is a near optimal system and does not have much room for improvement.
- On the other hand a system operating far from this fundamental bound can be improved (mainly through coding techniques).

4.44. Shannon's nonconstructive proof for his coding theorem

- Shannon introduces a method of proof called **random coding**.
- Instead of looking for the best possible coding scheme and analyzing its performance, which is a difficult task,
  - all possible coding schemes are considered
    - by generating the code randomly with appropriate distribution
  - and the performance of the system is averaged over them.

- Then it is proved that if $R < C$, the average error probability tends to zero.

- Again, Shannon proved that
  - as long as $R < C$,
  - at any arbitrarily small (but still positive) probability of error,
  - one can find (there exist) at least one code (with sufficiently long block length n) that performs better than the specified probability of error.

- If we used the scheme suggested and generate a code at random, the code constructed is likely to be good for long block lengths.

- No structure in the code. Very difficult to decode

4.45. Practical codes:

- In addition to achieving low probabilities of error, useful codes should be "simple", so that they can be encoded and decoded efficiently.

- Shannon's theorem does not provide a practical coding scheme.

- Since Shannon's paper, a variety of techniques have been used to construct good error correcting codes.
  - The entire field of coding theory has been developed during this search.

- Turbo codes have come close to achieving capacity for Gaussian channels.