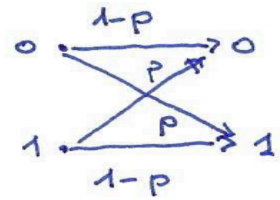
Definition 2.3.5 Given a channel with input alphabet $\mathcal{X}$ output alphabet $\mathcal{Y}$ and transition probabilities $p(y|x) \geq 0$; $\sum_{y \in \mathcal{Y}} p(y|x) = 1$, its <u>capacity</u> is given by

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

Example 2.3.2. Binary symmetric channel



$0 \leq p \leq 1$

Transition matrix (column ↑, row →)

$$p(y|x) = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) + \sum_{x=0}^{1} p(x) \sum_{y=0}^{1} p(y|x) \log p(x|x)$$

$$= H(Y) + p \log p + (1-p) \log (1-p) = H(Y) - h(p)$$

$$h(p) := -p \log p - (1-p) \log (1-p)$$

$$H(Y) = -\sum_{y=0}^{1} P_Y(y) \log P_Y(y) \; ; \qquad P_Y(y) = \sum_{x=0}^{1} p(y|x)\, P_X(x)$$

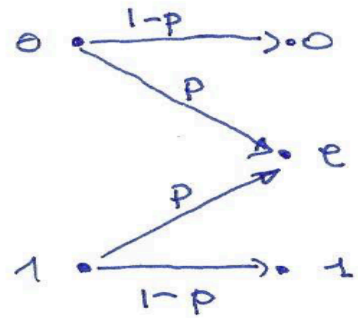
$$P_Y(0) = p(0|0)\, P_X(0) + p(0|1) P_X(1) = (1-p) P_X(0) + p\, P_X(1)$$

$$P_Y(1) = p(1|0)\, P_X(0) + p(1|1)\, P_X(1) = p\, P_X(0) + (1-p) P_X(1)$$

$$H(Y) \leq 1 \; ; \quad H(Y) = 1 \iff P_Y(0) = \tfrac{1}{2} = P_Y(1) \iff P_X(0) = P_X(1) = \tfrac{1}{2}$$

$$C = \max_{\pi_X} H(Y) - h(p) = 1 - h(p) \longleftarrow \pi_X = \left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$

Exemple 2.3.3 Binary erasure channel



$$p(y|x) = \begin{pmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{pmatrix}$$

$$H(Y|X=0) = H(Y|X=1) = h(p) \Rightarrow \boxed{I(X;Y) = H(Y) - h(p)}$$

$$H(Y) = -P_Y(0)\log P_Y(0) - P_Y(e)\log P_Y(e) - P_Y(1)\log P_Y(1)$$

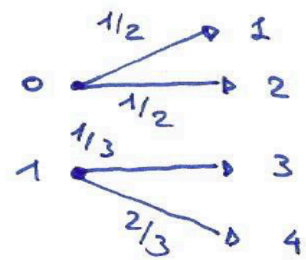
$$P_Y(0) = (1-p)P_X(0) \;;\quad P_Y(e) = p\,P_X(0) + p\,P_X(1) = p \;;\quad P_Y(1) = (1-p)P_X(1)$$

$$H(Y) = -(1-p)P_X(0)\left(\log(1-p) + \log P_X(0)\right) - p\log p - (1-p)P_X(1)\left(\log(1-p) + \log P_X(1)\right)$$

$$= h(p) + (1-p)\,h(\pi) \qquad \pi := P_X(0)$$

$$C = \max_{\pi_X} H(Y) - h(p) = (1-p)\max_{\pi_X} h(\pi) = 1-p \quad \Leftarrow \quad \pi_X = \left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$

Example 2.3.4 Noisy channel with non-overlapping outputs



$$P(Y|X) = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & \frac{2}{3} \end{pmatrix}$$

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

$$H(Y|X=1) = \tfrac{1}{3} + \log 3$$

$$H(Y|X) = P_X(0) + P_X(1)\left(\log 3 - \tfrac{2}{3}\right)$$

$$P_Y(1) = \frac{P_X(0)}{2} \; ; \quad P_Y(2) = \frac{P_X(0)}{2} \; ; \quad P_Y(3) = \frac{P_X(1)}{3} \; ; \quad P_Y(4) = 2\frac{P_X(1)}{3}$$

$$H(Y) = -P_X(0)\log\frac{P_X(0)}{2} - \frac{P_X(1)}{3}\log\frac{P_X(1)}{3} - \frac{2}{3}P_X(1)\log\left(\frac{2}{3}P_X(1)\right)$$

$$= -P_X(0)\log P_X(0) + P_X(0) - P_X(1)\log P_X(1) + P_X(1)\log 3 - \frac{2}{3}P_X(1)$$

$$I(X;Y) = h(\pi) + P_X(1)\log 3 - \frac{2}{3}P_X(1) + P_X(0) - P_X(0) - P_X(1)\log 3 + \frac{2}{3}P_X(1)$$

$$= h(\pi) \leq 1 \qquad \pi := P_X(0)$$

$$C = \max_{\pi_X} h(\pi) = 1 \Leftarrow \pi_X = \{\tfrac{1}{2}, \tfrac{1}{2}\}$$

Definition 2.3.6 An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$

consists of

1. an index set $\{1, 2, \dots, M\}$;
2. an encoding function $X^{(n)} : \{1, 2, \dots, M\} \to \mathcal{X}^{(n)}$ yielding codewords $X^{(n)}(w) = X_1(w) X_2(w) \dots X_n(w) \in \mathcal{X}^{(n)}$ where $w \in \{1, 2, \dots, M\}$;
3. a decoding function $g : \mathcal{Y}^{(n)} \to \{1, 2, \dots, M\}$ which assigns a guess to each received $Y^{(n)} \in \mathcal{Y}^{(n)}$.

Definition 2.3.7 Let

$$\lambda_i = \text{Prob}\left(g(Y^{(n)}) \neq w \mid X^{(n)} = X^{(n)}(w)\right)$$

$$= \sum_{y^{(n)} \in \mathcal{Y}^{(n)} : g(y^{(n)}) \neq w} p(y^{(n)} \mid x^{(n)}(w))$$

be the conditional probability of error upon sending the index w.

The maximal probability of error for an (M, n)-code is

$$\lambda^{(n)} = \max_{w \in \{1, 2, \dots, M\}} \lambda_w$$

Definition 2.3.8 The rate R of an (M,n) code is

$$R := \frac{\log M}{n} \quad \text{bits per channel use}$$

A rate R is <u>achievable</u> if there exists a sequence of $(2^{Rn}, n)$ codes such that the maximal probability of error $d^{(n)} \xrightarrow[n]{} 0$.

Definition 2.3.9 The capacity of a discrete memoryless channel is the supremum of all achievable rates:

$$C = \sup_R \{R \text{ achievable}\}$$

<u>Remark</u>: The capacity of a discrete memoryless channel is the highest number of bits that one can send per use of the noisy channel.

Remark : given codewords $x^{(n)}$ chosen i.i.d with single choice Shannon entropy $H(X)$ there will be roughly $\boxed{2^{nH(X)}}$ typical input words (see AEP in Example 2.1.3).

By sending the codewords $x^{(n)}$ through the discrete memoryless channel, one obtains output words $y^{(n)}$ that are also i.i.d. Indeed,

$$p(y^{(n)}) = \sum_{x^{(n)}} p(y^{(n)}|x^{(n)})\, p(x^n) = \prod_{j=1}^{n} \Big[\sum_j p(y_i|x_j)\, p(x_j) = \prod_{i=1}^{n} p(y_i)$$

Therefore there will be $\boxed{2^{nH(Y)}}$ typical output words.

Since $p(x^{(n)}, y^{(n)}) = p(y^{(n)}|x^{(n)})\, p(x^{(n)}) = \prod_{i=1}^{n} p(y_i, x_j)$,

there will be $\boxed{2^{nH(X,Y)}}$ jointly typical pair of words.

By decoding $y^{(n)}$ with w if $y^{(n)}$ is jointly typical with $x^{(n)}(w)$, an error will occur with probability

$$\boxed{\frac{2^{nH(X,Y)}}{2^{n(H(X)+H(Y))}} = 2^{-nI(X;Y)}}$$

Definition 2.3.10 The set $A_\varepsilon^{(n)}$ of jointly typical sequences $\{(x^{(n)}, y^{(n)})\}$ with respect to the distribution $p(x,y)$ is the set of sequences such that

$$A_\varepsilon^{(n)} = \Big\{ (x^{(n)}, y^{(n)}) \in \mathcal{X}^{(n)} \times \mathcal{Y}^{(n)} : \Big| -\frac{1}{n} \log p(x^{(n)}) - H(X) \Big| < \varepsilon,$$

$$\Big| -\frac{1}{n} \log p(y^{(n)}) - H(Y) \Big| < \varepsilon,$$

$$\Big| -\frac{1}{n} \log p(x^{(n)}, y^{(n)}) - H(X,Y) \Big| < \varepsilon \Big\}, \text{ where}$$

$$p(x^{(n)}, y^{(n)}) = \prod_{j=1}^{n} p(x_j, y_j).$$

Theorem 2.3.1 joint AEP

Let $(x^{(n)}, y^{(n)})$ be sequences of length n drawn i.i.d. according to $p(x^{(n)}, y^{(n)}) = \prod_{j=1}^{n} p(x_j, y_j)$. Then, 1) $\text{Prob}\big((x^{(n)}, y^{(n)}) \in A_\varepsilon^{(n)}\big) \xrightarrow{n} 1$; 2) $\#(A_\varepsilon^{(n)}) \le 2^{n(H(X,Y)+\varepsilon)}$;

3) if $\tilde{x}^{(n)}, \tilde{y}^{(n)}$ are independent with the marginals of $p(x^{(n)}, y^{(n)})$, then

$$\text{Prob}\big((\tilde{x}^{(n)}, \tilde{y}^{(n)}) \in A_\varepsilon^{(n)}\big) \le 2^{-n(I(X;Y)-3\varepsilon)}; \quad (1-\varepsilon)\, 2^{-n(I(X,Y)+3\varepsilon)} \le \text{Prob}\big((\tilde{x}^{(n)}, \tilde{y}^{(n)}) \in A_\varepsilon^{(n)}\big).$$

Proof: from the weak law of large numbers

$$\begin{cases} -\frac{1}{n}\log p(x^{(n)}) \xrightarrow{P} H(X) \\ -\frac{1}{n}\log p(y^{(n)}) \xrightarrow{P} H(Y) \\ -\frac{1}{n}\log p(x^{(n)}, y^{(n)}) \xrightarrow{P} H(X,Y) \end{cases}$$

Then, given the sets

$$B^{(n)}_{1,\varepsilon} = \left\{ x^{(n)} \in \mathcal{X}^{(n)} : \left| -\frac{1}{n}\log p(x^{(n)}) - H(X) \right| > \varepsilon \right\}$$

$$B^{(n)}_{2,\varepsilon} = \left\{ y^{(n)} \in \mathcal{Y}^{(n)} : \left| -\frac{1}{n}\log p(y^{(n)}) - H(Y) \right| > \varepsilon \right\}$$

$$B^{(n)}_{3,\varepsilon} = \left\{ (x^{(n)}, y^{(n)}) \in \mathcal{X}^{(n)} \times \mathcal{Y}^{(n)} : \left| -\frac{1}{n}\log p(x^{(n)}, y^{(n)}) - H(X,Y) \right| > \varepsilon \right\}$$

there exists $\bar{n}$ such that, for $n > \bar{n}$, $\boxed{\text{Prob}(B^{(n)}_{1,\varepsilon}) < \frac{\varepsilon}{3}, \ \text{Prob}(B^{(n)}_{2,\varepsilon}) < \frac{\varepsilon}{3}, \ \text{Prob}(B^{(n)}_{3,\varepsilon}) < \frac{\varepsilon}{3}}$.

Notice that $\boxed{A^{(n)}_{\varepsilon} = \mathcal{X}^{(n)} \times \mathcal{Y}^{(n)} \setminus \left(B^{(n)}_{1,\varepsilon} \cup B^{(n)}_{2,\varepsilon} \cup B^{(n)}_{3,\varepsilon} \right)}$.

Then, $\boxed{\text{Prob}(A^{(n)}_{\varepsilon}) = 1 - \text{Prob}\left(\bigcup_{i=1}^{3} B^{(n)}_{i,\varepsilon} \right) \geq 1 - \sum_{i=1}^{3} \text{Prob}\left(B^{(n)}_{i,\varepsilon} \right) \geq 1 - \varepsilon}$

Notice that if $(x^{(n)}, y^{(n)}) \in A_\varepsilon^{(n)}$

$$2^{-n(H(X)+\varepsilon)} \leq p(x^{(n)}) \leq 2^{-n(H(X)-\varepsilon)}$$

$$2^{-n(H(Y)+\varepsilon)} \leq p(y^{(n)}) \leq 2^{-n(H(Y)-\varepsilon)}$$

$$2^{-n(H(X,Y)+\varepsilon)} \leq p(x^{(n)}, y^{(n)}) \leq 2^{-n(H(X,Y)-\varepsilon)}$$

- $1 = \sum_{(x^{(n)}, y^{(n)}) \in \mathcal{X}^{(n)} \times \mathcal{Y}^{(n)}} p(x^{(n)}, y^{(n)}) \geq \sum_{(x^{(n)}, y^{(n)}) \in A_\varepsilon^{(n)}} p(x^{(n)}, y^{(n)}) \geq \#(A_\varepsilon^{(n)})\, 2^{-n(H(X,Y)+\varepsilon)}$
- $\#(A_\varepsilon^{(n)}) \leq 2^{n(H(X,Y)+\varepsilon)}$

- $1 - \varepsilon \leq \sum_{(x^{(n)}, y^{(n)}) \in A_\varepsilon^{(n)}} p(x^{(n)}, y^{(n)}) \leq \#(A_\varepsilon^{(n)})\, 2^{-n(H(X,Y)-\varepsilon)}$
- $\#(A_\varepsilon^{(n)}) \geq (1-\varepsilon)\, 2^{n(H(X,Y)-\varepsilon)}$

- $\mathrm{Prob}(\tilde{X}^{(n)}, \tilde{Y}^{(n)} \in A_\varepsilon^{(n)}) = \sum_{(x^{(n)}, y^{(n)}) \in A_\varepsilon^{(n)}} p(x^{(n)})\, p(y^{(n)}) \leq \#(A_\varepsilon^{(n)})\, 2^{-n(H(X)+H(Y)-2\varepsilon)} \leq 2^{-n(I(X;Y)-3\varepsilon)}$
- $\mathrm{Prob}(\tilde{X}^{(n)}, \tilde{Y}^{(n)} \in A_\varepsilon^{(n)}) \geq \#(A_\varepsilon^{(n)})\, 2^{-n(H(X)+H(Y)+2\varepsilon)} \geq (1-\varepsilon)\, 2^{-n(I(X;Y)+3\varepsilon)}$

Remark : Shannon's noisy channel theorem asserts that the channel capacity can be achieved so that, whatever the memoryless discrete noisy channel, one could send C bits per channel use in the limit of a large number of channel uses.

The ingredient in Shannon's proof are:

1. Allowing an arbitrarily small but non-zero probability of error.
2. Using the channel many times in order to get the law of large numbers into effect.
3. Calculating the average of the probability of error over a random choice of codebooks.

Remark : in the proof given below, one decodes by joint typicality and then looks for a code-word jointly typical with the received sequence. If there is only one that is decided to be the transmitted code-word. The probability that any other code-word looks jointly typical with the received sequence is 2^{-nI} so the errors can be kept low if we do with fewer than 2^{nI} codewords.

Theorem 2.3.2 : Channel Coding Theorem

1) All rates below capacity C are achievable. Namely, if $R<C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $d^{(n)} \xrightarrow[n]{} 0$.

2) Conversely, any sequence of $(2^{nR}, n)$ codes with $d^{(n)} \xrightarrow[n]{} 0$ must have $R \leq C$.

Proof of part 1).

Each code $\mathcal{C}$ is identified by a matrix $[X_i(w)]_{\substack{i=1,2,\cdots,n \\ w=1,2,\cdots,2^{nR}}}$

$$\mathcal{C} = \begin{bmatrix} X_1(1) & X_2(1) \cdots X_n(1) \\ \vdots & \\ X_1(2^{nR}) & X_2(2^{nR}) \cdots X_n(2^{nR}) \end{bmatrix}$$

— Assume each matrix entry to be generated according to the probability $p(x)$ and independently : $\boxed{p(x^{(n)}(w)) = \prod_{j=1}^{n} p(x_j(w))}$

- The probability of a given code $\mathcal{C}$ is then $\boxed{\text{Prob}(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{j=1}^{n} p(x_j|w)}$

- A word w is chosen from $\mathcal{W} = \{1, 2, \dots, 2^{nR}\}$ according to the uniform distribution: $\boxed{\text{Prob}(w) = 2^{-nR} \quad \forall \; w \in \mathcal{W}}$.

- The code-word $x^{(n)}(w)$ from the randomly generated code $\mathcal{C}$ is sent over the channel.

- The receiver receives a sequence $y^{(n)}$ with probability

$$\boxed{p(y^{(n)} \mid x^{(n)}(w)) = \prod_{j=1}^{n} p(y_j \mid x_j(w))}.$$

- The receiver guesses the word sent by typical set decoding:

$\hat{w}$ was sent if $\begin{cases} x^{(n)}(\hat{w}) \text{ and } y^{(n)} \text{ are jointly typical} \\ \text{there is no other } \bar{w} \in \mathcal{W} \text{ such that } x^{(n)}(\bar{w}) \text{ and } y^{(n)} \text{ are jointly typical} \end{cases}$

- An error has occurred if either $x^{(n)}(\hat{w})$ and $y^{(n)}$ are not jointly typical or more $x^{(n)}(w)$ jointly typical with $y^{(n)}$ exist.

- Average probability of error: let ξ denote the error occurring by wrong decoding. The average probability of error is:

$$\boxed{P_{prob}^{av}(\xi) := \sum_{\xi} Prob(\xi) \frac{1}{2^{nR}} \sum_{w \in \mathcal{W}} d_w^{(n)}(\xi)}$$

$$d_w^{(n)}(\xi) = \sum_{y^{(n)} : g(y^{(n)}) \neq w} p(y^{(n)} | x_\xi^{(n)}(w))$$

- Simple example of random coding: $\mathcal{W} = \{0,1\}$; $\mathcal{C} = \{0,1\}$; $n = 2$

$$\xi_1 = \begin{cases} 0 \to 00 \\ 1 \to 00 \end{cases} ; \quad \xi_2 = \begin{cases} 0 \to 00 \\ 1 \to 01 \end{cases} ; \quad \xi_3 = \begin{cases} 0 \to 00 \\ 1 \to 10 \end{cases} ; \quad \xi_4 = \begin{cases} 0 \to 00 \\ 1 \to 11 \end{cases}$$

$$\xi_5 = \begin{cases} 0 \to 01 \\ 1 \to 00 \end{cases} ; \quad \xi_6 = \begin{cases} 0 \to 01 \\ 1 \to 01 \end{cases} ; \quad \xi_7 = \begin{cases} 0 \to 01 \\ 1 \to 10 \end{cases} ; \quad \xi_8 = \begin{cases} 0 \to 01 \\ 1 \to 11 \end{cases}$$

$$\xi_9 = \begin{cases} 0 \to 10 \\ 1 \to 00 \end{cases} ; \quad \xi_{10} = \begin{cases} 0 \to 10 \\ 1 \to 01 \end{cases} ; \quad \xi_{11} = \begin{cases} 0 \to 10 \\ 1 \to 10 \end{cases} ; \quad \xi_{12} = \begin{cases} 0 \to 10 \\ 1 \to 11 \end{cases}$$

$$\xi_{13} = \begin{cases} 0 \to 11 \\ 1 \to 00 \end{cases} ; \quad \xi_{14} = \begin{cases} 0 \to 11 \\ 1 \to 01 \end{cases} ; \quad \xi_{15} = \begin{cases} 0 \to 11 \\ 1 \to 10 \end{cases} ; \quad \xi_{16} = \begin{cases} 0 \to 11 \\ 1 \to 11 \end{cases}$$

- Because of the random generation of the code-words, to each $w \in \mathcal{W}$ one associates all $x^{(n)} \in \mathcal{X}^{(n)}$. Then,

$$\sum_{\mathcal{E}} \text{Prob}(\mathcal{E})\, d_w^{(n)}(\mathcal{E}) \quad \boxed{\text{does not}} \quad \text{depend on } w \text{ and}$$

$$\boxed{P_{\text{Prob}}^{w}(\mathcal{E}) = \sum_{\mathcal{E}} \text{Prob}(\mathcal{E})\, \frac{1}{2^{nR}} \sum_{w \in \mathcal{W}} d_w^{(n)}(\mathcal{E}) = \sum_{\mathcal{E}} \text{Prob}(\mathcal{E})\, d_1^{(n)}(\mathcal{E})}$$

- Let $E_w = \left\{ (x^{(n)}(w), y^{(n)}) \in A_\varepsilon^{(n)} \right\}$ be the subset of pairs where $y^{(n)}$ is the string received when $x^{(n)}(1)$ is sent through the channel.

Then, $E_1^c = \left\{ (x^{(n)}(1), y^{(n)}) \notin A_\varepsilon^{(n)} \right\}$ and $E_{w \neq 1}$ are the subsets from each errors originate. Then,

$$P_{\text{Prob}}^{w}(\mathcal{E}) = P_{\text{Prob}}^{w}\left(E_1^c \cup \bigcup_{w \neq 1} E_w \right) \leq P_{\text{Prob}}^{w}\left(E_1^c\right) + \sum_{w=2}^{2^{nR}} P_{\text{Prob}}^{w}\left(E_w\right)$$

- By the random generation of the code words and the joint AEP $\boxed{P_{\text{Prob}}^{w}\left(E_1^c\right) \leq \varepsilon}$