Entropies & Information Theory

LECTURE I

Nilanjana Datta

University of Cambridge,U.K.

See lecture notes on: http://www.qi.damtp.cam.ac.uk/node/223

Quantum Information Theory

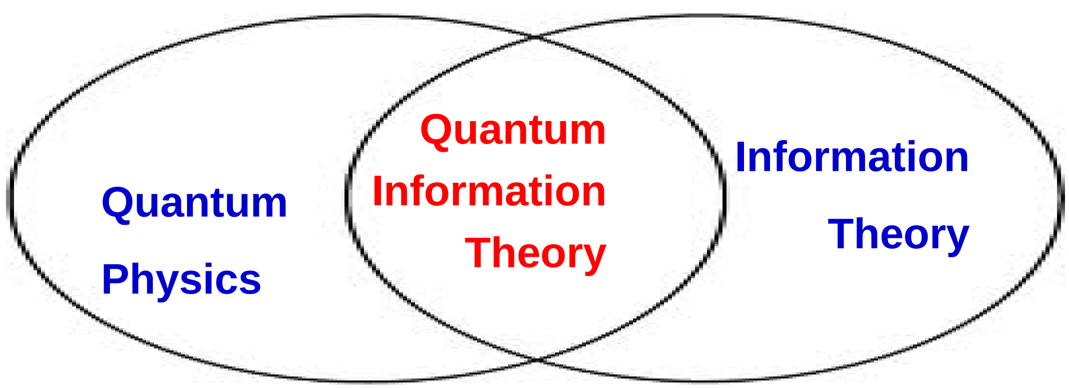
Born out of **Classical Information Theory**

Mathematical theory of storage, transmission & processing of information

Quantum Information Theory: how these tasks can be accomplished using

quantum-mechanical systems

as information carriers (e.g. photons, electrons,…)



UNIVERSITY OF CAMBRIDGE

The underlying quantum mechanics → distinctively **new features**

These can be **exploited** to:

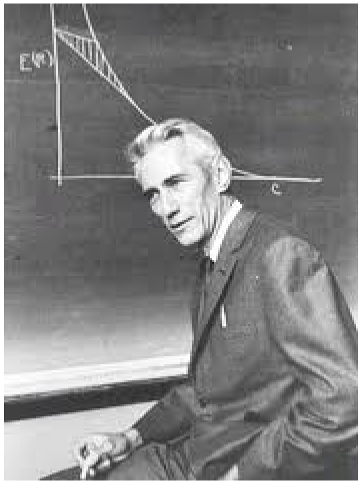
- improve the performance of certain information-processing tasks

as well as

- accomplish tasks which are impossible in the classical realm !

Classical Information Theory: 1948, Claude Shannon

- He posed 2 questions:



- **(Q1)** *What is the limit to which information can be reliably compressed ?*
  - **relevance:** there is often a physical limit to the amount of space available for storage information/data – e.g. in mobile phones

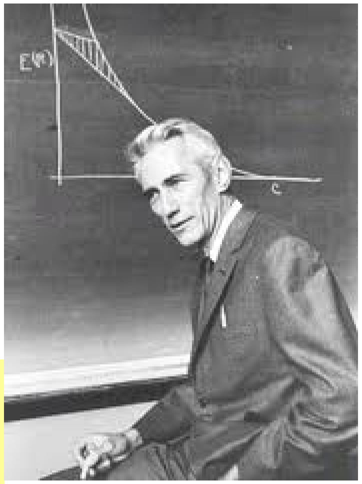
- *(Q2) What is the maximum amount of information that can be transmitted reliably per use of a communications channel ?*
  - **relevance:** biggest hurdle in transmitting info is presence of noise in communications channels, e.g. crackling telephone line,

- *information = data =signals= messages = outputs of a source*

UNIVERSITY OF CAMBRIDGE

Classical Information Theory: **1948,** Claude Shannon

- He posed 2 questions:



- *(Q1) What is the limit to which information can be reliably compressed ?*

- *(A1) Shannon's Source Coding Theorem:*

data compression limit = Shannon entropy of the source

- *(Q2) What is the maximum amount of information that can be transmitted reliably per use of a communications channel ?*

- *(A2) Shannon's Noisy Channel Coding Theorem:*

maximum rate of info transmission: given in terms of the mutual information

What is information?

- **Shannon:** information ⟷ uncertainty
- Information gain = decrease in uncertainty of an event
- measure of information ⟷ measure of uncertainty

Surprisal or Self-information:

- Consider an event described by a random variable (r.v.)

$X \sim p(x)$ (p.m.f); • $x \in J$ (finite alphabet)

- A measure of uncertainty in getting outcome x :

$$\gamma(x) := -\log p(x)$$ •

$\log \equiv \log_2$

- *a highly improbable outcome is surprising!*
- *rarer an event, more info we gain when we know it has occurred*
- only depends on $p(x)$ -- not on values x taken by X
- continuous; additive for independent events

Shannon entropy = average surprisal

- **Defn:** Shannon entropy $H(X)$ of a discrete r.v. $X \sim p(x)$, $x \in J$

$$H(X) = \mathrm{E}(\gamma(X)) = -\sum_{x \in J} p(x) \log p(x)$$

$$\log \equiv \log_2$$

- Convention: $0 \log 0 = 1$ $\quad \because \lim_{w \to 0} w \log w = 0$

(If an event has zero probability, it does not contribute to the entropy)

$H(X)$: a measure of uncertainty of the r.v. X

- also quantifies the amount of info we gain on average when we learn the value of X

$$H(X) \equiv H(p_X) = H\left(\{p(x)\}\right)$$

$$p_X = \{p(x)\}_{x \in J}$$

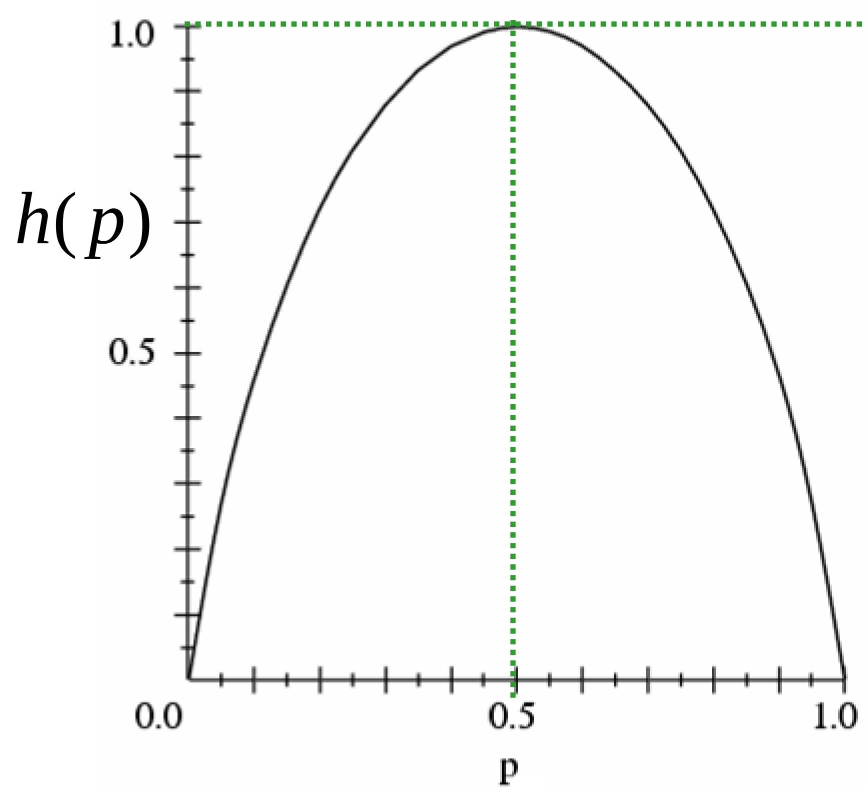
Example: Binary Entropy

$X \sim p(x) \qquad J \in \{0,1\} \qquad p(0) = p;\ p(1) = 1 - p;$

$$H(X) = -p \log p - (1-p)\log(1-p) \quad \equiv h(p)$$

Properties

- $p = 0 \Rightarrow x = 1$, $p = 1 \Rightarrow x = 0$ — $h(p) = 0$ no uncertainty
- $p = 0.5 : h(p) = 1$ maximum uncertainty
- Concave function of p
- Continuous function of p



UNIVERSITY OF CAMBRIDGE

Operational Significance of the Shannon Entropy

= *optimal rate of data compression for a classical memoryless (i.i.d.) information source*

Classical Information Source

- Outputs/signals : sequences of letters from a finite set J

J : *source alphabet*

(i) binary alphabet $J \in \{0,1\}$

(ii) telegraph English : 26 letters + a space

(iii) written English : 26 letters in upper & lower case + punctuation

UNIVERSITY OF CAMBRIDGE

Simplest example: a memoryless source

- successive signals: independent of each other
- characterized by a probability distribution $\{p(u)\}_{u\in J}$
- On each use of the source, a letter $u \in J$ emitted with prob $p(u)$

Modelled by a sequence of i.i.d. random variables

$$\boxed{U_1, U_2, \ldots, U_n} \qquad \boxed{U_i \sim p(u)} \qquad \boxed{u \in J}$$

$$\boxed{p(u) = P(U_k = u), \;\; u \in J \;\; \forall\, 1 \le k \le n.}$$

- Signal emitted by n uses of the source: $\underline{u} = (u_1, u_2, \ldots, u_n) = \underline{u}^{(n)}$

$$p(\underline{u}) = P(U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n) = p(u_1)\,p(u_2)\ldots p(u_n)$$

- Shannon entropy of the source: $$H(U) := -\sum_{u\in J} p(u)\log p(u)$$

(Q) Why is data compression possible?

(A) There is redundancy in the info emitted by the source

-- an info source typically produces some outputs more frequently than others:

In English text 'e' occurs more frequently than 'z'.

--during data compression one exploits this redundancy in the data to form the most compressed version possible

- Variable length coding:

-- *more frequently occurring signals (e.g 'e') assigned shorter descriptions (fewer bits) than the less frequent ones (e.g. 'z')*

- Fixed length coding:

-- *identify a set of signals which have high prob of occurrence: typical signals*

-- *assign unique fixed length (l) binary strings to each of them*

-- *all other signal (atypical) assigned a single binary string of same length (l)*

Typical Sequences

- **Defn:** Consider an i.i.d. info source :

$$U_1, U_2, \ldots U_n; \;\; p(u) \; ; \; u \in J$$

For any $\varepsilon > 0,$ sequences $\underline{u} := (u_1, u_2, \ldots u_n) \in J^n$ for which

$$2^{-n(H(U)+\varepsilon)} \le p(u_1, u_2, \ldots u_n) \le 2^{-n(H(U)-\varepsilon)},$$

where $H(U) =$ *Shannon entropy of the source*

are called $\varepsilon -$ typical sequences

$T_\varepsilon^{(n)} := \varepsilon -$ typical set = set of $\varepsilon -$ typical sequences

- **Note:** Typical sequences are almost equiprobable

$$\forall \; \underline{u} \in T_\varepsilon^{(n)}, \;\; p(\underline{u}) \approx 2^{-nH(U)}$$

UNIVERSITY OF CAMBRIDGE

$$\forall\ \underline{u} \in T_{\varepsilon}^{(n)}, \quad p(\underline{u}) \approx 2^{-nH(U)}$$

$U_1, U_2, \ldots U_n;$

$p(u)\ ;\ u \in J$

(Q) Does this agree with our intuitive notion of typical sequences?

(A) Yes! For an i.i.d. source : $U_1, U_2, \ldots U_n;\ \ U_i \sim p(u)\ ;\ u \in J$

A typical sequence $\underline{u} := (u_1, u_2, \ldots u_n)$ of length n,
is one which contains approx. $np(u)$ copies of u, $\forall\ u \in J$

- Probability of such a sequence is approximately given by

$$\approx \prod_{u \in J} p(u)^{np(u)} = \prod_{u \in J} 2^{np(u)\log p(u)} = 2^{\sum_{u \in J} p(u)\log p(u)}$$

$$= 2^{-nH(U)}$$

UNIVERSITY OF CAMBRIDGE

Properties of the Typical Set $T_\varepsilon^{(n)}$

- Let $\left|T_\varepsilon^{(n)}\right|$: number of typical sequences

  $P\left(T_\varepsilon^{(n)}\right)$: probability of the typical set

- **Typical Sequence Theorem:** Fix $\varepsilon > 0,$ then $\forall \delta > 0,$ and n large enough,

  - $$P\left(T_\varepsilon^{(n)}\right) > 1-\delta$$
  - $$(1-\delta)2^{n(H(U)-\varepsilon)} \le \left|T_\varepsilon^{(n)}\right| \le 2^{n(H(U)+\varepsilon)}$$

$$\Rightarrow J^n = T_\varepsilon^{(n)} \bigcup A_\varepsilon^{(n)}$$

atypical set

(disjoint union)

- *sequences in the atypical set rarely occur*

$$P\left(A_\varepsilon^{(n)}\right) \le \delta$$

- *typical sequences are almost equiprobable*

Operational Significance of the Shannon Entropy

- *(Q) What is the optimal rate of data compression for such a source?*

[min. # of bits needed to store the signals emitted per use of the source] *(for reliable data compression)*

- *Optimal rate is evaluated in the asymptotic limit* $n \to \infty$

  $n =$ *number of uses of the source*

- *One requires* $$\boxed{p_{error}^{(n)} \to 0 \; ; \; n \to \infty}$$

- *(A) optimal rate of data compression =* $H(U)$

Shannon entropy of the source

Compression-Decompression Scheme

Suppose $U_1, U_2, \ldots U_n; \; U_i \sim p(u) \; ; \; u \in J$ is an *i.i.d. information source*

Shannon entropy $H(U)$

- A compression scheme of rate R:

$$\mathcal{E}_n : \underline{u} := (u_1, u_2, \ldots u_n) \longrightarrow \underline{x} := (x_1, x_2, \ldots x_{m_n}) \in \{0,1\}^{m_n}$$

$\in J^n$

When is this a compression scheme?

$$\lim_{n \to \infty} \frac{m_n}{n} = R$$

- Decompression: $\mathcal{D}_n : \{0,1\}^{m_n} \longrightarrow J^n$

- Average probability of error: $p_{av}^{(n)} = \sum_{\underline{u}} p(\underline{u}) P\left(\mathcal{D}_n(\mathcal{E}_n(\underline{u})) \neq \underline{u}\right)$

- Compr.-decompr. scheme reliable if $p_{av}^{(n)} \to 0$ as $n \to \infty$

- Shannon's Source Coding Theorem:

Suppose $U_1, U_2, \ldots U_n;\ \ U_i \sim p(u)\ ;\ u \in J$ is an *i.i.d. information source*

Shannon entropy $H(U)$

- Suppose $R > H(U)$: then there exists a reliable compression scheme of rate R for the source.

- If $R < H(U)$ then any compression scheme of rate R will not be reliable.

- Shannon's Source Coding Theorem:

Suppose $U_1, U_2, \ldots U_n; \;\; U_i \sim p(u) \; ; \; u \in J$ is an *i.i.d. information source*

Shannon entropy $H(U)$

- Suppose $R > H(U)$: then there exists a reliable compression scheme of rate R for the source.

Sketch of proof

(achievability)

Shannon's Source Coding Theorem (proof contd.)

- If $R < H(U)$ then any compression scheme of rate R will not be reliable. **(converse)**

Proof follows from:

- **Lemma:** Let $\mathcal{S}^{(n)}$ be a set of sequences $\underline{u}^{(n)} := (u_1, u_2, \dots u_n)$ of length n of size $\left|\mathcal{S}^{(n)}\right| \le 2^{nR}$, where $R < H(U)$ is fixed. Each sequence $\underline{u}^{(n)}$ is produced with prob. $p(\underline{u}^{(n)})$

Then for any $\delta > 0$, and sufficiently large n,

$$\sum_{\underline{u}^{(n)} \in \mathcal{S}(n)} p(\underline{u}^{(n)}) \le \delta$$

$\Rightarrow$ if $\mathcal{S}^{(n)}$ is a set of at most 2^{nR} sequences with $R < H(U)$, then with a high probability the source will produce sequences which will not lie in this set.

Hence encoding 2^{nR} sequences $\not\longrightarrow$ reliable data compression

Entropies for a pair of random variables

- Consider a pair of discrete random variables

$X \sim p(x)\ ;\ x \in J_X$ and $Y \sim p(y)\ ;\ y \in J_Y$

Given their joint probabilities $P(X = x, Y = y) = p(x, y)\ ;$

& their conditional probabilities $P(Y = y \mid X = x) = p(y \mid x)\ ;$

- **Joint entropy:** $H(X,Y) := -\sum_{x \in J_X} \sum_{y \in J_Y} p(x,y)\ \log p(x,y)$

- **Conditional entropy:**

$$H(Y \mid X) := \sum_{x \in J_X} p(x) H(Y \mid X = x) = -\sum_{x \in J_X} \sum_{y \in J_Y} p(x,y)\ \log p(y \mid x)$$

- Chain Rule: $H(X,Y) = H(Y \mid X) + H(X)$

UNIVERSITY OF CAMBRIDGE

Entropies for a pair of random variables

- **Relative Entropy:** Measure of the "distance" between two probability distributions $p = \{p(x)\}_{x\in J}$; $q = \{q(x)\}_{x\in J}$

$$D(p \| q) := \sum_{x\in J} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

convention:

$$0 \log\left(\frac{0}{u}\right) = 0 \;\; ; \;\; u \log\left(\frac{u}{0}\right) = \infty \quad \forall u > 0$$

- $D(p \| q) \geq 0$
- $D(p \| q) = 0$ if & only if $p = q$

- **BUT** not a true distance
  - not symmetric;
  - does not satisfy the triangle inequality

UNIVERSITY OF CAMBRIDGE

Entropies for a pair of random variables

- **Mutual Information:** Measure of the amount of info one r.v. contains about another r.v. $X \sim p(x),\ Y \sim p(y)$

$$I(X,Y) := \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

$$\boxed{I(X:Y) = D(p_{XY} \,\|\, p_X\, p_Y)}$$

$$p_{XY} = \{p(x,y)\}_{x,y};\ p_X = \{p(x)\}_x;\ p_Y = \{p(y)\}_y$$

- Chain rules:

$$\begin{aligned} I(X:Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

UNIVERSITY OF CAMBRIDGE

Properties of Entropies

Let $X \sim p(x),\ Y \sim p(y)$ be discrete random variables: Then,

- $H(X) \geq 0,$ with equality if & only if X is deterministic
- $H(X) \leq \log|J|,$ if $x \in J$
- **Subadditivity:** $H(X,Y) \leq H(X) + H(Y),$
- **Concavity:** if p_X & p_Y are 2 prob. distributions,

$$H(\lambda p_X + (1-\lambda) p_Y) \geq \lambda H(p_X) + (1-\lambda) H(p_Y),$$

- $H(Y \mid X) \geq 0,$ or equivalently $H(X,Y) \geq H(Y),$
- $I(X:Y) \geq 0$ with equality if & only if X & Y are independent

UNIVERSITY OF CAMBRIDGE

- So far…….

- **Classical Data Compression:** answer to **Shannon's** 1^{st} **question**

(Q1) What is the limit to which information can be reliably compressed ?

(A1) Shannon's Source Coding Theorem:

data compression limit = Shannon entropy of the source

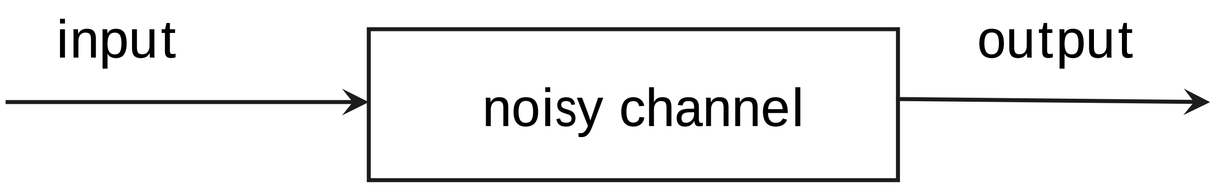
- **Classical entropies and their properties**

▪ Shannon's 2^{nd} question

- *(Q2) What is the maximum amount of information that can be transmitted reliably per use of a communications channel?*

The biggest hurdle in the path of efficient transmission of info is the presence of noise in the communications channel

- Noise distorts the information sent through the channel.



output $\neq$ input

- To combat the effects of noise: use error-correcting codes

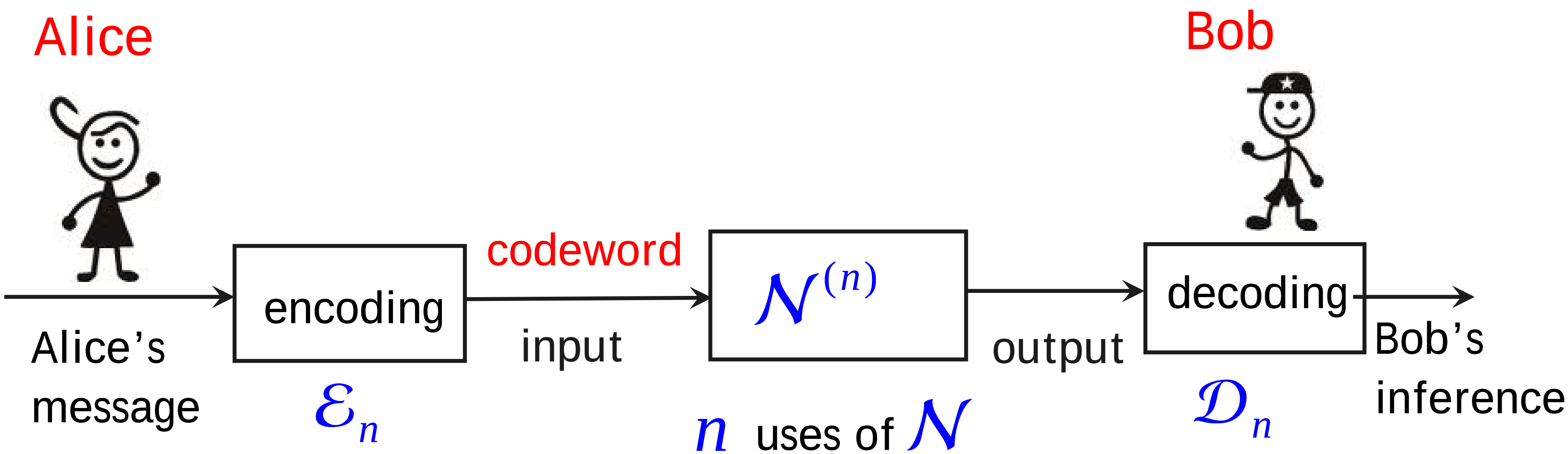
UNIVERSITY OF CAMBRIDGE

To overcome the effects of noise:

instead of transmitting the original messages,

-- the sender encodes her messages into suitable codewords

-- these codewords are then sent through (multiple uses of) the channel



- *Error-correcting code:* $C_n := (\mathcal{E}_n, \mathcal{D}_n):$

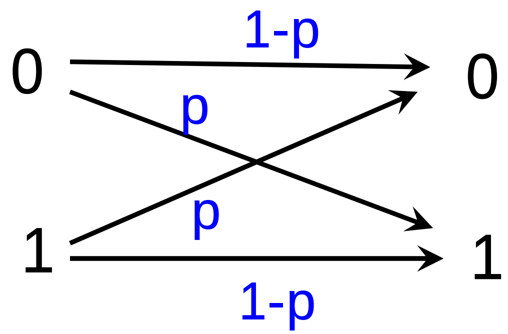
- The idea behind the encoding:

  - To introduce redundancy in the message so that upon decoding, Bob can retrieve the original message with a low probability of error:

  - The amount of redundancy which needs to be added – depends on the noise in the channel

Example

- Memoryless binary symmetric channel (m.b.s.c.)



- it transmits single bits
- effect of the noise: to flip the bit with probability p

Repetition Code

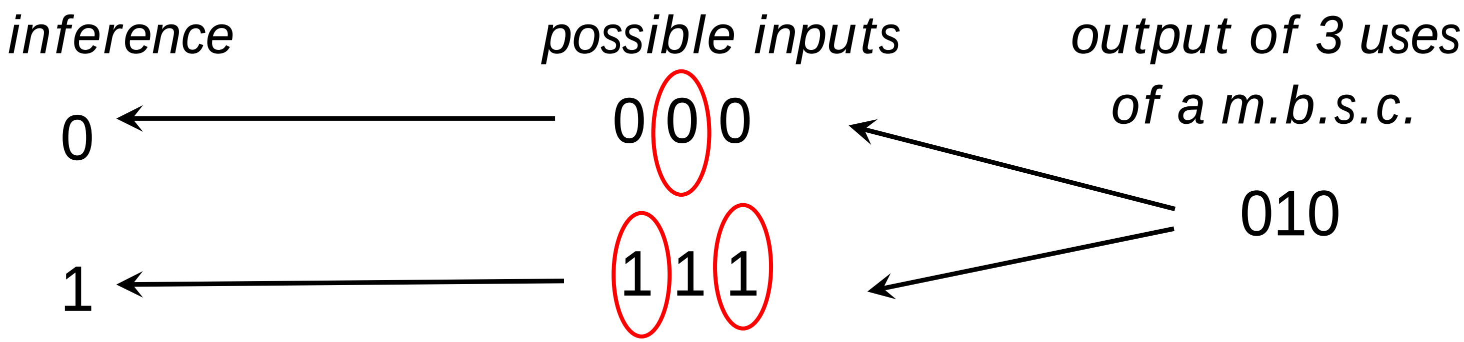
- Encoding: $0 \longrightarrow 000$

  $1 \longrightarrow 111$

codewords

- the 3 bits are sent through 3 successive uses of the m.b.s.c.
- Suppose | 000 $\xrightarrow[\text{m.b.s.c.}]{}$ 010 (Bob receives) | codeword
- Decoding : (*majority voting*) 010 $\longrightarrow$ 0 (Bob infers)

- Probability of error for the m.b.s.c. :
  - without encoding = p
  - with encoding = *Prob (2 or more bits flipped)* := q



- *Prove: $q < p$ if $p < 1/2$*

-- in this case encoding helps!

- (Encoding – Decoding) : Repetition Code.

- Information transmission is said to be reliable if:

-- the probability of error in decoding the output vanishes asymptotically in the number of uses of the channel

- Aim: to achieve reliable information transmission whilst optimizing the rate

  - the amount of information that can be sent per use of the channel

- The optimal rate of reliable info transmission: capacity

Discrete classical channel $\mathcal{N}$

$J_X =$ input alphabet; $J_Y =$ output alphabet

n uses of $\mathcal{N}$

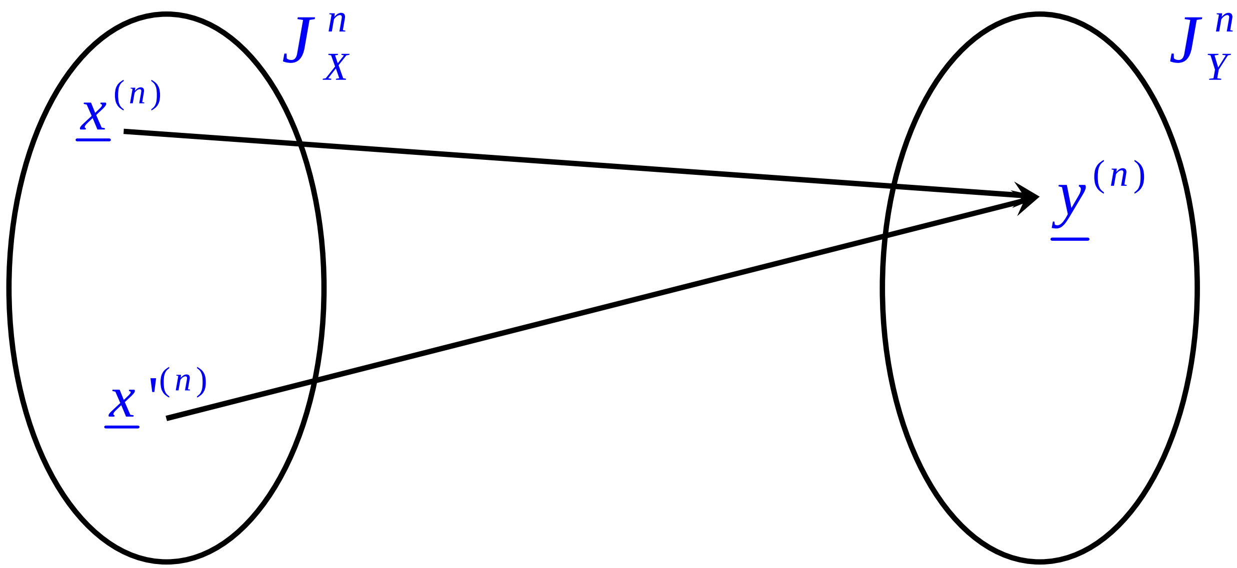
$$\xrightarrow{\text{input } \underline{x}^{(n)}} \boxed{\mathcal{N}^{(n)}} \xrightarrow{\text{output } \underline{y}^{(n)}}$$

$\underline{x}^{(n)} \in J_X^n$ $\quad$ $p(\underline{y}^{(n)} \mid \underline{x}^{(n)})$ $\quad$ $\underline{y}^{(n)} \in J_Y^n$

$p(y^{(n)} \mid x^{(n)})$

- conditional probabilities ;
- known to sender & receiver

Correspondence between input & output sequences is not 1-1



- Shannon proved: it is possible to choose a subset of input sequences--

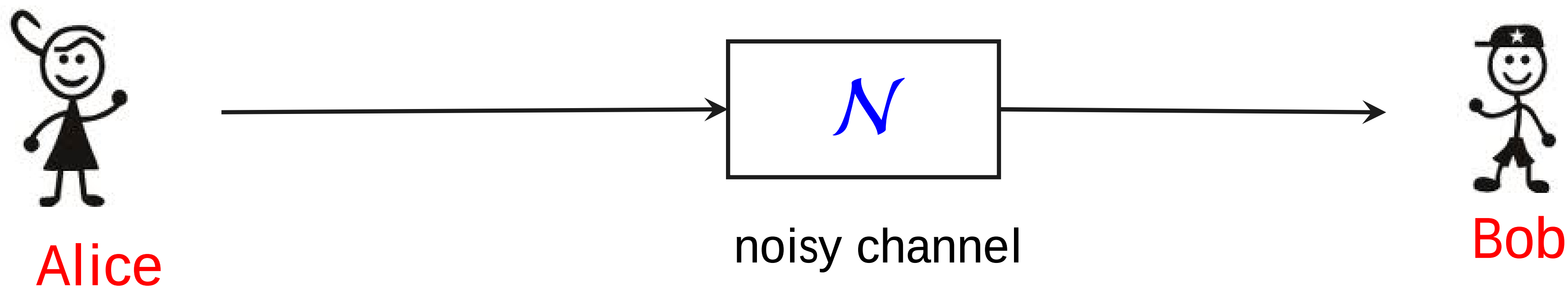
  such that there exists only :

  - 1 highly likely input corresponding to a given input

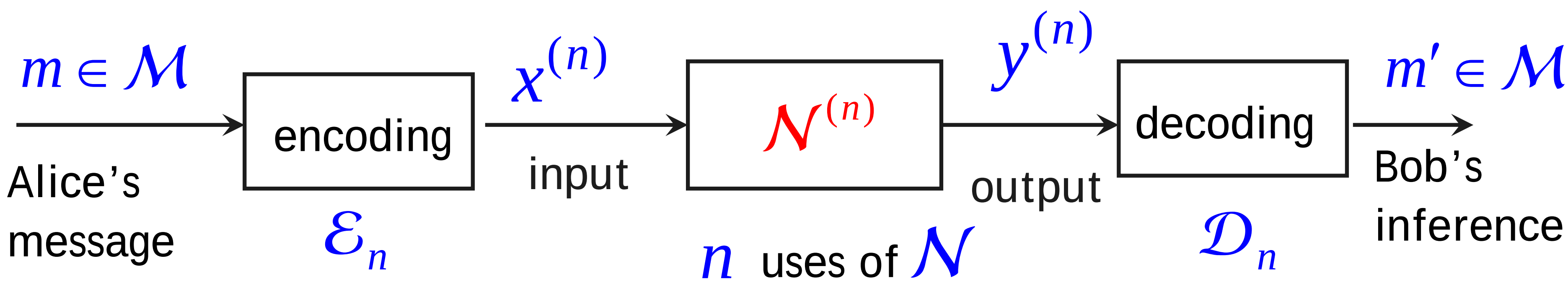
- Use these input sequences as codewords

UNIVERSITY OF CAMBRIDGE

Transmission of info through a classical channel



$\mathcal{M}$: finite set of messages

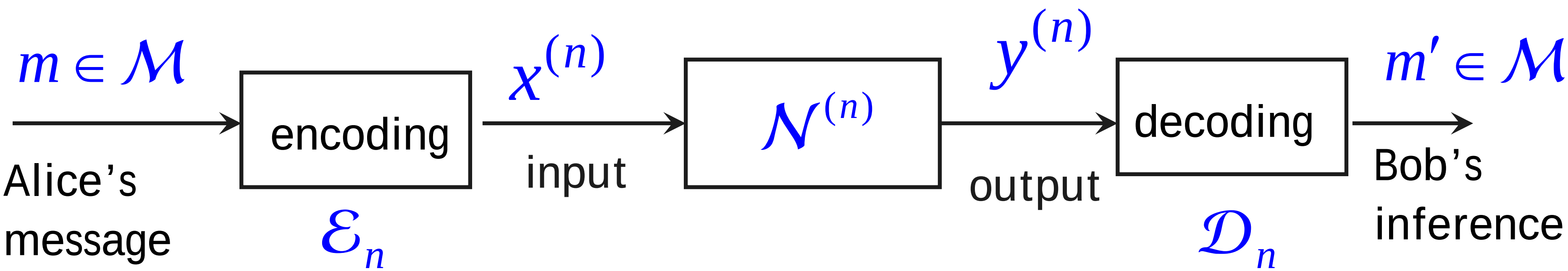


- *codeword:* $x^{(n)} = (x_1, x_2, \ldots, x_n);$
  output: $y^{(n)} = (y_1, y_2, \ldots, y_n);$

$$\mathcal{N}^{(n)} : p(y^{(n)} \mid x^{(n)})$$

- *Error-correcting code:* $C_n := (\mathcal{E}_n, \mathcal{D}_n):$

UNIVERSITY OF CAMBRIDGE



- *If* $m' \neq m$ *then an error occurs!*
- *Info transmission is reliable if: Prob. of error* $\rightarrow 0$ *as* $n \rightarrow \infty$
- *Rate of info transmission* = number of bits of message transmitted per use of the channel
- *Aim: achieve reliable transmission whilst maximizing the rate*
  - *Shannon: there is a fundamental limit on the rate of reliable info transmission ; property of the channel*
- *Capacity: maximum rate of reliable information transmission*

- Shannon in his Noisy Channel Coding Theorem:

-- obtained an explicit expression for the capacity of a memoryless classical channel

$$p(y^{(n)} \mid x^{(n)}) = \prod_{i=1}^{n} p(y_i \mid x_i)$$

Memoryless (classical or quantum) channels

- action of each use of the channel is identical and it is independent for different uses

-- i.e., the noise affecting states transmitted through the channel on successive uses is assumed to be uncorrelated.

- Classical memoryless channel: a schematic representation

$$X \sim p(x) \longrightarrow \boxed{\quad \mathcal{N} \quad} \longrightarrow Y$$

input x $p(y\,|\,x)$ output y

$x \in J_X$, $y \in J_Y$,

- channel: a set of conditional probs. $\{p(y\,|\,x)\}$

- Capacity $\boxed{C(\mathcal{N}) = \max_{\{p(x)\}} I(X:Y)}$

input distributions   *mutual information*

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

Shannon Entropy $$H(X) = -\sum_x p(x) \log p(x)$$

- Shannon's Noisy Channel Coding Theorem:
  - For a memoryless channel:

$$X \sim p(x) \xrightarrow[\text{input}]{} \boxed{\mathcal{N}} \xrightarrow[\text{output}]{} Y$$

$$p(y\,|\,x)$$

Optimal rate of reliable info transmission $\equiv$ capacity

$$\boxed{C(\mathcal{N}) = \max_{\{p(x)\}} I(X:Y)}$$

Sketch of proof