



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>

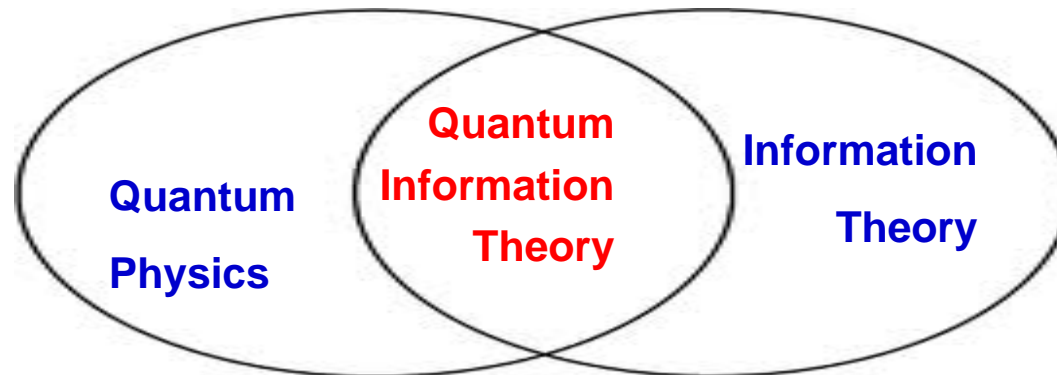
Quantum Information Theory

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,...)



The underlying
quantum mechanics



distinctively **new features**

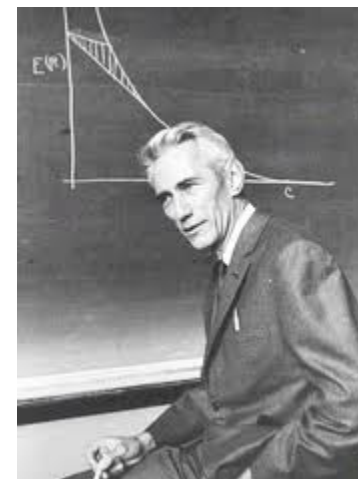
These can be **exploited** to:

- **improve** the **performance** of certain
information-processing tasks

as well as

- **accomplish** tasks which are
impossible in the **classical realm** !

- He posed 2 questions:
- (Q1) *What is the limit to which information can be reliably compressed?*
 - relevance: physical limit on storage
- (Q2) *What is the maximum amount of information that can be transmitted reliably per use of a communications channel?*
 - relevance: noise in communications channels
- *information = data = signals = messages = outputs of a source*

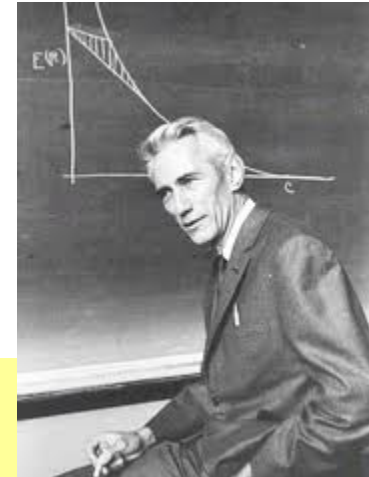


- He posed 2 questions:
- **(Q1)** *What is the limit to which information can be **reliably** compressed ?*

- **(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 \longleftrightarrow uncertainty
- Information gain = decrease in uncertainty of an event
- measure of information \longleftrightarrow 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 :

$$\gamma(x) := -\log p(x) \quad \bullet$$

$\log \equiv \log_2$

- *a highly improbable outcome is surprising!*

- 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. $X \sim p(x)$,
 $x \in J$

$$H(X) = E(\gamma(X)) = - \sum_{x \in J} p(x) \log p(x) \quad \log \equiv \log_2$$

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

(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

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

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

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

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**

$$U_1, U_2, \dots, U_n$$

$$U_i \sim p(u)$$

$$u \in J$$

$$p(u) = P(U_k = u), \quad u \in J \quad \forall 1 \leq k \leq n.$$

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

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

- **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 :

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

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

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

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

are called ε – typical sequences

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

- Note: Typical sequences are almost equiprobable

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

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 $n \rightarrow \infty$*
 $n =$ number of uses of the source

- *One requires*

$$p_{error}^{(n)} \rightarrow 0 ; n \rightarrow \infty$$

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

Shannon entropy of the source

Compression-Decompression Scheme

Suppose U_1, U_2, \dots, 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 :

$$\mathcal{E}_n : \underline{u} := (u_1, u_2, \dots, u_n) \xrightarrow{\hspace{2cm}} \underline{x} := (x_1, x_2, \dots, x_{m_n}) \in \{0, 1\}^{m_n}$$

$\in J^n$

When is this a compression scheme?

$$m_n = \lceil nR \rceil$$

- Decompression: $\mathcal{D}_n : \{0, 1\}^{m_n} \xrightarrow{\hspace{2cm}} J^n$ ●
- Average probability of error: $p_{av}^{(n)} = \sum_{\underline{u}} p(\underline{u}) P(\mathcal{D}_n(\mathcal{E}_n(\underline{u})) \neq \underline{u})$
- Compr.-decompr. scheme **reliable** if $p_{av}^{(n)} \rightarrow 0$ as $n \rightarrow \infty$

- Shannon's Source Coding Theorem:

Suppose U_1, U_2, \dots, 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**.

Entropies for a pair of random variables

- Consider a pair of discrete random variables

$$X \sim p(x) ; x \in J_X \quad \text{and} \quad Y \sim p(y) ; y \in J_Y$$

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

& their **conditional probabilities** $P(Y = y | X = x) = p(y | x) ;$

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

- **Conditional entropy:**

$$H(Y | X) := \sum_{x \in J_X} p(x) H(Y | X = x) \qquad H(Y | X) \geq 0$$

$$= \sum_{x \in J_X} p(x) \left[- \sum_{y \in J_Y} p(y | x) \log p(y | x) \right] = - \sum_{x \in J_X} \sum_{y \in J_Y} p(x, y) \log p(y | x)$$

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

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

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

- $D(p \parallel q) \geq 0$
- $D(p \parallel q) = 0$ if & only if $p = q$
 - not symmetric;
 - does not satisfy the triangle inequality
- **BUT** not a true distance

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

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

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

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

- Chain rules:

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

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

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

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

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,
$$H(\lambda p_X + (1-\lambda)p_Y) \geq \lambda H(p_X) + (1-\lambda)H(p_Y),$$
- $H(Y | X) \geq 0$, or equivalently $H(X, Y) \geq H(Y)$, ●
- $I(X : Y) \geq 0$ with equality if & only if X & Y
are independent

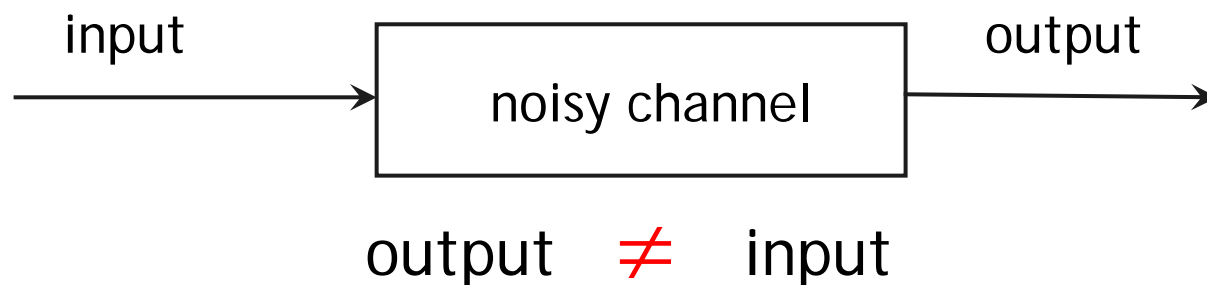
- So far.....
- Classical Data Compression: answer to Shannon's 1st question
(Q1) What is the limit to which information can be reliably compressed ?
- (A1) Shannon's Source Coding Theorem:*
data compression limit = Shannon entropy of the source
- Classical entropies and their properties

- Shannon's 2nd 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.



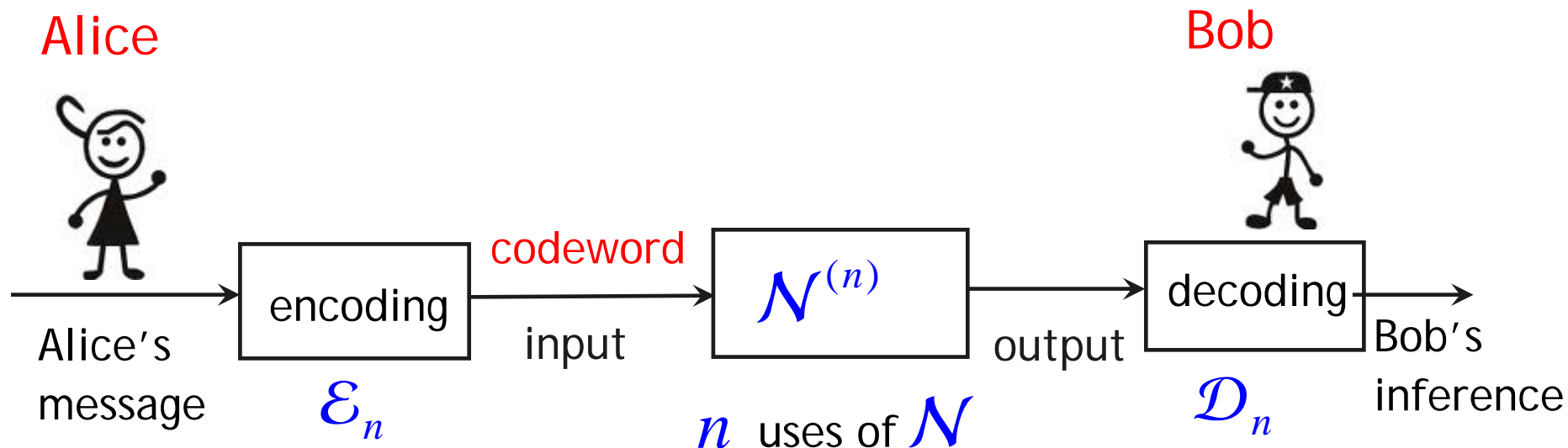
- To combat the effects of noise: use error-correcting codes

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**

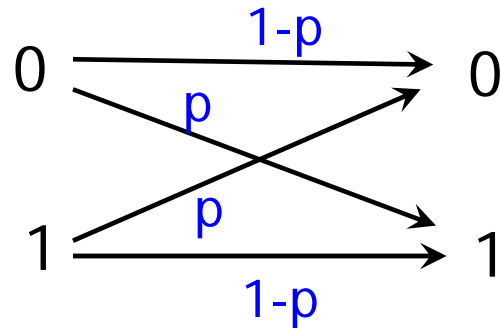


- **Error-correcting code:** $\mathcal{C}_n := (\mathcal{E}_n, \mathcal{D}_n)$:

- 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

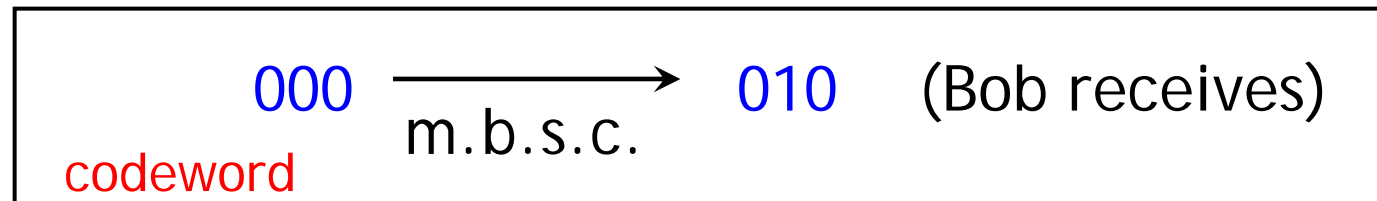
Repetition Code

- Encoding: $0 \longrightarrow 000$
 $1 \longrightarrow 111$

codewords

- the **3 bits** are sent through **3 successive uses** of the m.b.s.c.

- Suppose



- Decoding : (*majority voting*) $010 \longrightarrow 0$ (Bob infers)

- 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**

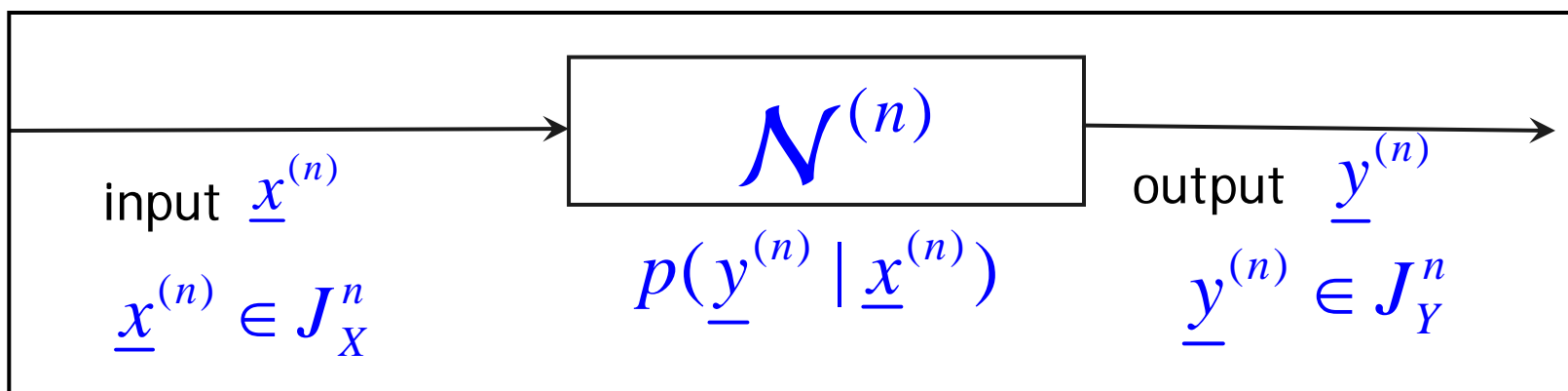
- the amount of **information** that can be sent **per use** of the channel

- The **optimal rate** of **reliable** info transmission: **capacity**

Discrete classical channel \mathcal{N}

J_X = input alphabet; J_Y = output alphabet

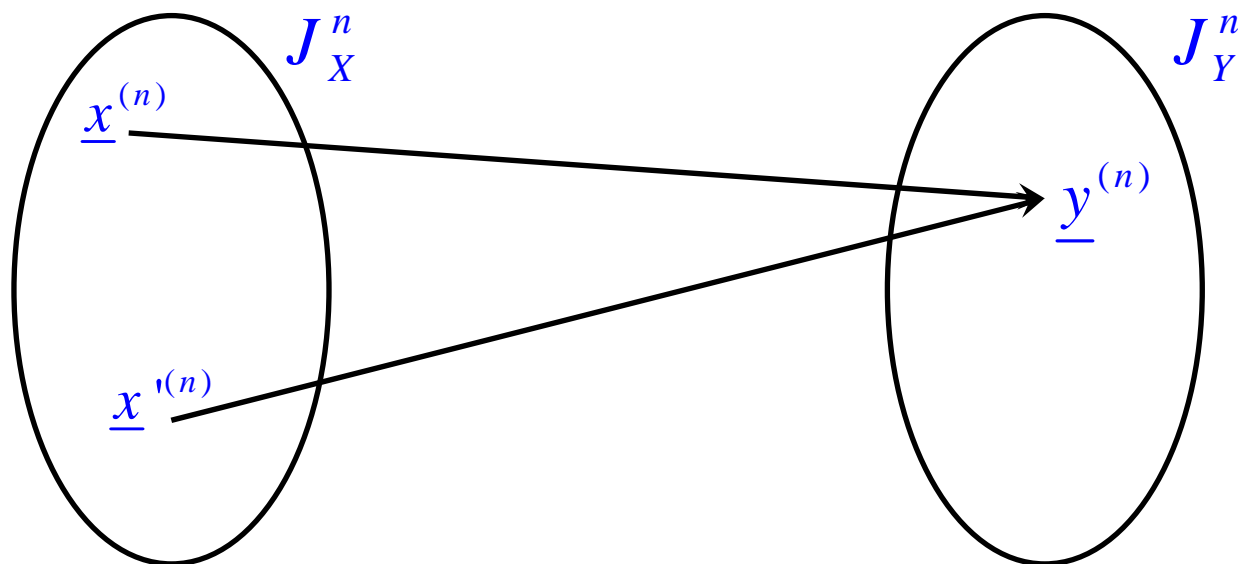
n uses of \mathcal{N}



$$p(\underline{y}^{(n)} | \underline{x}^{(n)})$$

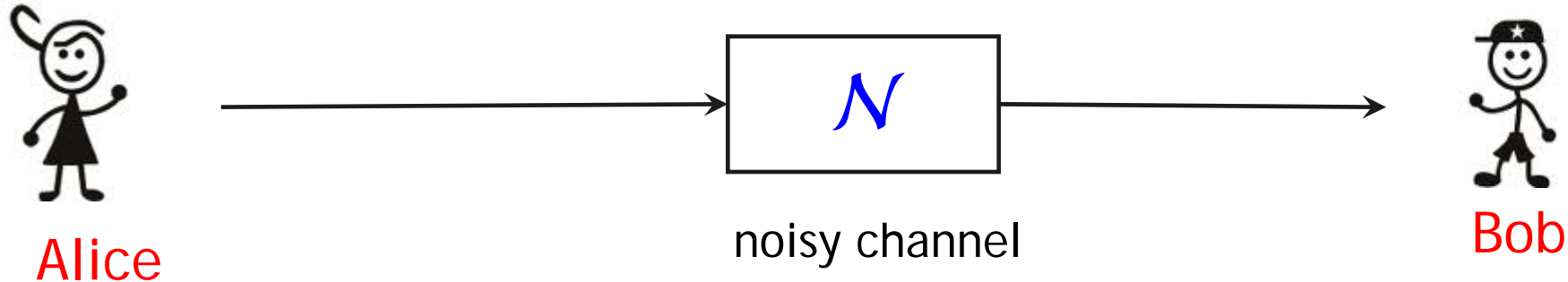
- conditional probabilities ;
- known to sender & receiver

Correspondence between input & output sequences is **not** 1-1

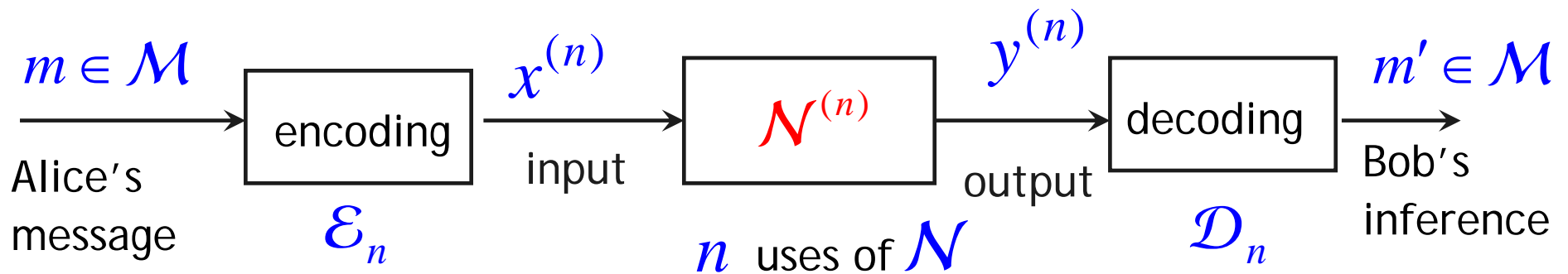


- **Shannon proved:** it is possible to choose a subset of input sequences--
such that there exists only :
 - 1 highly likely input corresponding to a given input
- Use these input sequences as codewords

Transmission of info through a classical channel



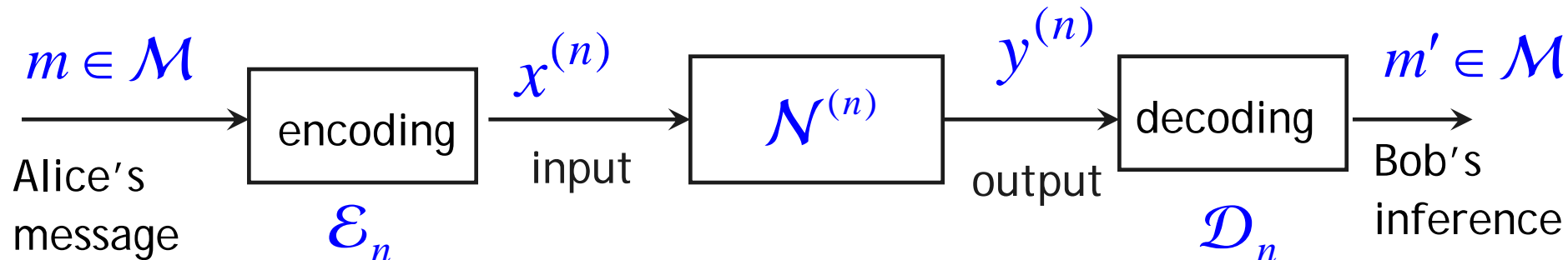
\mathcal{M} : finite set of messages



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

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

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



- 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

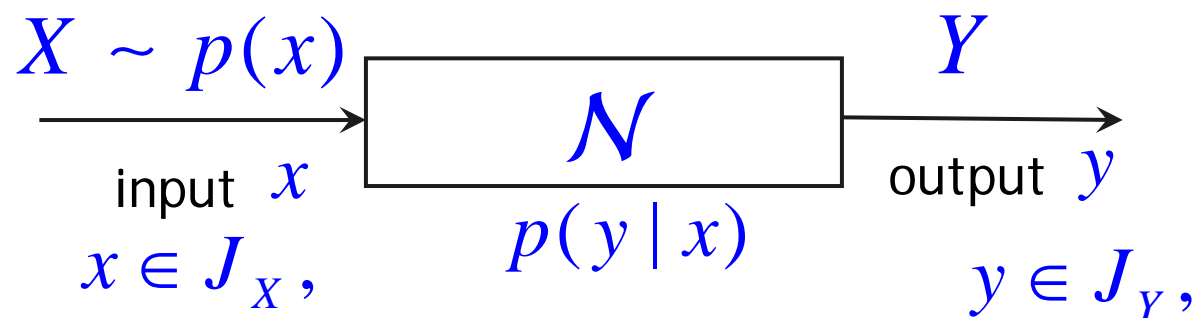
- **Shannon** in his **Noisy Channel Coding Theorem**:
 - obtained an explicit expression for the capacity of a **memoryless classical channel**

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

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



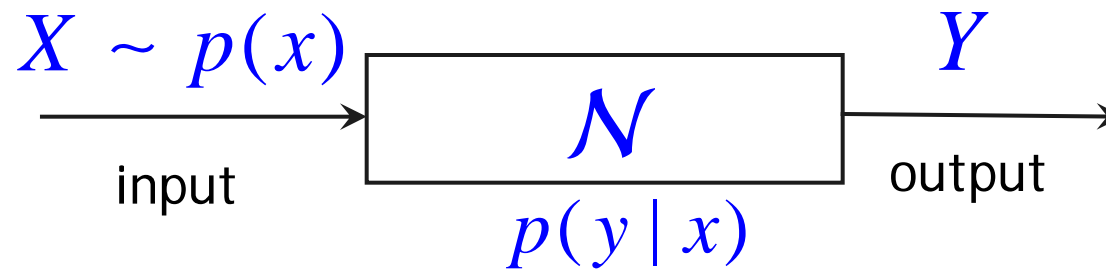
- **channel**: a set of conditional probs. $\{p(y|x)\}$

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



Optimal rate of reliable info transmission \equiv capacity

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

● Sketch of proof

- Supposedly [von Neumann](#) asked [Shannon](#) to call this quantity entropy, saying

“You should call it ‘[entropy](#)’ for two reasons; [first](#), the function is [already in use](#) in thermodynamics; [second](#), and most importantly, most people don’t know what entropy really is, & if you use ‘entropy’ in an argument, [you will win every time](#).”

Relation between Shannon entropy & thermodynamic entropy

$$S := k \log \Omega$$

total # of
microstates

- Suppose the r^{th} microstate occurs with prob. p_r

 consider ν replicas of the system

- Then on average $\nu_r \approx [\nu p_r]$ replicas are in the r^{th} state

- Total # of microstates
$$\Omega = \frac{\nu!}{\nu_1! \nu_2! \dots \nu_r!} \approx \frac{\nu^\nu}{\nu_1^{\nu_1} \nu_2^{\nu_2} \dots \nu_r^{\nu_r}} \quad (\text{by Stirling's})$$

- Th. dyn. entropy of compound system of ν replicas

$$S_\nu = -k\nu \sum_r p_r \log p_r$$

$$S = S_\nu / \nu = -k \sum_r p_r \log p_r = k H(\{p_r\})$$

Shannon
entropy