Introduction to Information Theory and Coding ENEE5304
Lecture Outline

Lecture 1

- Explain the course objectives
- List the subjects to be covered
- Provide a general description of a digital communication system
- Model the additive white Gaussian noise and its effect on error rate in transmission
- Introduce the term: system reliability
- Introduce the term: system sfficiency

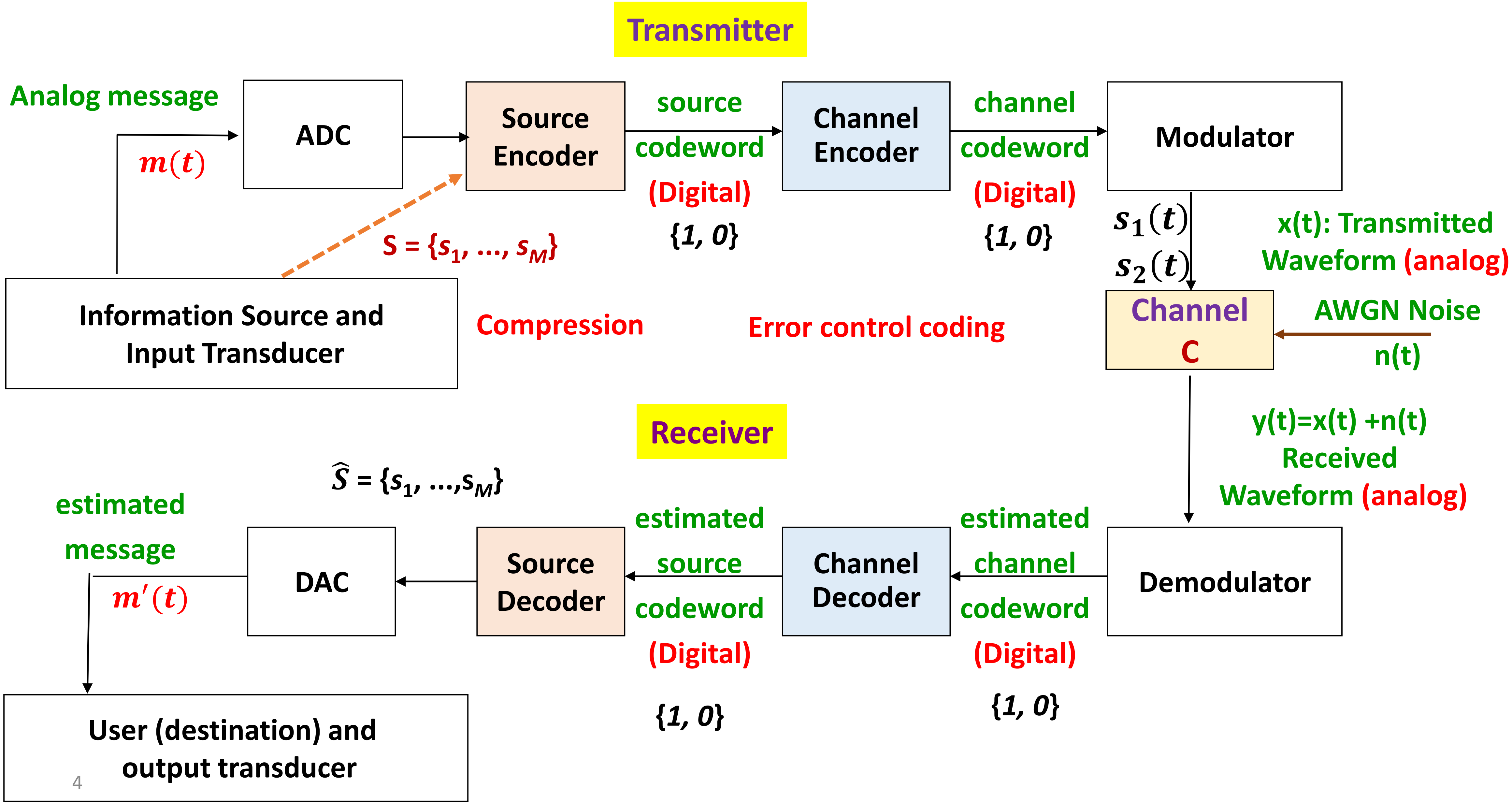
Information Theory ENEE5304

- **Course Objective**: The aim of this course is to introduce the **undergraduate students** to the fundamental concepts in information theory and coding and to indicate where and how the theory can be applied. Focus will be on interpretation of results. Try to avoid complex proofs of some theorems.
- Developed and Formulated by C. E. Shannon in 1948
- Fundamental to understanding and characterizing the performance of communication systems.
- Originally intended to study communication systems, then evolved to encompass other sorts of applications such as the stock market, probability, economics, investment, …
- Gave essential impacts on today’s digital technology
  - data compression
  - wired/wireless communication/broadcasting
  - cryptography, linguistics, bioinformatics, games, ...

Course Outline

- **Information Theory**: Uncertainty, Information, Entropy, Discrete Memory-less Sources, Extension of DMS, Markov Sources, Source-Coding Theorem, Data Compression, Prefix-Free Codes, Kraft Inequality, Huffman Coding, Lempel-Ziv Coding, Discrete Memoryless Channels (DMC), The Binary Symmetric Channel, Mutual Information, Capacity of the Discrete Memory-less Channel, Capacity of the Gaussian Channel, Channel Coding Theorem, Information Capacity Theorem.
- **Error-Control Coding**: Block Codes, Linear Codes, Hamming Codes, Generator Matrix, Parity-Check Matrix, Syndrome, Cyclic Redundancy Check. Basics of automatic repeat request.
- **Convolutional Codes**: Convolutional Encoder, General Rate 1/n Constraint Length-K Code, Tree, Finite-State Machine , and Trellis Representation of Convolutional Codes, Maximum Likelihood Decoding of a Convolutional Code, Viterbi Decoding Algorithm, Free Distance of a Convolutional Code.

A Basic Communication System Block Diagram



What is Information Theory about?

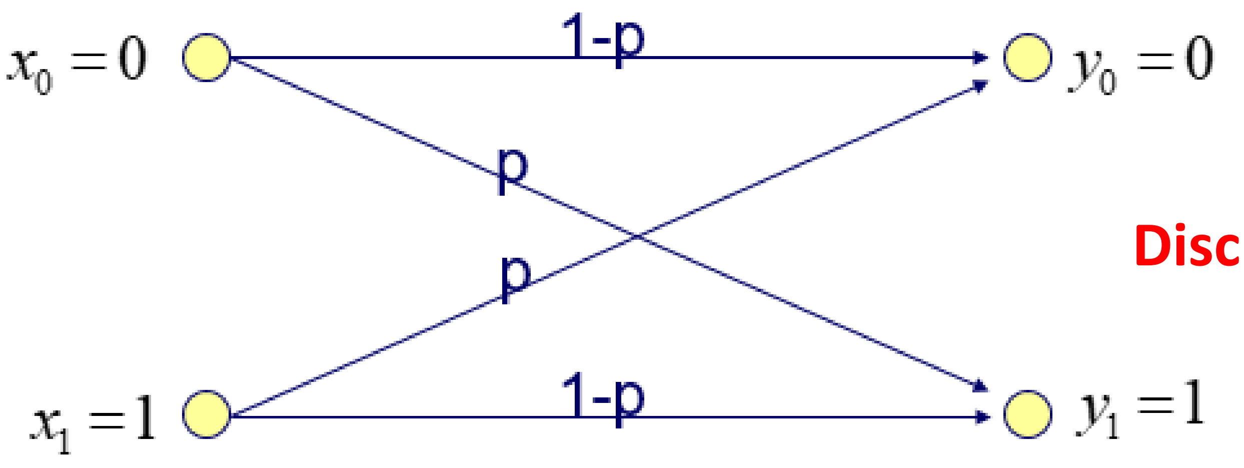
- **Information theory** answers two fundamental questions:
  - Given a **source,** *how much can we compress the data? Is there any limit*? **(Entropy H)**
  - Given a **channel**, how noisy can the channel be, or how much parity bits are necessary to minimize error in decoding?
  - What is the maximum rate of communication? **(Channel Capacity C)**
  - In early days, it was thought that increasing transmission rate over a channel increases the error rate.
  - Shannon showed that **this is not necessarily true as long as rate is below Channel Capacity**.

Modulation and Error Probability

- **Binary digits from the channel encoder are assigned electrical pulses for transmission over the channel.**
- **Transmitted pulses are corrupted by AWGN**
- **Noise will cause transmission error**

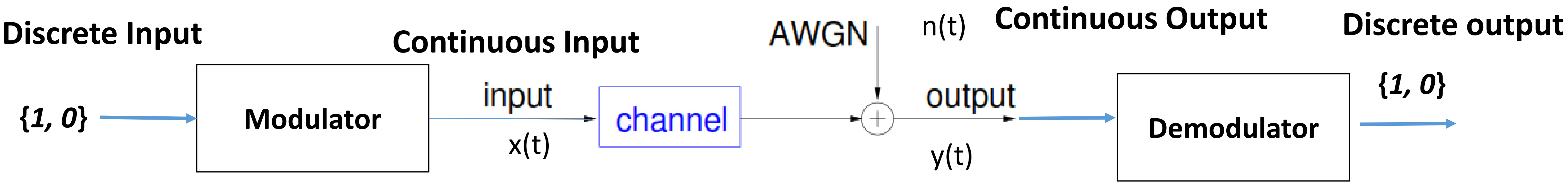
Digit 0 -> $s_2(t)$

Discrete Input



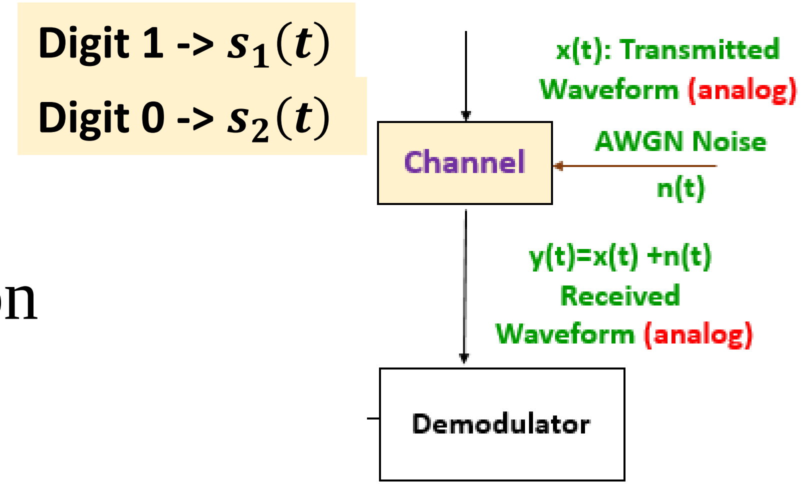
Discrete Output

Digit 1 -> $s_1(t)$



Communication System: Additive White Gaussian Noise

- **Additive White Gaussian Noise** is a basic noise model used in communication systems to mimic the effect of many random processes that occur in nature.
- This noise comes from many natural noise sources, such as the thermal vibrations of atoms in conductors (referred to as thermal noise), shot noise, black-body radiation from the earth and other warm objects, and from celestial sources such as the Sun.
- The central limit theorem of probability theory indicates that the summation of many random processes will tend to have distribution called Gaussian or Normal.
  - Transmitted signal: $x(t)$;
  - Channel Output: $y(t) = x(t) + n(t)$;
  - The pdf of $n(t)$ follows the Gaussian distribution
  - The power spectral density is a constant over a wide range of the frequency spectrum



Communication System: Optimum Binary Receiver Performance

- In a digital data transmission, the receiver has to decide which symbol was transmitted such that the probability of making errors in minimized. The receiver which satisfies this criterion is called an **optimum receiver.**

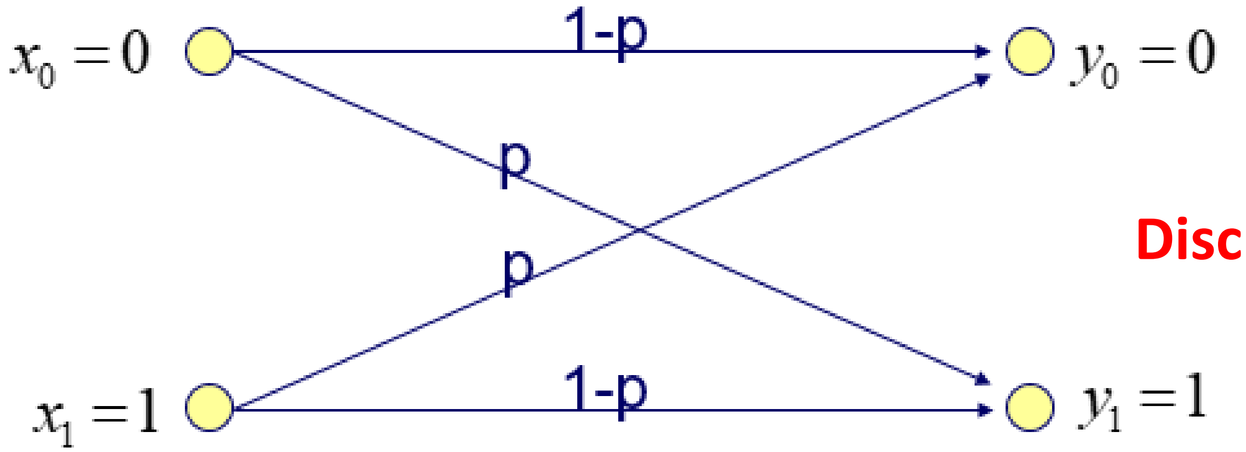
- Bit Error Probability (in the binary case): $p = Q\left(\sqrt{\frac{\int_0^{\tau}(s_1(t)-s_2(t))^2 dt}{2N_0}}\right)$

- τ: **binary symbol duration**
- N_0: **AWGN power**

Digit 0 -> $s_2(t)$

Discrete Input

Digit 1 -> $s_1(t)$

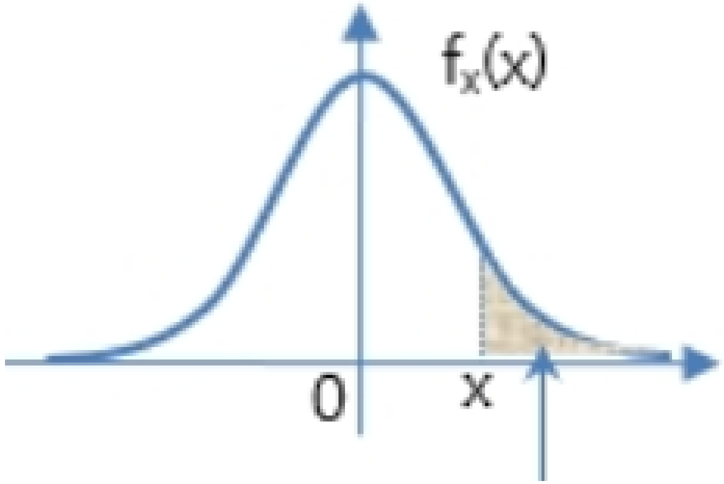


Discrete Output

Bit-error probability and data rate

Motivating Example: Binary PSK

- $s_1(t) = \mathrm{A}\cos(2\pi f_0 t)$; $0 \le t \le \tau;\ \tau = kT_0;$ Representing digit 1
- $s_2(t) = -\mathrm{A}\cos(2\pi f_0 t)$; ; $0 \le t \le \tau$; Representing digit 0
- $$\mathrm{p} = \boldsymbol{Q}\left(\sqrt{\frac{\int_0^\tau (\boldsymbol{s_1(t)} - \boldsymbol{s_2(t)})^2 \boldsymbol{dt}}{2\boldsymbol{N}_0}}\right) = Q\left(\sqrt{\frac{A^2\tau}{N_0}}\right) = \boldsymbol{Q}\left(\sqrt{\frac{\boldsymbol{A^2}}{\boldsymbol{R_b N_0}}}\right)$$
- **How to minimize the error probability?**
  - Increase the signal power (by increasing A); quite obvious
  - Reduce the data rate $\boldsymbol{R_b}$.
  - as $R_b \uparrow,\ x\ of\ Q(x)\ \downarrow and, therefore,\ P_b^* \uparrow.$
- Q(.) is the complementary Gaussian distribution function.



$\boldsymbol{p = Q(x)}$

Bit-error and block error probabilities

1-bit $\quad p = Q\left(\sqrt{\dfrac{A^2}{R_b N_0}}\right) \rightarrow 0 \ \ as \ \ R_b \rightarrow 0, or\ power\ (A) \rightarrow \infty$

τ	τ	τ	τ
bit	bit	bit	bit

Block of k bits $\rightarrow$ **Encoding Scheme** $\rightarrow$ **Channel with capacity C**

Block error probability of error $\rightarrow$ 0 for a finite data date R_b and a finite power

Remark: **Information theory promises that the probability of error can be made arbitrarily small (for a finite rate and a finite power) as long as the transmission rate is below a Channel Capacity.**

Efficiency and Reliability of a Digital Communication System

Lecture 2

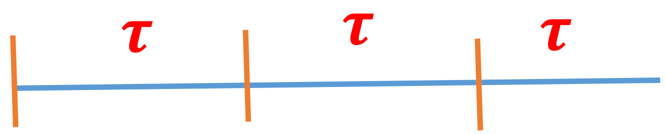
Lecture Outline

- Distinguish between bit error and block error probabilities in a digital communication system
- Define the efficiency of a digital communication system
- Explain the difference between fixed and variable length codes
- Define the reliability of a digital communication system

Modulation and Error Probability

- In a digital data transmission, the receiver has to decide which symbol was transmitted such that the probability of making errors in minimized. The receiver which satisfies this criterion is called an **optimum receiver.**

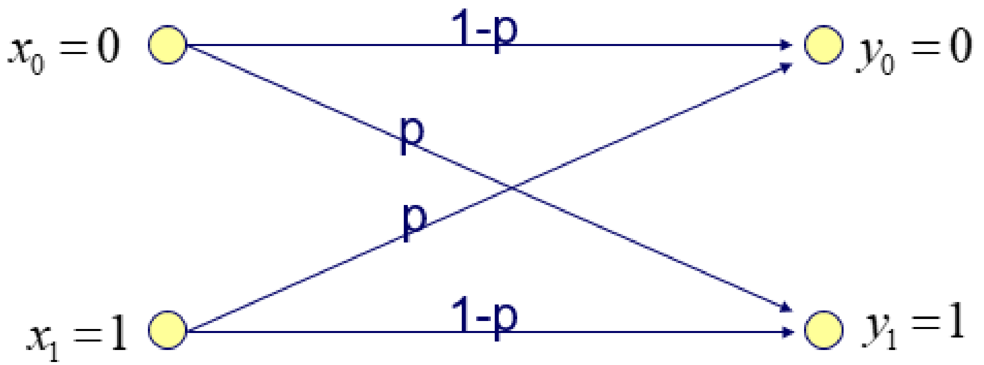
- Bit Error Probability (in the binary case): $p = Q\left(\sqrt{\frac{\int_0^{\tau}(s_1(t)-s_2(t))^2 dt}{2N_0}}\right)$



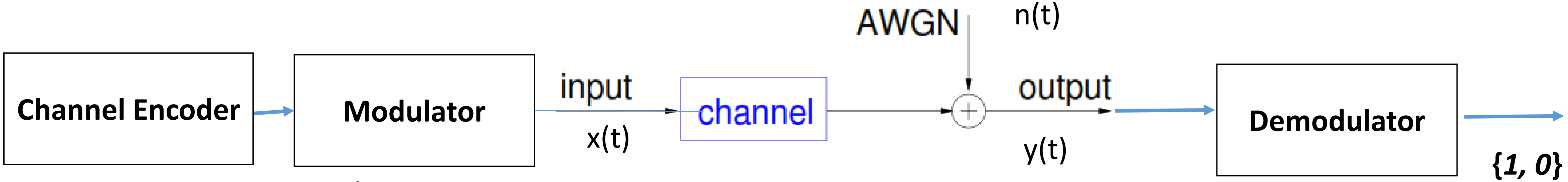
Digit 0 -> $s_2(t)$

Discrete Input

Digit 1 -> $s_1(t)$



Discrete Output

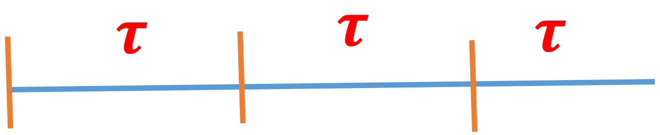


{*1, 0*} **Discrete Input** **Continuous Input** **Continuous Output** {*1, 0*} **Discrete output**

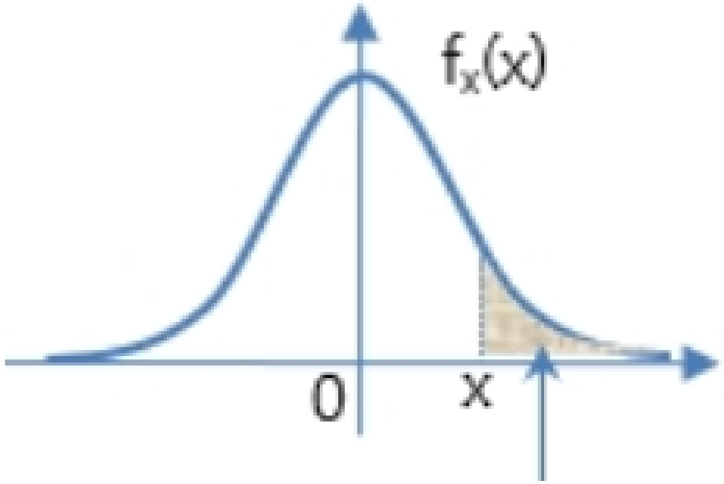
Bit-error probability and data rate

Motivating Example: Binary PSK

- $s_1(t) = \text{A}\cos(2\pi f_0 t)\,; \quad 0 \le t \le \tau;\ \tau = kT_0;$ Representing digit 1
- $s_2(t) = -\,\text{A}\cos(2\pi f_0 t)\,;;\, 0 \le t \le \tau;$ Representing digit 0
- $$\text{p} = \boldsymbol{Q}\left(\sqrt{\frac{\int_0^{\tau}\left(\boldsymbol{s_1(t)}-\boldsymbol{s_2(t)}\right)^2 \boldsymbol{dt}}{2\boldsymbol{N}_0}}\right) = Q\left(\sqrt{\frac{A^2\tau}{N_0}}\right) = \boldsymbol{Q}\left(\sqrt{\frac{\boldsymbol{A^2}}{\boldsymbol{R_b N_0}}}\right)$$



- **How to minimize the error probability?**
  - Increase the signal power (by increasing A); quite obvious
  - Reduce the data rate $\boldsymbol{R_b}$.
  - as $R_b \uparrow,\ x\ of\ Q(x)\ \downarrow and, therefore,\ P_b^* \uparrow.$
- Q(.) is the complementary Gaussian distribution function.



$\boldsymbol{p = Q(x)}$

Bit-error and block error probabilities

1-bit

$$p = Q\left(\sqrt{\frac{A^2}{R_b N_0}}\right) \rightarrow 0 \;\; as \;\; R_b \rightarrow 0, or\ power\ (A) \rightarrow \infty$$

τ	τ	τ	τ
bit	bit	bit	bit

Block of k bits

→ **Encoding Scheme** → **Channel with capacity C**

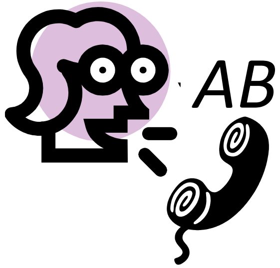
Block error probability of error $\rightarrow$ 0 for a finite data date R_b and a finite power

Remark: Information theory promises that the probability of error can be made arbitrarily small (for a finite rate and a finite power) as long as the transmission rate is below a Channel Capacity.

Problem One: Reliability

Communication is not always reliable.

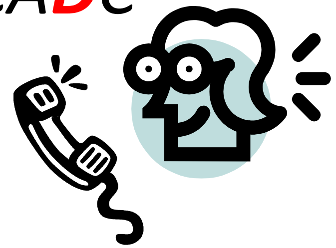
- **transmitted information ≠ received information**



ABCABC

Noisy Channel

*ABCA**D**C*



- Errors of this kind are unavoidable in real communication.
- In the usual conversation, we sometimes overcome these errors by
  - **Repeating the sentences**
  - **Using phonetic codes.**

ABC → **A**pple, **B**anana, **C**harlie → 

Phonetic Code

$\underbrace{\text{A}}_{\text{the real information}}\underbrace{\text{pple}}_{\text{redundant information for correcting possible errors}}$

- A phonetic code adds redundant characters (**parity characters**)
- The redundant part helps correcting possible errors.

→ *Use this mechanism over 0-1 data, and we can detect and correct errors?*

Redundancy to Improve Reliability

Q. **Can we add "redundant bits" to binary data?**

A. **Yes. One possibility is to use parity bits.**

A parity bit is: a binary digit, which is added to make the number of 1's in the data message even.

- 00101 → 001010 (two 1's → two 1's)
- 11010 → 110101 (three 1's → four 1's)

One parity bit may tell you that there are odd numbers of errors. But not more than that, i..e., **Error Detection (odd number of bits)**

Example: Receive 001010 (even number of bits) ⇒ accept (received = transmitted)

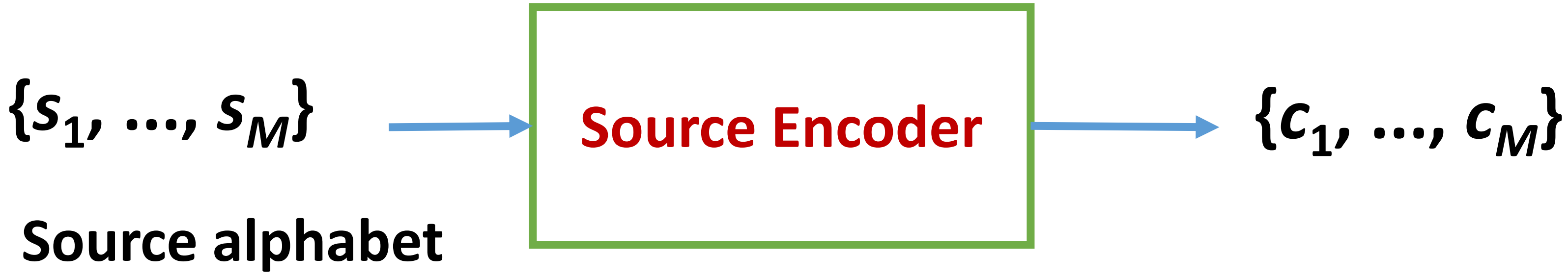
Example: Receive 001011 (odd number of bits) ⇒ Reject (one bit in error)

Example: Receive 001001 (even number of bits) ⇒ accept even though 2 bits in error

Note: Error detection is employed in the data link layer of computer networks. There, Cyclic Redundancy Check (CRC) error detection codes are used. We shall consider that later in the course

Problem Two: Efficiency

- Given a source S. Source encoder assigns binary digits for each source symbol such that
- the average number of digits/symbol is minimum (efficient representation)
- the code is uniquely decodable



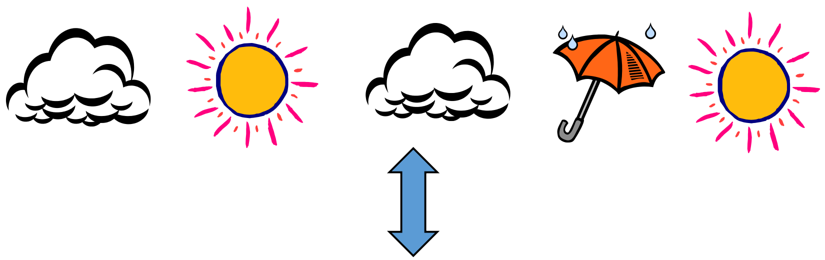
Problem Two: Efficiency

Example: We need to record the weather of a given city every day.

- Weather = {sunny, cloudy, rainy}; three possible states.
- We can use only "0" and "1", cannot use blank spaces.
- The source alphabet **M=3.**

weather	codeword
sunny	00
cloudy	01
rainy	10

- 2-bit record everyday (**equal length code**) ; $m = \lceil \log(3) \rceil$; => m=2
- **M=3** symbols need 2 binary digits
- (100 days, need 200 bits)



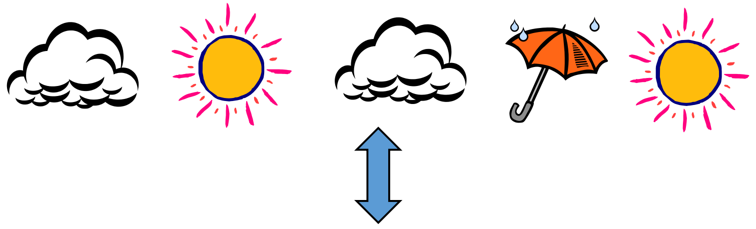
0100011000

The source consists of 3 messages each one is mapped into a sequence of binary digits (source codewords)

Can we shorten the representation?

A Better Code: Variable Length Code

weather	code A	code B
sunny	00	00 (2 digits)
cloudy	01	01 (2 digits)
rainy	10	1 (1 digit)



code A...0100011000

code B...010001100

Code B gives a shorter representation than Code A.

- Can we decode Code B correctly?
  - Yes, as far as the sequence is processed from the beginning.
- Is there a code which is more compact than code B?
  - Let us try that (→ next slide).
  - The probability distribution of the source need to be known

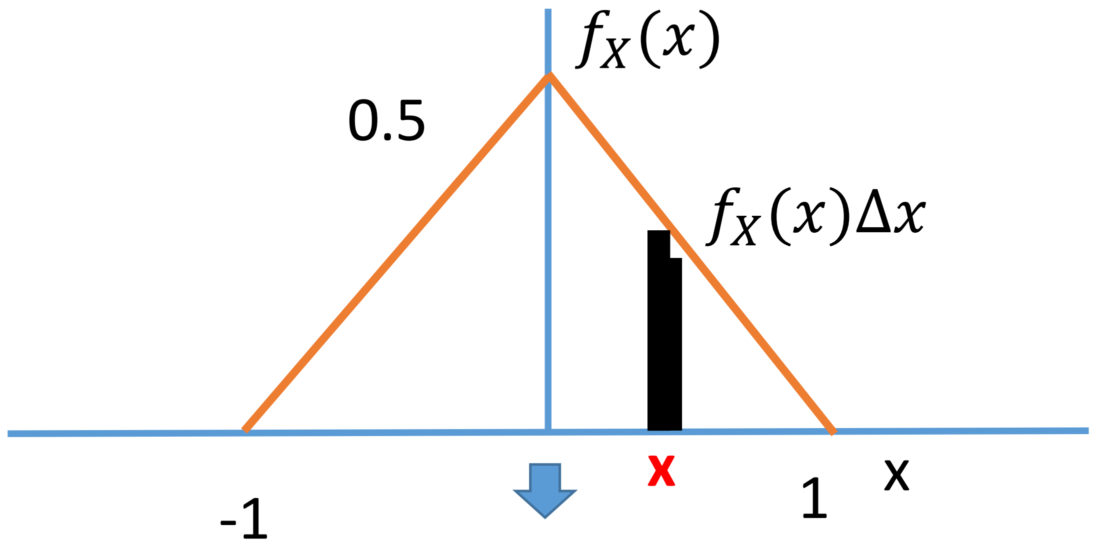
Mean and Variance of a Random Variable

Definition: The mean value or expected value or average value of a random variable X is defined as:

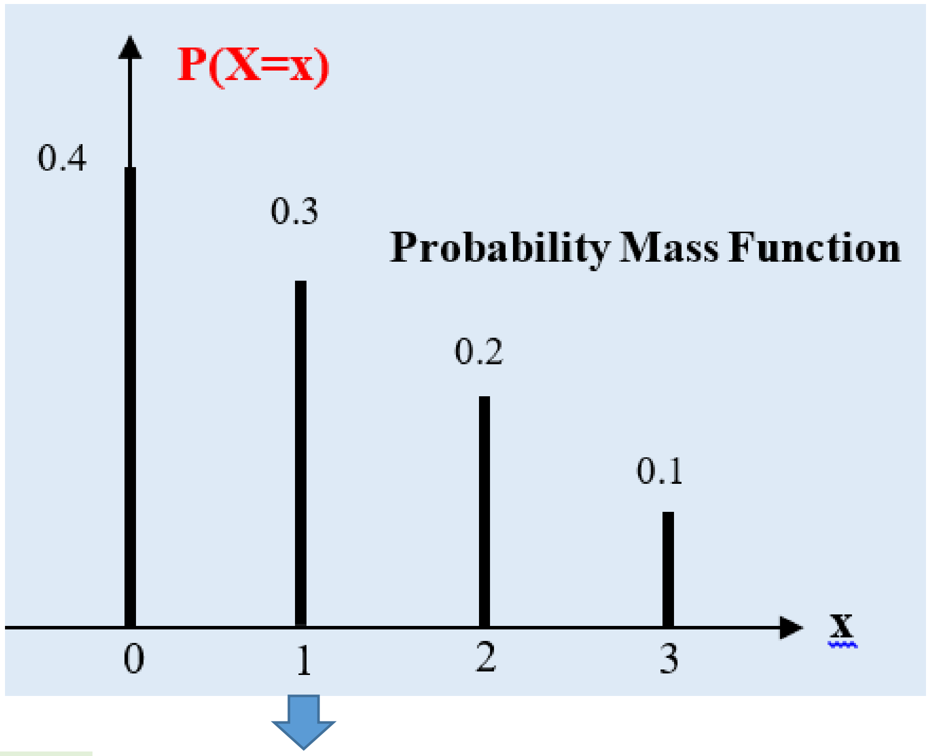
$$\mu_X = E\{X\} = \sum x_i \, P(X = x_i) \qquad \text{if X is discrete}$$

$$\mu_X = E\{X\} = \int_{-\infty}^{\infty} x \, f_X(x)\, dx \qquad \text{if X is continuous}$$

The mean is analogous to the center of mass of a weight distribution



Mean



E(X) = (0)(0.4)+(1)(0.3)+(2)(0.2)+(3)(0.1) = 1

Mean Point of equilibrium

Mean and Variance of a Random Variable

Definition: The **variance** of a random variable X is defined as:

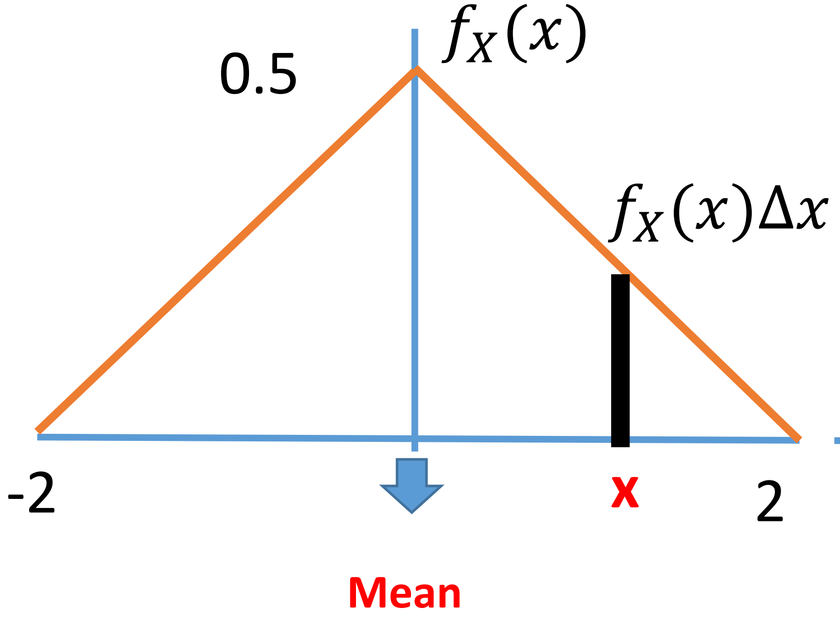
$$\sigma_X^2 = E\{(X-\mu_x)^2\} = \sum (X-\mu_x)^2 \; P(X=x_i) \quad \text{if X is discrete}$$

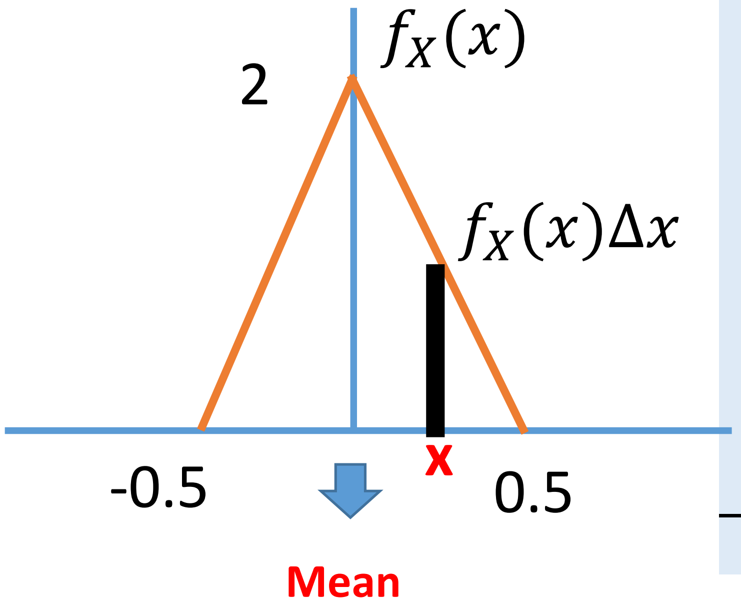
$$\sigma_X^2 = E\{(X-\mu_x)^2\} = \int_{-\infty}^{\infty} (x-\mu_x)^2 \; f_X(x)\,dx \quad \text{if X is continuous}$$

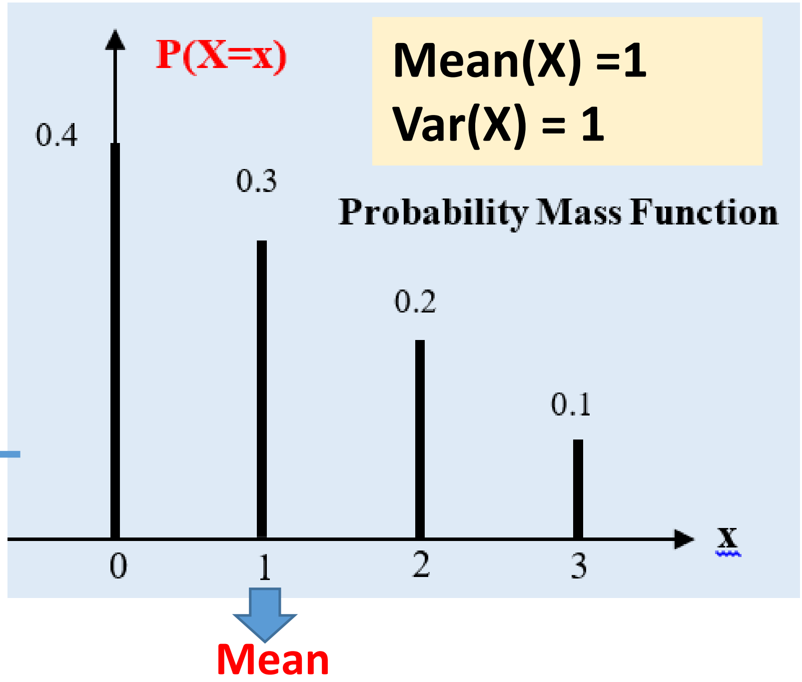
$$\sigma_X = \sqrt{\sigma_X^2} \quad \text{is the standard deviation}$$

The variance is analogous to the centralized moment of inertia

The variance is the measure of the spread of the distribution.







Average Length of Codes

Sometimes, events are not equally likely...

→ Probability comes into play

weather	probability	code A	code B	code C
sunny	0.5	00	00	1
cloudy	0.3	01	01	01
rainy	0.2	10	1	00

- **For Code A: 2.0 bit / event (always), (fixed length coding)**
- **Codes B and C are variable length source encoders.**
- **For Code B, (without a calculated knowledge)**

  $2\times0.5 + 2\times0.3 + 1\times0.2 =$ 1.8 bit / event (on the average)
- **For Code C, (educator's guess: Symbol probabilities exploited)**

  $1\times0.5 + 2\times0.3 + 2\times0.2 =$ 1.5 bit / event (on the average)

The Best Code

Question: Can we represent information with 1.1 binary digit/ per event (on the average)?

Answer: NO, To be investigated later in the course...

- **It is likely that there is a "limit" which we cannot get over.**
- **Shannon investigated the limit mathematically.**

  → For this event set, we need 1.485 or more bit per event.

weather	probability
sunny	0.5
cloudy	0.3
rainy	0.2

This is also **the average amount of information provided by the source**.

How do we arrive at the 1.485?

LATER

Discrete Memory-less Information Sources
Lecture Outline

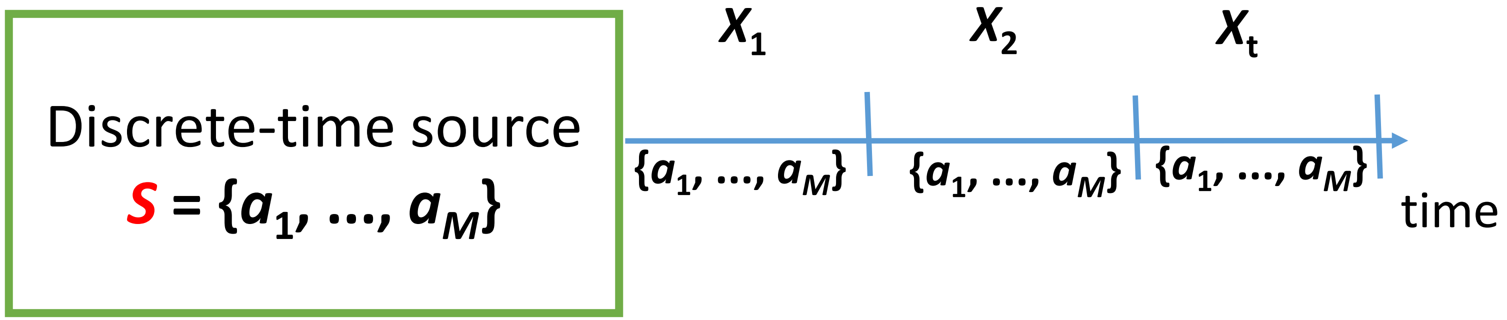
Lecture 3

- Two models are used describe discrete-time information sources
  - Discrete memory-less sources (DMS)
  - Markov sources; used to model sources with memory
- Markov sources are treated in the next lecture
- This lecture addresses DMS; two relevant concepts are introduced
  - Statistical Independence
  - Stationarity

Modeling Discrete Time Digital Information Sources

Two models are used to describe discrete information sources:

- Discrete memory-less sources (DMS)
- Markov information sources

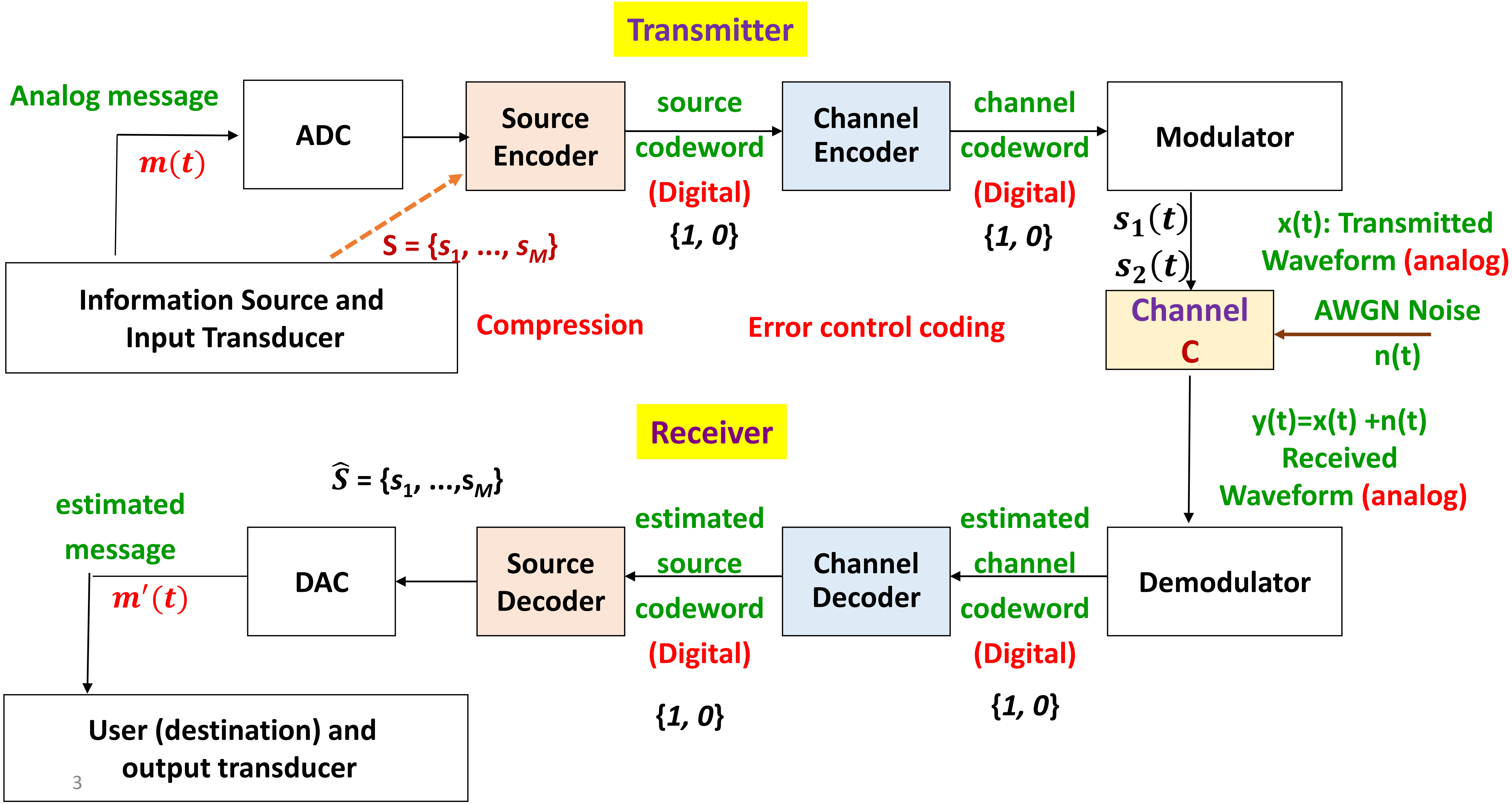


Assumptions on the source model:

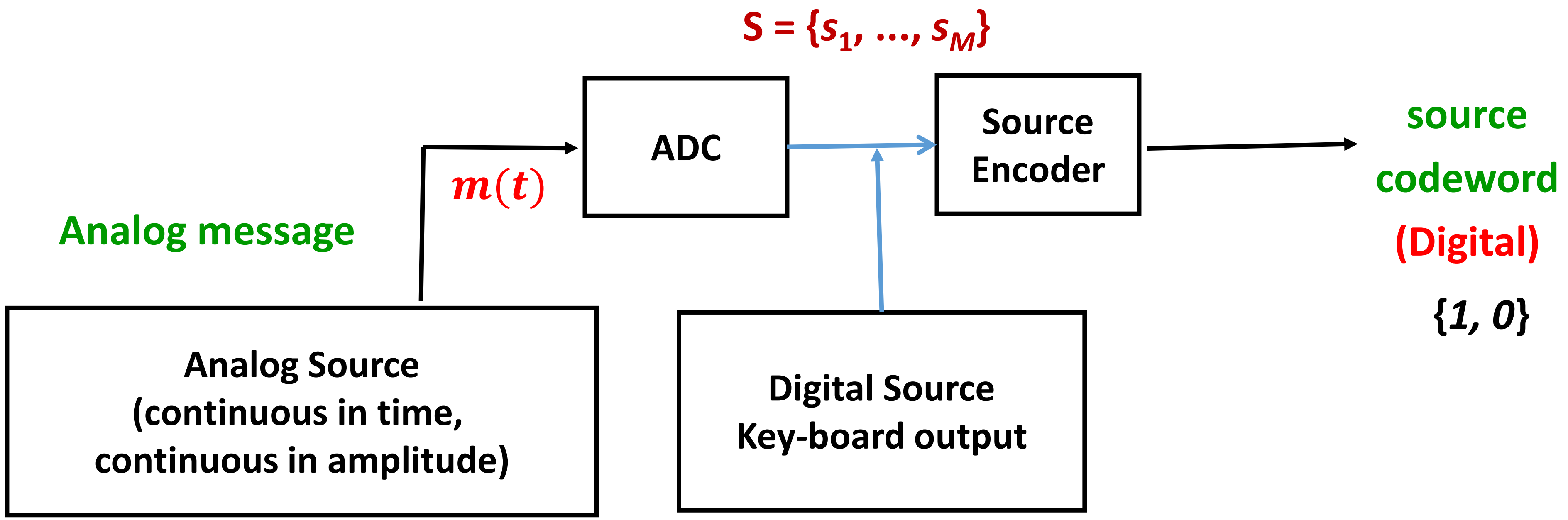
- **Discrete**: the set of possible symbols **S** is finite and countable. The number of elements in S is the size of the alphabet **|S|=M**
- The source generates one symbol from the set $S = \{a_1, \ldots, a_M\}$ each time unit. Hence the name **M-ary discrete-time information source.**

Remark: A continuous-time and/or analogue information sources can be converted into discrete source through sampling & quantization, as we have explained earlier.

A Basic Communication System Block Diagram: Revisited



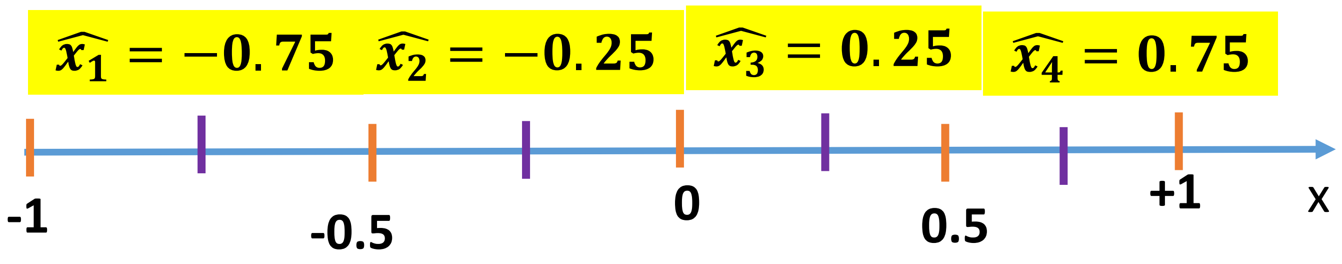
The Source Encoder

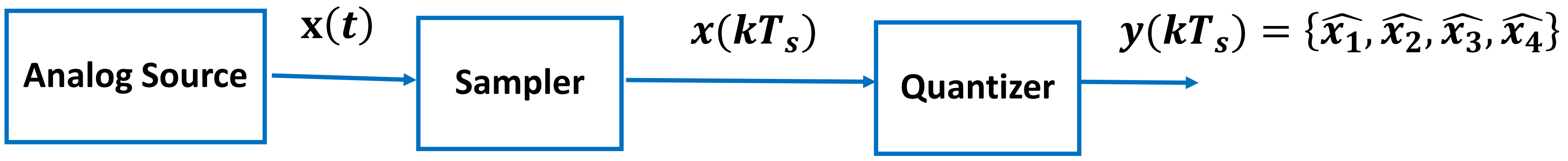


Quantization: the two-bit quantizer

- **Example**: The signal $\mathrm{x}(\mathrm{t}) = \cos(2\pi t)$ is sampled uniformly at a rate of 20 samples per second. The samples are applied to a four-level uniform quantizer with input-output characteristic

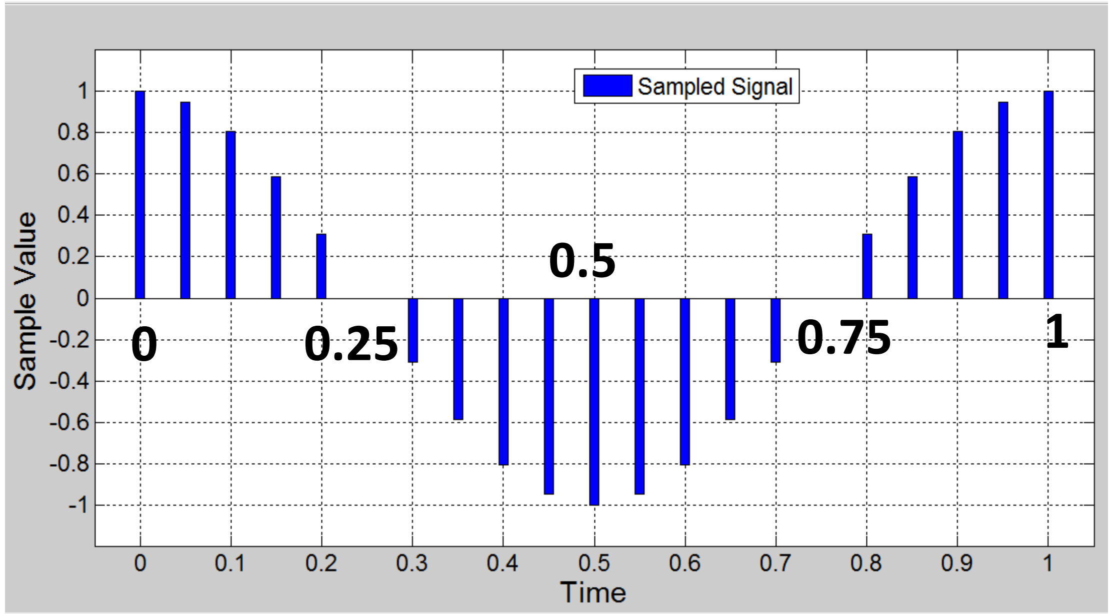
- $$y(kT_s) = \begin{cases} 0.75, & 0.5 < x < 1 \\ 0.25, & 0 < x < 0.5 \\ -0.25, & -0.5 < x < 0 \\ -0.75, & -1 < x < -0.5 \end{cases}$$





Quantization: the two-bit quantizer

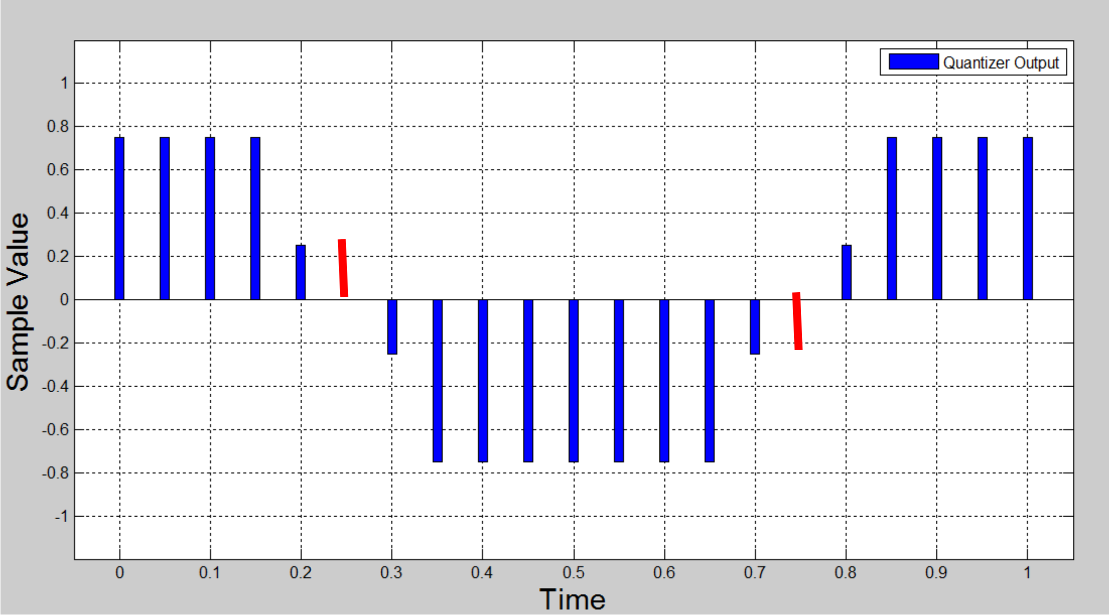
$$\mathbf{x}(t) = \cos(2\pi t)$$



Sampled Signal

$$T_s = 0.05$$

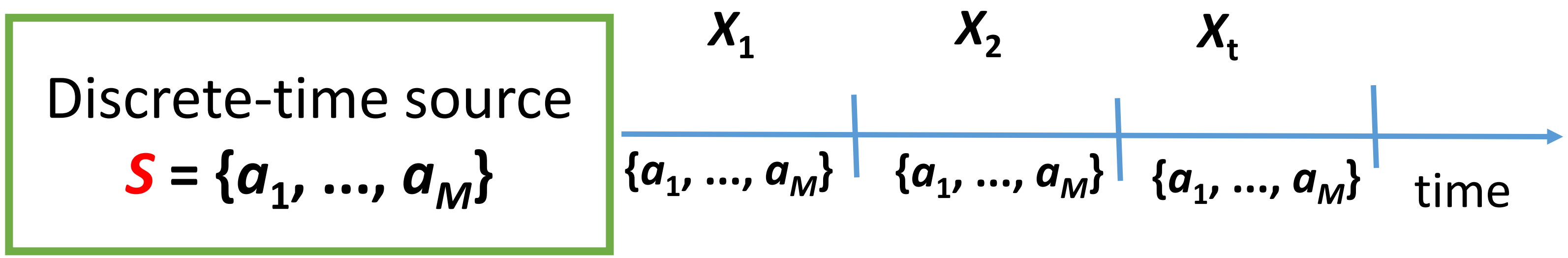
$$y(kT_s) = \{\widehat{-0.75}, -0.25, -0.25, 0.75\}$$



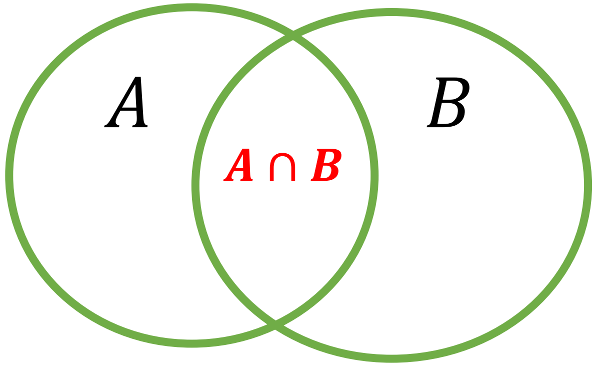
Quantized Signal

Discrete Time Digital Information Sources

The concept of statistical independence:



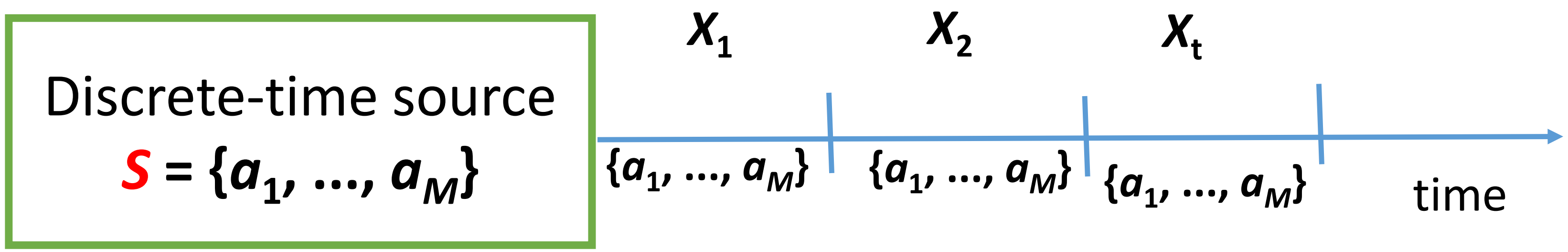
Two events A and B are said to be statistically independent when:



- $P(A \cap B) = P(A)P(B)$
- The conditional probability of A given B is given as:
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- For independent events,
- $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$
- $\boldsymbol{P(A|B) = P(A)}$; whether B is given or not, the probability of A remains the same.

Discrete Time Digital Information Sources

We apply the concept of statistical independence to the first model of discrete memory-less sources

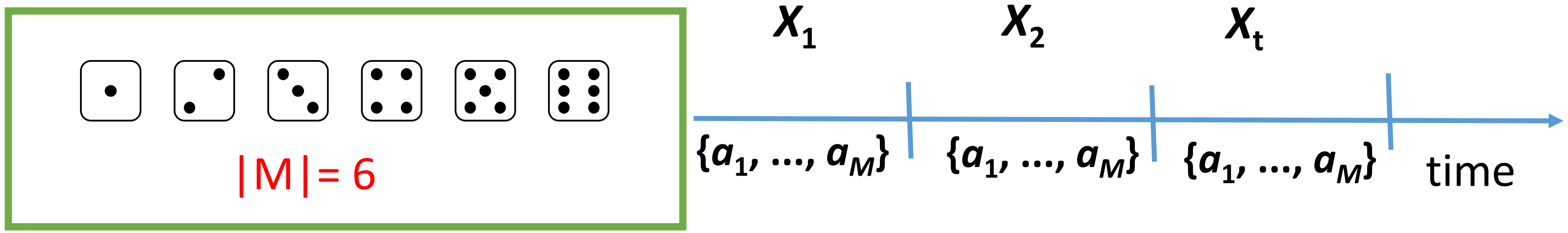


- $P(X_2 = x_2 | X_1 = x_1) = P(X_2 = x_2)$; **independent source**
- Also,
- $P(X_2 = x_2 \cap X_1 = x_1) = P(X_2 = x_2)P(X_1 = x_1)$
- And, in general, for an independent source we have:
- $P(X_t = x_t \cap \cdots \cap X_2 = x_2 \cap X_1 = x_1) = P(X_t = x_t) \ldots P(X_2 = x_2)P(X_1 = x_1)$

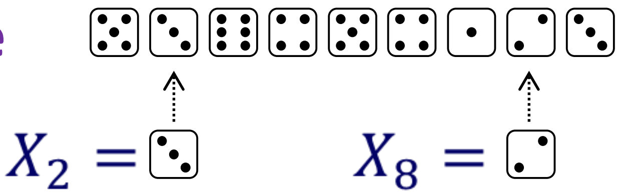
Discrete Time Digital Information Sources

- Assume a discrete-time digital information source X:
  - $\mathcal{X} = \{a_1, \ldots, a_M\}$… the set of symbols of X (alphabet of X)
    (X is said to be an M-ary information source.)
  - X_t : the symbol which X produces at time t. *Can assume any of M values*
  - The sequence $\boldsymbol{X_1, \ldots, X_n}$ is called a **message** produced by X (Here, the message consists of n symbols).

Example: Tossing a six-faced fair die 9 times independently



Let the message be

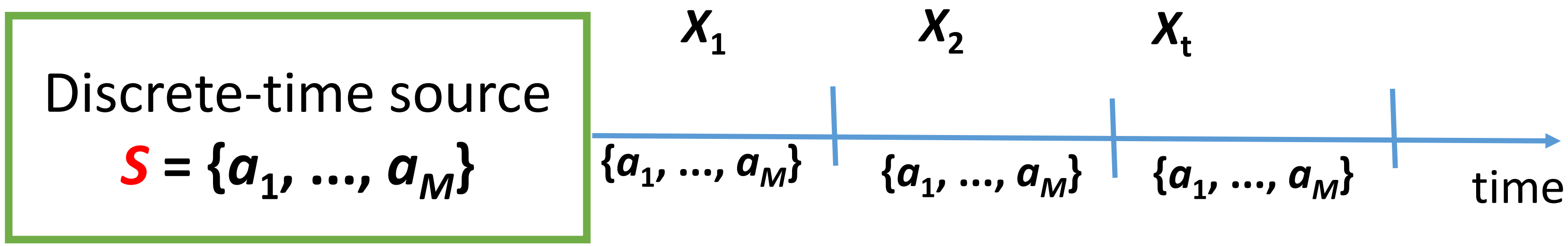


, then

Discrete Memoryless Sources (DMS)

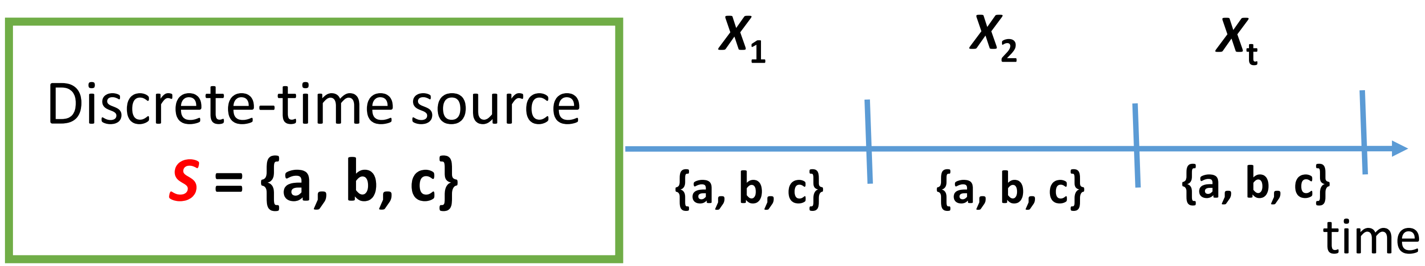
- A discrete memoryless and stationary information source satisfies the independence (memory-less) condition:
- **Memoryless condition**: $\boldsymbol{P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots . X_2 = x_2, X_1 = x_1) = P(X_t = x_t)}$
- **Memoryless condition:** "A symbol is chosen **independently** from past symbols."
- **Stationary condition**: The probability mass function is independent of time
- **For example,** $\boldsymbol{P(X_t = a_1) = P(X_1 = a_1)}$, for any time t, and so on

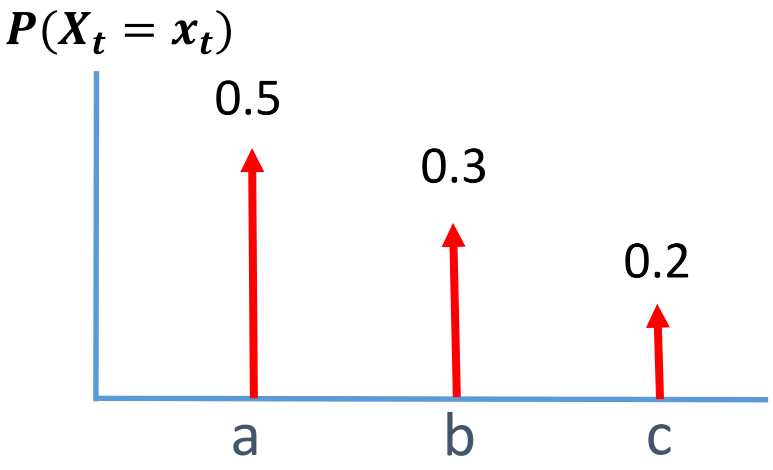
Stationarity: The probability distribution is time-invariant."



Discrete Memoryless Sources (DMS): Example

- **Example**: Consider a discrete memory-less source S which emits one of three possible symbols {a, b, c} every time unit with the following probabilities:
- P(a) = 0.5, P(b) = 0.3, P(c) = 0.2
- The probability mass function of the source is shown below.
- For a stationary source, this represents the pmf of $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_t}$
- $\mathrm{P}(X_2 = b) = 0.3$, $\mathrm{P}(X_1 = b) = 0.3$, $\mathrm{P}(X_{10} = b) = 0.3$
- $\mathrm{P}(X_2 = b \cap X_8 = a) = P(X_2 = b)P(X_8 = a) = (0.3)(0.5) = 0.15$

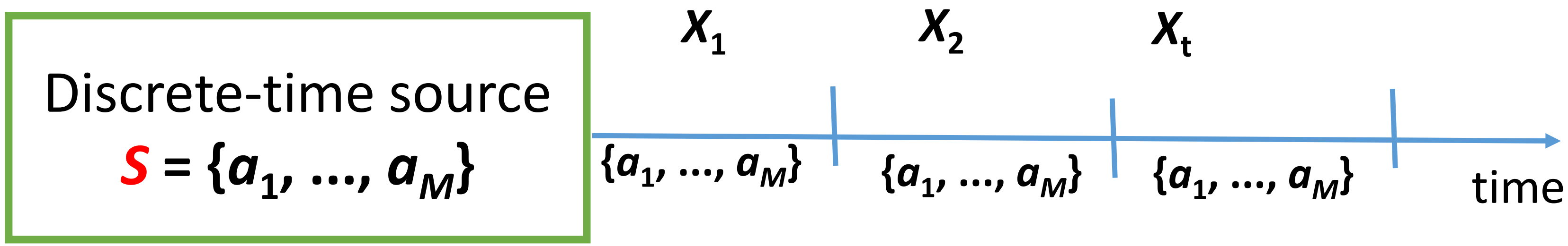




Sources with Memory

- A memoryless and stationary information source satisfies the independence condition:
- **Memoryless condition**: $P(X_t = x_t | X_{t-1} = x_{t-1}, \dots . X_2 = x_2, X_1 = x_1) = P(X_t = x_t)$
- **For a source with memory, past states affect the occurrence of future symbols, i.e.,**
- $P(X_t = x_t | X_{t-1} = x_{t-1}, \dots . X_2 = x_2, X_1 = x_1) \neq P(X_t = x_t)$
- This implies that the probability mass function is time-dependent.
- **For example,** $P(X_t = a_1) \neq P(X_{t-1} = a_1) \neq P(X_1 = a_1)$

The probability distribution is time-dependent

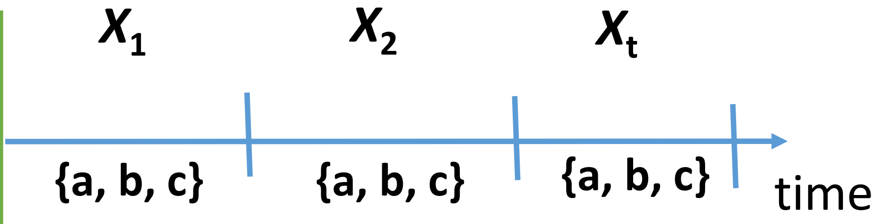


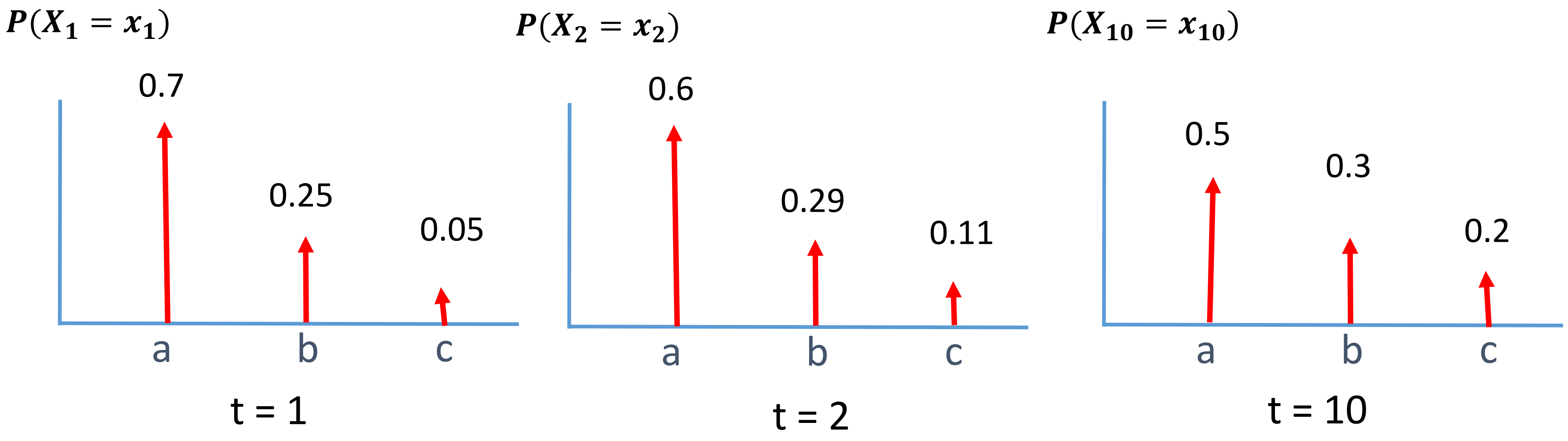
Sources with Memory

- **Sources with memory: The probability distribution is time-dependent**

Same set of alphabet as DMS

Discrete-time source
S = {a, b, c}





Sources with Memory

- **Example From English Language**:

In a given short story, one can find the following probabilities:

- P(o) = 0.063 , P(f) = 0.021, P(of) = 0.035493; P(x)= N_x/N
- Assuming independence: P(of) = P(o)P(f) = (0.063)(0.021) = 0.001323
- Note that **P(of) >> P(o)P(f)**
- Similar examples from the English language (sources with memory)
  - English text: $P_{X_t|X_{t-1}}(u|q) \gg P_{X_t|X_{t-1}}(u|u)$

Quality, Prerequisite
Continuum

Markov Sources
Lecture Outline

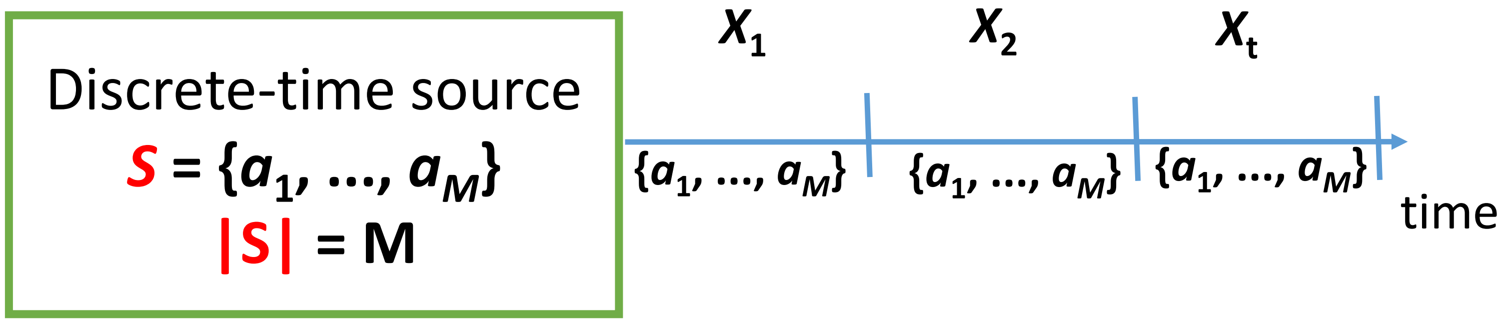
Lecture 4

- Two models describe discrete-time information sources:
  - Discrete memory-less sources (DMS); addressed in the previous lecture
  - Markov sources; used to model sources with memory
- Markov sources are the subject of this lecture. The lecture covers
  - The state diagram and the state equations of a simple Markov source.
  - Transient analysis of the Markov source
  - Steady-state solution of the stationary Markov source
  - Regular Markov sources

Modeling Discrete Time Digital Information Sources

Assumptions on the source model:

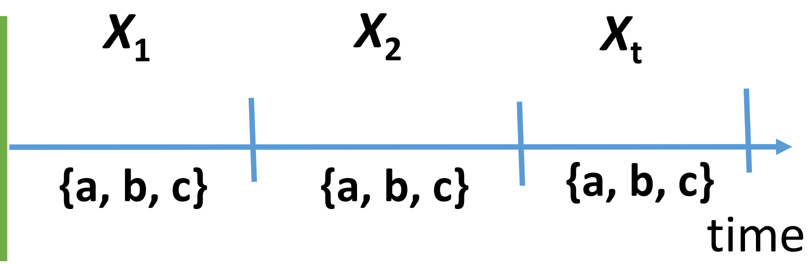
- **Discrete**: the set of possible symbols **S** is finite and countable. The number of elements in S is the size of the alphabet **|S|=M**
- The source generates one symbol from the set $\boldsymbol{S = \{a_1, \ldots, a_M\}}$ each time unit. Hence the name **M-ary discrete-time information source.**

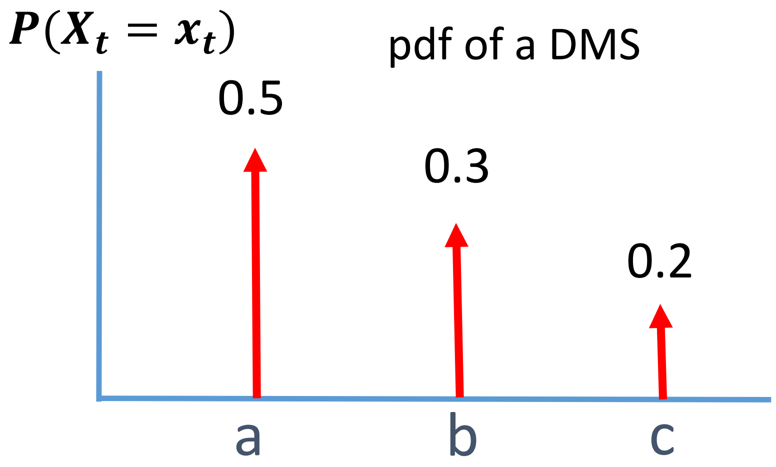


Sources with Memory

- A memoryless and stationary information source satisfies the independence condition:
- **Memoryless condition**: $P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots . X_2 = x_2, X_1 = x_1) = P(X_t = x_t)$
- **For a DMS source, the probability distribution is time-independent**
- **The random variables $X_1, X_2, \ldots, X_{t-1}, X_t$ are independent**
- **For a source with memory, past states affect the occurrence of future symbols, i.e.,**
- $P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots . X_2 = x_2, X_1 = x_1) \neq P(X_t = x_t)$
- This implies that the probability mass function is time-dependent.
- **For example,** $P(X_t = a_1) \neq P(X_{t-1} = a_1) \neq P(X_1 = a_1)$

Discrete-time source
S = {a, b, c}
$|S|$ = 3



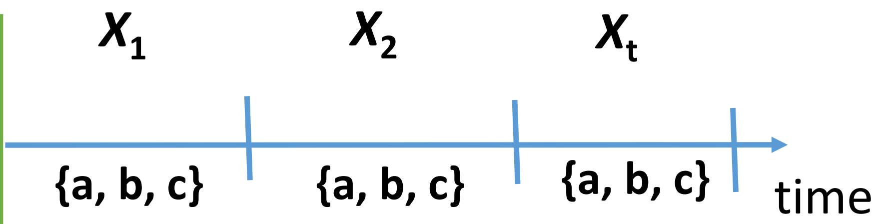


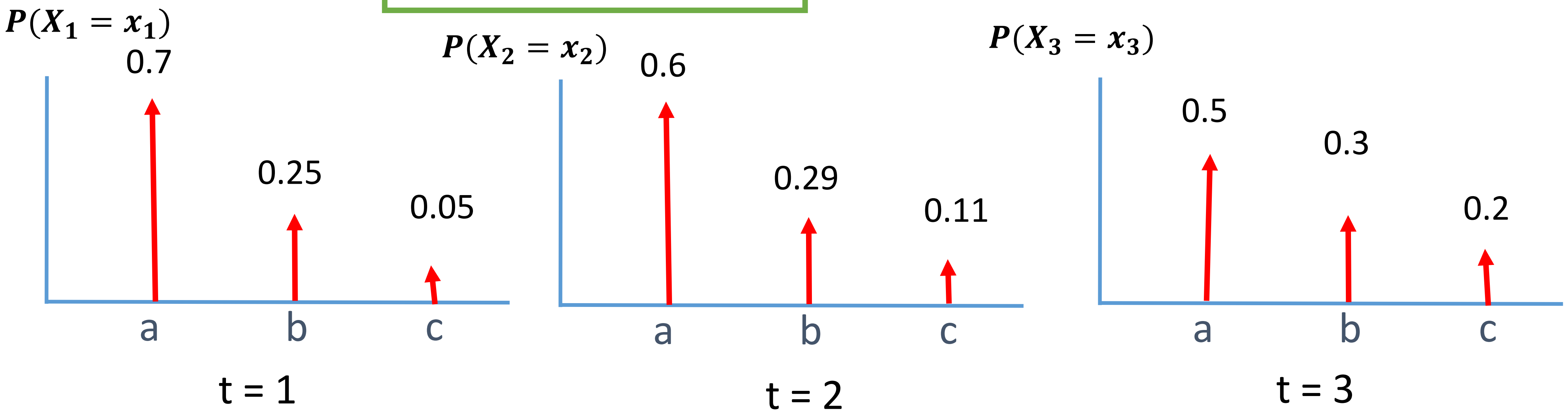
Sources with Memory

- **Sources with memory:**
  - **The probability distribution is time-dependent**
  - **The random variables $X_1, X_2, \dots, X_{t-1}, X_t$ are dependent**

Same set of alphabet as DMS

Discrete-time source
***S* = {a, b, c}**





Sources with Memory

- **Example From English Language**:

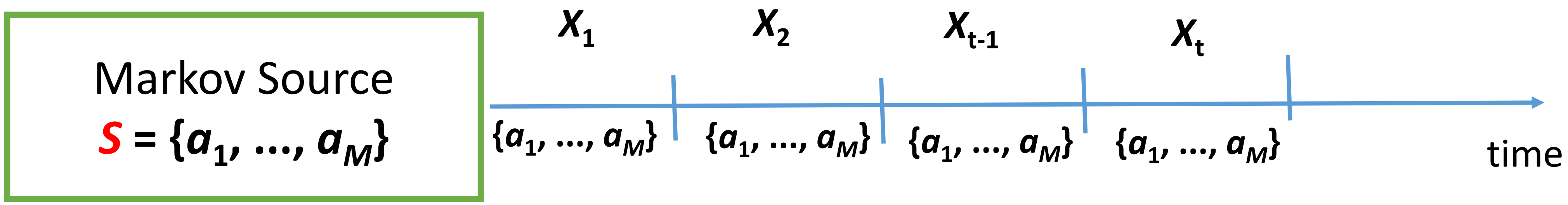
In a given short story, one can find the following probabilities:

- P(o) = 0.063 , P(f) = 0.021, P(of) = 0.035493; P(x)= N_x/N
- Assuming independence: P(of) = P(o)P(f) = (0.063)(0.021) = 0.001323
- Note that **P(of) >> P(o)P(f);**
- **Languages are structured and letters are not randomly chosen in words**
- Similar examples from the English language (sources with memory)
  - English text: $P_{X_t|X_{t-1}}(u|q) \gg P_{X_t|X_{t-1}}(u|u)$

Quality, Prerequisite
Continuum

Sources with Memory: Markov Information Sources

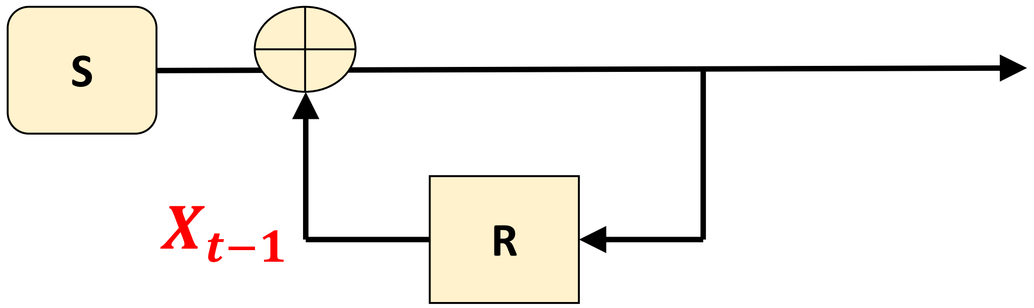
- Used to model information sources with memory.
- For an *m*-th order Markov source, the occurrence of the current symbol at time **t** depends on the past m symbols at **t-1, t-2, …, t-m**
- In a **simple Markov source**, the occurrence of the current symbol at time **t** depends only on the **previous symbol at time t-1**
- **Simple Markov Source to be discussed in this lecture,**
- $P(X_t = x_t | X_{t-1} = x_{t-1}, \dots . X_2 = x_2, X_1 = x_1) = P(X_t = x_t | X_{t-1} = x_{t-1})$



Example: Generation of a Simple Markov Source

The figure below shows how to generate a Markov source X_t. Let S be a discrete memoryless and stationary source with $P(0) = 0.2$, $P(1) = 0.8$

DMS with
P*(0) = *0.2
***P*(1) = 0.8**



1-bit register

$$X_t = S \oplus X_{t-1}$$

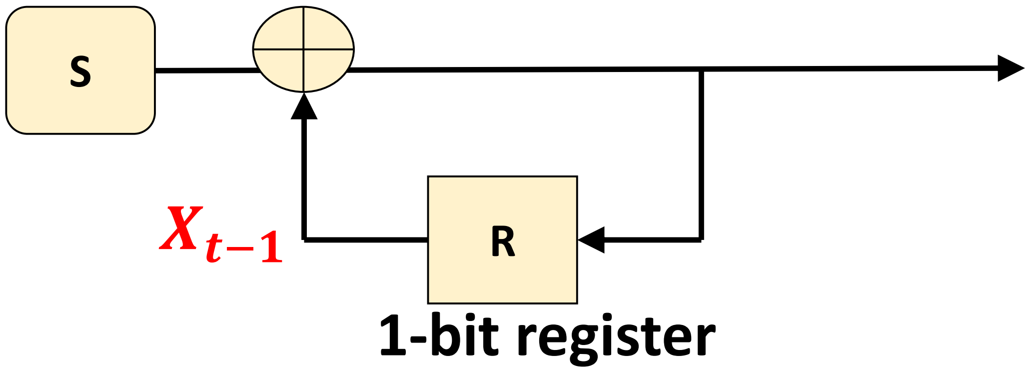
X_{t-1}	S	X_t
0	0	0
0	1	1
1	0	1
1	1	0

- The table shows the relationship between X_t, X_{t-1} and S.
- $P(X_t = 1) = P(X_{t-1} = 0 \cap S=1) + P(X_{t-1} = 1 \cap S=0)$
- From probability theory, we know that
- $P(A \cap B) = P(A)P(B|A)$;
- $P(X_t = 1) = P(X_{t-1} = 0)\, P(S=1/X_{t-1} = 0) + P(X_{t-1} = 1)\, P(S=0/X_{t-1} = 1)$
- *But S is an independent source, hence*
- $P(X_t = 1) = P(X_{t-1} = 0)\,(0.8) + P(X_{t-1} = 1)\,(0.2)$

Example: Generation of a Simple Markov Source

The figure below shows how to generate a Markov source X_t. Let S be a discrete memoryless and stationary source with *P*(0) = *0.2*, *P*(1) = 0.8

DMS with
P*(0) = *0.2
***P*(1) = 0.8**



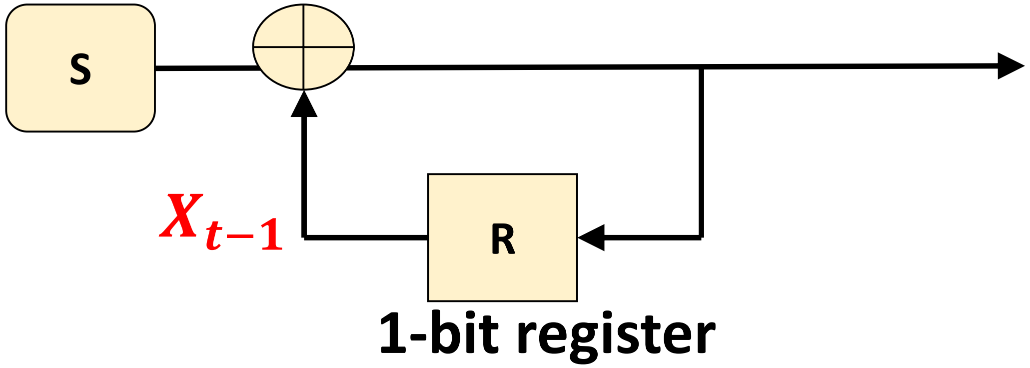
$$X_t = S \oplus X_{t-1}$$

X_{t-1}	S	X_t
0	0	0
0	1	1
1	0	1
1	1	0

- **Similarly, we have**
- **P(X_t =0) = P(X_{t-1} =0 ∩S=0) + P(X_{t-1} =1 ∩S=1)**
- **P(X_t =0) = P(X_{t-1} =0) P(*S=0*/X_{t-1} =0) + P(X_{t-1} =1) P(*S=1*/X_{t-1} =1)**
- ***But S is an independent source, hence***
- **P(X_t =0)=P(X_{t-1} =0) (0.2) + P(X_{t-1} =1) (*0.8*)**
- **Also, P(X_t = 0) = 1 - P(X_t = 1)**

The Simple Markov Source

DMS with
P*(0) = *0.2
***P*(1) = 0.8**



$$X_t = S \oplus X_{t-1}$$

Basic State Equations

$$\mathbf{P}(X_t = 1) = \mathbf{P}(X_{t-1} = 0)\,(0.8) + \mathbf{P}(X_{t-1} = 1)\,(0.2)$$

$$\mathbf{P}(X_t = 0) = \mathbf{P}(X_{t-1} = 0)(\,0.2) + \mathbf{P}(X_{t-1} = 1)\,(0.8)$$

Distribution at time t depends on the distribution at time t-1

State Representation of the Simple Markov Source

- In the previous slides, we have seen that X_t, S, and $X_{t-1} \in \{0, 1\}$.

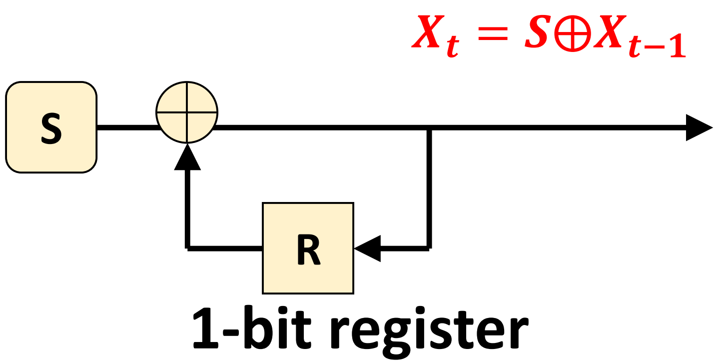
  The state equations are

  $P(X_t = 1) = P(X_{t-1} = 0)\,(0.8) + P(X_{t-1} = 1)\,(0.2)$

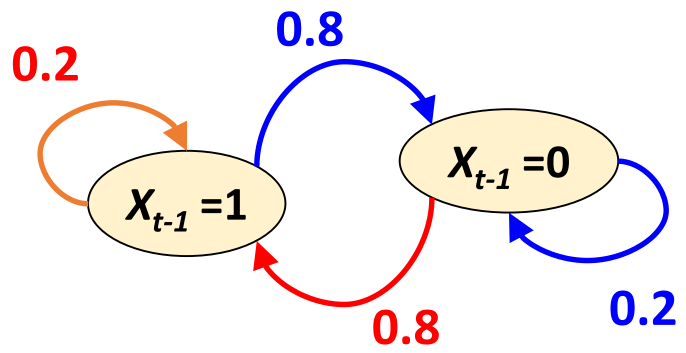
  $P(X_t = 0) = P(X_{t-1} = 0)\,(0.2) + P(X_{t-1} = 1)\,(0.8)$

- These equations can be represented in a state-diagram called the finite-state machine model.
- The arrows represent the transition probabilities from a given state to another state.

DMS with
P(0) =*0.2*
P(1) =0.8



Finite State Machine Model



Transient Analysis of the Simple Markov Source

- The state equations are
- $P(X_t = 1) = P(X_{t-1} = 0)\,(0.8) + P(X_{t-1} = 1)\,(0.2)$
- $P(X_t = 0) = P(X_{t-1} = 0)\,(0.2) + P(X_{t-1} = 1)\,(0.8)$
- Suppose that at t=0, system starts from state zero,
- i.e., **$P(X_{t-1} = 0) = 1$, so that $P(X_{t-1} = 1) = 0$.**
- With these initial conditions, we get
- $P(X_1 = 1) = P(X_{t-1} = 0)\,(0.8) + P(X_{t-1} = 1)\,(0.2)$

  $= (1)\,(0.8) + (0)(0.2) = 0.8$
- $P(X_1 = 0) = P(X_{t-1} = 0)(0.2) + P(X_{t-1} = 1)\,(0.8)$

  $= (1)(0.2) + (0)(0.8) = 0.2.$
- These values serve as initial conditions for the next time instance t = 2. The probabilities as a function of time are summarized in the table

t	$P(X_t = 1)$	$P(X_t = 0)$
0	0	1
1	0.8	0.2
2	0.32	0.68
3	0.608	0.392
4	0.4352	0.5648
5	0.53888	0.46112
6	0.476672	0.523328
7	0.5139968	0.4860032
8	0.49160192	0.50839808
∞	0.5	0.5

Steady-State Solution of the Simple Markov Source

- The state equations are
- $P(X_t = 1) = P(X_{t-1} = 0)\ (0.8) + P(X_{t-1} = 1)\ (0.2)$
- $P(X_t = 0) = P(X_{t-1} = 0)(\ 0.2) + P(X_{t-1} = 1)\ (0.8)$
- At steady-state, we have $P(X_t = 1) = P(X_{t-1} = 1) = \alpha$; time-independent
- $P(X_t = 0) = P(X_{t-1} = 0) = \beta$;
- Therefore,
- $P(X_t = 1) = P(X_{t-1} = 0)\ (0.8) + P(X_{t-1} = 1)\ (0.2)$

  $\alpha \qquad = \beta\ (0.8) + \alpha\ (0.2)$

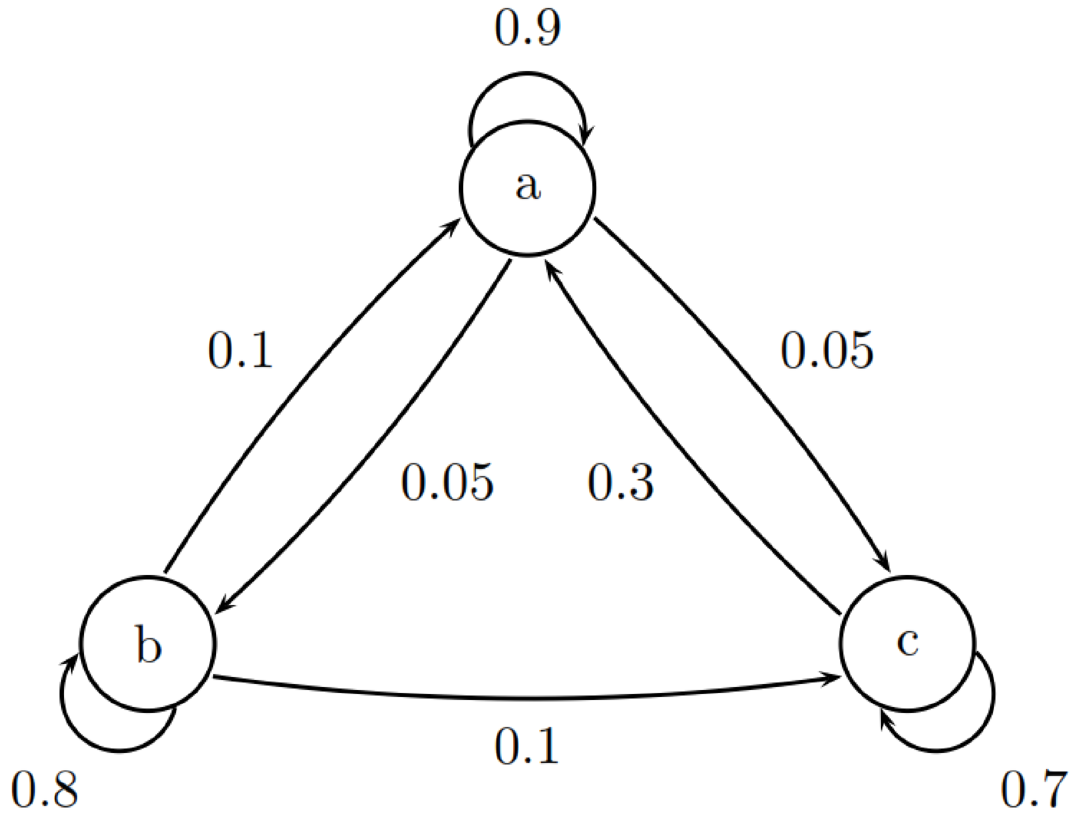
  $0.8\ \alpha = (0.8)\ \beta$
- Hence, **$\alpha = \beta = 0.5$**

Example: a three-state simple Markov source

Consider the stationary Markov source with three states of order 1 and transition probabilities as shown in the figure.

- Write down the state equations.
- Write down the steady-state state equations.
- Find the steady-state probabilities of the three states

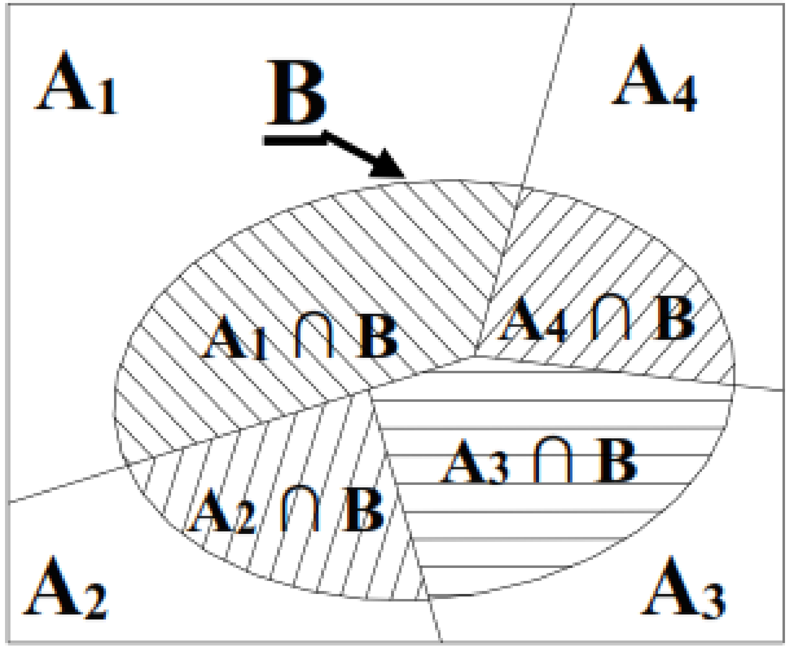
$$
\begin{array}{lll}
P(a|a) = 0.9 & P(b|a) = 0.05 & P(c|a) = 0.05 \\
P(a|b) = 0.1 & P(b|b) = 0.8 & P(c|b) = 0.1 \\
P(a|c) = 0.3 & P(b|c) = 0 & P(c|c) = 0.7
\end{array}
$$



Theorem of Total Probability

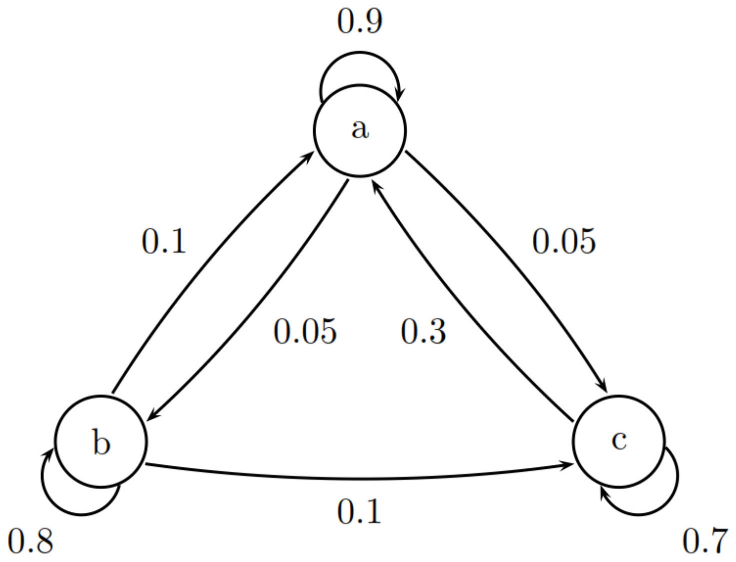
- In this example, we make use of the theorem of total probability.
- Let A_1, A_2, …, A_n be a set of events defined over S such that:
- $S = A_1 \cup A_2 \cup \ldots \cup A_n$; $A_i \cap A_j = \emptyset$ for $i \neq j$, and $P(Ai) > 0$ for $i = 1, 2, 3, \ldots n$.
- For any event (B) defined on S,

$$\mathbf{P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3)}$$



Example: a three-state simple Markov source

For the source shown on the previous slide, we can write the following state equations.

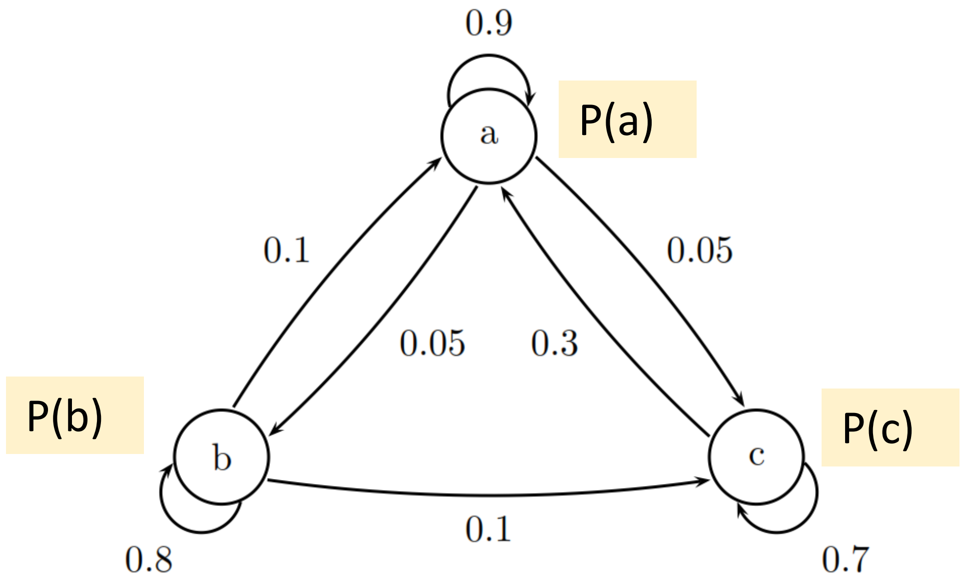


- $P(X_t = a) = P(X_{t-1} = a)\, P(X_t = a / X_{t-1} = a) + P(X_{t-1} = b)\, P(X_t = a / X_{t-1} = b)$
  $+ P(X_{t-1} = c)\, P(X_t = a / X_{t-1} = c)$
- $P(X_t = b) = P(X_{t-1} = a)\, P(X_t = b / X_{t-1} = a) + P(X_{t-1} = b)\, P(X_t = b / X_{t-1} = b)$
  $+ P(X_{t-1} = c)\, P(X_t = b / X_{t-1} = c)$
- $P(X_t = c) = P(X_{t-1} = a)\, P(X_t = c / X_{t-1} = a) + P(X_{t-1} = b)\, P(X_t = c / X_{t-1} = b)$
  $+ P(X_{t-1} = c)\, P(X_t = c / X_{t-1} = c)$
- Substituting the transition probabilities into the state equations above, we get
- $P(X_t = a) = P(X_{t-1} = a)\,(0.9) + P(X_{t-1} = b)\,(0.1) + P(X_{t-1} = c)\,(0.3)$
- $P(X_t = b) = P(X_{t-1} = a)\,(0.05) + P(X_{t-1} = b)\,(0.8) + P(X_{t-1} = c)\,(0)$
- $P(X_t = c) = P(X_{t-1} = a)\,(0.05) + P(X_{t-1} = b)\,(0.1) + P(X_{t-1} = c)\,(0.7)$

Example: a three-state simple Markov source

Steady-state solution

- **Note that the probabilities at time t are dependent on the probabilities at time (t-1).**
- **In the steady-state case, we have**
- **$P(X_{t-1}=a) = P(X_t=a) = P(a)$; $P(X_{t-1}=b)=P(X_t=b)=P(b)$; $P(X_{t-1}=c)=P(X_t=c)=P(c)$**
- **The state equations now become**
- **$P(a)=P(a)\,(0.9) + P(b)\,(0.1) + P(c)\,(0.3)$**
- **$P(b)=P(a)\,(0.05) + P(b)\,(0.8) + P(c)\,(0)$**
- **$P(c)=P(a)\,(0.05) + P(b)\,(0.1) + P(c)\,(0.7)$**



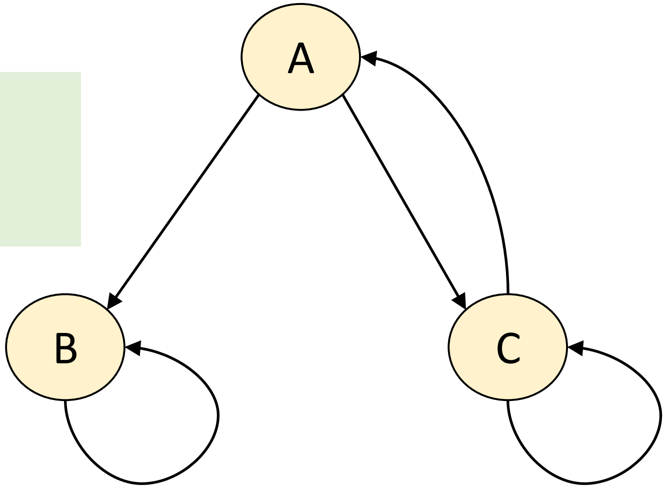
- **Solving the above equations, we get**
- **$P(a)=4/6$; $P(b)=1/6$; $P(c)=1/6$ (the following steady-state probabilities**

Two Important Properties of Markov Sources

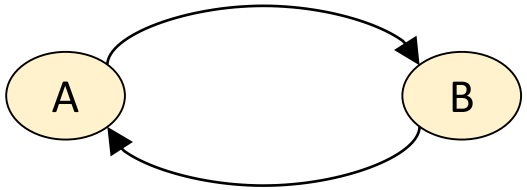
Irreducible Markov Source

- **Any state is accessible from any other state in a finite number of steps**

this example is NOT irreducible
If we start at B, we cannot reach either A or C



aperiodic Markov source: Source does not have a periodic behavior



Periodic Source

irreducible + aperiodic = regular (also known as ergodic).

Ergodic (Regular) Markov Process

Definition: A finite-state Markov chain is ergodic (regular) if all states are accessible from all other states and if all states are aperiodic, i.e., have period 1.

An important fact about ergodic Markov chains is that the chain has steady-state probabilities p(s) for all states.

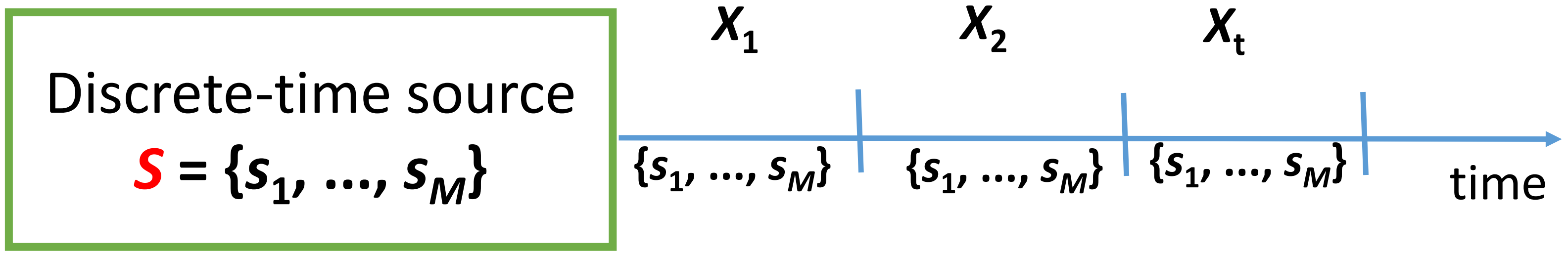
Measure of Information
Lecture Outline

Lecture 5

- Consider a discrete-time finite-alphabet source S of size M with a given probability distribution over its symbols.
- **In this lecture, we will try to answer the following questions:**
  - How do we measure the information produced by the source S?
  - What is the amount of information contained in each symbol?
  - What is the average amount of information per symbol in S?

The Source Entropy

- **Main Theme**: Consider a discrete-time finite-alphabet source *S of size M*



with a probability distribution over its symbols given by

$$P(s = s_m) = p_m \ , \ \mathrm{m} = 1, 2, \ .. \ , \mathrm{M} \quad \textit{and} \quad \sum_{m=1}^{\mathrm{M}} p_m = 1$$

Symbol	s_1	s_2	…	s_M
Probability	p_1	p_2	…	p_M

Question to be answered in this lecture:

- How do we measure the amount of information produced by the source?

Uncertainty, Information, and Entropy

- **Question**: What does the word "**information**" mean?
- There is no exact definition !!!
- Information in a message is meaningful only if the recipient is able to interpret it (For example, A chemist may a explain complex chain of reactions to kinder-garden students or he may present the same work to a group of specialists).
- Information is also about something which adds to your knowledge
- **Motivation for defining information**: Consider the following three sentences

  1) The sun will rise tomorrow from the east. (certain event; none of us will be surprised)

  2) The average grade in this class will be 85 and no one will fail the course (it is unlikely; some of you will be surprised)

  3) No attendance is required in this course, no exams will be given, and all students will receive A (almost improbable event; all of you will be surprised).

Information and Uncertainty

- **Information in a message is a measure of surprise or unpredictability**
  - sentence 1 has low information content (high predictability)
  - sentence 2 has higher information content (less predictable)
  - sentence 3 has even higher information content (it is an unlikely event).
- **Information content of an event is related to the uncertainty of that event**
- **Uncertainty is defined as the inverse of probability**
- **The less expected the event is (smaller probability), the more information it contains.**
- **Shannon's answer is:** The information content of a message is simply the number of 1s and 0s needed to represent it.
- **Hence, the elementary unit of information is a binary unit: a bit**
- One of the basic postulates of information theory is that **information can be treated like a measurable physical quantity** (such as density or length) with units in **bits**

Uncertainty, Information, and Entropy

Two conditions on the information measure

- **First Condition:** The self information of event A may be related to the inverse of P(A)

$$No\ Suprize \Rightarrow No\ Information$$

$$Information\ in\ Event\ (A) \propto \frac{1}{Probability\ of\ (A)}$$

- **Second Condition: If A is a surprise event and B is another independent surprise event, then the total information of a simultaneous event A and B is:**

$$Information\ in\ (A \cap B) = Information\ in\ (A) + Information\ in\ (B)$$

- **The logarithmic function satisfies the above two conditions**

$$I(s_m) = log_2\left(\frac{1}{p_m}\right);\ \text{bits}$$ **Self Information of Symbol s_m**

Properties of Information

$I(s_m) = log_2\left(\frac{1}{p_m}\right)$; Information in each symbol (units in bits)

Symbol	s_1	s_2	…	s_M
Probability	p_1	p_2	…	p_M
Information	$Log_2(1/p_1)$	$Log_2(1/p_2)$	…	$Log_2(1/p_M)$

Log(ab) = Log(a) + Log(b)

1. $I(s_m) = 0$ for $p_m = 1$

2. $I(s_m) \geq 0$ for $0 \leq p_m \leq 1$

3. $I(s_k) > I(s_i)$ for $p_k < p_i$

4. $I(s_k \cap s_i) = I(s_k) + I(s_i)$, if s_k and s_i statist. indep.

$$= log\left(1/P(s_k \cap s_i)\right) = log\left(1/P(s_k)\right) + log\left(1/P(s_i)\right)$$

1: A certain event (p = 1) contains no information (log(1) = 0)

2. Information is nonnegative (since $0 < x < 1$), then $\frac{1}{x} > 1 \Rightarrow log\left(\frac{1}{x}\right) > 0$)

3. The smaller the prob. of an event is, the more information it carries

3. Info in the intersection of two independent events = sum of information

* *Custom is to use logarithm of base 2*

The Average Information per Source Symbol
Source Entropy

- **The average information per source symbol, is the expected value of the random variable *I*.**

$$E(I)=\sum_{i=1}^{M} p_i I_i = \sum_{i=1}^{M} p_i \log_2(1/p_i) \quad \text{bits/symbol}$$

$$E(I)=H(S)=\sum_{i=1}^{M} p_i \log_2(1/p_i) \quad \text{Source Entropy}$$

Symbol	s_1	s_2	…	s_M
Probability	p_1	p_2	…	p_M
Information	$\mathrm{Log}_2(1/p_1)$	$\mathrm{Log}_2(1/p_2)$	…	$\mathrm{Log}_2(1/p_M)$

$$E(I)=\sum_{i=1}^{M} p_i I_i = \sum_{i=1}^{M} p_i \log_2(1/p_i)$$
$$= -\sum_{i=1}^{M} p_i \log_2(p_i)$$

- **This is known as: Entropy of Source S**
- **If all symbols are equally probable p_i = 1/M**

$$H(X)=\sum_{i=1}^{M}\frac{1}{M}\log_2 M = \log_2 M$$

Examples of Entropy Computation

- **Toss a Coin, S = {H, T}, P(H) = P(T) = 0.5**

$$H(S) = -0.5\log_2(0.5) - 0.5\log_2(0.5) = 1 \quad bit/symbol$$

- **Rolling a fair die, S = {1, 2, 3, 4, 5, 6}, P(si) = 1/6**

$$H(S) = -6[\frac{1}{6}\log_2(\frac{1}{6})] = 2.585 \quad bit/symbol$$

- **A biased die, P(1) = 0.9, P(s) = 0.02, s=(2, 3, 4, 5, 6)**

$$H(S) = -0.9\log_2 0.9 - 5[0.02\log_2 0.02] = 0.701 \quad bit/symbol$$

$$\mathbf{H(S)} = -\sum_{i=1}^{M} \boldsymbol{p_i log_2(p_i)}$$

- **Note that the entropy of the fair die is higher than that of the biased die. Why?**
- **The fair die has higher uncertainty than the biased one; hence higher entropy**

Average Information Content in English Language

Example 1: Calculate the average information in bits/character in English assuming each letter is equally likely

$$\boldsymbol{H(S) = \sum_{i=1}^{M} p_i log_2(1/p_i) = -\sum_{i=1}^{M} p_i log_2(p_i)}$$

$$H = -\sum_{i=1}^{26} \frac{1}{26} \log_2\left(\frac{1}{26}\right)$$

$$= 4.7\, bit\, s/c\, har$$

Average Information Content in English Language

Example 2: Calculate the average information in bits/character in English.

Since characters do not appear with the same frequency, we may use the following approximate probabilities

- ***P = 0.10*** **for a, e, o, t**
- ***P = 0.07*** **for h, i, n, r, s**
- ***P = 0.02*** **for c ,d ,f ,l, m, p, u, y**
- ***P = 0.01*** **for b, g, j, k, q, v, w, x, z**

$$H(S) = \sum_{i=1}^{M} p_i \log_2(1/p_i) = -\sum_{i=1}^{M} p_i \log_2(p_i)$$

$$H = -\begin{bmatrix} 4 \times 0.1\log_2(0.1) + 5 \times 0.07\log_2(0.07) \\ +8 \times 0.02\log_2(0.02) + 9 \times 0.01\log_2(0.01) \end{bmatrix}$$
$$= 4.17\ bits|character$$

The Source Entropy

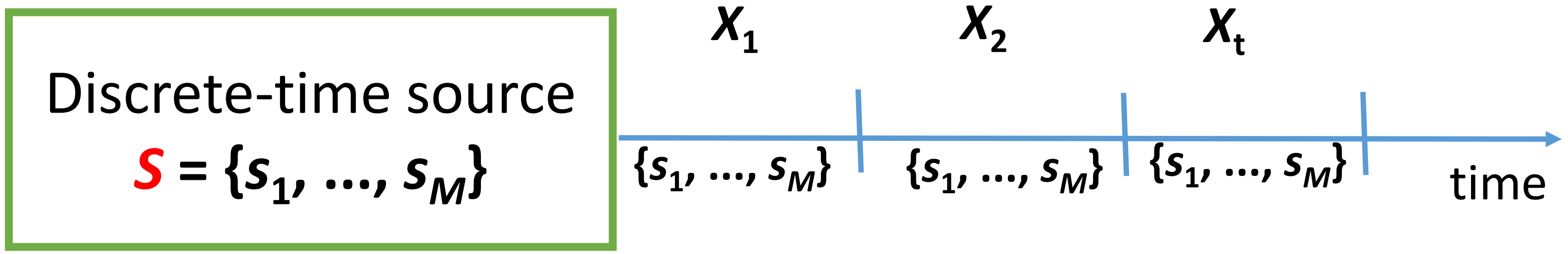
Lecture Outline

Lecture 6

- Define the source entropy
- Study the entropy of the binary source
- Prove that: $0 \leq H(S) \leq \log_2 M$

The Source Entropy

- **Main Theme**: Consider a discrete-time finite-alphabet source *S of size M*



with a probability distribution over its symbols given by

$$P(s = s_m) = p_m \;,\; \mathrm{m} = 1, 2, \,..\,, \mathrm{M} \quad and \quad \sum_{m=1}^{\mathrm{M}} p_m = 1$$

- **The information content of each symbol is**
- $I(s_m) = log_2\left(\frac{1}{p_m}\right)$; bits

Symbol	s_1	s_2	…	s_M
Probability	p_1	p_2	…	p_M
Information	$\text{Log}_2(1/p_1)$	$\text{Log}_2(1/p_2)$	…	$\text{Log}_2(1/p_M)$

The Average Information per Source Symbol
Source Entropy

- **The average information per source symbol, is the expected value of the random variable *I*.**

$$E(I) = \sum_{i=1}^{M} p_i I_i = \sum_{i=1}^{M} p_i \log_2(1/p_i) \quad \text{bits/symbol}$$

$$E(I) = H(S) = \sum_{i=1}^{M} p_i \log_2(1/p_i) \quad \text{Source Entropy}$$

Symbol	s_1	s_2	…	s_M
Probability	p_1	p_2	…	p_M
Information I	$\mathrm{Log}_2(1/p_1)$	$\mathrm{Log}_2(1/p_2)$	…	$\mathrm{Log}_2(1/p_M)$

- **This is known as: Entropy of Source S**
- **If all symbols are equally probable $p_i = 1/M$**

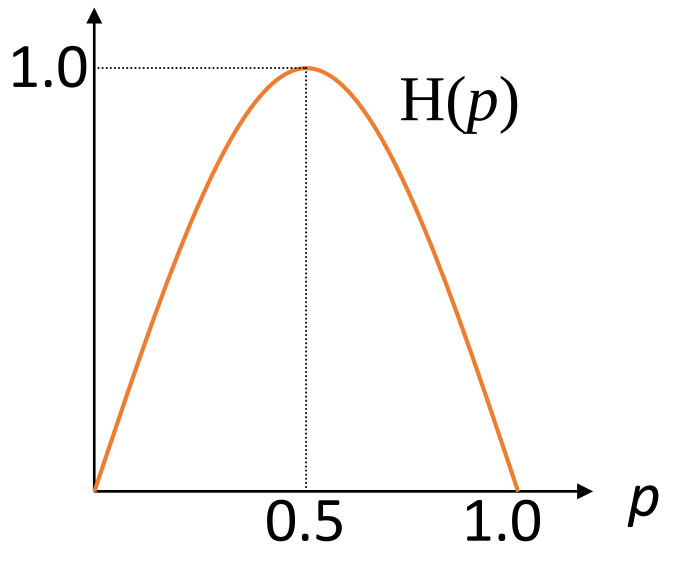
$$H(X) = \sum_{i=1}^{M} \frac{1}{M} \log_2 M = \log_2 M$$

Entropy is interpreted as:

- Measure of information in the source
- Measure of uncertainty in the source

Entropy of the Random Binary Source

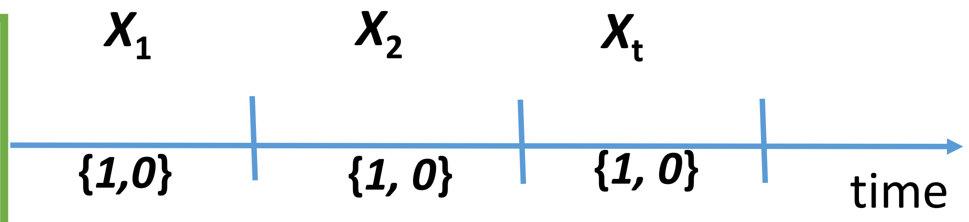
- Consider a random binary source S with probability assignment over its symbols as: P(S=1) = p, P(S=0)= 1-p. The entropy of the source is:
- $\boldsymbol{H(p) = -p\log_2 p - (1-p)\log_2(1-p)}$ ***bits/symbol***
- The binary entropy as a function of p is plotted below
- Note: $\lim_{p\to 0}(p)\log(p) = \lim_{p\to 1}(p)\log(p) = 0$; VERIFY



(binary entropy function)

- H(p) = 0 at p = 0 and at p = 1 (one event is certain)
- H(p) is maximum **(=1)** when p = ½ (symbols are equally probable, and hence uncertainty is maximum

Discrete-time source
S = {1, 0}



Properties of the Entropy Function

Lemma: For an *M*-ary information source *S*,

$$0 \le H(S) \le \log_2 M$$

- min $H(S) = 0$ (one symbol occurs with prob. 1, the others with 0)
- max $H(S) = \log_2 M$ (when all symbols are equally likely, i.e., when $P(s_i = \frac{1}{M})$
- **Proof** : min $H(S) = 0$.
- When one probability = 1 and the rest are zeros, we can make use of the limits: $\boldsymbol{lim_{p\to 0}(p)log(p) = lim_{p\to 1}(p)log(p) = 0}$

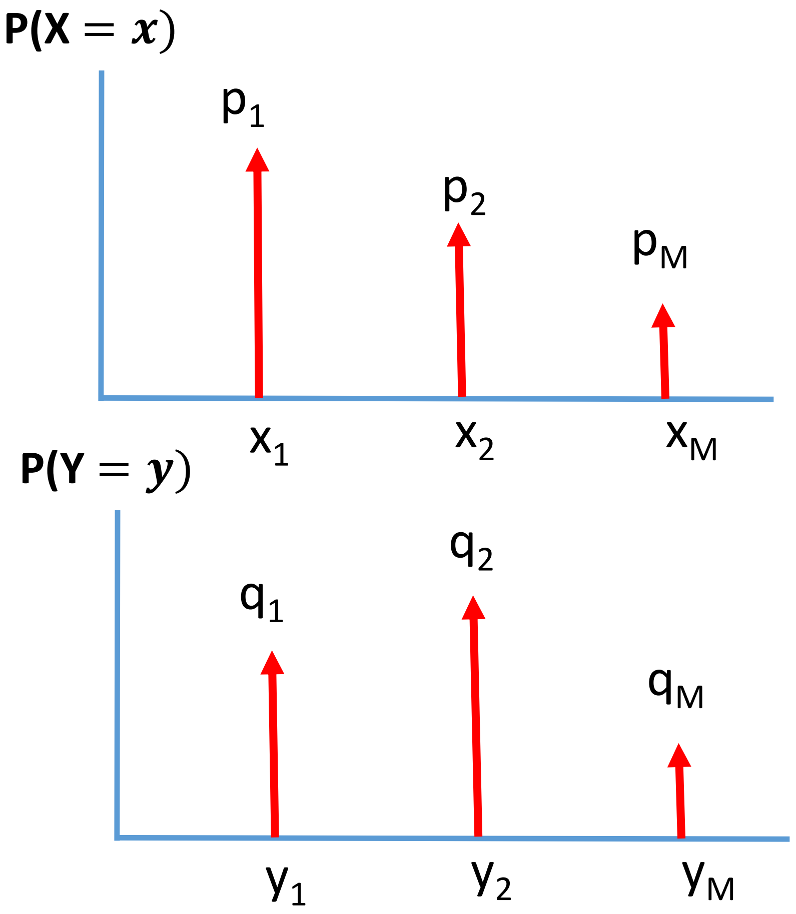
$$H(S) = \sum_{i=1}^{M} p_i \log_2(1/p_i) = -\sum_{i=1}^{M} p_i \log_2(p_i)$$

Properties of the Entropy Function

- **Here, we show that entropy is maximum when source probabilities are equal (p_i = 1/M)**

  We prove that in two steps:

- Define the **relative entropy** D(X, Y) between two distributions X and Y as
- $D(X,Y) = \sum_{j=1}^{M} p_j \log\left(\frac{p_j}{q_j}\right)$
- **First Step, we show that $D(X,Y) \geq 0$**
- **X is a random variable with distribution p_j (the given pmf)**
- **Y is a reference random variable with distribution q_j**
- **Rewrite D(X,Y) as:**
- $D(X,Y) = \sum_{j=1}^{M} p_j \log\left(\frac{p_j}{q_j}\right) = -\sum_{j=1}^{M} p_j \log\left(\frac{q_j}{p_j}\right)$
- $-D(X,Y) = \sum_{j=1}^{M} p_j \log\left(\frac{q_j}{p_j}\right)$

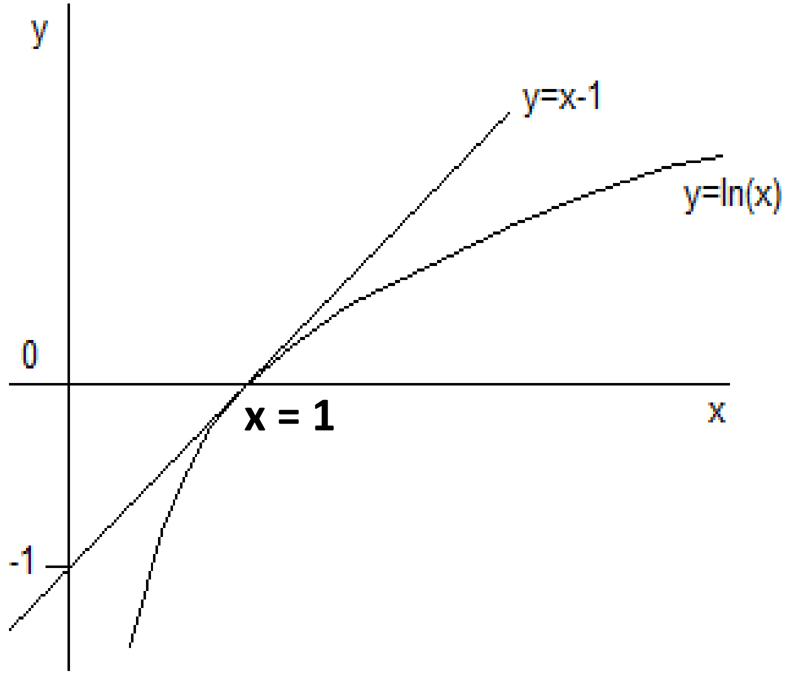


Properties of the Entropy Function

Since log(x) ≤ (x - 1) we have:

$$-D(X,Y) = \sum_{j=1}^{M} p_j \log\left(\frac{q_j}{p_j}\right) \leq \sum_{j=1}^{M} p_j \left(\frac{q_j}{p_j} - 1\right)$$

$$\leq \sum_{j=1}^{M} q_j - \sum_{j=1}^{M} p_j = (1) - (1) = 0$$



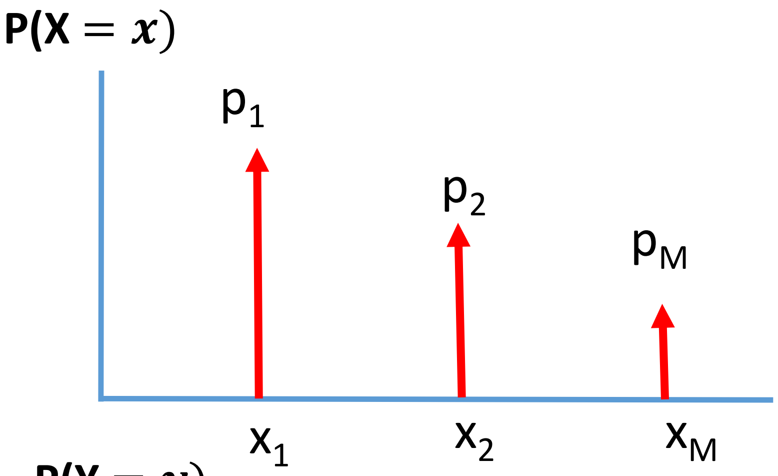
- $-D(X,Y) \leq 0$
- **Therefore $D(X,Y) \geq 0$**
- **Equality (i.e., $D(X,Y) = 0$) when $q_j = p_j$.**
- **This is the first step in the proof**

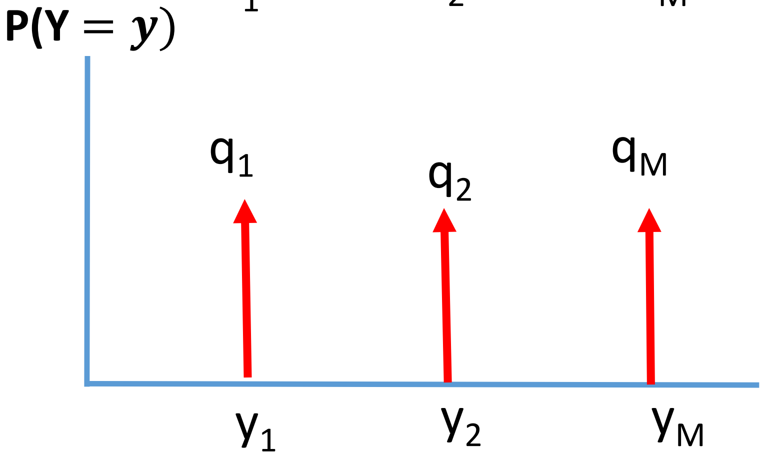
Properties of the Entropy Function

- **Second step: Now let Y be a uniform distribution , then q_j = 1/M since j ranges from 1 to M.**

$$D(X,Y) = \sum_{j=1}^{M} p_j \log\left(\frac{p_j}{q_j}\right) = \sum_{j=1}^{M} p_j \log p_j - \sum_{j=1}^{M} p_j \log q_j$$

$$= -H(X) - \sum_{j=1}^{M} p_j \log(1/M) = -H(X) - \log(1/M) \sum_{j=1}^{M} p_j$$





- $D(X,Y) = \log(M) - H(X) \geq 0$
- **Note that:** $\sum_j p_j = 1$
- **Therefore, since D(X,Y) ≥ 0 ,** $\boldsymbol{H(X) \leq \log(M)}$

Entropy of a Discrete Memory-less Source
Lecture Outline

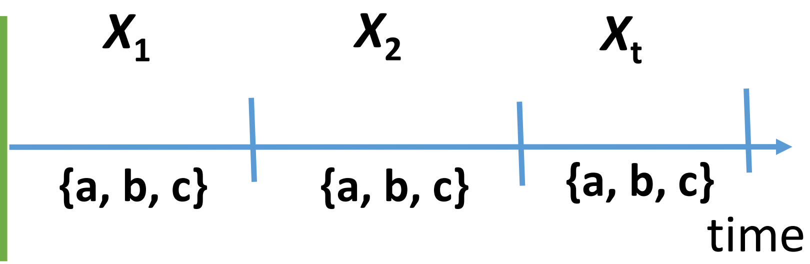
Lecture 7

- **Find the entropy of a discrete memory-less source (DMC)**
- **Define the n’th order extension of a DMS information source.**
- **Evaluate the first, second,… and n’th order entropies of a DMS**
- **Find the relationship between the entropy per symbol and the entropy per message.**

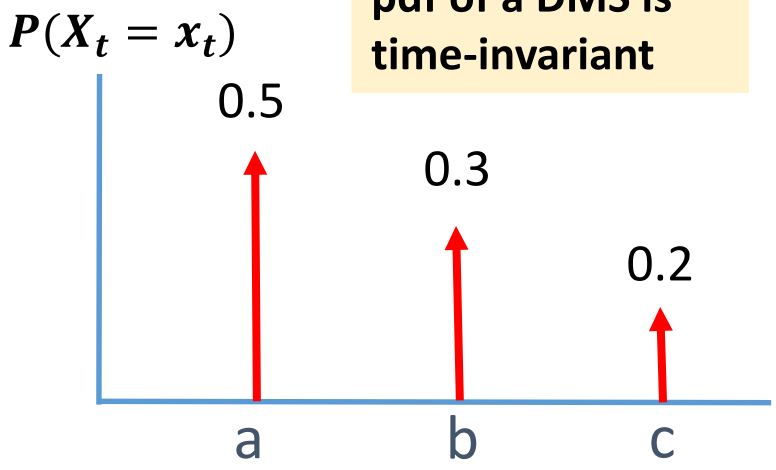
Discrete-time Information Sources

- **Assumptions on the source model:**
  - **Discrete**: the set of possible symbols **S** is finite and countable.
  - **Discrete-time**: The source generates one symbol from the set ***S* = {*a*$_1$, ..., *a*$_M$}** each time unit.
  - A memoryless and stationary information source satisfies the independence condition:
- **Two models:**
  - **Discrete memoryless sources**: $P(X_t = x_t | X_{t-1} = x_{t-1}, \dots . X_2 = x_2, X_1 = x_1) = P(X_t = x_t)$
  - **Sources with memory:** $P(X_t = x_t | X_{t-1} = x_{t-1}, \dots . X_2 = x_2, X_1 = x_1) \neq P(X_t = x_t)$; **Markov Sources**
- **For a DMS source, the probability distribution is time-independent**
- **The random variables $X_1, X_2, \dots, X_{t-1}, X_t$ are independent**

Discrete-time source
S **= {a, b, c}**

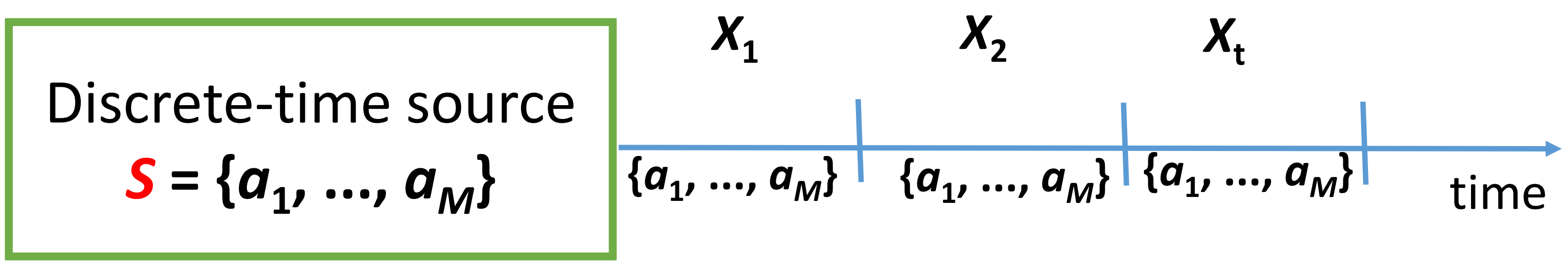


pdf of a DMS is time-invariant



The Source Entropy

- **Main Theme**: Consider a discrete-time finite-alphabet source *S of size M*



with a probability distribution over its symbols given by

$$P(s = a_m) = p_m \ , \ \mathrm{m} = 1, 2, \ .. \ , \mathrm{M} \quad and \quad \sum_{m=1}^{\mathrm{M}} p_m = 1$$

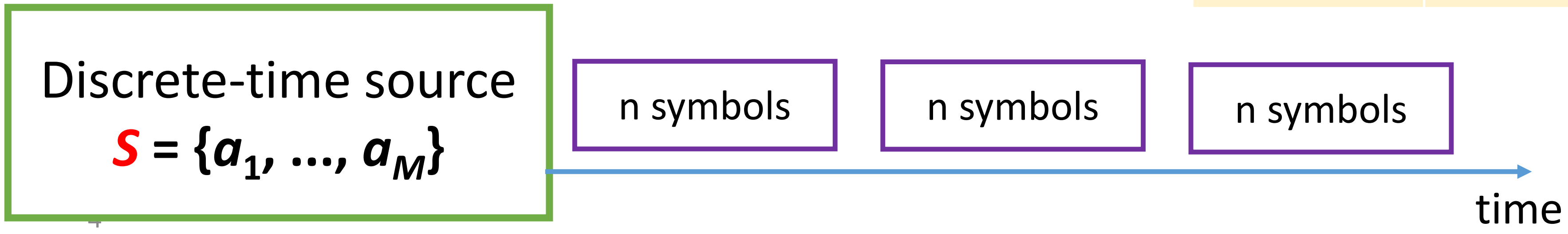
- **The information content of each symbol is**
- $I(s_m) = log_2\left(\frac{1}{p_m}\right)$; bits

Symbol	a_1	a_2	…	a_M
Probability	p_1	p_2	…	p_M
Information	$\mathrm{Log}_2(1/p_1)$	$\mathrm{Log}_2(1/p_2)$	…	$\mathrm{Log}_2(1/p_M)$

The Average Information per Source Symbol
Source Entropy

- The entropy of S is given as:
- $H(S) = \sum_{i=1}^{M} -p_i \log_2 p_i \qquad \text{(bit/symbol)}$
- So far, we have two interpretation for the entropy
  a. The average amount of information in the source
  b. It is a measure of uncertainty in the source

Symbol	s_1	s_2	…	s_M
Probability	p_1	p_2	…	p_M
Information	$\text{Log}_2(1/p_1)$	$\text{Log}_2(1/p_2)$	…	$\text{Log}_2(1/p_M)$



Extension of Information Sources

- Consider a source S with symbol probability distribution
$$P(a_i) = p_i; i = 1, 2, \dots, M$$
- The n'th order extension of the source, denoted S^n, consists of messages of n-symbols drawn from S.
- Any message $m_j = \{x_1, x_2, \dots, x_n\};\ j = 1, 2, 3, \dots, M^n;\ x_k = \{a_1, a_2, \dots, M\}$
- The probability of any message m_j is:

n symbols
$\{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$

- $\mathrm{P}(m_j) = P\{x_1, x_2, \dots, x_n\}; = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \dots P(x_n|x_1, \dots, x_{n-1})$
- $\mathrm{P}(m_j) = P\{x_1, x_2, \dots, x_n\}; = P(x_1)P(x_2)P(x_3)P(x_n);$ *For a DMS*

Below, is an example of a second order extension (Here, the message consists of two symbols)

A **message** of S^n is a **block of *n* symbols**

$\boxed{S} \longrightarrow$ 1 0 0 1 0 0 0 1 1 1

$\boxed{S^2} \longrightarrow$ 10 01 00 01 11

M = {0, 1}; the original alphabet.

***M*²** = {00, 01, 10, 11}; extended alphabet or number of possible messages (4).

Entropy per source symbol and entropy per message

- Consider a source S with symbol probability distribution

$$P(a_i) = p_i; i = 1, 2, \dots, M$$

- The source entropy is $\mathrm{H(S)} = -\sum_{i=1}^{M} p_i log p_i$ bits/symbol
- If a message m_j consists of n symbols, then the entropy of the extended source S^n is:

$$\boldsymbol{H(S^n) = -\sum_{i=1}^{M^n} P_j log P_j} \quad \text{bits/message}$$

$$\mathrm{P}(m_j) = P\{x_1, x_2, \dots, x_n\}$$

We need to find the relationship between H(S) and H(S^n) for both of

- **Discrete memoryless sources (DMS)**
- **Markov sources**

$\boldsymbol{S} \longrightarrow \mathrm{H(S)}$

$\boldsymbol{S}^{\mathrm{M}} \longrightarrow \boldsymbol{H(S^n)}$

First and Second Order Entropies of a DMS

- Example: Consider a DMS, S, which emits either a 1 or a 0 with the following probability: $P(0)$=0.8, $P(1)$=0.2.
- Find H(S) and H(S^2)
- Note that for a DMS: $P(x_1 x_2) = P(x_1)P(x_2)$; Statistical Independence

S

0	0.8
1	0.2

First Order Entropy

$H(S)$= –0.8log0.8 – 0.2log0.2 = **0.72 bits/symbol**

Second Order Entropy

S^2

00	0.64
01	0.16
10	0.16
11	0.04

$H(S^2)$= –0.64log0.64 – 0.16log0.16 –0.16log0.16 – 0.04log0.04 = 1.44 **bit/message**

$H(S^2) = 2\,H(S);$

For DMS

$H(S^n)$ = n $H(S)$

- **Information in one message = twice the information in one symbol.**
- **Amount of uncertainty in one message = twice the amount of uncertainty in one symbol**

$$Entropy\ per\ message = n(Entropy\ per\ symbol)$$

$$Entropy\ per\ symbol = \frac{Entropy\ per\ message}{n}$$

Proof for the Entropy of a DMS

Theorem: If *S* is a discrete memory-less and stationary source, then $H(S^n) = nH(S)$.

Sketch of the proof, for the case $n = 2$

Memoryless (i.e., independence)
$P(x_0, x_1) = P(x_0)P(x_1)$

$\log P(x_0)P(x_1) = \log P(x_0) + \log P(x_1)$

$$H_1(S^2) = -\sum_{x_0 \in M}\sum_{x_1 \in M} P(x_0, x_1)\log P(x_0, x_1)$$

$$= -\sum_{x_0}\sum_{x_1} P(x_0)P(x_1)\log P(x_0)P(x_1)$$

$$= -\sum_{x_0}\sum_{x_1} P(x_0)P(x_1)\log P(x_0) - \sum_{x_0}\sum_{x_1} P(x_0)P(x_1)\log P(x_1)$$

$$= -\sum_{x_0} P(x_0)\log P(x_0)\sum_{x_1} P(x_1) - \sum_{x_1} P(x_1)\log P(x_1)\sum_{x_0} P(x_0)$$

$$= -\sum_{x_0} P(x_0)\log P(x_0) - \sum_{x_1} P(x_1)\log P(x_1)$$

$$= H_1(S) + H_1(S)$$

$$H_1(S^2) = 2H_1(S)$$

the sum of $P(x_0)$ is 1

Entropy in a message of n symbols = n*Entropy of one Symbol

Entropy of a DMS

Summary

For the n’th order extension source (S^n), of a DMS (S),

$$\mathrm{P}\left(m_j\right) = P\{x_1, x_2, \dots, x_n\}; = P(x_1)P(x_1)\dots P(x_1)$$

$$H(S^n) = nH(S)$$

$$H(S) = \frac{H(S^n)}{n}$$

$$H(S) = constant\ independent\ of\ n.$$

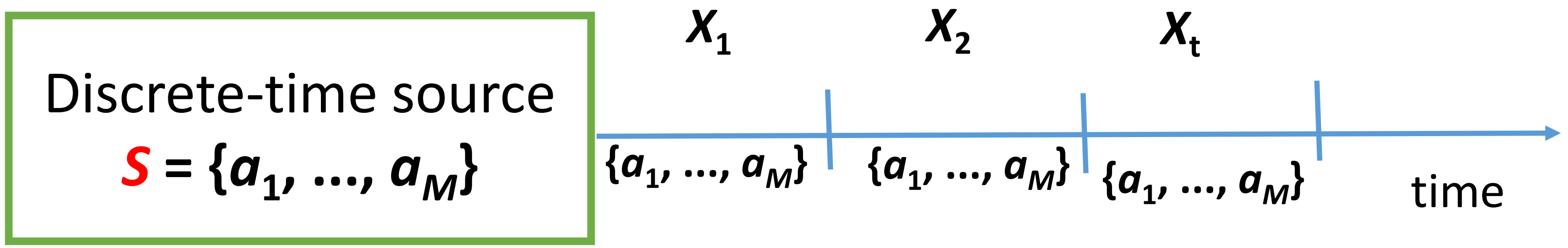
Entropy of a Simple Markov Source
Lecture Outline

Lecture 8

- **Find the first order entropy of a simple Markov source.**
- **Define the n'th extension of a Markov information source.**
- **Find the Entropy per source symbol and the entropy per message.**
- **Evaluate the first, second,… and n'th order entropies.**
- **Find the average (expected value) of the entropy.**

Discrete Memory-less Sources

- **Memoryless property**: $P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots . X_2 = x_2, X_1 = x_1) = P(X_t = x_t)$
- **For a DMS source, the probability distribution is time-independent**
- **The random variables $X_1, X_2, \ldots, X_{t-1}, X_t$ are independent**
- $P(X_2 = x_2 | X_1 = x_1) = P(X_2 = x_2)$; **independent source**
- $P(X_2 = x_2 \cap X_1 = x_1) = P(X_2 = x_2) P(X_1 = x_1)$
- And, in general, for an independent source we have:
- $P(X_t = x_t \cap \cdots \cap X_2 = x_2 \cap X_1 = x_1) = P(X_t = x_t) \ldots P(X_2 = x_2) P(X_1 = x_1)$



The Average Information per Source Symbol
Source Entropy

- **The entropy of S is given as:**
- $\mathrm{H}(S) = \sum_{i=1}^{M} -p_i \log_2 p_i$ $\quad$ **(bit/symbol)**
- **So far, we have two interpretation for the entropy**
  a. **The average amount of information in the source**
  b. **It is a measure of uncertainty in the source**
- **Information/message= n*information/symbol**

Symbol	s_1	s_2	...	s_M
Probability	p_1	p_2	...	p_M
Information	$\text{Log}_2(1/p_1)$	$\text{Log}_2(1/p_2)$	...	$\text{Log}_2(1/p_M)$

$$I(s_m) = log_2\left(\frac{1}{p_m}\right);$$

Discrete-time source
$S = \{a_1, \ldots, a_M\}$

1 symbol	1 symbol	1 symbol
n symbols	n symbols	n symbols

time

Entropy per symbol and entropy per message

Summary

For the n’th order extension source (S^n), of a DMS (S),

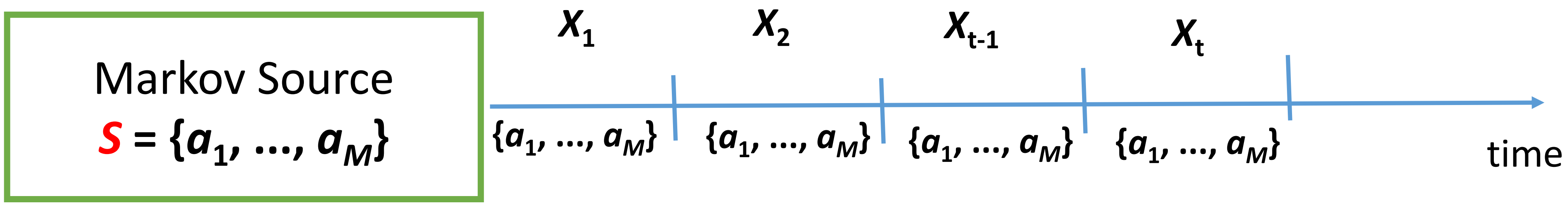
$$\mathrm{P}\left(m_j\right) = P\{x_1 \cap x_2 \cap \cdots \cap x_n\}; = P(x_1)P(x_1) \ldots P(x_1)$$

$$H(S^n) = nH(S)$$

$$\boldsymbol{H(S) = \frac{H(S^n)}{n}}; \; constant\ independent\ of\ n.$$

Sources with Memory: Markov Information Sources

- Used to model information sources with memory.
- In a **simple Markov source**, the occurrence of the current symbol at time **t** depends only on the **previous symbol at time t-1**
- **For a simple Markov source,**
- $P(X_t = x_t | X_{t-1} = x_{t-1}, \dots . X_2 = x_2, X_1 = x_1) = P(X_t = x_t | X_{t-1} = x_{t-1})$

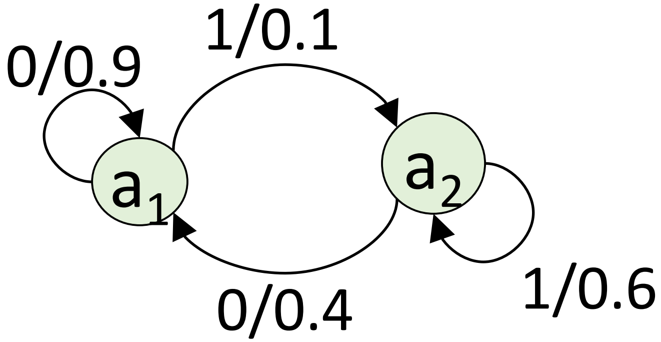


Ergodic (Regular) Markov Process

Definition: A finite-state Markov chain is ergodic (regular) if all states are accessible from all other states and if all states are aperiodic, i.e., have period 1.

An important fact about ergodic Markov chains is that the chain has steady-state probabilities p(s) for all states.

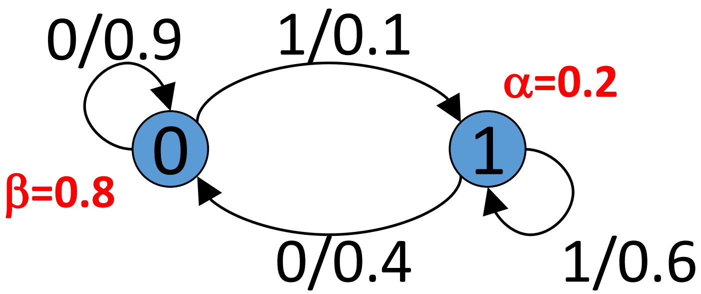
$$P(X_{t-1} = a_j) = P(X_t = a_j) = P(a_j); \text{ for all states } j$$



First and Second Order Entropy of a Markov Source

Consider a Markov source with two states as shown in the figure. It can be shown that the steady-state probabilities are:

$$\boldsymbol{P(X_t = 0), P(X_t = 1)) = (0.8, 0.2)}\textbf{; steady-state probabilities (verify)}$$



$\mathrm{H}(S) = \sum_{i=1}^{M} -p_i \log_2 p_i$

0	0.8·0.9 + 0.2·0.4 = 0.80
1	0.8·0.1 + 0.2·0.6 = 0.20
00	0.8·0.9·0.9 + 0.2·0.4·0.9 = 0.72
01	0.8·0.9·0.1 + 0.2·0.4·0.1 = 0.08
10	0.8·0.1·0.4 + 0.2·0.6·0.4 = 0.08
11	0.8·0.1·0.6 + 0.2·0.6·0.6 = 0.12

First order entropy

$H_1(S)$= –0.8log0.8 – 0.2log0.2

= 0.72 bits/symbol

$H_1(S^2)$ = –0.72log0.72 – 0.08log0.08 – 0.08log0.08 – 0.12log0.12

= 1.2914

$H_1(S^2)$ = 1.2914

$H_2(S)$ = $H_1(S^2)/2$ = 0.6457

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2)$$

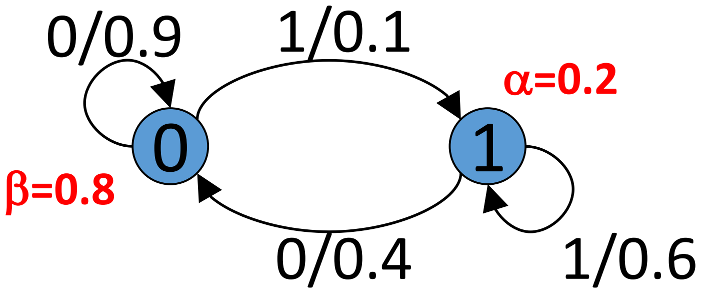
β= 0.4α + 0.9β

α= 0.6α + 0.1β;

(β= 0.8, α =0.2);

Steady State Equations

First and Second Order Entropy of a Markov Source



define the *n*-th order entropy of *S*

$$H_n(S) = \frac{H_1(S^n)}{n} = \frac{Entropy\ of\ a\ message}{number\ of\ symbols\ in\ message}$$

$$Entropy\ H = \lim_{n\to\infty} H_1(S^n)/n$$

$$H(S) = P(S_A)H(S/S_A) + P(S_B)H(S/S_B)$$

First Order Entropy

$H_1(S)$ **= 0.72 bits/symbol**

Second Order Entropy

$H(S^2)$ **= 1.2914/2 = 0.6457**

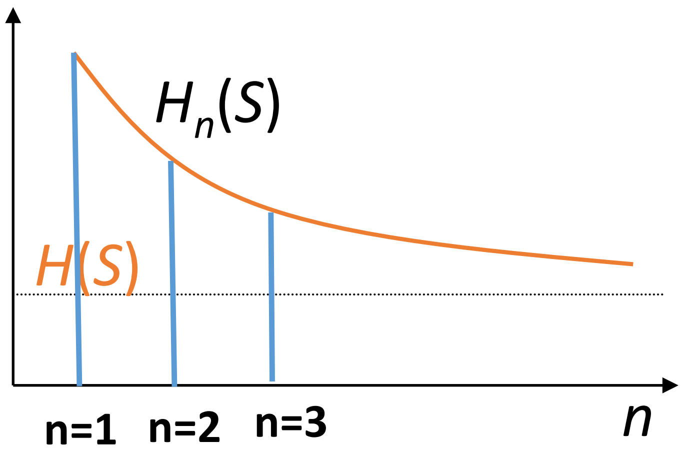
What happens whe n=3? What is $H_3(S)$?

$$H_1(S) > H_2(S) > H_3(S) > H_4(S) > \cdots > \mathrm{LIMIT}$$

The Entropy of Markov Sources

- For a Markov source, we have $H_1(S) > H_2(S) > \ldots H(S)$ (limit entropy)
- **Theorem:**

  The n-th order entropy approaches the limit entropy $H(S)$



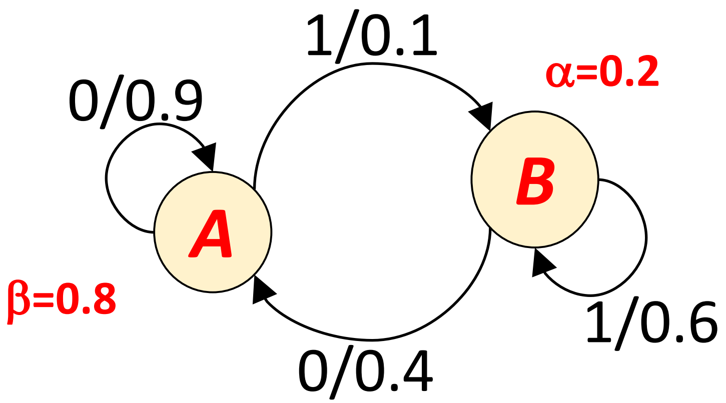
How to compute the limit entropy H(S) of a Markov source:

1. Determine the stationary probabilities of the states
2. Identify the outgoing probability of each state.
3. Compute entropies of each state (using those of Part 2)
4. Determine the weighted average of the state entropies.

$$H(S) = P(S_A)H(S/S_A) + P(S_B)H(S/S_B)$$

Example: Entropy of A Markov Source

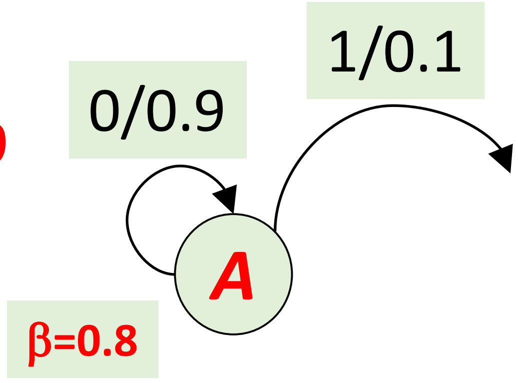
Consider the Markov source in the figure. Earlier, it was found that the stationary probabilities are $(\beta , \alpha) = (0.8, 0.2)$



When in state A, source emits 0 and 1 with probabilities: {P(0)=0.9, P(1)=0.1}

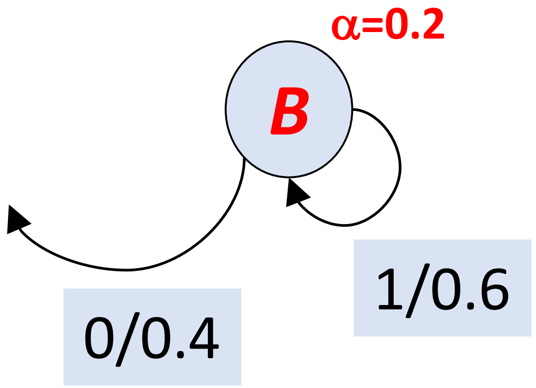
The source entropy is:

$$H(S/S_A) = -0.9\log 0.9 - 0.1\log 0.1 = 0.469$$



When in state B, source emits 0 and 1 with probabilities: {P(0)=0.4, P(1)=0.6}. The source entropy is

$$H(S/S_B) = -0.4\log 0.4 - 0.6\log 0.6 = 0.971$$



The expected value (mean value of the entropy)

$$H(S) = P(S_A)H(S/S_A) + P(S_B)H(S/S_B)$$

$$H(S) = 0.8 \times 0.469 \; + \; 0.2 \times 0.971 = 0.5694 \; bit/symbol$$