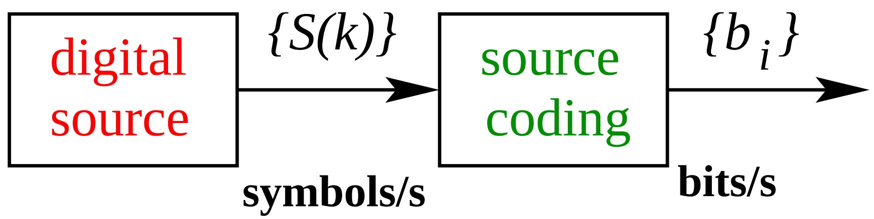
Revision of Lecture 2

- Source is defined by



1. Symbol set: $\mathcal{S} = \{m_i, 1 \le i \le q\}$
2. Probability of occurring of m_i: p_i, $1 \le i \le q$
3. Symbol rate: R_s [symbols/s]
4. Interdependency of $\{S(k)\}$ (memory or memoryless source)

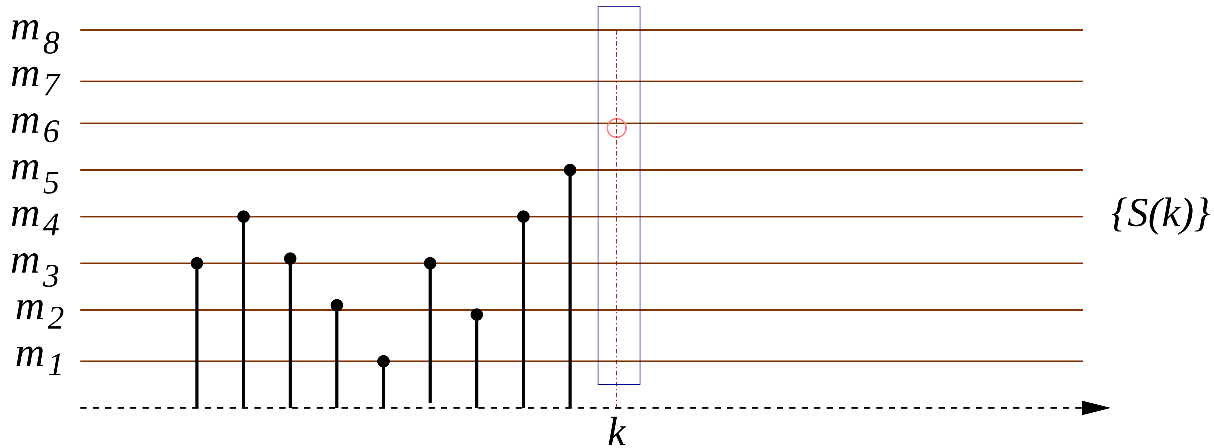
- We have completed discussion on memoryless source
  - Entropy $\surd$
  - Information rate $\surd$
  - Efficient coding (entropy encoding) $\surd$

- For source with memory
  - Entropy ?
  - Information rate ?
  - How to code ?

- Question: Two sources have same 1., 2. and 3. but one is memoryless, another has memory – Which has larger entropy/information rate ?

Sources with Memory

- Most real world sources exhibit memory, resulting in *correlated source signals*; this property is retained during sampling and quantisation
  - This implies that the signal exhibits some form of *redundancy*, which should be exploited when the signal is coded
  - For example, samples of speech waveform are correlated; redundancy in samples is first removed, as it can be predicted; the resulting residuals, almost memoryless or uncorrelated, can then be coded with far fewer bits



- $S(k) \in \{m_i, 1 \le i \le 8\}$, but given $S(k-1) = m_5$, $S(k-2) = m_4, \cdots$, unlikely $S(k) = m_1$ or m_2
  - there exists interdependency between $S(k)$ and previous samples $S(k-j)$, $j \ge 1$

Model Source Memory

- Here memory can be completely modelled by a stochastic probabilistic **Markov process**
  - Consider source with memory that emits a sequence of symbols $\{S(k)\}$ with "time" index k
  - First order Markov process: the current symbol depends only on the previous symbol, $p(S(k)|S(k-1))$
  - N-th order Markov process: the current symbol depends on N previous symbols, $p(S(k)|S(k-1), S(k-2), \cdots, S(k-N))$

- Alternatively, if $S(k)$ is influenced by $S(k-1)$ up to $S(k-N)$, then it may be modelled by **predictive** model

$$S(k) = f(S(k-1), \cdots S(k-N)) + \varepsilon(k)$$

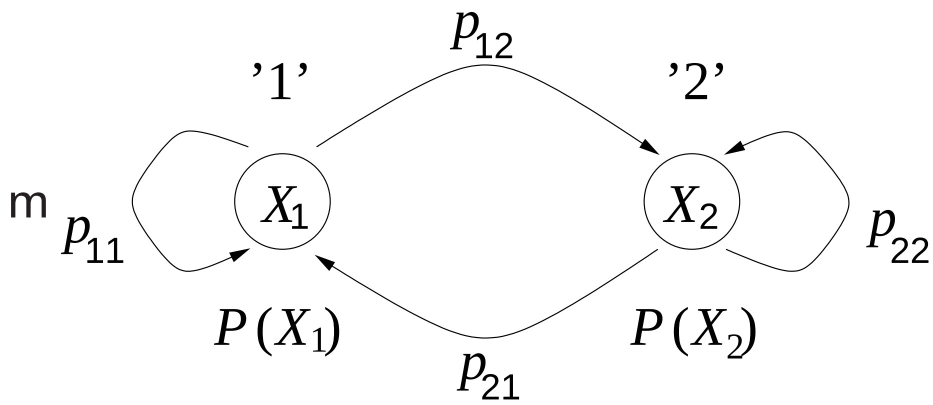
  - Prediction model $f(S(k-1), \cdots S(k-N))$ contains information of $S(k)$ that can be predicted by $S(k-1), \cdots S(k-N)$
  - Innovation $\varepsilon(k)$ contains new information of $S(k)$ that cannot be predicted by $S(k-1), \cdots S(k-N)$

Two-State First Order Markov Process

- Source $S(k)$ can only generate two symbols, $X_1 = 1$ and $X_2 = 2$; their probability explicitly depends on the previous state (i.e. $p(S(k)|S(k-1))$)

  Transition probability matrix

$$\Gamma = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$



- Probabilities of occurrence (prior probabilities) for states X_1 and X_2: $P_1 = P(X_1)$ and $P_2 = P(X_2)$ (i.e. $p(S(0) = 1) = P(X_1)$ and $p(S(0) = 2) = P(X_2)$)

- Transition probabilities: transition probabilities from state X_1 are given by the conditional probabilities $p_{12} = P(X_2|X_1)$ and $p_{11} = P(X_1|X_1) = 1 - P(X_2|X_1)$, etc. (i.e. $p(S(k) = j|S(k-1) = i) = p_{ij}$)

Entropy for 2-State 1st Order Markov Source

- Entropy H_i for state X_i, $i = 1, 2$:

$$H_i = -\sum_{j=1}^{2} p_{ij} \cdot \log_2 p_{ij} = -p_{i1} \cdot \log_2 p_{i1} - p_{i2} \cdot \log_2 p_{i2} \quad \text{(bits/symbol)}$$

  – describes average information carried by the symbols emitted in state X_i

- The overall entropy H includes the probabilities P_1, P_2 of the states X_1, X_2:

$$H = \sum_{i=1}^{2} P_i H_i = -\sum_{i=1}^{2} P_i \sum_{j=1}^{2} p_{ij} \cdot \log_2 p_{ij} \quad \text{(bits/symbol)}$$

  – For a highly correlated source, it is likely to remain in a state rather than to change, and *entropy decreases as correlation increases*

- Information rate $R = R_s \cdot H$ (bits/second)

Entropy for q-State 1st Order Markov Source

- For q-state 1st-order Markov source with q symbols $X_i = i$, $1 \le i \le q$, symbol entropy H_i for state X_i:

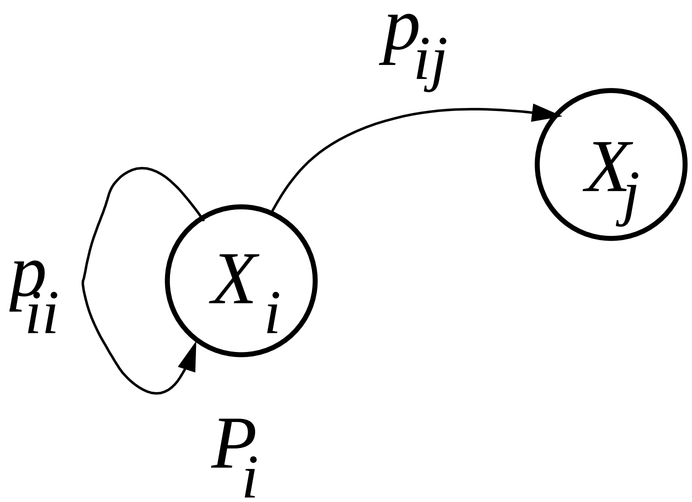
$$H_i = -\sum_{j=1}^{q} p_{ij} \cdot \log_2 p_{ij} \quad \text{(bits/symbol)}$$

where p_{ij} is transition probability from X_i to X_j

- Source entropy is obtained by averaging all symbol entropies with corresponding prior symbol probabilities

$$H = \sum_{i=1}^{q} P_i H_i = -\sum_{i=1}^{q} P_i \sum_{j=1}^{q} p_{ij} \cdot \log_2 p_{ij} \text{ (bits/symbol)}$$

where P_i is the probability of occurrence (prior probability) of state X_i



Transition probability matrix

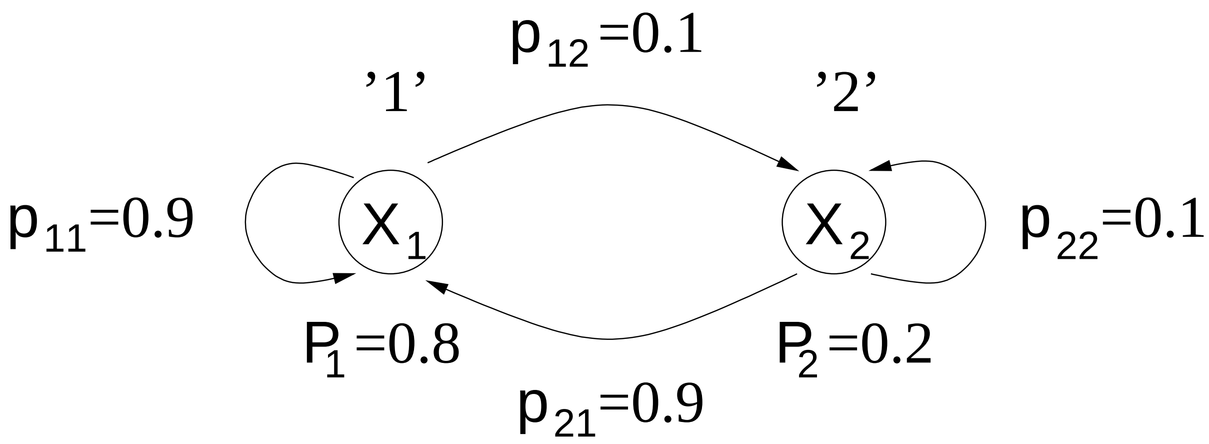
$$\mathbf{\Gamma} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1q} \\ p_{21} & p_{22} & \cdots & p_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ p_{q1} & p_{q2} & \cdots & p_{qq} \end{bmatrix}$$

- With a symbol rate R_s symbols/second, the average source information rate R is

$$R = R_s \cdot H \qquad \text{(bits/second)}$$

A 2-State 1st Order Markov Source – Problem

- Consider the following state diagram with associated probabilities:



- **Q1**: What is the source entropy?

- **Q2**: What is the average information content in message sequences of length 1, 2, and 3 symbols, respectively, constructed from a sequence of X_1 and X_2?

A 2-State 1st Order Markov Source – Solution

- **A1**: The source entropy is given by $H = -0.8 \cdot (0.9 \log_2 0.9 + 0.1 \log_2 0.1)$ $-0.2 \cdot (0.9 \log_2 0.9 + 0.1 \log_2 0.1) = 0.4690$ (bits/symbol)

- **A2** Average information for
  - 1-symbol sequence: $H^{(1)} = -0.8 \log_2 0.8 - 0.2 \log_2 0.2 = 0.7219$ (bits/symbol)
  - 2-symbols sequence: $P('11') = P_1 \cdot p_{11} = 0.72$; $P('12') = P_1 \cdot p_{12} = 0.08$; $P('21') = P_2 \cdot p_{21} = 0.18$; $P('22') = P_2 \cdot p_{22} = 0.02 \longrightarrow$ average 1.190924 bits for 2-symbol sequence, hence $H^{(2)} = 1.190924/2 = 0.5955$ (bits/symbol)
  - 3-symbols sequence: $P('111') = P('11') \cdot p_{11} = 0.648$; $P('112') = P('11') \cdot p_{12} = 0.072$; etc. $\longrightarrow H^{(3)} = 0.5533$ (bits/symbol)

- Consider sequence length of more symbols, which exhibits more memory dependency of the source, and therefore the average information or entropy *decreases*; e.g. $H^{(20)} = 0.4816$ bits/symbol

- In the limit: $H^{(k)} \longrightarrow H$ for message sequence length $k \longrightarrow \infty$

A2 Solution Explained

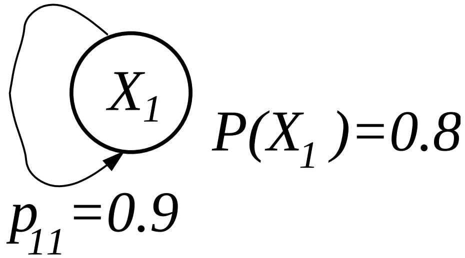
- One-symbol sequences: either "1" or "2" with $P(\text{"1"}) = 0.8$ and $P(\text{"2"}) = 0.2$

  Hence average information content (bits or bits/symbol as it is just one symbol)

$$-P(\text{"1"})\log_2 P(\text{"1"}) - P(\text{"2"})\log_2 P(\text{"2"})$$

$$= -0.8\log_2 0.8 - 0.2\log_2 0.2 = 0.7219 \text{ (bits/symbol)}$$

- Two-symbol sequences: "11", "12", "21" or "22"
  - Consider "11": $P(\text{"11"}) = 0.8 \times 0.9 = 0.72$
  - Average information contents (bits) for 2-symbol sequence:



$$-P(\text{"11"})\log_2 P(\text{"11"}) - P(\text{"12"})\log_2 P(\text{"12"}) - P(\text{"21"})\log_2 P(\text{"21"}) - P(\text{"22"})\log_2 P(\text{"22"})$$

$$= -0.72\log_2 0.72 - 0.08\log_2 0.08 - 0.18\log_2 0.18 - 0.02\log_2 0.02$$

$$= 0.3412304 + 0.2915084 + 0.4453076 + 0.1128771 = 1.1909235 \text{ (bits)}$$

Compare Memory and Memoryless Sources

- Two sources with
  1. **Same** symbol set: $\mathcal{S} = \{m_i, 1 \leq i \leq q\}$
  2. **Same** probability of occurring of m_i: p_i, $1 \leq i \leq q$
  3. **Same** symbol rate: R_s [symbols/s]
  4. One has **memory**, i.e. $\{S(k)\}$ has interdependency; the other is **memoryless**, i.e. $\{S(k)\}$ is independent

- Entropy of memoryless source, $H^{(\mathrm{ml})}$, and entropy of memory source, $H^{(\mathrm{m})}$

$$H^{(\mathrm{ml})} \gg H^{(\mathrm{m})}$$

  – Entropy, a fundamental physical quantity of the source, quantifies average information conveyed per symbol

- Thus, information rate of memoryless source, $R^{(\mathrm{ml})}$, and information rate of memory source, $R^{(\mathrm{m})}$

$$R^{(\mathrm{ml})} \gg R^{(\mathrm{m})}$$

  – Information rate, a fundamental physical quantity of the source, tells you how many bits/s of information the source really needs to send out

How not to Code Memory Source

- For memoryless source, entropy coding allows us to code $\{S(k)\}$ most efficiently
  - Data rate R_b is as small as possible, close to source information rate $R^{(\mathrm{ml})}$

- For source with same 1. symbol set, same 2. set of probabilities of occurrence, and same 3. symbol rate, but has memory, i.e. $\{S(k)\}$ is not independent
  - How should we carry our source coding to convert the symbol sequence $\{S(k)\}$ to the bit sequence $\{b_i\}$?

- Code memory source $\{S(k)\}$ directly by entropy coding ? Really bad idea !
  - Do so you only get "1-symbol-sequence entropy" $H^{(1)}$, i.e. close to "equivalent" memoryless source (with same 1., 2. and 3.) entropy $H^{(\mathrm{ml})} = H^{(1)}$
  - So your data rate R_b gets close to $R_s \cdot H^{(1)}$, but $H^{(1)} \gg H^{(\mathrm{m})}$, i.e. far far larger true source entropy $H^{(\mathrm{m})}$
  - Hence your data rate $R_b \gg R^{(\mathrm{m})} = R_s \cdot H^{(\mathrm{m})}$, i.e. you send at rate far far larger than true source information rate $R^{(\mathrm{m})}$

Comments on Markov Source Model

- Markov process is a most complete model to describe sources with memory; it is a *probabilistic* model

- Most widely used Markov process is 1st order Markov process, where
  - $P_i = P(X_i)$ is probability of occurrence of state X_i; image starting an experiment with time index t, at the beginning or $t = 0$, you can find that the process $S(0)$ starts from state X_i with probability P_i; hence P_i is *a priori* probability
  - Transition probability p_{ij} describes the probability of the process changing from state X_i to X_j, hence is *conditional* probability $p(S(t) = X_j | S(t-1) = X_i) = p_{ij}$

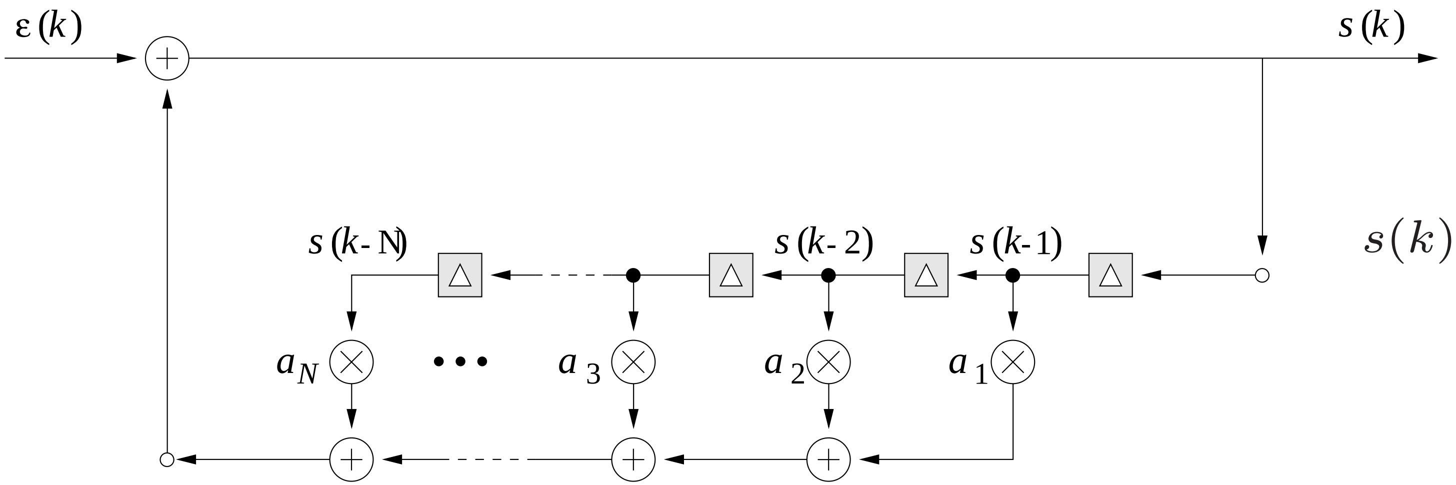
- To describe source with memory longer than 1, higher order Markov process is needed, but this is much more difficult to use
  - In practice, simplified parametric model is often used to describe source with higher-order memory, i.e.
  - Use **conditional mean** $E[s(t)|s(t-1), s(t-2), \cdots, s(t-N)]$ of **realisation** (observation) $s(t)$ to "replace" probabilities of stochastic process $S(t)$

Predictive Models

- An Nth order predictive model with parameter vector $\boldsymbol{a}$:

$$\begin{aligned} s(k) =& E[s(k)|s(k-1), s(k-2), \cdots, s(k-N)] + \varepsilon(k) \\ =& f(s(k-1), s(k-2), \cdots, s(k-N); \boldsymbol{a}) + \varepsilon(k) \end{aligned}$$

- For example, qth order linear autoregressive (AR) model:



$$s(k) = \sum_{j=1}^{N} a_j s(k-j) + \varepsilon(k)$$

  - Aim is to get residual sequence $\{\varepsilon(k)\}$ uncorrelated and zero-mean
  - This parametric model is widely used, for example, in speech source coding (transmit a_j and $\varepsilon(k)$ instead of $s(k)$) – **Why does this?**

Summary

- How to model sources with memory – Markov model and predictive model
  - How to compute entropy and information rate for sources with memory, at least for 1st-order Markov sources

- Most importantly, we know for two sources, with

1. **Same** symbol set: $\mathcal{S} = \{m_i, 1 \le i \le q\}$
2. **Same** probability of occurring of m_i: p_i, $1 \le i \le q$
3. **Same** symbol rate: R_s [symbols/s]
4. One has **memory**; the other is **memoryless**

- Entropy of memoryless source, $H^{(\mathrm{ml})}$, and entropy of memory source, $H^{(\mathrm{m})}$

$$H^{(\mathrm{ml})} \gg H^{(\mathrm{m})}$$

- Information rate of memoryless source, $R^{(\mathrm{ml})}$, and information rate of memory source, $R^{(\mathrm{m})}$

$$R^{(\mathrm{ml})} \gg R^{(\mathrm{m})}$$

- Thus, code memory source $\{S(k)\}$ directly with entropy coding is inefficient