Information and Communication Theory

Lecture 2

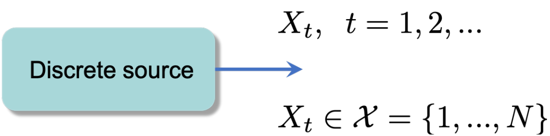
Markov Sources

Mário A. T. Figueiredo

DEEC, Instituto Superior Técnico, University of Lisbon, **Portugal**

2023

Discrete Sources

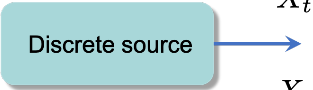


- Memoryless assumption is dropped.
- Sequence of random variables: discrete-time stochastic process.
- Full characterization: for any $L \in \mathbb{N}$ and any $\{t_1, ..., t_L\}$

$$f_{X_{t_1},...,X_{t_L}}(x_1, ..., x_L) = \mathbb{P}(X_{t_1} = x_1, ..., X_{t_L} = x_L)$$

  must be known.
- Without some structure, essentially impossible in general.

Stationary Sources



$X_t, \quad t = 1, 2, ...$

$X_t \in \mathcal{X} = \{1, ..., N\}$

- Stationary source: for any $L \in \mathbb{N}$ and any $\{t_1, ..., t_L\}$,

$$f_{X_{t_1}, ..., X_{t_L}}(x_1, ..., x_L) = f_{X_{t_1+s}, ..., X_{t_L+s}}(x_1, ..., x_L),$$

for any shift $s \in \mathbb{Z}$ such that all $t_1 + s \geq 1$, ..., $t_L + s \geq 1$.

- Example, with $\mathcal{X} = \{a, b, c, d\}$, $L = 3$,

$$f_{X_2, X_5, X_7}(b, c, a) = f_{X_{32}, X_{35}, X_{37}}(b, c, a) = f_{X_1, X_4, X_6}(b, c, a)$$

...in other notation:

$$\begin{aligned} \mathbb{P}(X_2 = b, X_5 = c, X_7 = a) &= \mathbb{P}(X_{32} = b, X_{35} = c, X_{37} = a) \\ &= \mathbb{P}(X_1 = b, X_4 = c, X_6 = a) \end{aligned}$$

Memoryless Sources

$$X_t, \quad t = 1, 2, \ldots$$

$$X_t \in \mathcal{X} = \{1, \ldots, N\}$$

- Memoryless source: for any $L \in \mathbb{L}$ and any $\{t_1, \ldots, t_L\}$,

$$f_{X_{t_1}, \ldots, X_{t_L}}(x_1, \ldots, x_L) = \prod_{i=1}^{L} f_{X_{t_i}}(x_i)$$

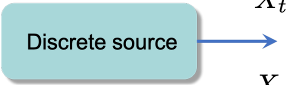
  ...that is, symbols are independent.

- Example: $f_{X_2, X_5, X_7}(b, c, a) = f_{X_2}(b)\, f_{X_5}(c)\, f_{X_7}(a)$
- Memoryless stationary source:

$$f_{X_{t_1}, \ldots, X_{t_L}}(x_1, \ldots, x_L) \overset{\text{(memoryless)}}{=} \prod_{i=1}^{L} f_{X_{t_i}}(x_i) \overset{\text{(stationary)}}{=} \prod_{i=1}^{L} f_{X_1}(x_i)$$

- Example: $f_{X_2, X_5, X_7}(b, c, a) = f_{X_1}(b)\, f_{X_1}(c)\, f_{X_1}(a)$

Markov Sources



$X_t, \quad t = 1, 2, \ldots$

$X_t \in \mathcal{X} = \{1, \ldots, N\}$

- Markov (or Markovian) source: for any $t \in \mathbb{N}$,

$$f_{X_{t+1}|X_t,\ldots,X_1}(x_{t+1}|x_t, \ldots, x_1) = f_{X_{t+1}|X_t}(x_{t+1}|x_t)$$

- In other notation

$$\mathbb{P}\big(X_{t+1} = x_{t+1}|X_t = x_t, \ldots, X_1 = x_1\big) = \mathbb{P}\big(X_{t+1} = x_{t+1}|X_t = x_t\big)$$

- Time-invariant Markov source: for any t and any $a, b \in \mathcal{X}$

$$f_{X_{t+1}|X_t}(b|a) = f_{X_2|X_1}(b|a)$$

- Also required: the initial distribution: $f_{X_1}(x) = \mathbb{P}(X_1 = x)$.

Markov Sources: Probability Transition Matrix

- Time-invariant Markov source: for any t and any $a, b \in \mathcal{X}$

$$f_{X_{t+1}|X_t}(b|a) = f_{X_2|X_1}(b|a) = P_{a,b} \qquad \mathbf{P} = \begin{bmatrix} P_{1,1} & \cdots & P_{1,N} \\ \vdots & \ddots & \vdots \\ P_{N,1} & \cdots & P_{N,N} \end{bmatrix}$$

- Stochastic matrix (a.k.a. Markov matrix):

$$P_{a,b} \geq 0, \text{ for all } a, b \in \{1, ..., N\} \text{ and } \sum_{b=1}^{N} P_{a,b} = 1.$$

- Non-consecutive conditionals: Chapman-Kolmogorov equations,

$$\begin{aligned} f_{X_{t+1}|X_{t-1}}(b|a) &= \sum_{x_t} f_{X_{t+1}|X_t}(b|x_t) f_{X_t|X_{t-1}}(x_t|a) \\ &= \sum_{x_t} P_{a,x_t} P_{x_t,b} = (\mathbf{P}^2)_{a,b} \end{aligned}$$

...generalizing:

$$f_{X_{t+L}|X_t}(b,a) = (\mathbf{P}^L)_{a,b}$$

Higher-Order Markov Sources

- This slide uses a more compact notation: simply $p(\cdot) = f_X(\cdot)$.
- Order-n Markov source: for any $t \in \mathbb{N}$,

$$p(x_{t+1}|\underbrace{x_t, x_{t-1}, ..., x_1}_{\text{all the past}}) = p(x_{t+1}|\underbrace{x_t, ..., x_{t-n+1}}_{n \text{ previous}})$$

- Consider $p(x_{t+1}, \underbrace{x_t, ..., x_{t-n+2}}_{\substack{\text{conditionally} \\ \text{deterministic}}} | x_t, x_{t-1}, ..., x_{t-n+1})$.
- Lifting: defining $z_t = (x_t, x_{t-1}, ..., x_{t-n+1}) \in \mathcal{X}^n$,

$$p(x_{t+1}, x_t, ..., x_{t-n+2}|x_t, x_{t-1}, ..., x_{t-n+1}) = p(z_{t+1}|z_t)$$

  ... the lifted source is order-1 Markov

- Probability transition matrix of the lifted source: $\mathbf{P} \in N^n \times N^n$

Higher-Order Markov Sources

- Example of order-2 Markov source, with $\mathcal{X} = \{1, 2\}$, and the following conditional probabilities

$p(x_{t+1} \mid x_t, x_{t-1})$	x_{t+1}	
(x_t, x_{t-1})	1	2
$(1, 1)$	0.1	0.9
$(1, 2)$	0.6	0.4
$(2, 1)$	0.3	0.7
$(2, 2)$	1	0

- After lifting, $z_t = (x_t, x_{t-1})$

$p(z_{t+1} \mid z_t)$	$z_{t+1} = (x_{t+1}, x_t)$			
$z_t = (x_t, x_{t-1})$	$(1, 1)$	$(1, 2)$	$(2, 1)$	$(2, 2)$
$(1, 1)$	0.1	0	0.9	0
$(1, 2)$	0.6	0	0.4	0
$(2, 1)$	0	0.3	0	0.7
$(2, 2)$	0	1	0	0

- Probability transition matrix of the lifted source: $\mathbf{P} \in 2^2 \times 2^2 = 4 \times 4$.

Markov Models of English

- Uniform distribution over $\mathcal{X} = \{A, B, ..., Z, \_\}$ (N=27)

  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

- Memoryless model w/ estimated probabilities.

  OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

- Order-1 Markov model.

  ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TO BE SEACE

- Order-2 Markov model.

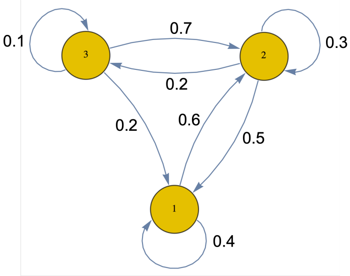
  IN NO IS LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTRURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Markov Sources: Graph Representation

- Consider the probability transition matrix

$$\mathbf{P} = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.7 & 0.1 \end{bmatrix}$$

- ...its graph representation (node = symbol = state) is



Markov Sources: Computing Probabilities

- The pair $(\mathbf{P}, \mathbf{f}_{X_1})$ provide a complete characterization of the source.

- Probability of a sequence of consecutive symbols starting at $t = 1$:

$$f_{X_1,X_2,\ldots,X_L}(x_1, x_2, \ldots, x_L) = f_{X_1}(x_1) P_{x_1,x_2} P_{x_2,x_3} \cdots P_{x_{L-1},x_L}$$

Example: $\mathbb{P}(X_1 = 4, X_2 = 1, X_3 = 8, X_4 = 5) = \mathbb{P}(X_1 = 4) P_{4,1} P_{1,8} P_{8,5}$.

- For non-consecutive symbols, just marginalize. Example:

$$\begin{aligned} f_{X_2,X_5,X_7}(8,3,9) &= \sum_{x_1,x_3,x_4,x_6} f_{X_1,X_2,X_3,X_4,X_5,X_6,X_7}(x_1, 8, x_3, x_4, 3, x_6, 9) \\ &= \sum_{x_1,x_3,x_4,x_6} f_{X_1}(x_1) P_{x_1,8} P_{8,x_3} P_{x_3,x_4} P_{x_4,3} P_{3,x_6} P_{x_6,9} \end{aligned}$$

Markov Sources: Symbol/State Distribution

- Distribution at time $t+1$:

$$f_{X_{t+1}}(x_{t+1}) = \sum_{x_t \in \mathcal{X}} f_{X_{t+1},X_t}(x_{t+1}, x_t) \qquad \text{(marginalization)}$$

$$= \sum_{x_t \in \mathcal{X}} \underbrace{f_{X_{t+1}|X_t}(x_{t+1}|x_t)}_{P_{x_t,x_{t+1}}} f_{X_t}(x_t) \qquad \text{(Bayes)}$$

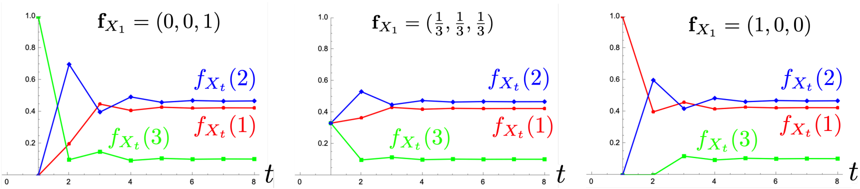
- In matrix notation (recall that $(\mathbf{A}\,\mathbf{v})_j = \sum_i A_{j,i} v_i$)

$$\mathbf{f}_{X_{t+1}} = \begin{bmatrix} f_{X_{t+1}}(1) \\ \vdots \\ f_{X_{t+1}}(N) \end{bmatrix} = \begin{bmatrix} P_{1,1} & \cdots & P_{N,1} \\ \vdots & \ddots & \vdots \\ P_{1,N} & \cdots & P_{N,N} \end{bmatrix} \begin{bmatrix} f_{X_t}(1) \\ \vdots \\ f_{X_t}(N) \end{bmatrix} = \mathbf{P}'\mathbf{f}_{X_t}$$

- Generalizing: $\mathbf{f}_{X_{t+1}} = \underbrace{\mathbf{P}'\mathbf{P}'\cdots\mathbf{P}'}_{t \text{ times}}\mathbf{f}_{X_1} = (\mathbf{P}')^t\,\mathbf{f}_{X_1} = (\mathbf{P}^t)'\,\mathbf{f}_{X_1}$

Markov Sources: Stationary Distribution

- Consider $\mathbf{P}$ from slide 10 and three different initial distributions



- Clearly, the distribution f_{X_t} converges to the same limit
- Stationary distribution: fixed point of its evolution $(\mathbf{f}_{X_{t+1}} = \mathbf{f}_{X_t})$

$$\mathbf{f}_{X_{t+1}} = \mathbf{P}'\mathbf{f}_{X_t} = \mathbf{f}_{X_t} \quad \Leftrightarrow \quad \mathbf{f}_{X_t} \text{ is eigenvector of } \mathbf{P}' \text{ with eigenvalue } 1$$

- Notation: $\boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbf{P}'\boldsymbol{\mu}$
- Example: for the matrix $\mathbf{P}$ in slide 10, $\boldsymbol{\mu} = [49,\ 54,\ 12]^T/115$.

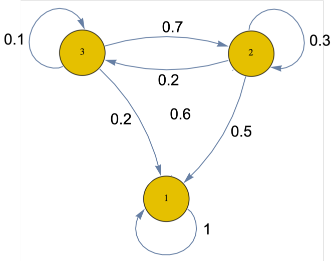
Irreducible and Aperiodic Sources

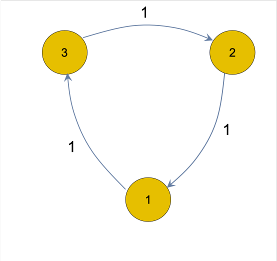
- Irreducible Markov process: for any $x, y \in \mathcal{X}$,

$$\text{there exists } L \in \mathbb{N} \text{ such that } (\mathbf{P}^L)_{x,y} > 0,$$

  ...i.e., it is possible to go from any state to any state, in a finite number of steps, with non-zero probability.

- Aperiodic Markov process: if, for any x, $\gcd\{L : (\mathbf{P}^L)_{x,x} > 0\} = 1$.
- Examples: a non-irreducible and a non-aperiodic source:





Perron-Frobenius Theorem

- If a Markov process is irreducible and aperiodic, then
  - ✓ matrix $\mathbf{P}'$ has a simple eigenvalue 1.
  - ✓ for any initial distribution $\mathbf{f}_{X_1}$,

$$\lim_{t\to\infty} \mathbf{f}_{X_t} = \lim_{t\to\infty} (\mathbf{P}')^t \mathbf{f}_{X_1} = \boldsymbol{\mu}, \text{ where } \boldsymbol{\mu} = \mathbf{P}'\boldsymbol{\mu}$$

- An irreducible and aperiodic source is stationary if and only if $\mathbf{f}_{X_1} = \boldsymbol{\mu}$

Entropy Rate

- Random source/process $X = (X_1, X_2, ..., X_t, ...)$

- The entropy rate is (if the limit exists)

$$H(X) = \lim_{t \to \infty} \frac{H(X_1, X_2, ..., X_t)}{t}$$

- Particular case: stationary memoryless source:

$$H(X) = \lim_{t \to \infty} \frac{H(X_1, X_2, ..., X_t)}{t} = \lim_{t \to \infty} \frac{t\, H(X_1)}{t} = H(X_1)$$

Conditional Entropy Rate

- The conditional entropy rate is (if the limit exists)

$$H'(X) = \lim_{t\to\infty} H(X_t|X_{t-1}, ..., X_1)$$

- Particular case: memoryless (a) and stationary (b) source:

$$H'(X) = \lim_{t\to\infty} H(X_t|X_{t-1}, ..., X_1) \overset{(a)}{=} \lim_{t\to\infty} H(X_t) \overset{(b)}{=} H(X_1)$$

- Time-invariant (b), irreducible, aperiodic Markov (a) source:

$$\begin{aligned} H'(X) &= \lim_{t\to\infty} H(X_t|X_{t-1}, ..., X_1) \overset{(a)}{=} \lim_{t\to\infty} H(X_t|X_{t-1}) \\ &\overset{(b)}{=} \lim_{t\to\infty} \sum_x H(X_2|X_1 = x) f_{X_t}(x) = \sum_x H(X_2|X_1 = x)\mu_x \\ &= -\sum_x \sum_y \mu_x P_{x,y} \log P_{x,y} \end{aligned}$$

Entropy Rates of Stationary Processes

- If X is stationary, $H'(X)$ exists:

  $H(X_t|X_{t-1}, ..., X_2, X_1) \leq H(X_t|X_{t-1}, ..., X_2) = H(X_{t-1}|X_{t-2}, ..., X_1)$

  i.e., $H(X_t|X_{t-1}, ..., X_1)$ is a decreasing non-negative sequence, thus it converges.

- Cesáro mean theorem: $\lim_{t\to\infty} a_t = a \;\Rightarrow\; \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} a_t = a$

- If X is stationary, $H(X) = H'(X)$:

$$\begin{aligned} H(X) &= \lim_{t\to\infty} \frac{H(X_1, X_2, ..., X_t)}{t} \\ &= \lim_{t\to\infty} \frac{1}{t}\sum_{n=1}^{t} H(X_n|X_{n-1}, ..., X_1) && \text{(chain rule)} \\ &= \lim_{t\to\infty} H(X_n|X_{n-1}, ..., X_1) && \text{(Cesáro mean)} \\ &= H'(X) \end{aligned}$$

Markov Models of English: Entropy Rates

- Uniform distribution over $\mathcal{X} = \{A, B, ..., Z, \_\}$ (N=27):
  $H(X) = \log_2 27 \simeq 4.75$ bits/symbol

  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

- Memoryless model w/ estimated prob.: $H(X) \simeq 4.07$ bits/symbol

  OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

- Order-1 Markov model: $H(X) \simeq 3.36$ bits/symbol

  ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TO BE SEACE

- Order-2 Markov model: $H(X) \simeq 2.77$ bits/symbol

  IN NO IS LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTRURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Recommended Reading

- T. Cover and J. Thomas, "Elements of Information Theory", John Wiley & Sons, 2006 (Chapter 4).