Background and examples AIP Part 2 (5SSB0)

Tj.J. Tjalkens

March 23, 2016

Entropy

Entropy quantifies the amount of information contained in a (sequence of) random variables.
It obtains its operational meaning through Shannon's Source coding theorem that states that the entropy (rate) is the achievable best possible compression of the data.

Formulas

- Definition.

$$H(X) = H(P_X) = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{1}{P_X(x)}.$$

- Joint entropy:

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y) \log_2 \frac{1}{P_{X,Y}(x, y)}.$$

- Conditional entropy on fixed condition:

$$H(X|Y=y) = \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \log_2 \frac{1}{P_{X|Y}(x|y)}.$$

- Conditional entropy:

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y=y) \\ &= \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{X,Y}(x,y) \log_2 \frac{1}{P_{X|Y}(x|y)}. \end{aligned}$$

- Chain rule of entropies:

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

Markov sources

Markov models (or sources) are advanced probabilistic data models that can capture real world data dependancies (and memory) quite well.

Introduction

Let $\{X_i\}_{i=-\infty}^{\infty}$ be a Markov source.

- It is described by a finite set of states, $\mathcal{S} = \{1, 2, \cdots, J\}$,
- and a source letter alphabet, $\mathcal{X} = \{a_1, a_2, \cdots, a_K\}$.
- Furthermore for every state $s \in \mathcal{S}$ we have a probability vector for the source letters, $(P(x|s) \stackrel{\Delta}{=} \Pr\{X = x|S = s\} : x \in \mathcal{X})$.
- Also, there is a probability vector that describes the next state $S' = s'$ given the current state $S = s$ and the next output $X = x$.
- We denote this by $(T(s'|s, x) \stackrel{\Delta}{=} \Pr\{S' = s'|S = s, X = x\} : s' \in \mathcal{S})$. If at time instant i the source is in a state s then the next source output $X_{i+1} = x$ is produced and the source enters state $S_{i+1} = s'$ with a probability $Q(x, s'|s) \stackrel{\Delta}{=} P(x|s)T(s'|s, x)$.

So we see here an important property of Markov sources:

the behavior of the source, i.e. the probabilities of selecting an output and a next state, given the current state does not depend on past states or outputs.

$$\Pr\{X_{i+1} = x, S_{i+1} = s' | S_i = s, X^{i-1} = x^{i-1}, S^{i-1} = s^{i-1}\} = Q(x, s'|s).$$

A Markov source exhibits a stationary behavior under certain conditions.

- We shall consider only those sources for which the probabilities $Q(x, s'|s)$ are time invariant.
- Also the source must be such that from every state $s \in \mathcal{S}$ we must be able to reach, in a finite number of steps (time instances), any other state $s' \in \mathcal{S}$ with non-zero probability. In particular, this means that for every state $s \in \mathcal{S}$ it must be possible to return to s in a finite number of steps.
- The recurrence time of a state s is the number of steps needed to return to s when started in s. This recurrence time is a random variable.
- A Markov source is said to be *connected* if from any state s in the source, any other state s' in the source can be reached in a finite number of steps.
- All states in a connected source have the same period m. The source is called *periodic* with period m if $m > 1$.
- If $m = 1$ then the source is called ergodic.
- If a periodic source has, at some time t, a state distribution $\underline{q}(t)$ equal to the stationary state distribution $\underline{q}$, then this state distribution holds for all t and the source is still stationary. (See the next pages.)

Stationary and ergodic sources

We shall mainly consider stationary and ergodic sources. We define the state transition probability matrix $[W_{ji}]$ as

$$\begin{aligned}\forall i \in \mathcal{S}, j \in \mathcal{S} : W_{ji} &\triangleq \Pr\{S_{t+1} = i | S_l = j\}, \\ &= \sum_{x \in \mathcal{X}} Q(x, i|j)\end{aligned}$$

Because $Q(x, s'|s)$ is time invariant, t can be an arbitrary time instant.

Let $\underline{q}(t) \triangleq (q_1(t), q_2(t), \cdots, q_J(t))$ be the state distribution at time t, i.e. $q_s(t) \triangleq \Pr\{S_t = s\}$. It is obvious that the state distribution at time $t+1$ depends on $\underline{q}(t)$ and $[W_{ji}]$ as

$$\underline{q}(t+1) = \underline{q}(t)[W_{ji}].$$

Without proof we state the following important result for stationary and ergodic sources that says that there exists a unique limiting state distribution $\underline{q}(\infty) = (q_1(\infty), q_2(\infty), \cdots, q_J(\infty))$ independent from $\underline{q}(1)$ such that

$$\lim_{t\to\infty} \underline{q}(t) = \underline{q}(\infty).$$

Moreover, this state distribution $\underline{q}(\infty)$ is also the solution of

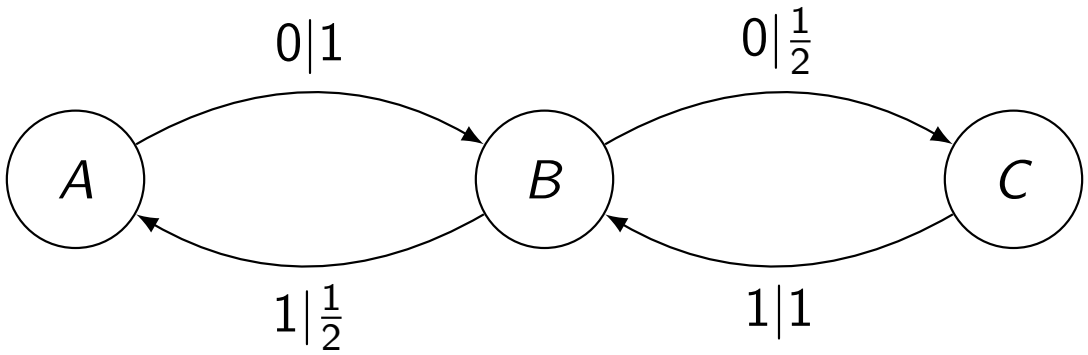
$$\underline{q}(\infty)[W_{ji}] = \underline{q}(\infty).$$

Because of this $\underline{q}(\infty)$ is also known as the stationary state distribution, also simply denoted by $\underline{q} = (q_1, q_2, \cdots, q_J)$.

Technically the source was started infinitely far in the past, "$t = -\infty$". This means that $\underline{q}(1) = \underline{q}(\infty)$. Because we usually only care for the subsequence $X_1, X_2, \cdots$ this means that $\underline{q}(t) = \underline{q}(\infty)$ for all relevant t, $t = 1, 2, \cdots$. This implies that the source is really stationary. This is also the case, for our purposes, when the source is started at time $t = 1$ with state distribution $\underline{q}(1) = \underline{q}(\infty)$.

Explanation by example

Consider the following periodic Markov source (the period is 2).



Suppose that $S_0 = A$, which means that the source is in state A at time 0. The following table describes the state of the source in the next few time instances.

time	states
1	B
2	A or C, each with probability $\frac{1}{2}$.
3	B
4	A or C, each with probability $\frac{1}{2}$.
5	B
6	A or C, each with probability $\frac{1}{2}$.
$\vdots$	$\vdots$

We see that the state distribution $\underline{q}(t)$ alternates between $(\frac{1}{2}, 0, \frac{1}{2})$ and $(0, 1, 0)$. So it does not converge!

We can still solve for the stationary state distribution as follows.

$$q_A = \frac{1}{2} q_B$$
$$q_C = \frac{1}{2} q_B$$
$$1 = q_A + q_B + q_C$$

And thus we find

$$q_A = q_C = \frac{1}{4}$$
$$q_B = \frac{1}{2}$$

Suppose that the state distribution at time $t = 0$ is equal to the stationary state distribution, i.e. $\underline{q}(0) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. It is easy to check with

$$\underline{q}(t+1) = \underline{q}(t)[W_{ji}]$$

that $\underline{q}(1) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ and so on!
We can conclude that the source is now stationary!

Entropy rate and Markov source

- $P(X_i = x, S_{i+1} = s' | S_i = s)$.
  Given the current state $S_i = s$, the current output $X_i = x$ and the next state $S_{i+1} = s'$ depend only on the current state.
-

$$P(X_i = x, S_{i+1} = s' | S_i = s) =$$
$$P(X_i = x | S_i = s) \times \quad \text{(output probability)}$$
$$P(S_{i+1} = s' | S_i = s, X_i = x). \quad \text{(next state probability)}$$

- If $P(S_{i+1} = s' | S_i = s, X_i = x)$ only contains zeros and ones, then the next state is uniquely determined by the current state and the current output. This is called an *unifilar* source.

- *Entropy rate*

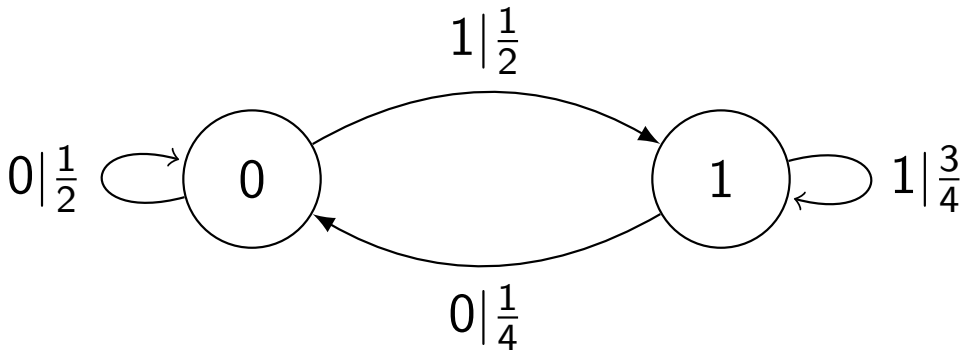
$$H_\infty(X) = \lim_{L\to\infty} \frac{1}{L} H(X^L).$$

- Unifilar source gives

$$\begin{aligned} H_\infty(X) &= H(X|S) \\ &= \sum_{s\in\mathcal{S}} q_s H(X|S=s) \end{aligned}$$

Example

Consider the first order Markov source.



In this figure the states are depicted by circles with their label in the center. Arrows from one state (s) to another (or the same) state (s') depict non zero $Q(x, s'|s)$ values. The arrows are labeled by the source output x and the probability $Q(x, s'|s)$ as:

$$x|Q(x, s'|s).$$

For this source we can compute the probability of the sequence $X_1, X_2, X_3, X_4 = 1, 1, 0, 0$ given that $S_1 = 0$ as follows. From $S_1 = 0$ and $X_1 = 1$ we know $S_2 = 1$. Then we have $(S_2 = 1,\ X_2 = 1) \rightarrow S_3 = 1$ and $(S_3 = 1, X_3 = 0) \rightarrow S_4 = 0$. So we compute

$$\begin{aligned}\Pr\{X_1,X_2, X_3, X_4 = 1, 1, 0, 0 | S_1 = 0\} &= \Pr\{X_1 = 1 | S_1 = 0\} \times \\ &\Pr\{X_2 = 1 | S_2 = 1\} \Pr\{X_3 = 0 | S_3 = 1\} \Pr\{X_4 = 0 | S_4 = 0\} = \\ &\frac{1}{2}\frac{3}{4}\frac{1}{4}\frac{1}{2} = \frac{3}{64}.\end{aligned}$$

Stationary state distribution:

$$q_0 = \frac{1}{2} q_0 + \frac{1}{4} q_1$$
$$(q_1 = \frac{1}{2} q_0 + \frac{3}{4} q_1)$$
$$1 = q_0 + q_1$$

So, $q_0 = \frac{1}{3}$, $q_1 = \frac{2}{3}$.

Entropy per state:

$$
\begin{aligned}
H(X|S=1) &= \sum_{x\in\mathcal{X}} P(X=x|S=1)\log_2 \frac{1}{P(X=x|S=1)} \\
&= \frac{1}{4}\log_2 4 + \frac{3}{4}\log_2 \frac{4}{3} \\
&= h(\frac{1}{4}) = 0.8113.
\end{aligned}
$$

where $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$.

$$
H(X|S=0) = h(\frac{1}{2}) = 1.
$$

Entropy rate: (source is unifilar)

$$
\begin{aligned}
H_\infty(X) &= H(X|S) \\
&= \sum_{s\in\mathcal{S}} q_s H(X|S=s) \\
&= \frac{1}{3}h(\frac{1}{2}) + \frac{2}{3}\cdot h(\frac{1}{4}).
\end{aligned}
$$

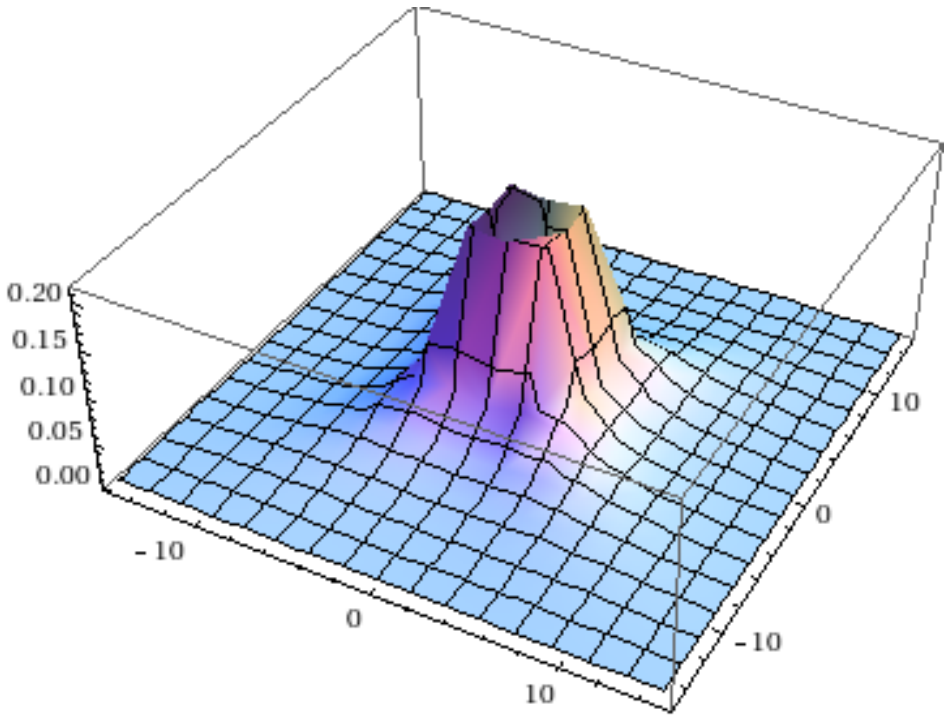
Part A

In this chapter we show by example that the Maximum Likelihood (ML) approach does not work (directly) when we wish to estimate a data model. The ML over estimates a model. In general it will select the most complex model that fits the data and thus it models the noise (or randomness) of the data as a structural part. This degrades the performance of the resulting predictions/estimations.
Methods such as the BIC introduce a cost term that depends on the number of parameters in the model. Such a method can be derived from the Laplace approximation.

Laplace approximation

The function f is defined on the two dimensional real space and

$$f(x, y) = \frac{x^2 + y^2}{x^4 + y^4 + 1}.$$



The function f

Find a local maximum of f.
Hint: in finding a local maximum it can be useful to assume an additional condition $x = y$.

We find $x = y = \pm 2^{-1/4}$ so e.g. $x = y = 0.8409$. The corresponding maximum is $2^{-1/2} = 0.7071$.

Approximate

$$Z_f = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy,$$

using the Laplace approximation.

We know that f has a maximum at $x = y = 2^{-1/4}$. So we calculate at this x and y the following.

$$\begin{aligned} f_x &= 0, \\ f_y &= 0, \\ f_{xx} &= -2, \\ f_{xy} = f_{yx} &= 0, \\ f_{yy} &= -2. \end{aligned}$$

We write f for $f(x,y)$ evaluated in $x = y = 2^{-1/4}$.

$$\frac{\partial}{\partial x} \ln f(x,y) = \frac{f_x}{f} = 0,$$
$$\frac{\partial}{\partial y} \ln f(x,y) = \frac{f_y}{f} = 0,$$
$$\frac{\partial^2}{\partial x^2} \ln f(x,y) = \frac{f \cdot f_{xx} - (f_x)^2}{f^2} = -2\sqrt{2},$$
$$\frac{\partial^2}{\partial x \partial y} \ln f(x,y) = \frac{f_y \cdot f_x - f_x \cdot f_y}{f^2} = 0,$$
$$\frac{\partial^2}{\partial y \partial x} \ln f(x,y) = \frac{f_y \cdot f_x - f_x \cdot f_y}{f^2} = 0,$$
$$\frac{\partial^2}{\partial y^2} \ln f(x,y) = \frac{f \cdot f_{yy} - (f_y)^2}{f^2} = -2\sqrt{2}.$$

The Hessian is

$$H = \begin{bmatrix} -2\sqrt{2} & 0 \\ 0 & -2\sqrt{2} \end{bmatrix}.$$

And so

$$A = -H = \begin{bmatrix} 2\sqrt{2} & 0 \\ 0 & 2\sqrt{2} \end{bmatrix}.$$

$$-\frac{1}{2}(z - z_0)A(z - z_0) = -\sqrt{2}\left((x - 2^{-1/4})^2 + (y - 2^{-1/4})^2\right)$$

So, the approximation is

$$g(x, y) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}\left((x-2^{-1/4})^2+(y-2^{-1/4})^2\right)}.$$

The integral of $g(x, y)$ is

$$f(2^{-1/4}, 2^{-1/4})\sqrt{\frac{(2\pi)^2}{\det(A)}} = \frac{\pi}{2}.$$

Part B

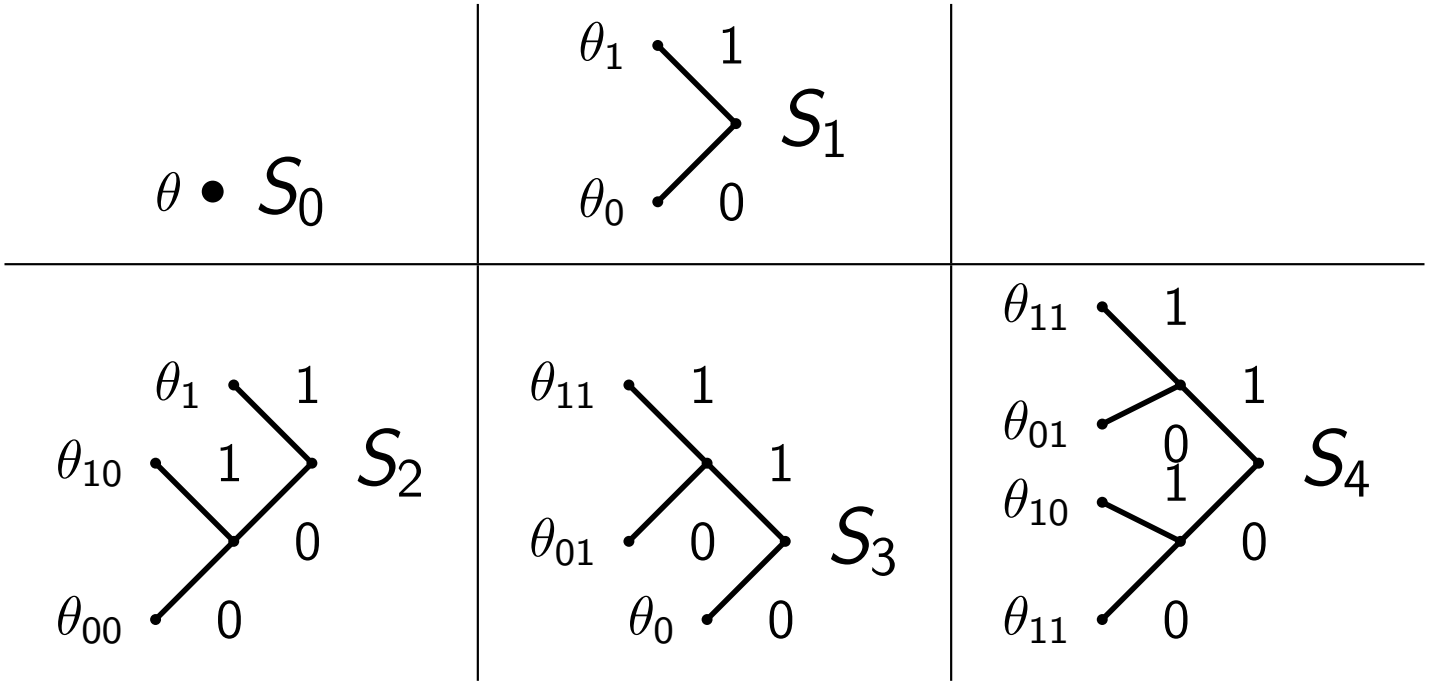
In this part we show that the Bayesian mixture approach for binary memoryless model also results in a parameter cost term of $\frac{1}{2}\log N$.
We aply this mixture approach to finite order Markov models and to Context-Tree models. We discuss and efficient method to calculate a Bayesian mixture over a class of Context-Tree models. Finally we discuss the determination of the a-posteriori model probability calculation with this mixture.

Model a-posteriori probabilities for tree sources

We observe a binary sequence $x^{15} = 110\,111\,101\,100\,001$.
Assume that this sequence is preceded by two zeros as the initial context.
Consider a context model $\mathcal{S}$ of depth 2 and the "CTW prior" $P(S_i)$ given as:

$$\Delta_2(S) = 2|S| - 1 - |\{s \in S : |s| = 2\}|,$$
$$P(S_i) = 2^{-\Delta_2(S_i)}$$

The following five tree structures are possible models, S_0, S_1, S_2, S_3, and S_4:



Compute the "CTW" probability of this sequence,

$$P_w^{\lambda}(x^{15}) = \sum_{i=0}^{4} P(S_i) P(x^{15}|S_i).$$

We consider a binary context tree of depth 2. We must collect the number of zeros and ones in eacht of the contexts $\lambda, 0, 1, 00, 10, 01$, and 11. We find

s	$N(s0\|x^{15}) = a$	$N(s1\|x^{15}) = b$
λ	6	9
0	3	4
1	3	5
00	2	2
10	1	2
01	0	3
11	3	2

From this we calculate the $P_e(a,b)$'s and the P_w^s's.

s	$P_e(a,b)$	P_w^s
00	$2.3438 \cdot 10^{-2}$	$2.3438 \cdot 10^{-2}$
10	$6.25 \cdot 10^{-2}$	$6.25 \cdot 10^{-2}$
01	$3.125 \cdot 10^{-1}$	$3.125 \cdot 10^{-1}$
11	$1.1719 \cdot 10^{-2}$	$1.1719 \cdot 10^{-2}$
0	$2.4414 \cdot 10^{-3}$	$1.9531 \cdot 10^{-3}$
1	$1.3733 \cdot 10^{-3}$	$2.5177 \cdot 10^{-3}$
λ	$8.3596 \cdot 10^{-6}$	$6.6347 \cdot 10^{-6}$

The requested probability is $p_w^\lambda = 6.6347 \cdot 10^{-6}$.

Determine the a-posteriori model probabilities for these five models.

For every model we must compute $2^{-\Delta_2(S)} \prod_{s \in S} P_e(a_s, b_s)/P_w^{\lambda}$.

S	$2^{-\Delta_2(S)}$	$P(x^{15}\|S)$	$P(S\|x^{15})$
S_0	2^{-1}	$8.3596 \cdot 10^{-6}$	0.630
S_1	2^{-3}	$3.3528 \cdot 10^{-6}$	0.063
S_2	2^{-3}	$2.0117 \cdot 10^{-6}$	0.038
S_3	2^{-3}	$8.9407 \cdot 10^{-6}$	0.168
S_4	2^{-3}	$5.3644 \cdot 10^{-6}$	0.101

Part C

Using ideas from the field of Universal Data Compression we see a reason why $\frac{1}{2} \log N$ is the correct correction (cost) factor for auto-regressive and discrete Markov models. It is the amount of information we obtain over a parameter value through the observation of a data sequence.
This leads to a notion of Descriptive Complexity and the separation of the description of data into two parts: the structure (logarithmic in the sequence length) and randomness (linear in the sequence length). In this way we argue that the shortest total description length must come from a model that describes the largest part of the data in the structure and thus gives the "best" explanation of the data.

Consider the binary sequence x^{4n} of length $4n$, for $n = 1, 2, 3, \ldots$. Given is

$$x^{4n} = 001100110011\ldots$$

So, x^{4n} is a repetition of 0011.
M_0 is the i.i.d. memoryless model with parameter θ_0. Here θ_0 describes the probability of a one, or

$$\Pr\{X_i = 1\} = \theta_0 (= 1 - \Pr\{X_i = 0\}).$$

Likewise is M_1 the first-order Markov model with parameters $\theta_{1,0}$ and $\theta_{1,1}$, defined as

$$\theta_{1,0} = \Pr\{X_i = 1 | X_{i-1} = 0\},$$
$$\theta_{1,1} = \Pr\{X_i = 1 | X_{i-1} = 1\}.$$

The $(\frac{1}{2}, \frac{1}{2})$ Beta prior of a parameter Θ (seen as a random variable) is given as

$$P_\Theta(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}.$$

Determine the probability of the sequence x^{4n} given the model M_0 with known parameter θ_0. So determine $P(x^{4n}|M_0, \theta_0)$.

The sequence of length $4n$ contains $2n$ zeros and $2n$ ones, so

$$P(x^{4n}|M_0, \theta_0) = (\theta_0(1-\theta_0))^{2n}.$$

Assume the $(\frac{1}{2}, \frac{1}{2})$ Beta prior of a parameter Θ_0 and determine the probability of the sequence x^{4n} given the model M_0. So determine $P(x^{4n}|M_0)$.

From the lecture notes we know the expression, so again with $a = 2n$ zeros and $b = 2n$ ones we have

$$P(x^{4n}|M_0) = \frac{\frac{1}{2} \cdot \frac{3}{2} \cdots \frac{2a-1}{2} \frac{1}{2} \cdot \frac{3}{2} \cdots \frac{2b-1}{2}}{(a+b)!}$$
$$= \frac{\left(\frac{1}{2} \cdot \frac{3}{2} \cdots \frac{4n-1}{2}\right)^2}{(4n)!}$$

Assume the $(\frac{1}{2}, \frac{1}{2})$ Beta priors for both parameters $\Theta_{1,0}$ and $\Theta_{1,1}$ and determine the probability of the sequence x^{4n} given the model M_1. So determine $P(x^{4n}|M_1)$. Additionally, it is given that the initial state of M_1 is "state one", indicating that the symbol preceding the sequence x^{4n} was a "1".

In the first order model we count the occurrences of zeros and ones in a state, indicated by the preceding symbol. It is easy to see that there are $2n$ (preceding) zeros, i.e. we visit state zero $2n$ times. Of the $2n$ times we saw n new zeros and n new ones. The same holds for the state one.

$$P(x^{4n}|M_1) = P(x_2x_3x_6x_7\ldots|M_1, \text{state} = 0)\times P(x_1x_4x_5x_8\ldots|M_1, \text{state} = 1)$$

$$P(x_2x_3x_6x_7\ldots|M_1, \text{state} = 0) = \frac{\left(\frac{1}{2}\cdot\frac{3}{2}\cdots\frac{2n-1}{2}\right)^2}{(2n)!}$$

$$P(x_1x_4x_5x_8\ldots|M_1, \text{state} = 1) = \frac{\left(\frac{1}{2}\cdot\frac{3}{2}\cdots\frac{2n-1}{2}\right)^2}{(2n)!}$$

$$P(x^{4n}|M_1) = \left(\frac{\left(\frac{1}{2}\cdot\frac{3}{2}\cdots\frac{2n-1}{2}\right)^2}{(2n)!}\right)^2$$

Consider the MDL principle. What model, M_0 or M_1, would you chose as a function of n? Motivate your answer.
Hint: Remember that the codeword length is determined by the cost of the model and the cost of the data, given the model. This second part is mainly determined by the entropy of the model given the most likely parameter values.

It is easy to see that the codeword length for model M_0 approximately equals $L_0 \approx \frac{1}{2}\log_2 4n + 4n$. The term $\frac{1}{2}\log_2 4n$ represents the cost of the single parameter and the $4n$ bits come from the fact that we find a model with parameter value $\theta_0 = \frac{1}{2}$.
For the first order model the parameter cost is $\log 4n$ (two parameters) and the entropy remains 1 bit per symbol, so the codeword length becomes $L_1 \approx \log_2 4n + 4n$.
So the MDL principle will select, with high probability, M_0.

Consider also a second order Markov model M_2 with corresponding parameters. Again consider the MDL principle. What model, M_0, M_1, or M_2, would you chose as a function of n? Motivate your answer.

The entropy of the second order model is zero because the two preceding symbols determine the next symbol uniquely. We have four parameters so $L_2 \approx 2 \log_2 4n$. This is almost immediately smaller than the other two lengths so MDL quickly selects M_2.