Part B

Bayesian model estimation and the Context-tree model selection

Bayesian model estimation

Additional reading

Bishop §14.1: Bayesian Model Averaging
Bishop §14.4: Tree-based Models

Bayesian model estimation

Additional notation

i^{th} order Markov Model	$\mathcal{M}_i$	The state is determined by the previous i symbols.
Parameter vector	θ_i	This vector describes all probabilities $P(x_n \mid x_{n-i}, x_{n-i+1}, \ldots, x_{n-1})$.
Parameter element	$\theta_{i,s}$	$\theta_{i,s} = P(x_n \mid x_{n-i}^{n-1} = s)$.
Model state	s	s is a binary sequence of length i.

Bayesian model estimation

Example 2: [Revisit first lecture]
Let $\mathcal{M}_i$ be the i-th order binary Markov model (source).

Then $\Theta_i = [0,1]^{2^i}$.

Beta distribution for prior $p(\theta_i|\mathcal{M}_i)$, with $\alpha = \beta = \frac{1}{2}$ (Jeffreys prior).

$$p(\theta_i|\mathcal{M}_i) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{2^i} \prod_{s\in\{0,1\}^i} \theta_{i,s}^{\alpha-1}(1-\theta_{i,s})^{\beta-1}$$

$$= \frac{1}{\pi^{2^i}} \prod_{s\in\{0,1\}^i} \theta_{i,s}^{-1/2}(1-\theta_{i,s})^{-1/2}$$

Bayesian model estimation

$$p(x^N|\mathcal{M}_i) = \int_{\Theta_i} p(\theta_i|\mathcal{M}_i)p(x^N|\mathcal{M}_i,\theta_i)\, d\theta_i$$

$$= \frac{1}{\pi^{2^i}} p(x_1,\ldots,x_i)$$

$$\int_{\Theta_i} \prod_{s\in\{0,1\}^i} \theta_{i,s}^{n(s1|x^N)-1/2}(1-\theta_{i,s})^{n(s0|x^N)-1/2}\, d\theta_i$$

$$= \frac{p(x_1,\ldots,x_i)}{\pi^{2^i}}$$

$$\prod_{s\in\{0,1\}^i} \int_{[0,1]} \theta_{i,s}^{n(s1|x^N)-1/2}(1-\theta_{i,s})^{n(s0|x^N)-1/2}\, d\theta_{i,s}$$

$$= p(x^i) \prod_{s\in\{0,1\}^i} \frac{\Gamma(n(s0|x^N)+\frac{1}{2})\Gamma(n(s1|x^N)+\frac{1}{2})}{\pi\Gamma(n(s0|x^N)+n(s1|x^N)+1)}$$

Bayesian model estimation

So we must study the behaviour of

$$P_e(a,b) \triangleq \frac{\Gamma(a+\frac{1}{2})\Gamma(b+\frac{1}{2})}{\pi\Gamma(n+1)}$$

$$a \triangleq n(0|x^N)$$

$$b \triangleq n(1|x^N)$$

It is a memoryless sub-sources of the Markov source. x^N is generated i.i.d. with parameter θ.

The actual probability of x^N under this source is

$$p(x^N|\mathcal{M},\theta) = (1-\theta)^a\theta^b$$

Bayesian model estimation

We can write

$$P_e(a,b) = \frac{\frac{1}{2}\frac{3}{2}\cdots(a-\frac{1}{2})\cdot\frac{1}{2}\frac{3}{2}\cdots(b-\frac{1}{2})}{(a+b)!}$$

Again with the help of Stirling's approximation we can derive, for large a and b the following. (Excercise). Note: $a+b=N$.

$$\log_2 \frac{p(x^N|\mathcal{M},\theta)}{P_e(a,b)} \leq \frac{1}{2}\log_2 N + \frac{1}{2}\log_2\frac{\pi}{2}$$

Actually, we can prove that for all $a \geq 0$ and $b \geq 0$

$$\log_2 \frac{p(x^N|\mathcal{M},\theta)}{P_e(a,b)} \leq \frac{1}{2}\log_2 N + 1$$

Bayesian model estimation

Back to the i-th order Markov source.

$$p(x^N|\mathcal{M}_i, \theta_i) = p(x^i) \prod_{s \in \{0,1\}^i} \theta_{i,s}^{n(s1|x^N)} (1 - \theta_{i,s})^{n(s0|x^N)}$$

$$p(x^N|\mathcal{M}_i) = p(x^i) \prod_{s \in \{0,1\}^i} P_e(n(s1|x^N), n(s0|x^N))$$

Bayesian model estimation

With the previous bound we find

$$\log_2 \frac{p(x^N|\mathcal{M}_i,\theta_i)}{p(x^N|\mathcal{M}_i)} =$$

$$\log_2 \frac{p(x^i)\prod_{s\in\{0,1\}^i}\theta_{i,s}^{n(s1|x^N)}(1-\theta_{i,s})^{n(s0|x^N)}}{p(x^i)\prod_{s\in\{0,1\}^i}P_e(n(s1|x^N),n(s0|x^N))}$$

$$= \sum_{s\in\{0,1\}^i} \log_2 \frac{\theta_{i,s}^{n(s1|x^N)}(1-\theta_{i,s})^{n(s0|x^N)}}{P_e(n(s1|x^n),n(s0|x^N))}$$

$$\leq \sum_{s\in\{0,1\}^i} \frac{1}{2}\log_2 n(s|x^N) + 1 \overset{*1}{\leq} \frac{2^i}{2}\log_2 \frac{N-i}{2^i} + 2^i$$

(*1): here we use Jensen's inequality.

Bayesian model estimation

So we conclude that for any parameter vector $\boldsymbol{\theta_i}$ we have (approximately!)

$$\log_2 p(x^N|\mathcal{M}_i) \approx \log_2 p(x^N|\mathcal{M}_i, \theta_i) - \frac{2^i}{2}\log_2 \frac{N-i}{2^i} - 2^i$$

Maximum Likelihood parameters (and $N \gg \max\{2^i, 2^j\}$)

$$\log_2 \frac{p(\mathcal{M}_i|x^N)}{p(\mathcal{M}_j|x^N)} \approx \log_2 \frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)} + \log_2 \frac{p(x^N|\mathcal{M}_i, \hat{\theta}_i)}{p(x^N|\mathcal{M}_j, \hat{\theta}_j)}$$

$$-\frac{2^i - 2^j}{2}\log_2 N$$

So, again we observe the parameter cost!

Context trees

Recap: Memoryless binary source: one parameter $\theta = \Pr\{X = 1\}$

Recap: Markov order-k: one parameter per state. There are 2^k states. The k symbols $x_{i-k}, \ldots, x_{i-1}$ form the context of the symbol x_i.

Real world models: Length of context depends on its contents.
e.g. Natural language (English, Dutch, $\cdots$): if context starts with $x_{i-1} =$'q' then no more symbols are needed.

Context trees

A tree source is a nice concept to describe such sources.

A tree source consists of a set $\mathcal{S}$ of suffixes that together form a tree.
To each suffix (leaf) s in the tree there corresponds a parameter θ_s.

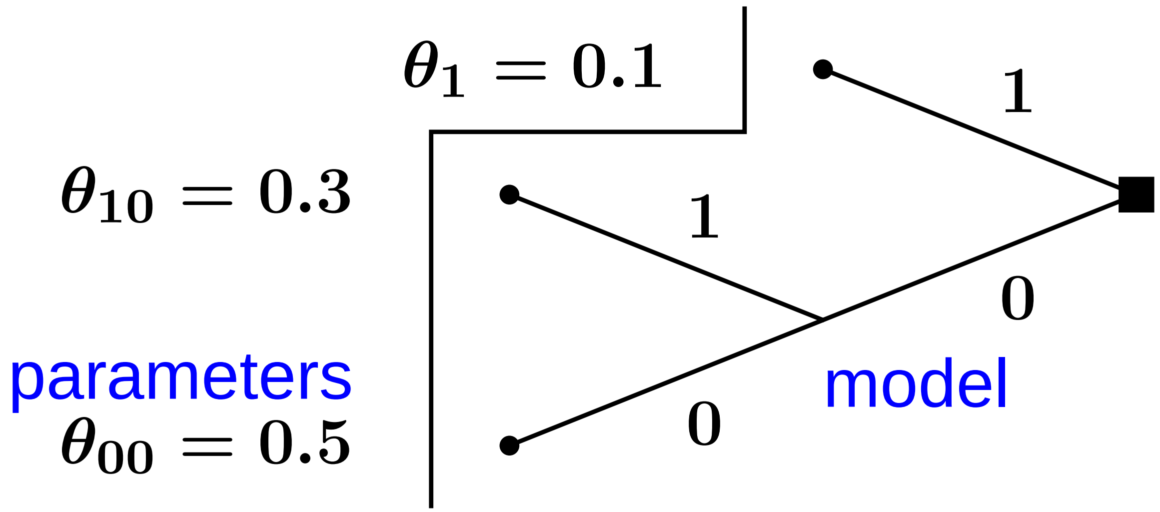
Some more notation: By $x^N_{|s}$ we denote the sub-sequence of symbols from x^N that are preceded by the sequence s.

Example: $x^8 = 01011010$; $s = 01$; then
$x^8_{|01} = x_3 x_5 x_8 = 010$.

Context trees

Example 3: Let $\mathcal{S} \triangleq \{00, 10, 1\}$ and $\theta_{00} = 0.5, \theta_{10} = 0.3$, and $\theta_1 = 0.1$ then

$$\Pr\{X_i = 1 | \cdots, x_{i-2} = 0, x_{i-1} = 0\} = 0.5,$$
$$\Pr\{X_i = 1 | \cdots, x_{i-2} = 1, x_{i-1} = 0\} = 0.3,$$
$$\Pr\{X_i = 1 | \cdots, x_{i-1} = 1\} = 0.1.$$



Context trees

Just as before (“Bayesian model estimation”) we must estimate the sequence probabilities of the memoryless subsources that correspond to the leaves of the tree (states of the source).

Let the full sequence be x^N and the subsequence for state s be written as $x^N_{|s}$. Before we wrote

$$P_e(a,b) = \frac{\Gamma(a+\frac{1}{2})\Gamma(b+\frac{1}{2})}{\pi\Gamma(a+b+1)}$$

Context trees

We shall now use the shorthand notation for the estimated probability of the subsequence generated in state s given the full sequence x^i:

$$P_e(a_s, b_s) = \frac{\Gamma(a_s + \frac{1}{2})\Gamma(b_s + \frac{1}{2})}{\pi\Gamma(a_s + b_s + 1)}$$

$$a_s = n(s0|x^N) = n(0|x^N_{|s})$$

$$b_s = n(s1|x^N) = n(1|x^N_{|s})$$

Context trees

Example 4: Let $\mathcal{S} = \{00, 10, 1\}$.

$$
\begin{array}{cccc|ccccccc}
\cdots & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
 & & & & & 00 & & & 00 & & \\
 & & & & 10 & & & 10 & & & \\
 & & & & & & 1 & & & 1 & 1 \\
 & \text{past} & & & & & & & & &
\end{array}
$$

$$
p(0100110|\cdots 110) = \underbrace{P_e(00)}_{10}\,\underbrace{P_e(11)}_{00}\,\underbrace{P_e(010)}_{1}
$$

$$
= \frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{2}\cdot\frac{1}{4}\cdot\frac{3}{6} = \frac{9}{1024}
$$

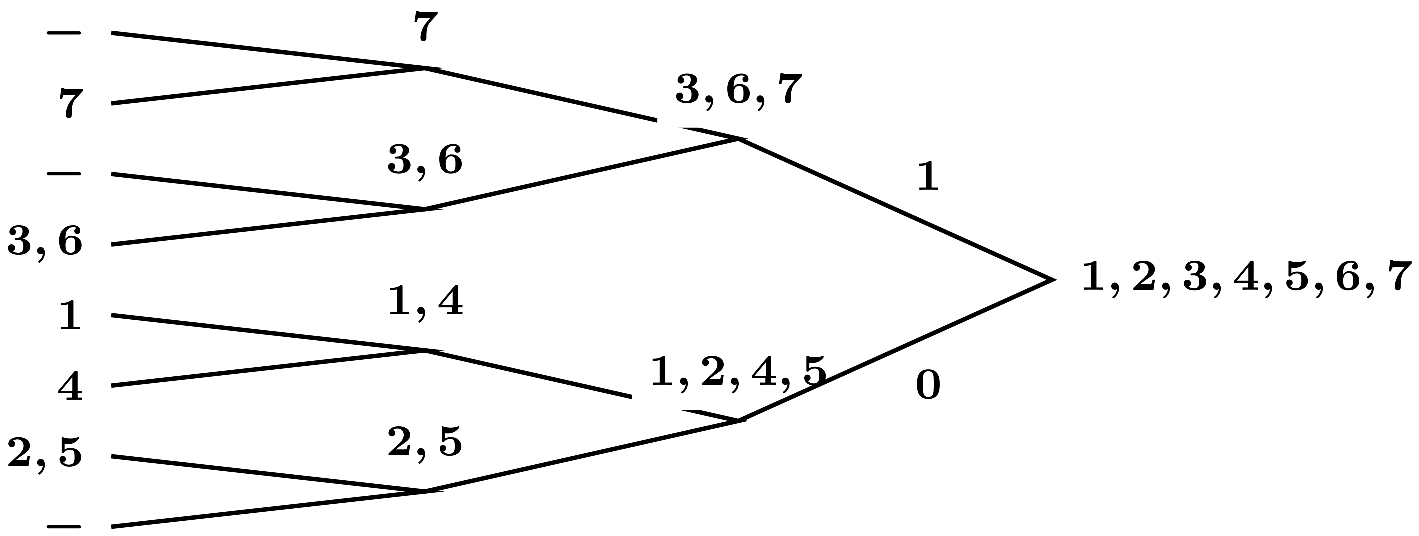
See “Bayesian Estimation”

$$
\log_2 \frac{p(x^N|\mathcal{S},\theta)}{\prod_{s\in\mathcal{S}} P_e(a_s, b_s)} \le \frac{|\mathcal{S}|}{2}\log_2\frac{N}{|\mathcal{S}|} + |\mathcal{S}|
$$

Context trees

Problem: We do not know $\mathcal{S}$!

Context tree (of depth D)



				1	2	3	4	5	6	7
$\cdots$	1	1	0	0	1	0	0	1	1	0
		past		source sequence						

In every node s use $a_s = n(s0|x^N)$ and $b_s = n(s1|x^N)$.

Context trees

We shall assign a probability to the subsequence $x^N_{|s}$ for every state s in the context tree.

We shall do this in such a way that in the root of the tree we assign a probability to the whole sequence x^N that is a mixture of all possible tree sources.

We use the following observations to build, recursively, this probability.

The probability we build is written as follows

$$P^s_w = P_w(x^N_{|s}),$$

where P^s_w is the shorthand notation we shall use.

Later we return to this and make the notation more precise.

Context trees

Suppose s is a leaf:
All we know are a_s and b_s so we assign the subsequence probability

$$P_w^s = P_e(a_s, b_s).$$

Now if s is an internal node, we have two options for the subsequence probability.

1: $P_e(a_s, b_s)$.
2: $P_w^{0s} P_w^{1s}$.

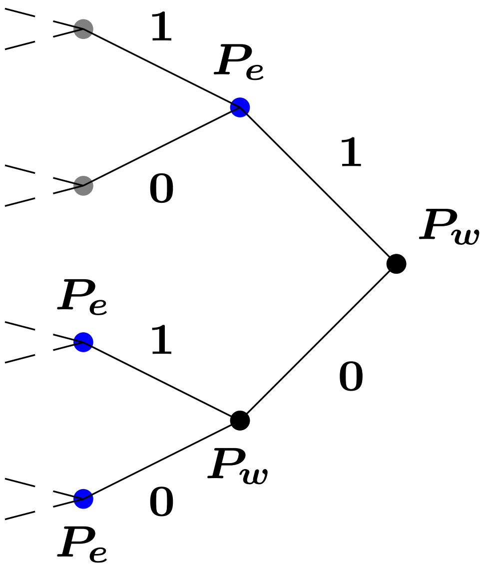
We must make a choice or better even, mix these options.
So we set

$$P_w^s = \frac{P_e(a_s, b_s) + P_w^{0s} P_w^{1s}}{2}.$$

Context trees

Analysis.

Let $\mathcal{S} = \{00, 10, 1\}$ and we use a context tree with depth $D > 2$.
We look at the P_w's for different nodes.
For the nodes $s \in \mathcal{S}$ we consider (in the analysis) only the P_e's.



$$P_w^{00} \geq \frac{1}{2} P_e(a_{00}, b_{00})$$

$$P_w^{10} \geq \frac{1}{2} P_e(a_{10}, b_{10})$$

$$P_w^{1} \geq \frac{1}{2} P_e(a_{1}, b_{1})$$

Context trees

Now we consider nodes nearer to the root and take only the $P_w^{0s} P_w^{1s}$ part.

$$
\begin{aligned}
P_w^0 &\geq \frac{1}{2} P_w^{00} P_w^{10} \\
&\geq \frac{1}{8} P_e(a_{00}, b_{00}) P_e(a_{10}, b_{10}) \\
P_w^\lambda &\geq \frac{1}{2} P_w^0 P_w^1 \\
&\geq \frac{1}{32} P_e(a_{00}, b_{00}) P_e(a_{10}, b_{10}) P_e(a_1, b_1)
\end{aligned}
$$

Here λ denotes the root of the tree.

Context trees

For general trees (or suffix sets) $\mathcal{S}$ we find

$$P_w^{\lambda} \geq 2^{1-2|\mathcal{S}|} \prod_{s \in \mathcal{S}} P_e(a_s, b_s)$$

So

$$\log_2 P_w^{\lambda} \geq \log_2 p(x^N|\mathcal{S}, \theta) - \left(2|\mathcal{S}| - 1 + \frac{|\mathcal{S}|}{2} \log_2 N + |\mathcal{S}|\right).$$

Context trees

For general trees (or suffix sets) $\mathcal{S}$ we find

$$P_w^\lambda \geq 2^{1-2|\mathcal{S}|} \prod_{s \in \mathcal{S}} P_e(a_s, b_s)$$

So

$$\log_2 P_w^\lambda \geq \log_2 p(x^N|\mathcal{S}, \theta) - \left(2|\mathcal{S}| - 1 + \frac{|\mathcal{S}|}{2} \log_2 N + |\mathcal{S}|\right).$$

Real sequence probability

Context trees

For general trees (or suffix sets) $\mathcal{S}$ we find

$$P_w^\lambda \geq 2^{1-2|\mathcal{S}|} \prod_{s \in \mathcal{S}} P_e(a_s, b_s)$$

So

$$\log_2 P_w^\lambda \geq \log_2 p(x^N|\mathcal{S}, \theta) - \left(2|\mathcal{S}| - 1 + \frac{|\mathcal{S}|}{2} \log_2 N + |\mathcal{S}| \right).$$

Cost of describing the tree

Context trees

For general trees (or suffix sets) $\mathcal{S}$ we find

$$P_w^\lambda \geq 2^{1-2|\mathcal{S}|} \prod_{s\in\mathcal{S}} P_e(a_s, b_s)$$

So

$$\log_2 P_w^\lambda \geq \log_2 p(x^N|\mathcal{S},\theta) - \left(2|\mathcal{S}| - 1 + \frac{|\mathcal{S}|}{2}\log_2 N + |\mathcal{S}|\right).$$

Cost of the parameters

Context trees

For general trees (or suffix sets) $\mathcal{S}$ we find

$$P_w^\lambda \geq 2^{1-2|\mathcal{S}|} \prod_{s \in \mathcal{S}} P_e(a_s, b_s)$$

So

$$\log_2 P_w^\lambda \geq \log_2 p(x^N|\mathcal{S}, \theta) - \left(2|\mathcal{S}| - 1 + \frac{|\mathcal{S}|}{2} \log_2 N + |\mathcal{S}|\right).$$

Contributions to the weighted probability are: Real sequence probability; Cost of describing the tree; Cost of the parameters

Context trees

This algorithm achieves the (asymptotically) optimal log-likelihood ratio (not only on the average but also individually for every data sequence).

$$\log \frac{p(x^N|\mathcal{S},\theta)}{P_w^\lambda} \leq 2|\mathcal{S}| - 1 + \frac{|\mathcal{S}|}{2}\log_2 N + |\mathcal{S}|.$$

Another essential property of the “Context-Tree Weighting” (CTW) algorithm is its efficient implementation. The number of trees squares with every increment of D and yet the amount of work is at most linear in $D \cdot N$.

Context trees

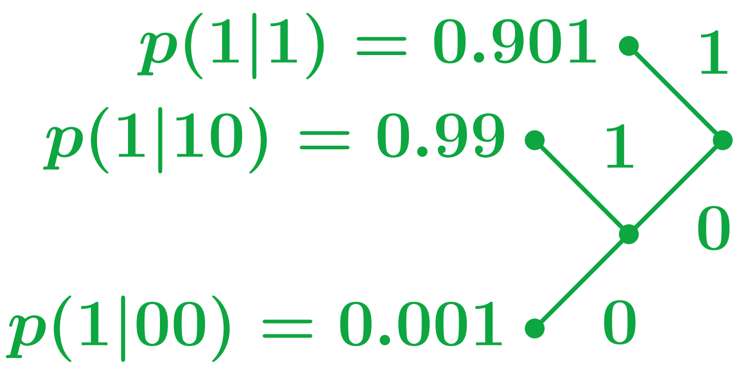
We can even write a stronger result when we realise that the method has no knowledge of a “real model”.
Let $\mathcal{S}_D$ be the set of all tree models with a maximal depth of atmost D.

$$\log P_w^{\lambda} \geq \max_{\mathcal{S} \in \mathcal{S}_D} \Big\{ \log p(x^N | \mathcal{S}, \theta) - \left(2|\mathcal{S}| - 1 + \frac{|\mathcal{S}|}{2} \log_2 N + |\mathcal{S}| \right) \Big\}.$$

This algorithm is an instantiation of the MDL principle. It finds (in the class $\mathcal{S}_D$) the model $\mathcal{S}$ that maximizes the sequence probability.

Context trees

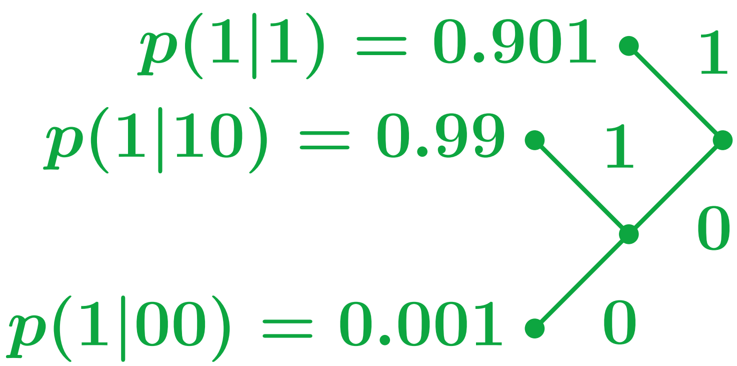
Example: Consider the following actual model.



We shall use the following models.

Context trees

Example: Consider the following actual model.



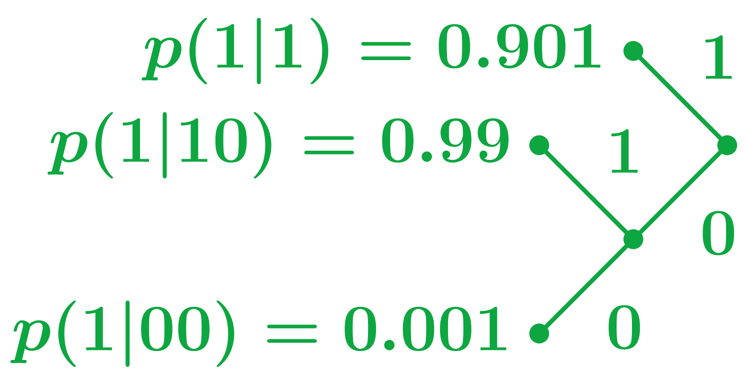
We shall use the following models.

$p(1)$ •

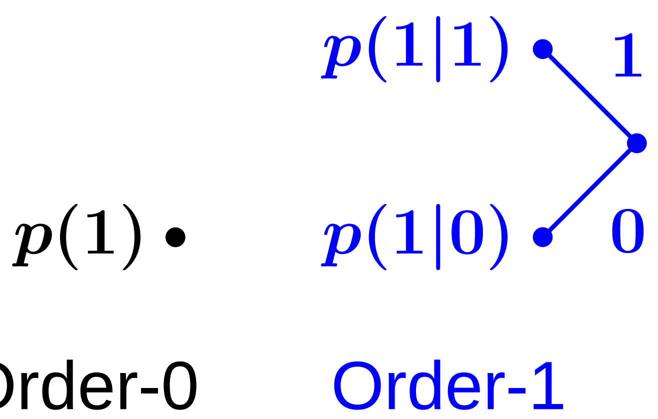
Order-0

Context trees

Example: Consider the following actual model.

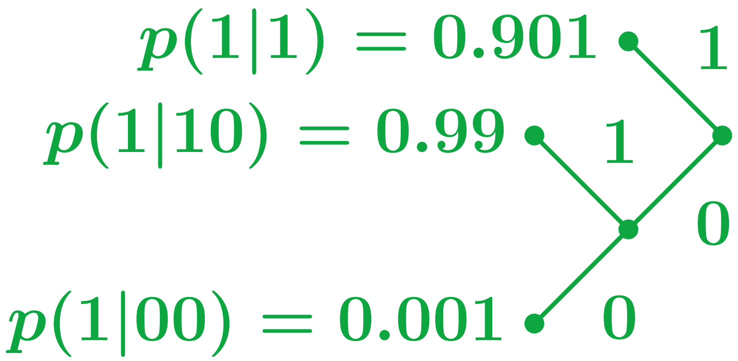


We shall use the following models.

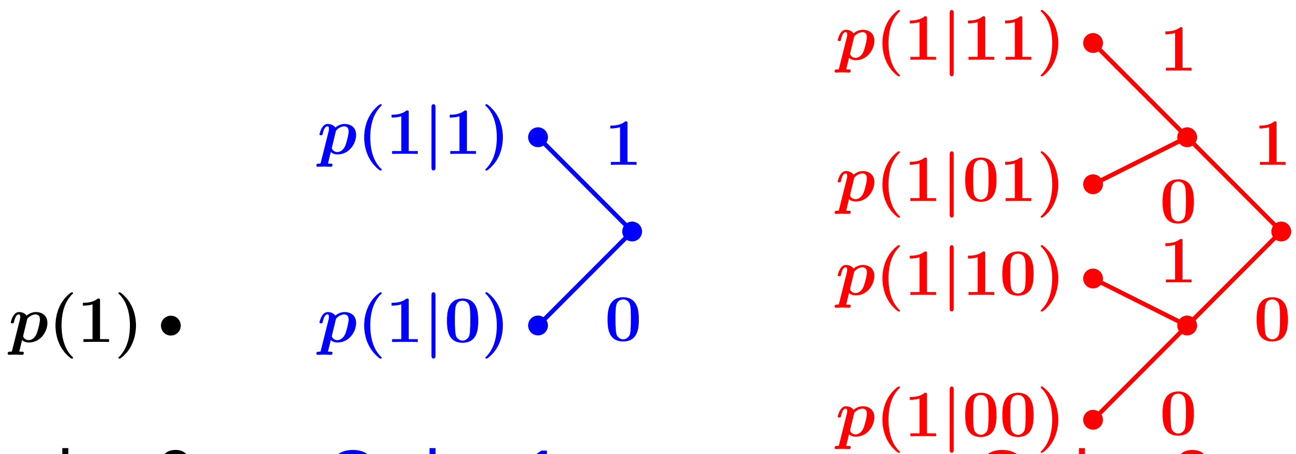


Context trees

Example: Consider the following actual model.



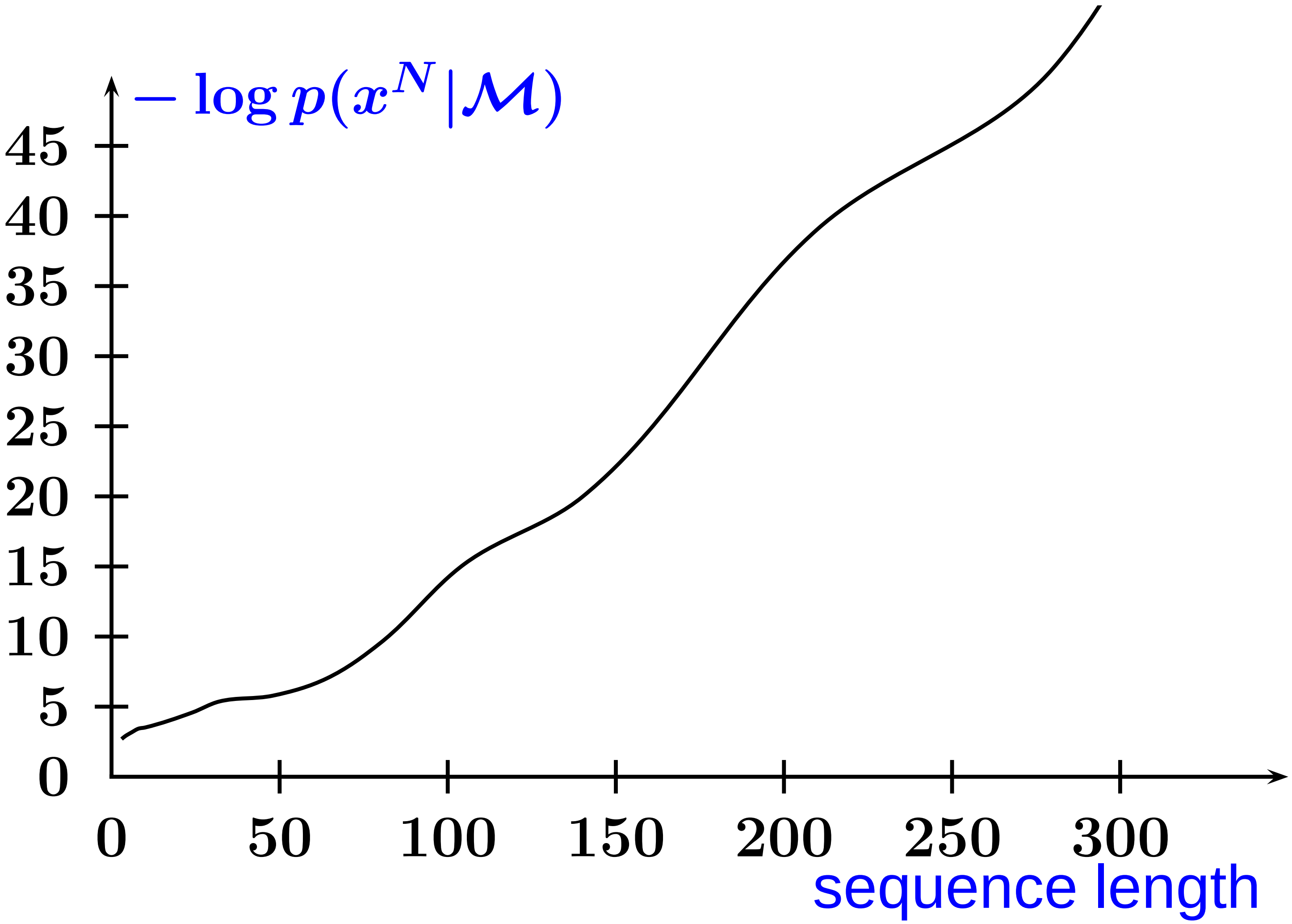
We shall use the following models.



Order-0 Order-1 Order-2

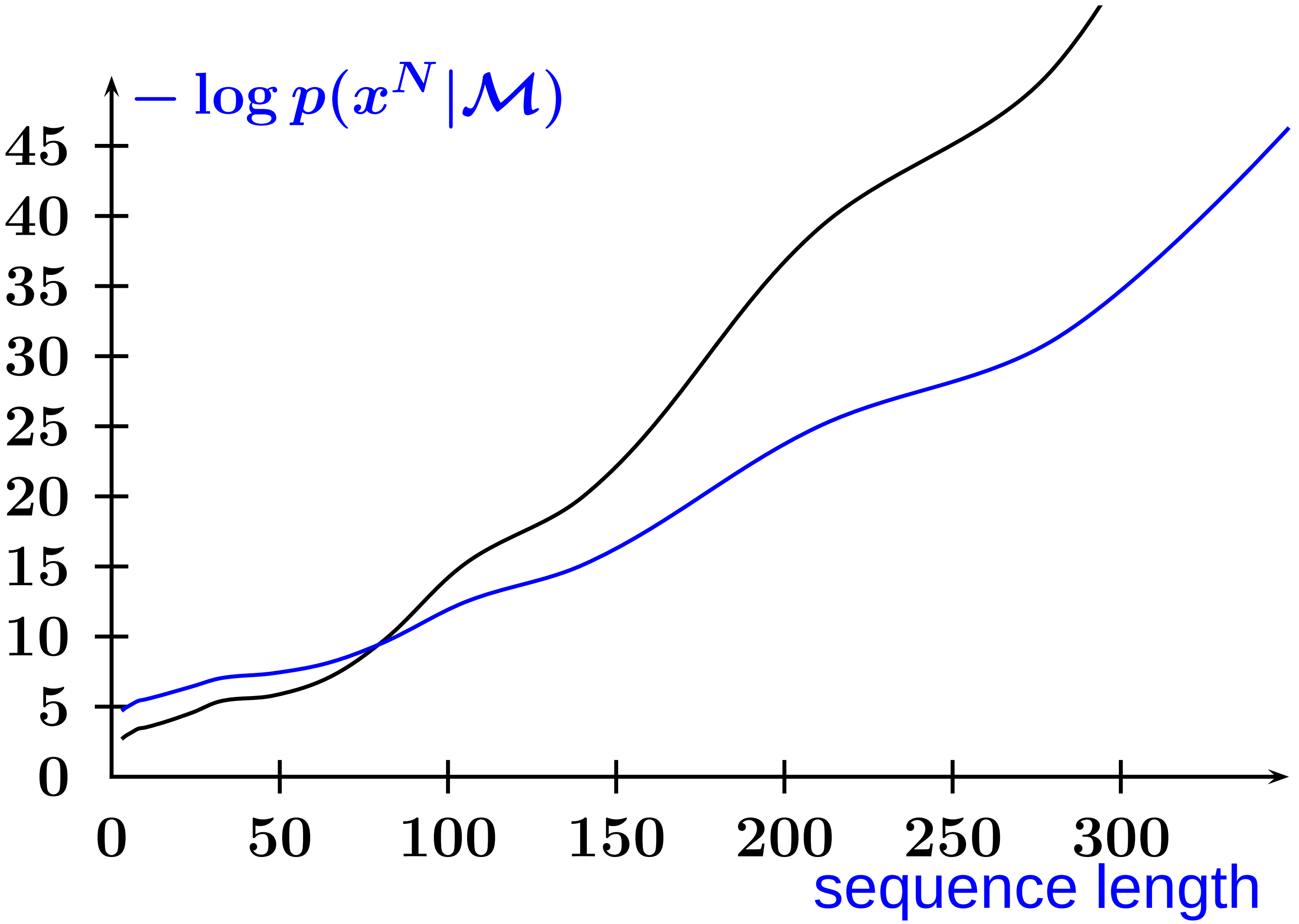
Context trees

The results for sequences of length upto $N = 350$ are shown graphically.



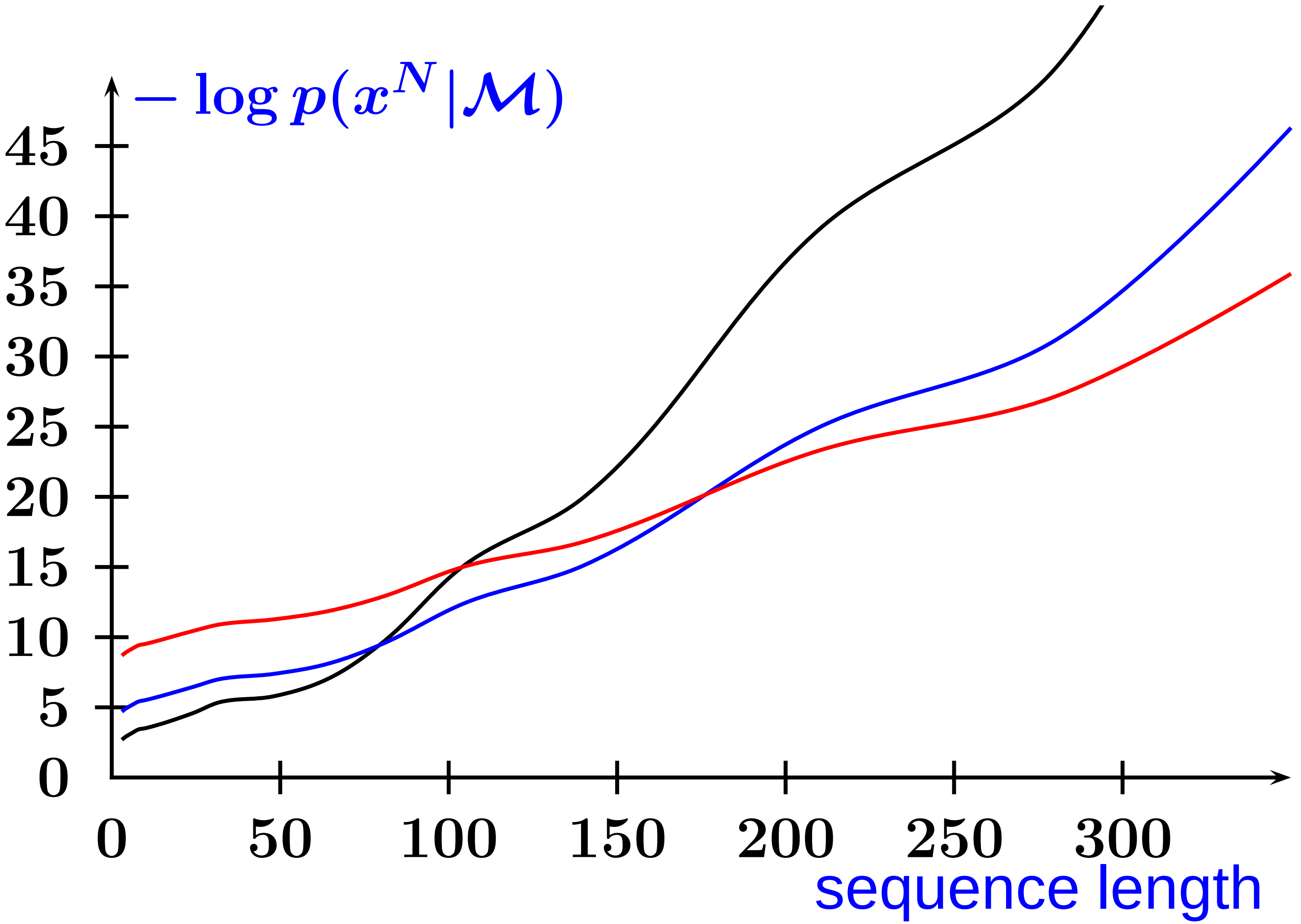
Context trees

The results for sequences of length upto $N = 350$ are shown graphically.



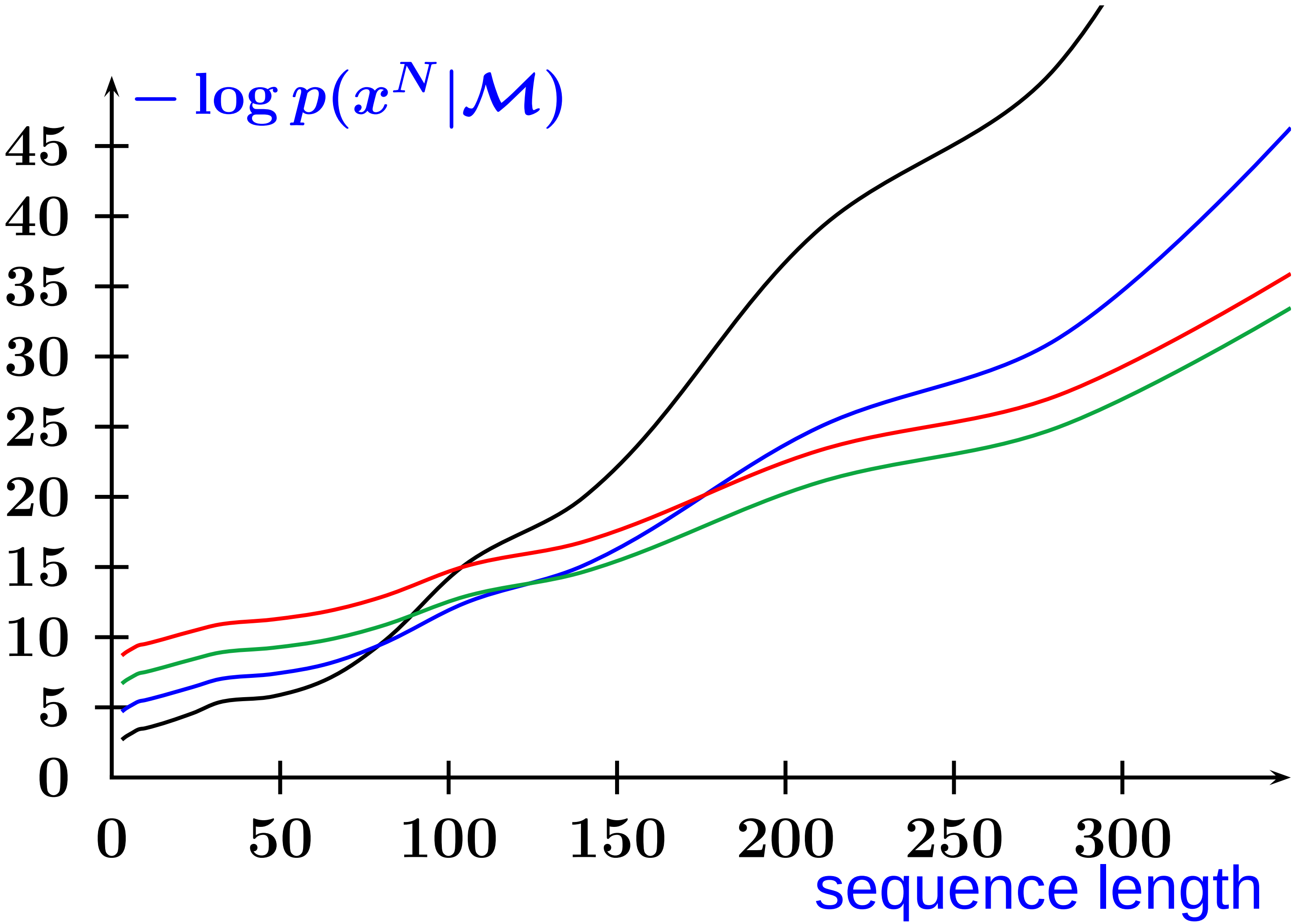
Context trees

The results for sequences of length upto $N = 350$ are shown graphically.



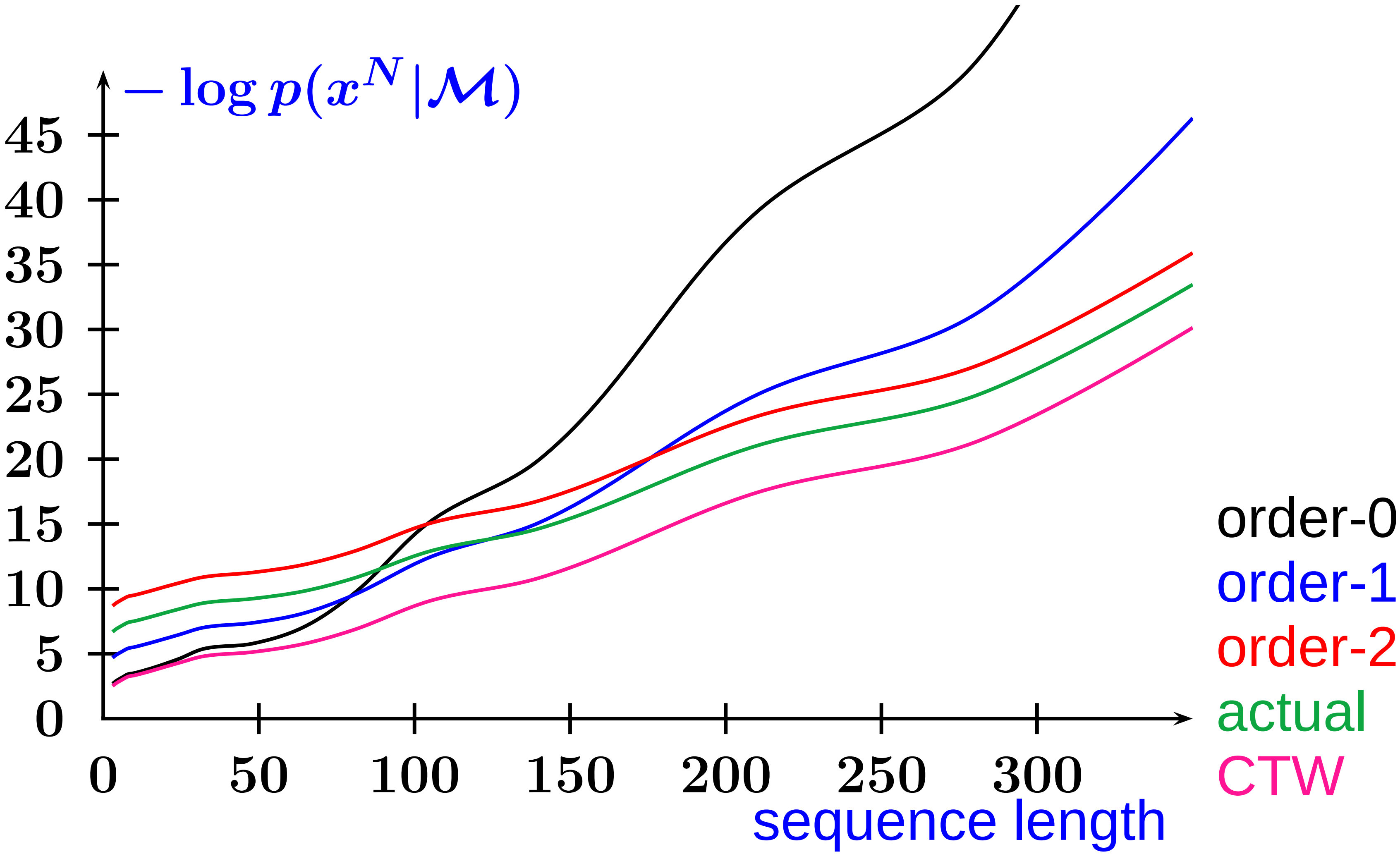
Context trees

The results for sequences of length upto $N = 350$ are shown graphically.



Context trees

The results for sequences of length upto $N = 350$ are shown graphically.



Context trees

We see that initially the memoryless (order-0) model performs even better than the actual model.

After about 80 symbols the order-1 model becomes better than both the order-0 and the actual model.

From 120 symbols on the actual model is better than the simpeler models.

The order-2 model is always worse than the actual model. It describes the same probabilities but has too many parameters.

But the CTW model outperforms all models over the whole range of sequence lengths!

Model posterior for Context trees

We shall now derive an expression, based on the previous method, for the a-posteriori model probability. We consider only binary data but the approach also works for arbitrary alphabets.

First we repeat our notation.

A model is described by a complete suffix set $\mathcal{S}$.

The suffix set can be seen as the set of leaves of a binary tree. Our model class is the set of all complete binary trees whose depth is not more than D, for a given D. We write $\mathcal{S}_D$ for the model class. So we say that $\mathcal{S} \in \mathcal{S}_D$.
The depth of a tree is the length of the longest path from the root to a leaf.

Model posterior for Context trees

Every model $\mathcal{S}$ has a set of parameters θ_s, one for every state $s \in \mathcal{S}$ of the model. θ_s gives the probability of a 1 given that the previous symbols were s.

$$\theta_s = \Pr\{X_t = 1 | X_{t-\ell}^{t-1} = s\}, \text{ where } \ell = |s|$$

Model posterior for Context trees

The probability of a sequence, given a model $\mathcal{S}$ with parameters θ_s, $s \in \mathcal{S}$ is

$$p(x^N|\mathcal{S},\theta) = \prod_{s\in\mathcal{S}} p(x^N_{|s}|\theta_s)$$

and

$$p(x^N_{|s}|\theta_s) = (1-\theta_s)^{n(0|x^N_{|s})}\theta_s^{n(1|x^N_{|s})}$$

Note (again) that $n(0|x^N_{|s}) = n(s0|x^N)$.

Actually, we silently assume that the first few symbols also have a "context". So we assume that there are some symbols preceding x^N.

Model posterior for Context trees

We must define some prior distributions. First the prior on the parameters.

We use the beta distribution. (In a non-binary case this generalizes to the Dirichlet distribution.) As done before we select the parameters in the beta distribution to be $\frac{1}{2}$.

So given a model $\mathcal{S}$ then for every $s \in \mathcal{S}$ we take

$$p(\theta_s|\mathcal{S}) = \frac{1}{\pi}\theta_s^{-\frac{1}{2}}(1-\theta_s)^{-\frac{1}{2}}$$

Model posterior for Context trees

This results in the following sequence probability, first assuming one state s only

$$p(x^N_{|s}) = \int_0^1 p(\theta_s|\mathcal{S})\theta_s^{n(s1|x^N)}(1-\theta_s)^{n(s0|x^N)}\,d\theta_s$$
$$= \frac{\Gamma(n(s0|x^N)+\frac{1}{2})\Gamma(n(s1|x^N)+\frac{1}{2})}{\pi\Gamma(n(s|x^N)+1)}$$

Now for any tree model $\mathcal{S}$ we find

$$p(x^N|\mathcal{S}) = \prod_{s\in\mathcal{S}}\int_0^1 p(\theta_s|\mathcal{S})p(x^N_{|s}|\theta_s)\,d\theta_s$$
$$= \prod_{s\in\mathcal{S}}\frac{\Gamma(n(s0|x^N)+\frac{1}{2})\Gamma(n(s1|x^N)+\frac{1}{2})}{\pi\Gamma(n(s|x^N)+1)}$$

Model posterior for Context trees

Next we need a prior on the tree models $\mathcal{S}$ in the set $\mathcal{S}_D$. We wish to use the efficient CTW method of weighting so we choose the corresponding prior.

First define

$$\Delta_D(S) \triangleq 2|\mathcal{S}| - 1 - |\{s \in \mathcal{S} : |s| = D\}|.$$

Then we take the prior

$$p(\mathcal{S}) = 2^{-\Delta_D(\mathcal{S})}$$

We prove that this is a proper prior probability.

Model posterior for Context trees

Obviously, $p(\mathcal{S}) > 0$ for all $\mathcal{S} \in \mathcal{S}_D$. We must show that it sums up to one.

We give a proof by induction.

Step 1: $D = 0$: $\mathcal{S}_0 = \{\lambda\}$, the memoryless source.

$$\Delta_0(\lambda) = 2 \cdot 1 - 1 - 1$$

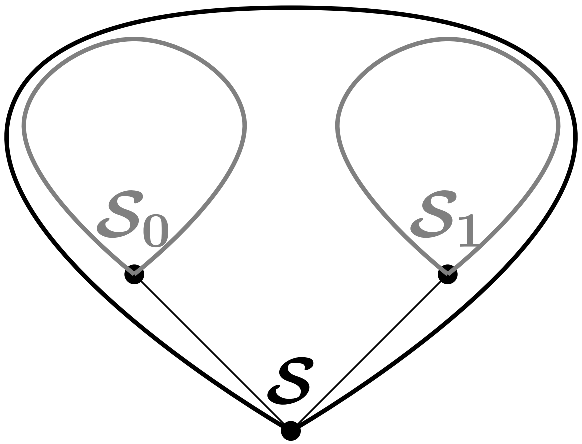
Where the last -1 comes from the fact that the single state of λ is at level $D = 0$ so $p(\lambda) = 1$.

Model posterior for Context trees

Induction: Assume it holds for $D \leq D^*$. Now if $\mathcal{S} \in \mathbb{S}_{D^*+1}$ then

- $\mathcal{S} = \lambda$, i.e. root node only.
- $\mathcal{S}$ contains two trees on level 1, say $\mathcal{S}_0 \in \mathbb{S}_{D^*}$ and $\mathcal{S}_1 \in \mathbb{S}_{D^*}$. We have

$$\Delta_{D^*+1}(\mathcal{S}) = 1 + \Delta_{D^*}(\mathcal{S}_0) + \Delta_{D^*}(\mathcal{S}_1)$$



Model posterior for Context trees

- We repeat: $\mathcal{S}$ contains two trees on level 1, say $\mathcal{S}_0 \in \mathbb{S}_{D^*}$ and $\mathcal{S}_1 \in \mathbb{S}_{D^*}$. We have

$$\Delta_{D^*+1}(\mathcal{S}) = 1 + \Delta_{D^*}(\mathcal{S}_0) + \Delta_{D^*}(\mathcal{S}_1)$$

$$\sum_{\mathcal{S} \in \mathbb{S}_{D^*+1}} 2^{-\Delta_{D^*+1}(\mathcal{S})} = 2^{-1} + \sum_{\mathcal{S}_0 \in \mathbb{S}_{D^*}} \sum_{\mathcal{S}_1 \in \mathbb{S}_{D^*}} 2^{-1-\Delta_{D^*}(\mathcal{S}_0)-\Delta_{D^*}(\mathcal{S}_1)}$$

$$= 2^{-1} + 2^{-1} \underbrace{\sum_{\mathcal{S}_0 \in \mathbb{S}_{D^*}} 2^{-\Delta_{D^*}(\mathcal{S}_0)}}_{=1} \underbrace{\sum_{\mathcal{S}_1 \in \mathbb{S}_{D^*}} 2^{-\Delta_{D^*}(\mathcal{S}_1)}}_{=1}$$

$$= 2^{-1} + 2^{-1} = 1$$

Model posterior for Context trees

We now show that the weighted sequence probability

$$p(x^N) = \sum_{\mathcal{S} \in \mathcal{S}_D} p(\mathcal{S}) p(x^N | \mathcal{S}),$$

is produced by the weighting procedure of CTW, so

$$p(x^N) = P_w^\lambda.$$

Model posterior for Context trees

We shall prove this using (mathematical) induction.

First assume $D = 0$: $\mathcal{S}_0 = \{\lambda\}$, so the only tree in the set consists of a root only. Therefor $\Delta_0(\lambda) = 0$. So,

$$\begin{aligned} p(x^N) &= p(\lambda)p(x^N|\lambda) \\ &= 2^0 P_e(n(0|x^N), n(1|x^N)) \\ &= P_w^\lambda, \end{aligned}$$

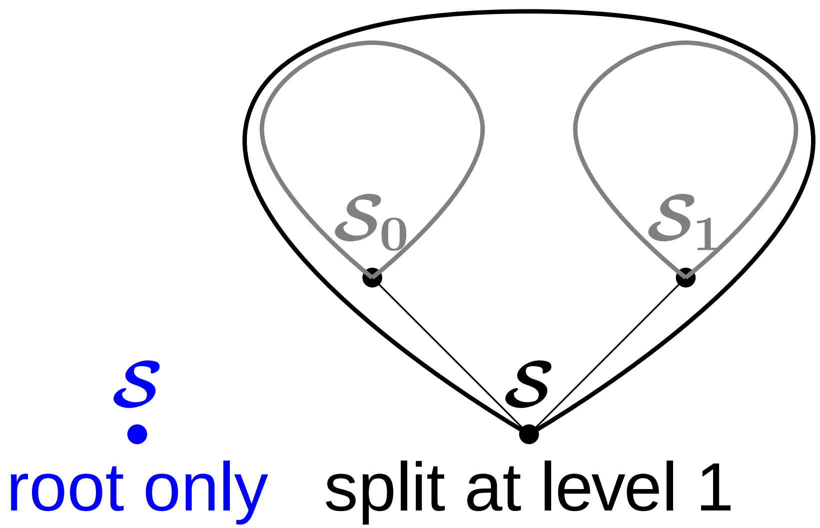
because λ is also a leaf and in a leaf $P_w = P_e$.

Model posterior for Context trees

Now assume that for all $D \leq D^*$

$$\sum_{\mathcal{S} \in \mathbb{S}_D} p(\mathcal{S}) p(x^N | \mathcal{S}) = P_w^{\lambda}$$

The tree $\mathcal{S}$ is either the root only or it consists of a root plus two trees, $\mathcal{S}_0$ and $\mathcal{S}_1$, on level one.



Model posterior for Context trees

$$\sum_{\mathcal{S} \in \mathbb{S}_{D^*+1}} p(\mathcal{S}) p(x^N|\mathcal{S}) =$$

$$= 2^{-1} P_e(n(0|x^N), n(1|x^N)) +$$

$$\sum_{\mathcal{S} \in \mathbb{S}_{D^*+1} : \mathcal{S} \neq \lambda} p(\mathcal{S}) p(x^N|\mathcal{S})$$

Model posterior for Context trees

$$\sum_{\mathcal{S} \in \mathbb{S}_{D^*+1} : \mathcal{S} \neq \lambda} p(\mathcal{S}) p(x^N | \mathcal{S}) =$$

$$\sum_{\mathcal{S}_0 \in \mathbb{S}_{D^*}} \sum_{\mathcal{S}_1 \in \mathbb{S}_{D^*}} \frac{1}{2} 2^{-\Delta_{D^*}(\mathcal{S}_0)} 2^{-\Delta_{D^*}(\mathcal{S}_1)} \times$$

$$p(x^N_{|0} | \mathcal{S}_0) p(x^N_{|1} | \mathcal{S}_1)$$

$$= \frac{1}{2} \sum_{\mathcal{S}_0 \in \mathbb{S}_{D^*}} 2^{-\Delta_{D^*}(\mathcal{S}_0)} p(x^N_{|0} | \mathcal{S}_0) \times$$

$$\sum_{\mathcal{S}_1 \in \mathbb{S}_{D^*}} 2^{-\Delta_{D^*}(\mathcal{S}_1)} p(x^N_{|1} | \mathcal{S}_1)$$

$$= \frac{1}{2} P_w^0 P_w^1$$

Model posterior for Context trees

And so we find

$$\sum_{\mathcal{S} \in \mathbb{S}_{D^*+1}} p(\mathcal{S}) p(x^N|\mathcal{S}) =$$

$$= \frac{1}{2} P_e(n(0|x^N), n(1|x^N)) + \frac{1}{2} P_w^0 P_w^1$$

$$= P_w^\lambda$$

Model posterior for Context trees

Thus we can compute the a-posteriori model probability.

$$p(\mathcal{S}|x^N) = \frac{p(\mathcal{S})p(x^N|\mathcal{S})}{p(x^N)}$$

Model posterior for Context trees

Thus we can compute the a-posteriori model probability.

$$p(\mathcal{S}|x^N) = \frac{p(\mathcal{S})p(x^N|\mathcal{S})}{p(x^N)}$$

Model posterior for Context trees

Thus we can compute the a-posteriori model probability.

$$p(\mathcal{S}|x^N) = \frac{2^{-\Delta_D(\mathcal{S})} p(x^N|\mathcal{S})}{p(x^N)}$$

Model posterior for Context trees

Thus we can compute the a-posteriori model probability.

$$p(\mathcal{S}|x^N) = \frac{2^{-\Delta_D(\mathcal{S})} \prod_{s \in \mathcal{S}} P_e(n(s0|x^N), n(s1|x^N))}{p(x^N)}$$

Model posterior for Context trees

Thus we can compute the a-posteriori model probability.

$$p(\mathcal{S}|x^N) = \frac{2^{-\Delta_D(\mathcal{S})} \prod_{s \in \mathcal{S}} P_e(n(s0|x^N), n(s1|x^N))}{P_w^\lambda}$$

Model posterior for Context trees

Thus we can compute the a-posteriori model probability.

$$p(\mathcal{S}|x^N) = \frac{2^{-\Delta_D(\mathcal{S})} \prod_{s \in \mathcal{S}} P_e(n(s0|x^N), n(s1|x^N))}{P_w^\lambda}$$

So, we can use the same computations as in the CTW.

An efficient way to find the Bayesian MAP model exists, but its discussion is not a part of this course.