Context-Tree estimation via algorithm Context and Penalized Maximum Likelihood

Aurélien Garivier, CNRS Telecom ParisTech

June 19, 2008

Outline

Outline

1. Context Tree Sources
   - Variable Length Memory
   - Definition and Properties

2. Context Tree estimation: Two Algorithms
   - Algorithm Context
   - Penalized Maximum Likelihood

3. Consistency results and perspectives

Need for adaptive memory

- Data compression:

t r y i n g _ v a n i l l a _ q u i e t

- Linguistic:

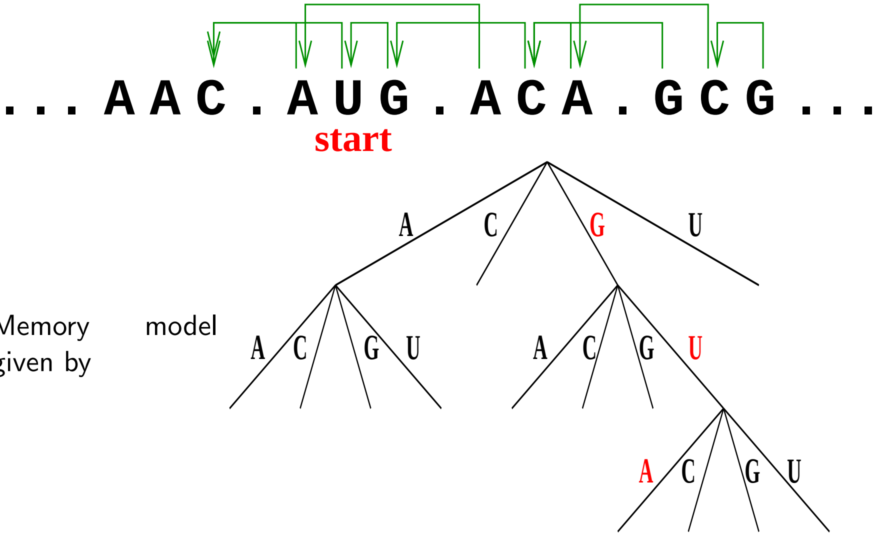
L o n g t e m p s , j e m e s u i s c o u c h é d e b o n n e h e u r e . P a r f o i s , . . .

- Renewal Processes:

1 0 0 1 0 1 0 0 0 0 1 1 0 0 1 . . .

- Music, biology, ingeneering, . . .

Biological sequences



Memory model given by

Large memory: limits of Markov models

- Data compression: $|A| = 2,\ k = 8 \implies \dim \Theta = 256$
- Biological sequences: $|A| = 4,\ k = 6 \implies \dim \Theta \approx 12000$
- Linguistic: $|A| = 3000,\ k = 10 \implies \dim \Theta = \dots$
- Renewal Processes: infinite memory

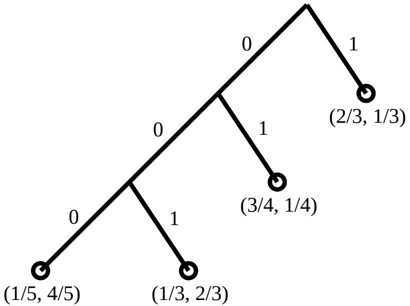
Need for more flexibility: larger memory only where necessary!

Context Tree Sources

A Context Tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$\begin{aligned} & P(X_1^4 = 00110 | X_{-1}^0 = 10) \\ = \quad & P(X_1 = 0 | X_{-1}^0 = 10) \\ \times \quad & P(X_2 = 0 | X_{-1}^1 = 100) \\ \times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) \\ \times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) \\ \times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) \end{aligned}$$



A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \dots, \theta_{|A|}^s \right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\} \in \mathbb{R}^{|T|(|A|-1)}$$

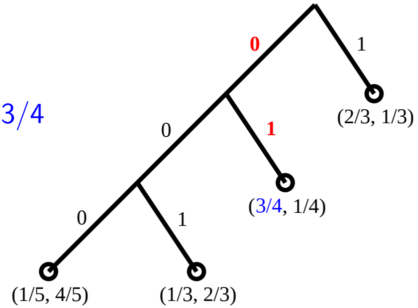
Context Tree Sources

A Context Tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned} = \quad & P(X_1 = 0 | X_{-1}^0 = 10) \qquad 3/4 \\ \times \quad & P(X_2 = 0 | X_{-1}^1 = 100) \\ \times \quad & P(X_3 = 1 | X_{-1}^2 = 1000) \\ \times \quad & P(X_4 = 1 | X_{-1}^3 = 10001) \\ \times \quad & P(X_5 = 0 | X_{-1}^4 = 100011) \end{aligned}$$



A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s \right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\} \in \mathbb{R}^{|T|(|A|-1)}$$

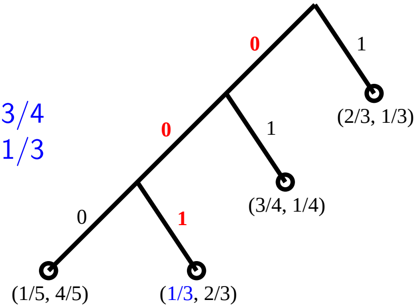
Context Tree Sources

A Context Tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{array}{rll} = & P(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & P(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & P(X_3 = 1 | X_{-1}^2 = 1000) & \\ \times & P(X_4 = 1 | X_{-1}^3 = 10001) & \\ \times & P(X_5 = 0 | X_{-1}^4 = 100011) & \end{array}$$



A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s\right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\} \in \mathbb{R}^{|T|(|A|-1)}$$

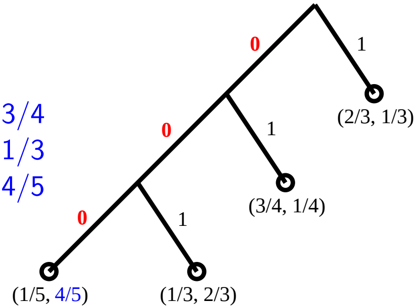
Context Tree Sources

A Context Tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{array}{llr} = & P(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & P(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & P(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\ \times & P(X_4 = 1 | X_{-1}^3 = 10001) & \\ \times & P(X_5 = 0 | X_{-1}^4 = 100011) & \end{array}$$



A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \ldots, \theta_{|A|}^s\right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\} \in \mathbb{R}^{|T|(|A|-1)}$$

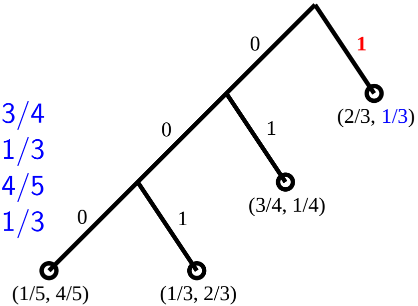
Context Tree Sources

A Context Tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{array}{lll} = & P(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & P(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & P(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\ \times & P(X_4 = 1 | X_{-1}^3 = 10001) & 1/3 \\ \times & P(X_5 = 0 | X_{-1}^4 = 100011) & \end{array}$$



A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \dots, \theta_{|A|}^s\right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\} \in \mathbb{R}^{|T|(|A|-1)}$$

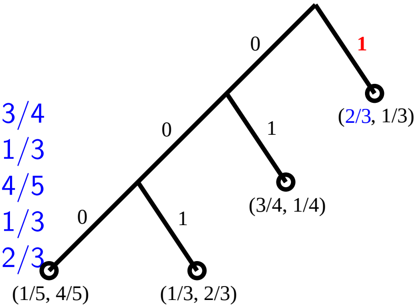
Context Tree Sources

A Context Tree Source or Variable Length Markov Chain is a Markov Chain whose order is allowed to depend on the past data.

$$T = \{1, 10, 100, 000\}$$

$$P(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$=$	$P(X_1 = 0 \| X_{-1}^0 = 10)$	$3/4$
$\times$	$P(X_2 = 0 \| X_{-1}^1 = 100)$	$1/3$
$\times$	$P(X_3 = 1 \| X_{-1}^2 = 1000)$	$4/5$
$\times$	$P(X_4 = 1 \| X_{-1}^3 = 10001)$	$1/3$
$\times$	$P(X_5 = 0 \| X_{-1}^4 = 100011)$	$2/3$



A stationary context tree source is parameterized by

$$\Theta_T = \left\{ \left(\theta_1^s, \dots, \theta_{|A|}^s\right) : s \in T, \sum_{i=1}^{|A|} \theta_i^s = 1 \right\} \in \mathbb{R}^{|T|(|A|-1)}$$

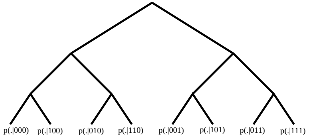
Definition, Existence and Uniqueness

- Formally, a context tree source P_T is defined by
  1. a Complete Suffix Dictionary T = a set of words on alphabet A such that:
  $$\forall x_{-\infty}^0 \in A^{\mathbb{Z}_-}, \exists! L \in \mathbb{N} : x_{-L}^0 \in T;$$
  2. a family of $|T|$ conditional distributions $\{P_T(\cdot|s) : s \in T\}$.
  $$\forall x_{-\infty}^0 \in A^{\mathbb{Z}_-}, P_T(\cdot|x_{-\infty}^0) = P_T(\cdot|x_{-L}^0).$$
- **Theorem [Fernandez-Galves '02]**: if $\sum_{a\in A} \inf_{s\in T} P(a|s) > 0$ and if the
$$\beta_k = \max_{a\in A} \sup \left\{ |P(a|s) - P(a|t)| : (s,t) \in T^2 \text{ and } s_{-k+1}^0 = t_{-k+1}^0 \right\}$$
are summable, then there exists a unique stationary context tree source with the given conditional probabilities.

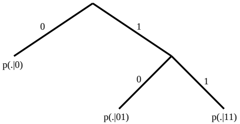
CTS versus Markov Chains

- Markov chains of order r are context tree sources corresponding to a complete tree of depth r.
  Markov chain of order 3

$$M = \begin{pmatrix} p\left(.|000\right) \\ p\left(.|100\right) \\ \vdots \\ p\left(.|111\right) \end{pmatrix} \quad \Rightarrow$$



- Finite context tree sources of depth d are Markov Chains of order d.



$$\Rightarrow \quad M = \begin{pmatrix} p\left(.|0\right) \\ p\left(.|0\right) \\ p\left(.|01\right) \\ p\left(.|11\right) \end{pmatrix}$$

$\Longrightarrow$ much more flexibility: large number of models per dimension.

Likelihood and ML estimates

- Expression of the likelihood:

$$P_T(x_1^n|x_{-\infty}^0) = \prod_{i=1}^{n} P_T(x_i|x_{i-L_i}^{i-1}) = \prod_{s\in T}\prod_{i\in I_s} P_T(x_i|s),$$

  where $I_s = \{i \in \{1,\dots,n\} : x_{i-|s|}^{i-1} = s\}$.
- Maximum likelihood estimate in model T: for all $s \in T$,

$$\hat{P}_T(\cdot|s) = \frac{N(sa)}{N(s)},$$

  where $N(s) = \sum_{i=1}^n \mathbb{1}_{X_{i-|s|}^{i-1}=s} = |I_s|$.

Outline

Algorithm Context: Description

- Introduced by Rissanen in 1981.
- For each $s \in A^*$, compute

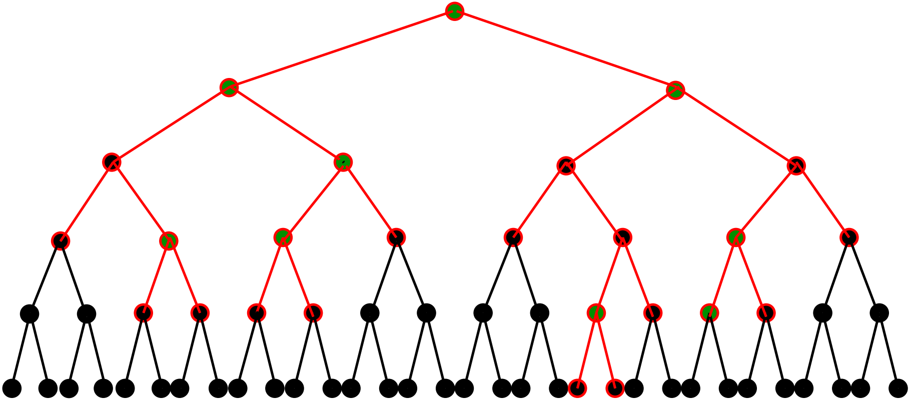
$$\delta(s) = \max_{a \in A} \left\| \hat{P}(\cdot|s) - \hat{P}(\cdot|as) \right\| .$$

- Keep all $t \in A^*$ such that

$$\exists u \in A^* : \delta(us) \geq \epsilon(n)$$

as internal nodes of $\hat{T}_C$. Thus, $\hat{T}_C$ is made of all active nodes, their ancestors and immediate children.

Algorithm Context: Illustration



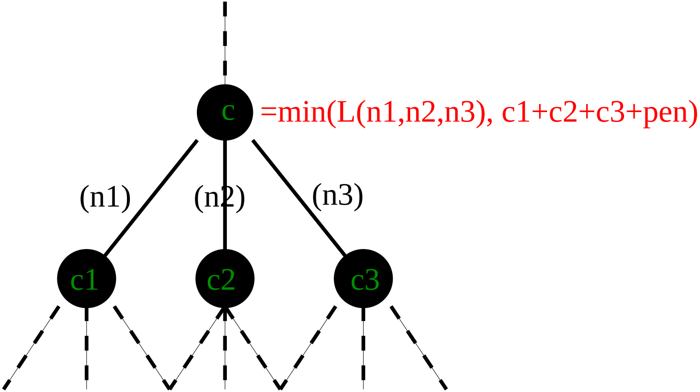
active nodes - estimated tree $\hat{T}_C$

PML: Description

- Choose
$$\hat{T}_{pml} = \arg\max_{T} \log \hat{P}_T(x_1^n | x_{-\infty}^0) + pen(n, T),$$
where $pen(n, T)$ is a penalty function (growing with n and $|T|$).
- MDL: universal code length, BIC penalty
$$pen(n, T) = \frac{|T|(|A| - 1)}{2} \log n.$$
- MDL: Krichevski-Trofimov mixture
$$\hat{T}_{KT} = \arg\max_{T} \log KT_T(x_1^n | x_{-\infty}^0).$$

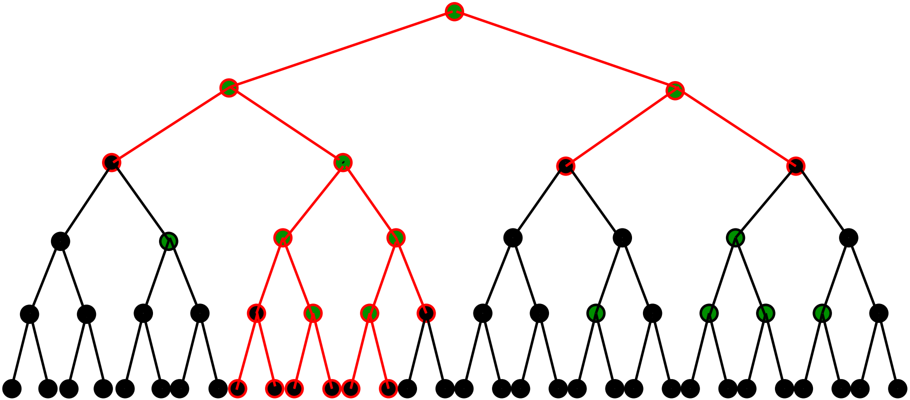
Effective computation

A node s is active if coding $x_{I(s)}$ is cheaper with than without memory



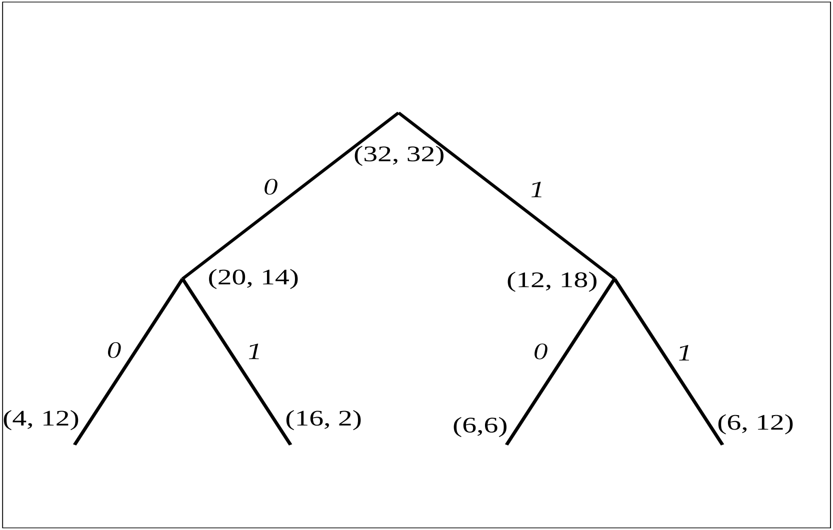
Construction of $\hat{T}_{pml}$: Starting from the root, keep only active nodes as internal nodes.

PML: Illustration

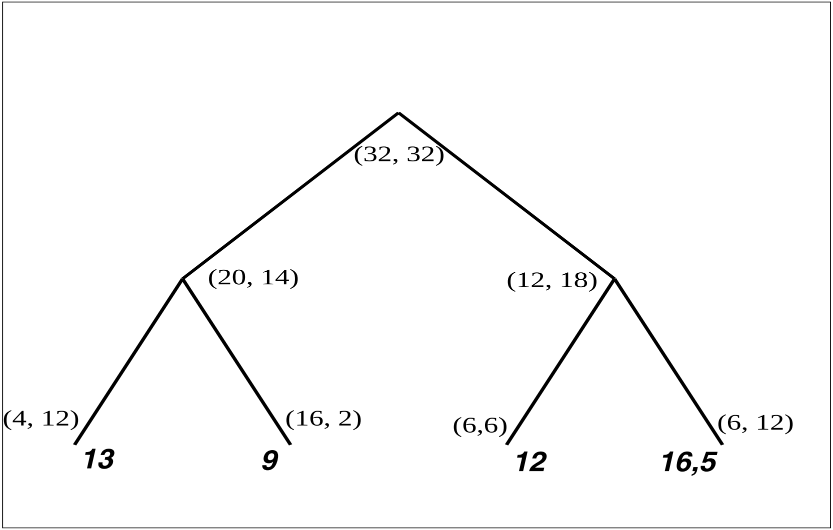


active nodes - estimated tree $\hat{T}_{pml}$

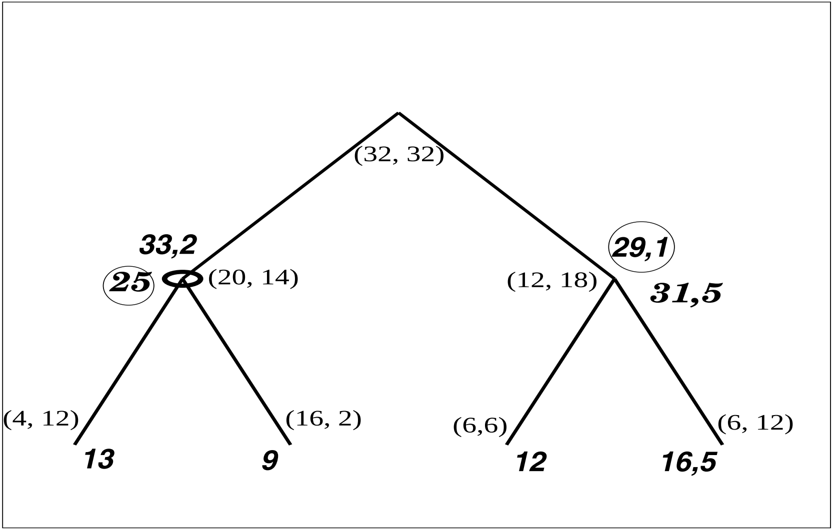
PML: Example



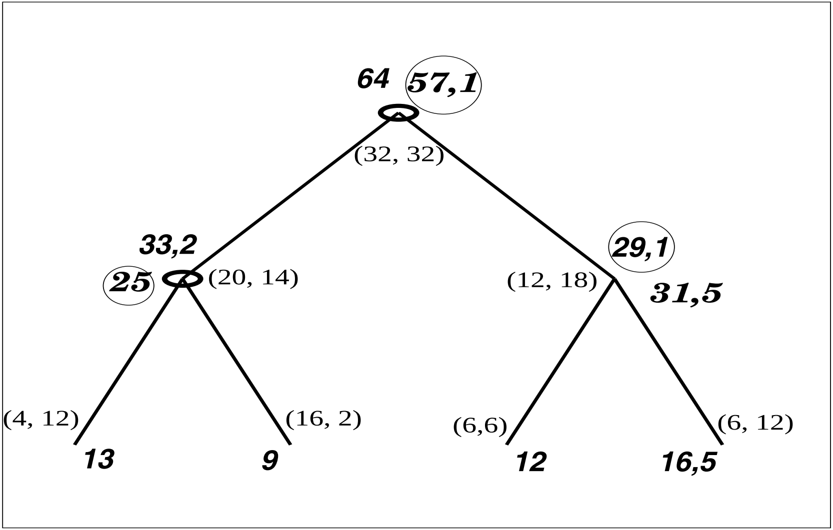
PML: Example



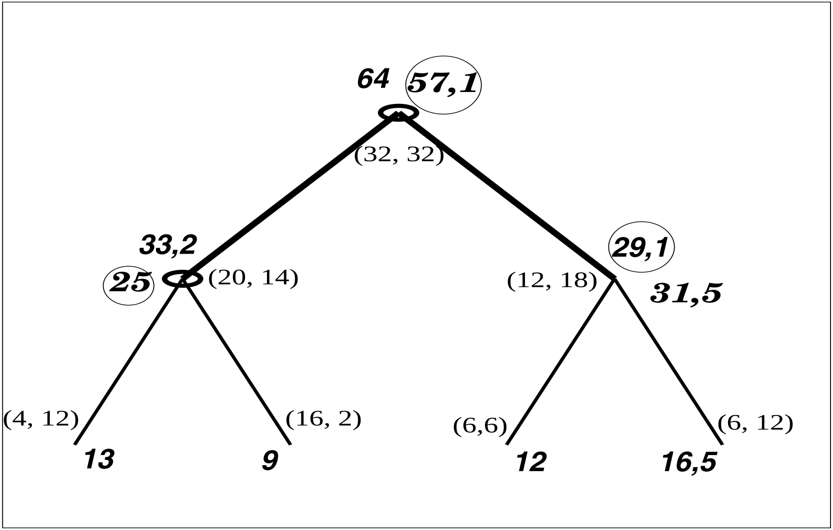
PML: Example



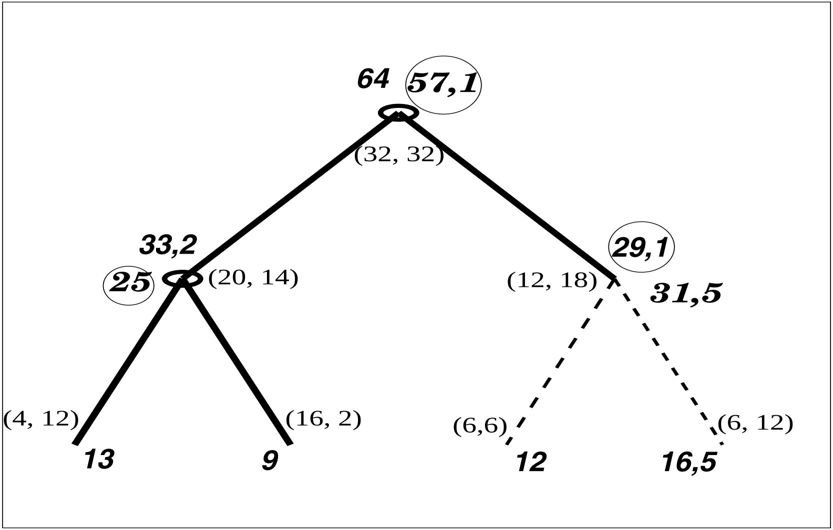
PML: Example



PML: Example



PML: Example

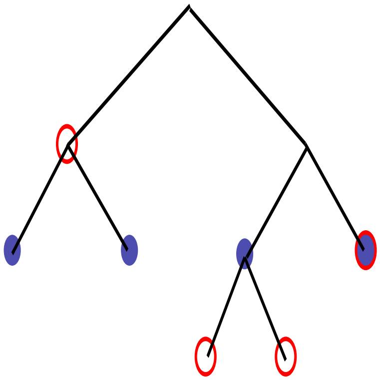


Outline

Under- and Over-estimation

Two possible errors:

1. under-estimation:
   $\exists s \in T_0 : s \notin \hat{T}$
   $\Longrightarrow$ easily avoided (large deviations), exponential rates
2. over-estimation:
   $\exists s \in \hat{T} : s \notin T_0$
   $\Longrightarrow$ more difficult, no exponential rate [Finesso '92]



Asymptotic results

- **Theorem [Rissanen '81, ...]**: For a finite tree T_0, if $\epsilon(n) = C \log(n)/n$, then as n goes to inifinity

$$P(\hat{T}_C \neq T_0) \to 0.$$

- **Theorem [Csiszár and Talata '06, Garivier '06]**: If K is a positive integer and if $\hat{T}_{pml}$ is maximizer of the penalized maximum likelihood among all trees with depth $D(n) = o(\log n)$, then

$$\hat{T}_{pml}^{|K} = T_0^{|K}$$

eventually almost surely as $n \to \infty$. For a finite tree T_0, there is no need to restrict the maximization.

- Non-asymptotic probability of estimation errors? [Galves, Maume-Deschamps, Leonardi] give non-asymptotic results, but relying under unpleasant conditions ($\forall a \in A, P(a|s) > \epsilon$).

Tools

- For the Context algorithm, need to control

$$\|\hat{P}(\cdot|s) - P(\cdot|s)\|.$$

- For the PML, need to control

$$KL\left(\hat{P}(\cdot|s), P(\cdot|s)\right).$$

- In both cases, amounts to study the maxima of the martingale

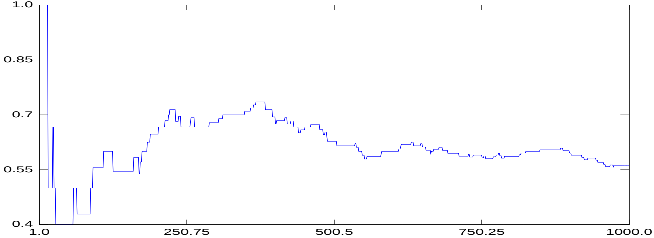
$$Z_t = \frac{1}{\sqrt{N_t(s)}} \sum_{u=1}^{t} (\mathbb{1}_{\{X_u=a} - P(a|s)\})\mathbb{1}_{\{X_{u-|s|}^{u-1}=s\}}.$$

- The asymptotic consistency of $\hat{T}_{pml}$ relies Csiszár's "typicality results" = uniform Law of Iterated Logarithm.

What happens in each node ?

For each possible context s, the ML estimate of the conditional distribution is given by

$$\forall a \in A, \hat{P}(a|s) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k=a\}}\mathbb{1}_{\{X_{k-|s|}^{k-1}=s\}}$$



Hoeffding / Bernstein bounds unsatisfactory (for $P(a|s)$ very small), need the large deviations bou[illegible]ut random number of summands.

Perspectives

- Are hypotheses like $P(a|s) > \epsilon$ really necessary?
- Better bounds for the martingales?
- How small may we choose $\epsilon(n)$ and $pen(n, T)$?
- In the inifinite case, is always (eventually almost surely) a subtree of T_0 selected?
- For prediction or compression, what estimated tree is the best?
  - if there are few observations?
  - if the context tree T_0 is infinite?
  - if the source P is not a context tree?

Perspectives

- Are hypotheses like $P(a|s) > \epsilon$ really necessary?
- Better bounds for the martingales?
- How small may we choose $\epsilon(n)$ and $pen(n, T)$?
- In the infinite case, is always (eventually almost surely) a subtree of T_0 selected?
- For prediction or compression, what estimated tree is the best?
  - if there are few observations?
  - if the context tree T_0 is infinite?
  - if the source P is not a context tree?

Thank you for your attention!

References

- Bühlmann, P. and Wyner, A. (1999) Variable length Markov chains, Ann. Statist. Volume 27, Number 2 (1999), 480-513.
- Csiszár, I. and Talata, Z. (2006) Context tree estimation for not necessarily finite memory processes, via BIC and MDL. IEEE Transactions on Information Theory 52(3): 1007-1016
- Garivier A. (2006) Consistency of the unlimited BIC Context Tree Estimator, IEEE-Transactions on Information Theory in October 2006 (see also Context Tree Weighting algorithm and Renewal Processes, IEEE-IT December 2006).
- Csiszár I.(2002) Large-scale typicality of Markov sample paths and consistency of MDL Order estimators. IEEE Transactions on Information Theory 48(6): 1616-1628
- Galves, A. and Leonardi, F. (2007) Exponential inequalities for empirical unbounded context trees. To appear in Progress in Probability, Birkhäuser. ArXiv: math/0710.5900
- Leonardi, F. (2007) On the rate of convergence of penalized likelihood context tree estimators. ArXiv: math/0701810
- Leonardi, F., Matioli, S.R., Armelin, H.A. and Galves, A. (2007) Detecting phylogenetic relations out from sparse context trees. ArXiv: math/0804.4279
- A. Galves, V. Maume-Deschamps, B.Schmitt Exponential inequalities for VLMC empirical trees. ESAIM Prob. Stat., (2008), 12, 119–229.