Context Tree Models and Renewal Processes

Aurélien Garivier

Université Paris Sud

Orsay

Outline

- Universal Coding
- Context Tree Weighting Algorithm
- Redundancy of CTW on Renewal Processes

Basics of Information Theory

- Let X be a stochastic process on the finite alphabet A, with stationary ergodic distribution $\mathbb{P}$.
- For $n \in \mathbb{N}^*$ and the coding function $C_n : A^n \to \{0,1\}^*$, the *average coding rate* is

$$\frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[l(C_n(x))\right] \geqslant \frac{1}{n}H_n(X) = \frac{1}{n}\mathbb{E}_{\mathbb{P}}\left[-\log \mathbb{P}(X_1^n)\right] \to H(X).$$

- Kraft inequality : $\sum_{x\in A^n} 2^{-l(C_n(x))} \leqslant 1$
- Arithmetic coding $\Rightarrow$ correspondence between coding functions and probability distributions.
- $-\log Q(x)$=*code length* for x with *coding distribution* Q.

Universal Coding

- $\mathbb{P}$ known only to belong to a class of sources $\mathcal{S} = \{\mathbb{P}_\theta : \theta \in \Theta\}$.
- Ex: Markov chains, general stationary ergodic processes.
- *Two- steps* codes : code θ and then $x|\theta$.
- *Mixture* codes : coding distribution = weighted average of the $(\mathbb{P}_\theta)_{\theta\in\Theta}$.
- Ex: memoryless sources $\Theta = \{\theta \in [0,1]^A : \sum_{a\in A} \theta_a = 1\}$.
  Krichevski-Trofimov ($\mathcal{KT}$) mixture

$$\mathcal{KT}(x_1^n) = \int_{\theta\in\Theta_A} P_\theta(x_1^n) \frac{\Gamma\left(\frac{|A|}{2}\right)}{\sqrt{|A|}\Gamma(\frac{1}{2})^{|A|}} \prod_{a\in A} \theta_a^{-1/2} \mathrm{d}\theta_a .$$

Redundancy

- *Pointwise* redundancy $R(C_n|P)(x) = l(C_n(x)) + \log \mathbb{P}(x)$
  Maximal redundancy $R^*(C_n|P) = \max_x R(C_n|P)(x)$.
  Minimax redundancy in class S :

$$R^*_n(\mathcal{S}) = \inf_{C_n} \sup_{\mathbb{P}\in\mathcal{S}} R^*(C_n|P)$$

- For parametric classes with k free parameters (like Markov Chains), $R^*_n(\mathcal{S}) = \frac{k}{2}\log n + O(1)$
- For the whole class of stationary ergodic processes, no universal rate (Shields '93).
- Ex: the $\mathcal{KT}$ mixture is a almost optimal since :
  $-\log \mathcal{KT}(x) \leqslant \inf_{\theta\in\Theta} -\log P_\theta(x) + \frac{1}{2}(|A|-1)\log n + |A|/2.$

Outline

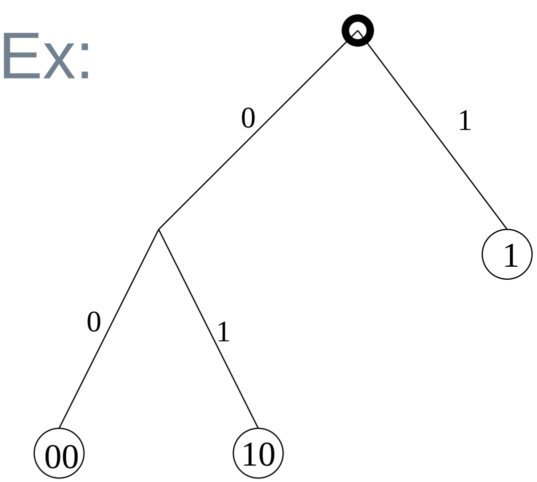
- Universal Coding
- Context Tree Weighting Algorithm
- Redundancy of CTW on Renewal Processes

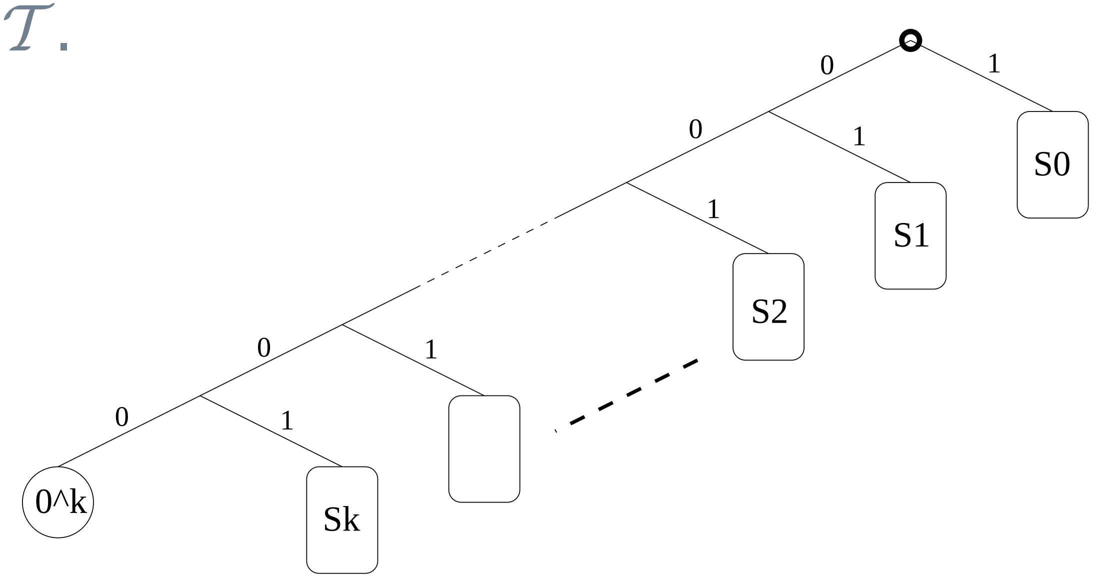
Complete suffix dictionnary

- $\mathcal{T}$ is a *Complete Suffix Dictionnary* (CSD) iff

$$\forall x_{-\infty}^{0} \in A^{\mathbb{Z}-}, \exists ! k \in \mathbb{N} : x_{-k}^{0} \in \mathcal{T}.$$

- For $x_{-\infty}^{0} \in A^{\mathbb{Z}-}$, we call $\mathcal{T}(x)$ its suffix in $\mathcal{T}$.
- Any CSD can be represented as a trie whose *leaves* are the elements of $\mathcal{T}$.
- Ex:





$$\mathcal{T} = \{00, 10, 1\} \qquad \mathcal{T} = \{0^k\} \cup \{10^j : 0 \le j < k\}$$

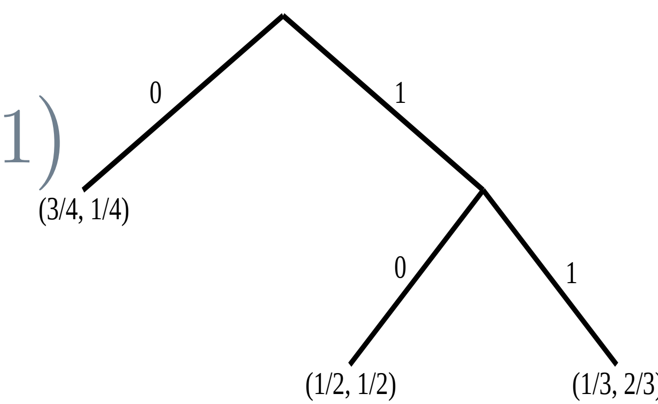
Context Tree Sources

- Let $\mathcal{T}$ be a CSD and $p = (p(.|w))_{w\in\mathcal{T}}$ be $|\mathcal{T}|$ probability distributions on A.
- The *Context tree source* $\mathbb{P}_{\mathcal{T},p}$ is the stationary distribution on $A^{\mathbb{Z}}$ defined by

$$\mathbb{P}_{\mathcal{T},p}\left(X_1 = x_1 | X_{-\infty}^0 = x_{-\infty}^0\right) = p\left(x_1 | \mathcal{T}\left(x_{-\infty}^0\right)\right).$$

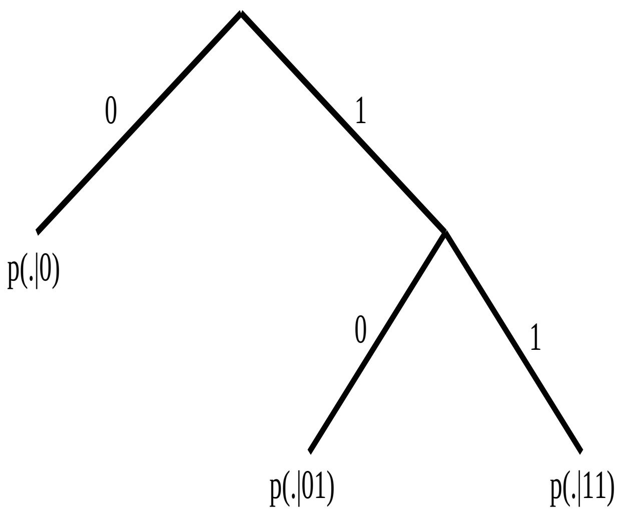
- Ex:

$$\mathbb{P}\left(X_1^4 = 1001 | X_{-\infty}^0 = \ldots 01\right)$$
$$= \frac{1}{2} \times \frac{1}{3} \times \frac{3}{4} \times \frac{1}{4}$$



Finite CTS are Markov chains

- The depth of the trie = Markovian order.



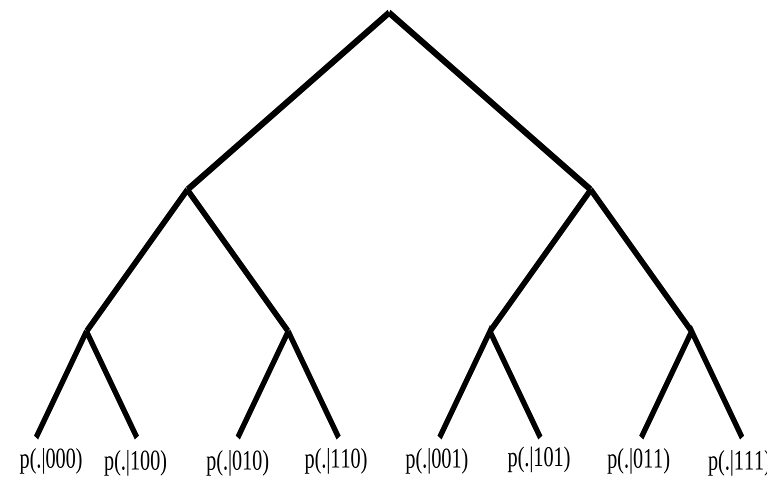
$$\rightarrow M = \begin{pmatrix} p\left(.|0\right) \\ p\left(.|0\right) \\ p\left(.|01\right) \\ p\left(.|11\right) \end{pmatrix}$$

- *Variable Length* Markov Chains : fewer free parameters for a given memory size.

Markov Chains are CTS

- Corresponding trie = the full tree of depth equal to the Markovian order :

$$M = \begin{pmatrix} p(.|000) \\ p(.|100) \\ \vdots \\ p(.|111) \end{pmatrix} \rightarrow$$



- $\Longrightarrow$ CTS have the approximation power of Markov chains to approach every stationary ergodic source.
- They are not more complicated to use.

Expression of the likelihood

- As $x = \bigodot_{s \in \mathcal{T}} \mathcal{T}(x, s)$, the likelihood is:

$$\begin{aligned} P_{\mathcal{T},p}(x_1^n | x_{-\infty}^0) &= \prod_{i=1}^{n} p(x_i | \mathcal{T}(x_{-\infty}^{i-1})) \\ &= \prod_{s \in \mathcal{T}} p_s\left(\mathcal{T}(x, s)\right) \end{aligned}$$

- Hence the expression of the Maximum Likelihood:

$$-\log \hat{P}_{\mathcal{T}}(x) = \sum_{s \in \mathcal{T}} H(\mathcal{T}(x, s))$$

$\mathcal{KT}$ mixture for a given model

- We define similarily:

$$\mathcal{KT}_{\mathcal{T}}(x_1^n|x_{-\infty}^0) \;=\prod_{s\in\mathcal{T}}\mathcal{KT}\left(\mathcal{T}(x,s)\right)$$

- **Theorem** : there is a constant C such that :

$$-\log_2 \mathcal{KT}_{\mathcal{T}}(x_1^n|x_{-\infty}^{-1}) \;\leqslant\; \inf_{\theta\in\Theta^{\mathcal{T}}} -\log_2 \mathbb{P}_{\mathcal{T},\theta}(x_1^n|x_{-\infty}^{-1}) + |\mathcal{T}|\frac{|A|-1}{2}\log\left(\frac{n}{|T|}\right) + C\,|\mathcal{T}|$$

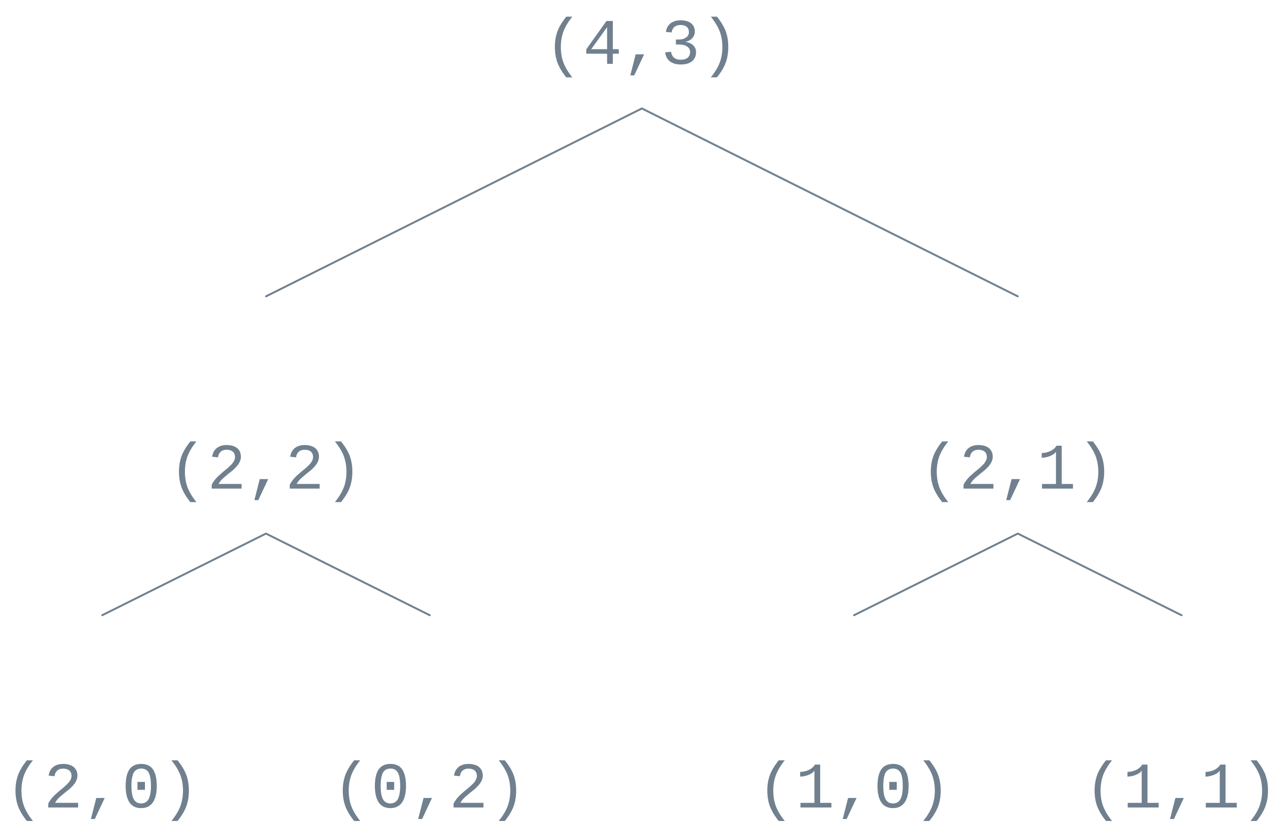
- $\Longrightarrow$ minimax redundancy.

Context Tree Weighting

- We only consider the *binary* case here.
- Prior π on the trees : there are $Catalan_s = \frac{1}{s+1}\binom{2s}{s}$ trees with $s+1$ leaves, thus we can choose:
$$\pi(\mathcal{T}) = 2^{-2|\mathcal{T}|+1}.$$
- We define the *double mixture*
$$\mathcal{CTW}(x) = \sum_{\mathcal{T}\in\mathbb{T}} \mathcal{KT}_{\mathcal{T}}(x)\pi(\mathcal{T}).$$
- This is a probability distribution on each A^n
$\Longrightarrow$ we can use arithmetic coding.
- Efficiency : oracle inequality
$$-\log \mathcal{CTW}(x) \leqslant \inf_{\mathcal{T},p} -\log P_{\mathcal{T},p}(x) + |\mathcal{T}|\log\left(\frac{n}{|\mathcal{T}|}\right) + 2|\mathcal{T}|$$

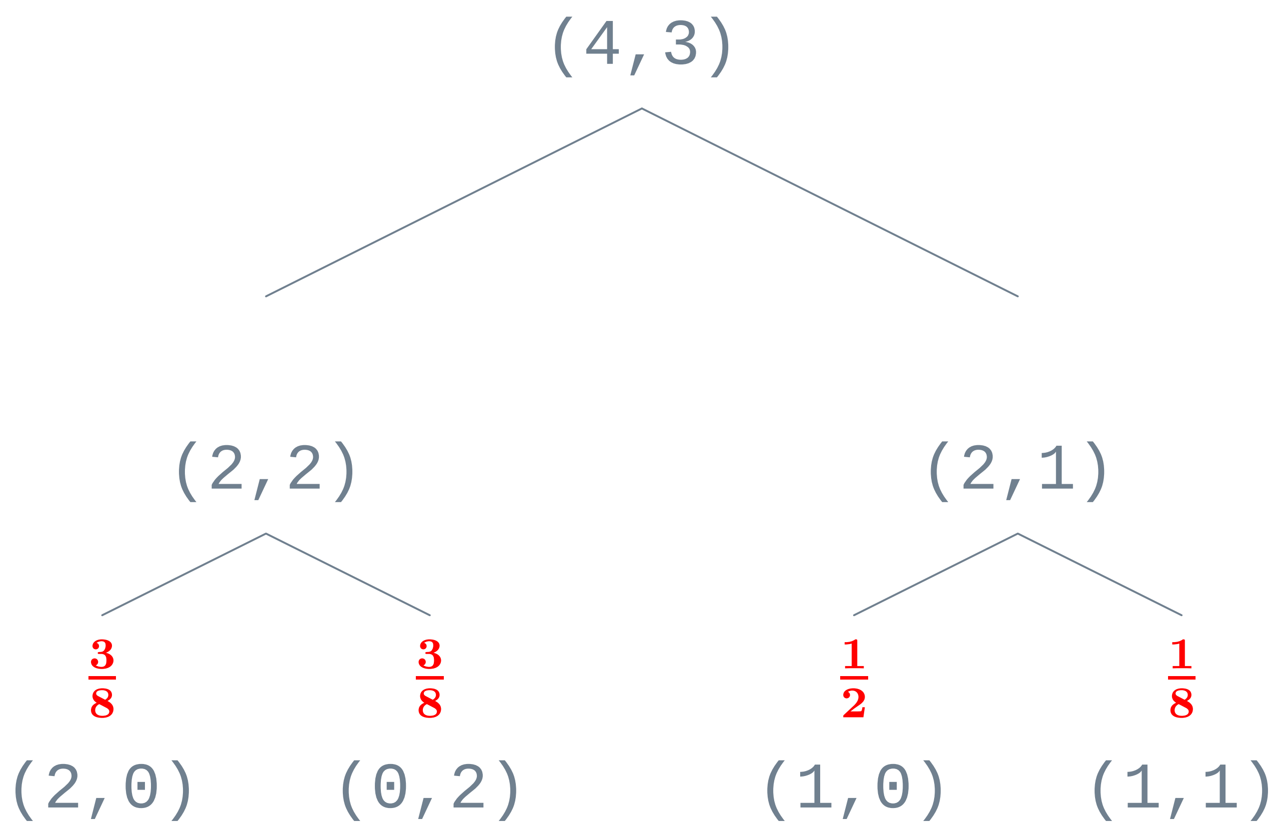
Algorithm to compute $\mathcal{CTW}(x)$

In each node, compute the arithmetical mean of a *selfcost* and a *subcost*.



Algorithm to compute $\mathcal{CTW}(x)$

In each node, compute the arithmetical mean of a *selfcost* and a *subcost*.



Algorithm to compute $\mathcal{CTW}(x)$

In each node, compute the arithmetical mean of a *selfcost* and a *subcost*.

(4,3)

$$\frac{1}{2}\left(\frac{3}{8}\times\frac{3}{8}+\frac{3}{128}\right)=\frac{21}{256} \qquad \frac{1}{2}\left(\frac{1}{2}\times\frac{1}{8}+\frac{1}{16}\right)=\frac{1}{16}$$

(2,2) (2,1)

$$\frac{3}{8} \qquad \frac{3}{8} \qquad \frac{1}{2} \qquad \frac{1}{8}$$

(2,0) (0,2) (1,0) (1,1)

Algorithm to compute $\mathcal{CTW}(x)$

In each node, compute the arithmetical mean of a *selfcost* and a *subcost*.

$$\frac{1}{2}\left(\frac{21}{256} \times \frac{1}{16} + \frac{5}{2058}\right) = \frac{31}{8192}$$

(4,3)

$$\frac{1}{2}\left(\frac{3}{8} \times \frac{3}{8} + \frac{3}{128}\right) = \frac{21}{256}$$

(2,2)

$$\frac{1}{2}\left(\frac{1}{2} \times \frac{1}{8} + \frac{1}{16}\right) = \frac{1}{16}$$

(2,1)

$\frac{3}{8}$ (2,0) $\frac{3}{8}$ (0,2) $\frac{1}{2}$ (1,0) $\frac{1}{8}$ (1,1)

Outline

- Universal Coding
- Context Tree Weighting Algorithm
- Redundancy of CTW on Renewal Processes

Renewal Processes

- X is a ***Renewal Process*** if it takes its values in $A = \{0, 1\}$ and if the distances between successive '1' in X are iid random variables on $\mathbb{N}^*$ with distribution Q.

$$1\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 1$$

$$1 \quad 4 \quad 1 \quad 2 \quad 5$$

- Similarily, ***Markovian Renewal processes*** are defined thru a Markovian kernel Q.

Properties

- A memoryless $\mathcal{B}(p)$ process is a RP with geometric $\mathcal{G}(p)$ renewal times;
- If Q is bounded, the RP is a Markov Chain (better : a CTS).
- If $x_1^n = 0^{t_0-1}1\ 0^{t_1-1}1\ 0^{t_2-1}1\ \ldots\ 0^{t_N-1}1\ 0^{t_{N+1}-1}$, and if the renewal distribution is Q, then letting $R_Q(t) = \sum_{u=t}^{\infty} Q(u)$ we have:

$$\mathbb{P}_Q^{\mathcal{R}}(x) = \left(\frac{1}{\mathbb{E}[Q]} R_Q(t_0)\right) \prod_{i=1}^{N} Q(t_i)\ R_Q\left(t_{N+1}\right).$$

Minimax Redondancy - Csiszár–Shields '96

Theorem : In the class $\mathcal{R}$ of Renewal Processes, there are two positive constants c et C such that $\forall n \in \mathbb{N}$:

$$c\sqrt{n} \leqslant R_n^*(\mathcal{R}) \leqslant C\sqrt{n}.$$

Theorem : In the class $\mathcal{MR}$ of Markovian Renewal Processes, there are two positive constants c et C such that $\forall n \in \mathbb{N}$:

$$cn^{2/3} \leqslant R_n^*(\mathcal{MR}) \leqslant Cn^{2/3}.$$

$\Longrightarrow$ First example of *intermediate complexity* classes.

Redundancy of $\mathcal{CTW}$ on RP - Garivier '04

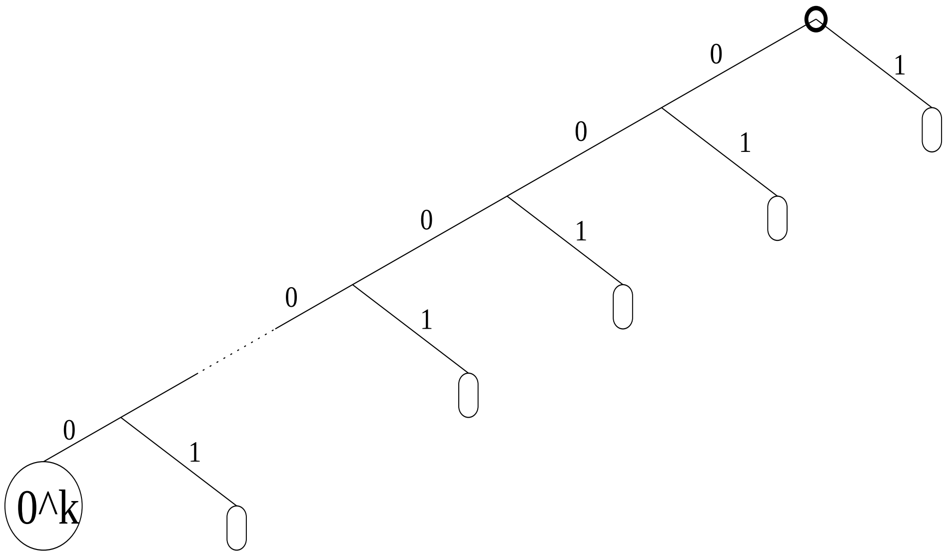
Theorem: There are two positive constants c et C such that for all $n \in \mathbb{N}$:

$$c\sqrt{n}\log n \leqslant R_n^*(\mathcal{CTW}, P.R.) \leqslant C\sqrt{n}\log n.$$

Theorem: There are two positive constants c et C telles que pour tout $n \in \mathbb{N}$:

$$cn^{2/3}\log n \leqslant R_n^*(\mathcal{CTW}, M.R.) \leqslant Cn^{2/3}\log n.$$

Outline for the upper-bound



The approximation tree sources "understands" $\mathbb{P}_Q$ until depth k.

- Approximation tree of depth $k = \sqrt{n}$;
- $-\log \mathcal{CTW}(x) \leqslant -\log \mathcal{KT}_T(x) + 2k + 1$;
- $-\log \mathcal{KT}(x) \leqslant -\log \hat{P}_{\mathcal{T}}(x) + \frac{2k+1}{2} \log n + 3k + 1$;
- $-\log \hat{P}_{\mathcal{T}}(x) \leqslant -\log \mathbb{P}_Q(x) - \log \mathcal{KT}(\mathcal{T}(0^k, x))$;
- $\mathcal{T}(0^k, x)$ contains at most $n/k = \sqrt{n}$ symbols '1', hence it can be coded with less than $\sqrt{n} \log n$ bits.

Consequences

- $\mathcal{CTW}$ is thus almost *adaptive* in this intermediate complexity, long-memory class.
- Extends to the Markovian case with $n^{2/3}$.
- Also shows that restricting depth *is* a serious limitation.
- We use very *unbalanced* trees : decisive advantage of Context Tree sources over Markov models.

Micro-bibliography

- **Universal modeling and coding** - Rissanen, Langdon, IEEE-IT 1981
- **Universal coding, Information, Prediction, and Estimation** -Rissanen, IEEE-IT 1984
- **The Context-Tree Weighting Method : basic Properties** - Willems, Shtarkov, Tjalkens IEEE-IT 1995
- **Redundancy Rates for Renewal and Other Processes** - Csiszár, Shields - IEEE-IT 1996
- **Variable length Markov chains** - Bühlmann, Wyner, Abraham, Annals of Statistics 1999
- **Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL.** - Csiszár, Talata (Budapest), IEEE-IT 2004