Bayesian Inference for Discrete Time Series via Tree Weighting

I. Kontoyiannis, A. Panotopoulou, M. Skoularidou
Department of Informatics, Athens Univ of Econ & Business

3-7 September 2012 IEEE Information Theory Workshop (ITW) 2012 Lausanne, Switzerland

Variable Length Markov Chains

Markov chain: $\{X_n\}$ with **alphabet** $A = \{0, 1, \ldots, m-1\}$

Memory length d:
$P(X_n|X_{n-1}, X_{n-2}, \ldots) = P(X_n|X_{n-1}, \ldots, X_{n-d})$

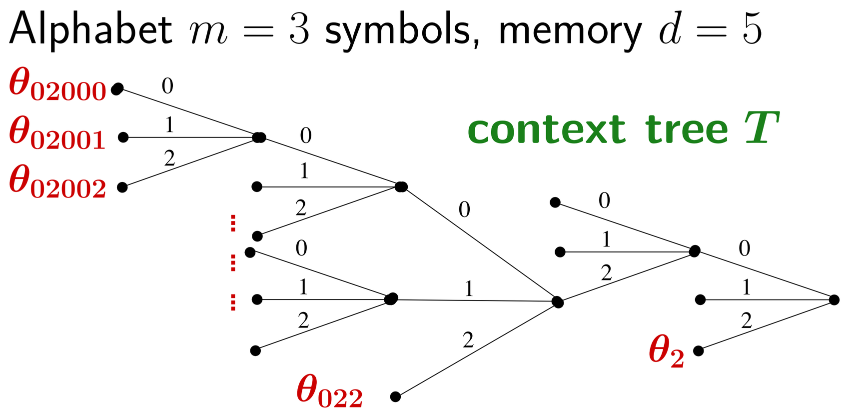
Distribution: To fully describe it need to specify m^d conditional distributions $P(X_n|X_{n-1}, \ldots, X_{n-d})$ one for each context $(X_{n-1}, \ldots, X_{n-d})$

Problem m^d grows very fast
e.g. $m = 8$ symbols & memory length $d = 10$ needs $\approx 10^9$ distributions

Idea Use *variable length contexts* described by a **context tree** T

VLMC Example

Alphabet $m = 3$ symbols, memory $d = 5$



Each past string $X_{n-1}, X_{n-2}, \ldots$ corresponds to a unique context on a leaf of the tree

The distr of X_n given the past is given by the distr of that leaf

E.g. $P(X_n = 1|X_{n-1} = 0, X_{n-2} = 2, X_{n-2} = 2, X_{n-3} = 1, \ldots) = \theta_{022}(1)$

The VLMC Likelihood

Likelihood Given a model (context tree) T and parameters $\theta = (\theta_s; s \in T)$ the likelihood of X_1^n is:

$$f(X_1^n|X_{-D+1}^0, \theta, T) = \Pi_{s\in T}\Pi_{j\in A}\theta_s(j)^{a_s(j)}$$

where $a_s(j) = \#$ times j follows s in X

VLMC Advantages

- E.g., above with memory length 5, instead of $3^5 = 243$ conditional distributions only need to specify 13!
- For an alphabet of size m and memory depth D there are m^D contexts $\Rightarrow$ potentially huge savings
- Determining the underlying context tree of an empirical time series is of great scientific and engineering interest

Motivation & Earlier Results

△ Our results are primarily motivated by:
- The results of Willems, Shtarkov, Tjalkens and co. on data compression via the CTW and related algorithms
- Basic questions of Bayesian inference for discrete time series

△ All our results can be seen as generalizations or extensions of results and algorithms in these earlier papers

△ Here we ignore this connection entirely and present everything from the point of view of Bayesian statistics

Bayesian Modeling

A NEW Prior on models
Given m, D, for each $\alpha \in (0,1)$ let

$$\pi_D(T) = \alpha^{|T|-1}\beta^{|T|-L_D(T)}$$

where $\beta = 1 - \alpha^{m-1}$; $|T| = \#$ leaves of T; and $L_D(T) = \#$ leaves at depth D

Prior on parameters
Given a model (context tree) T the parameters $\theta = (\theta_s; s \in T)$ are taken to be independent each with Dirichlet$(1/2, \ldots, 1/2)$ distr:

$$\pi((\theta_s; s \in T)|T) = \Pi_{s\in T}\frac{\Gamma(m/2)}{\pi^{m/2}}\Pi_{j\in A}\frac{1}{\sqrt{\theta_s(j)}}$$

Likelihood
Given T, θ, the likelihood of $X = X_1^n$ is:

$$f(X|\theta, T) = \Pi_{s\in T}\Pi_{j\in A}\theta_s(j)^{a_s(j)}$$

The Goal of Bayesian Inference

Determine the **posterior distributions**:

$$\pi(\theta, T|X) = \frac{\pi_D(T)\pi(\theta|T)f(X|\theta, T)}{f(X)}$$

$$\pi(T|X) = \frac{\int_\theta f(X|\theta, T)\pi(\theta|T)\,d\theta}{f(X)}$$

Main obstacle
Computation of the **marginal likelihood**:

$$f(X) = \sum_T \pi_D(T)\int_\theta f(X|\theta, T)\pi(\theta|T)\,d\theta$$

E.g. the number of models in the sum grows *doubly exponentially* in D

The Marginal Likelihood Algorithm

Given [*formerly known as CTW*]
Data $X = X_{-D+1}, \ldots, X_0, X_1, \ldots, X_n$
alphabet size m & max model depth D

△ **1.** [*Tree.*] Construct a tree with nodes corresponding to all contexts of length $1, 2, \ldots, D$ contained in X

△ **2.** [*Estimated probabilities.*]
At each node s compute the a_s and

$$P_{e,s} = \frac{\Pi_{j\in A}[(1/2)(3/2)\cdots(a_s(j) - 1/2)]}{(m/2)(m/2+1)\cdots(m/2+n-1)}$$

△ **3.** [*Weighted probabilities.*]
Let $\rho = 1 - \alpha^{m-1}$; at each node s compute

$$P_{w,s} = \begin{cases} P_{e,s}, & s \text{ a leaf} \\ \rho P_{e,s} + (1-\rho)\Pi_{j\in A}P_{w,sj}, & \text{o/w} \end{cases}$$

Theorem
The weighted probability $P_{w,root}$ given by the MLA at the root is exactly equal to the marginal likelihood

Note MLA computes a "doubly exponentially hard" quantity in $O(n \cdot D)$ time; one of the very few examples – the most complex and interesting one – of nontrivial Bayesian models for which the marginal likelihood is explicitly computable

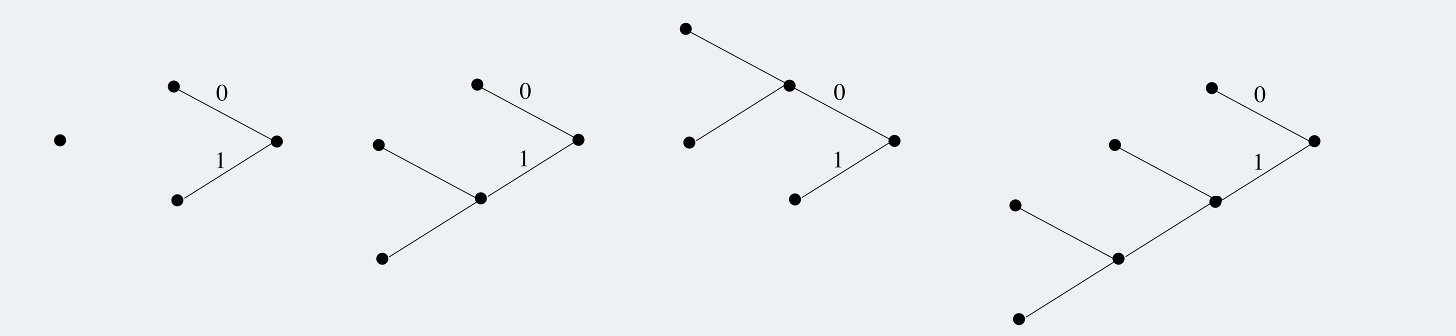
* MAPT & k-MAPT Algorithms *

Theorem Two analogous algorithms **MAPT** and **k-MAPT** *provably* compute the most likely and the k most likely models with respect to the posterior distribution $\pi(T|X)$ on model space

Experimental Results: IID Data

IID binary data $X_{-D+1}, \ldots, X_0, X_1, X_2, \ldots, X_n$
Distr Bern$(1/20)$, length $n = 50000$ bits

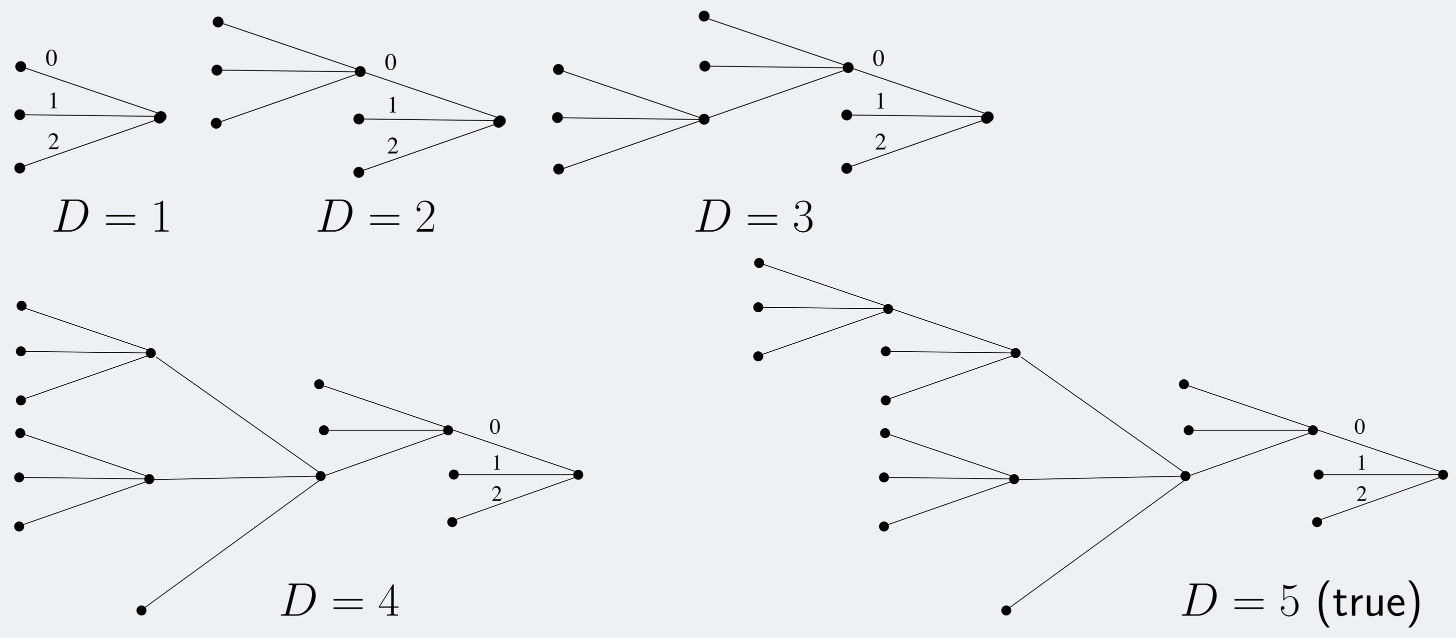
k-MAPT Find the top $k = 5$ models, with max depth $D = 15$



MAPT for the Earlier 5th-order VLMC

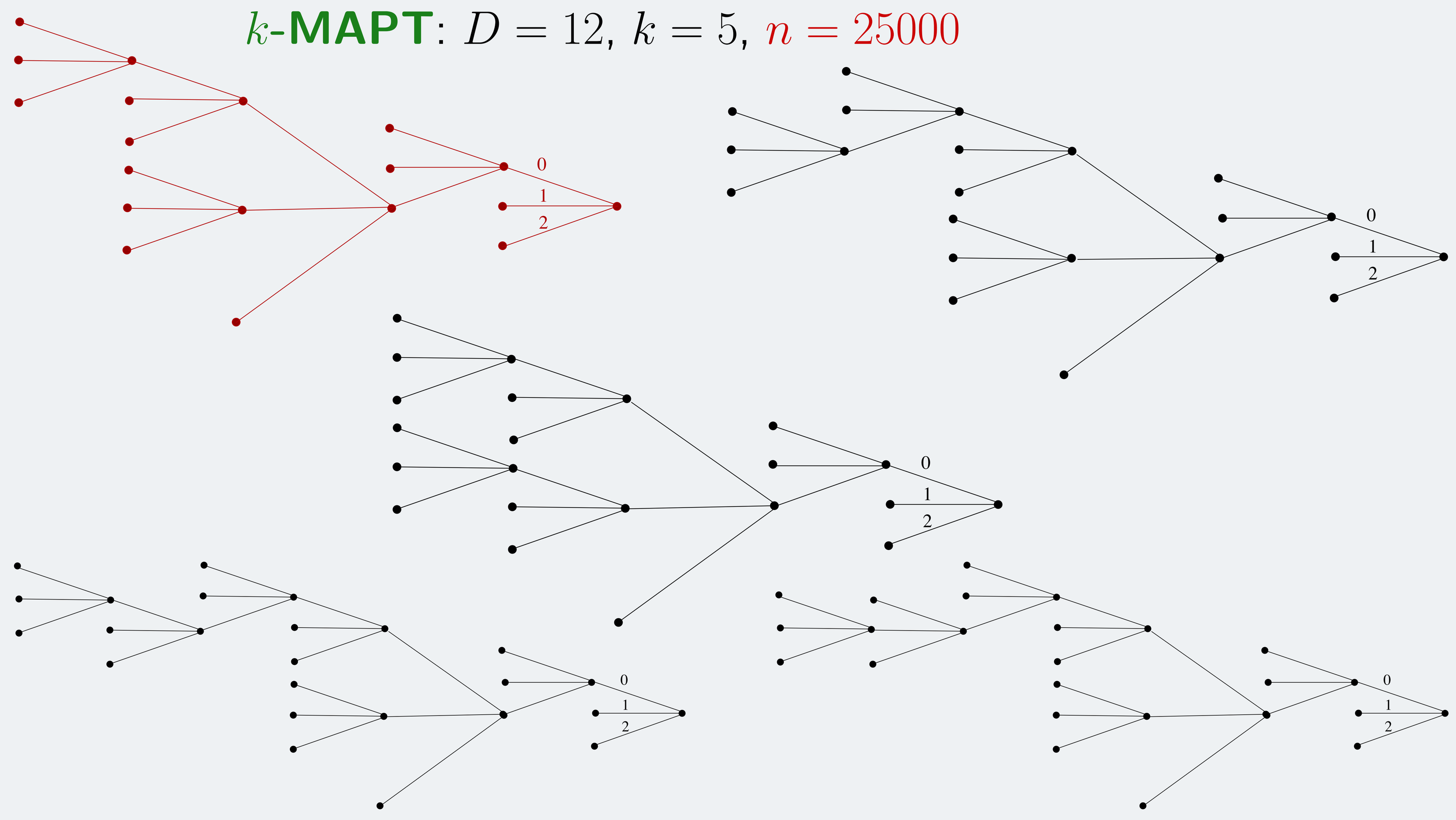
5th order VLMC data
$X_{-D+1}, \ldots, X_0, X_1, X_2, \ldots, X_n$, alphabet size $m = 3$
Distr VLMC as before (last MAP tree), data length $n = 80000$ symbols

MAPT
Find the MAP models with maximum depth $D = 1, 2, 3, \ldots$



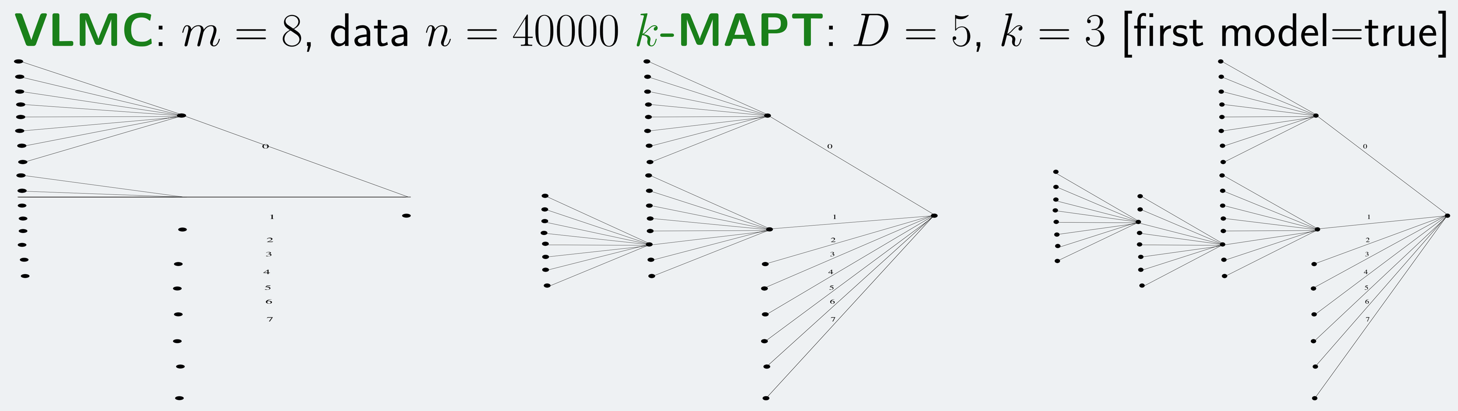
k-MAPT for same 5th-order VLMC

k-MAPT: $D = 12$, $k = 5$, $n = 25000$



k-MAPT for a 2nd-order 8-symbol VLMC

VLMC: $m = 8$, data $n = 40000$ **k-MAPT**: $D = 5$, $k = 3$ [first model=true]



Futher Results & Extensions

- Posterior model probabilities
- Simulated and real data...
- MCMC Exploring the full posterior
- Truly Bayesian entropy estimation