

Learning Hidden Markov Models to Fit Long-Term Dependencies

Pierre Dupont and Jérôme Callut

ERCIM workshop
26 October 2005

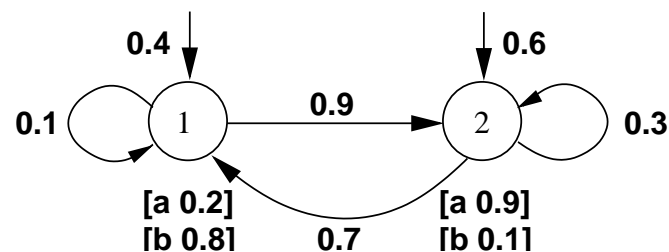


Computing science & engineering dept. (INGI)
UCL Machine Learning Group

Sequential process modeling

- Sequential data:*** sequences of *symbols* drawn from some unknown stochastic process, e.g. speech, process event logs, texts, ...
- Problem:*** given a sequential data sample, model the unknown generative process (target model)
- Applications:*** speech recognition, information extraction, biological sequence modeling, business process modeling, ...
- Models:*** Hidden Markov Models (HMM), Probabilistic Automata, Markov chains (MC), Petri nets, ...
- HMM learning:*** estimating ***model structure*** + parameters

HMM Learning problem



Given a sample : DNA fragments, amino acids, process logs such as Web usage logs, speech events, ...

Find a HMM which generates the observed sequences (and many more...)

- HMM topology
- emission and transitions probabilities

Objectives:

- to avoid the manual tuning of the model for each task adaptation
- to predict/analyze/classify new sequences of the same nature

Existing approaches and techniques

Different learning problems

- **P1** : Identify in the limit the target model
- **P2** : Bayesian framework: trade-off likelihood–model prior
- **P3** : Model selection in a discriminative setting

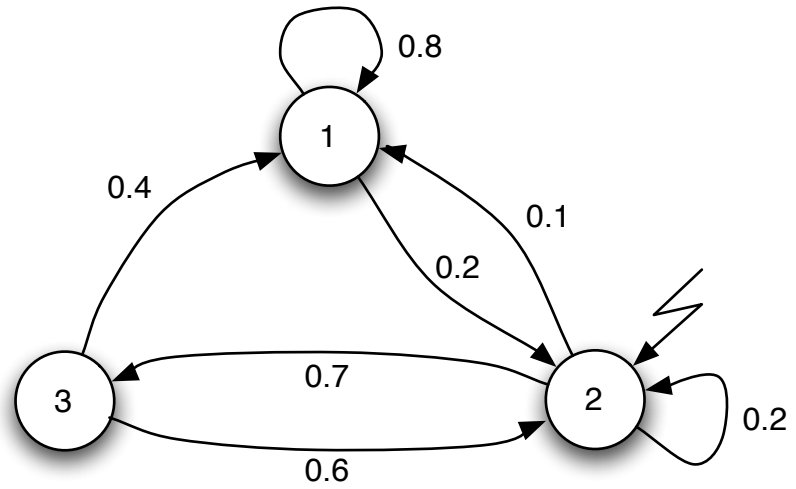
Existing techniques

1. EM on a fully connected graph: poor results in practice
2. Alergia/RLIPS [Carrasco, Oncina 1994]: identification of PDFAs
3. Bayesian state merging [Stolcke, 1994]: applied only on toy problems
4. State splitting [Ostendorf, 1997]: restricted to *left-to-right* topology
5. N-gram smoothed by back-off: very efficient but short-term modeling

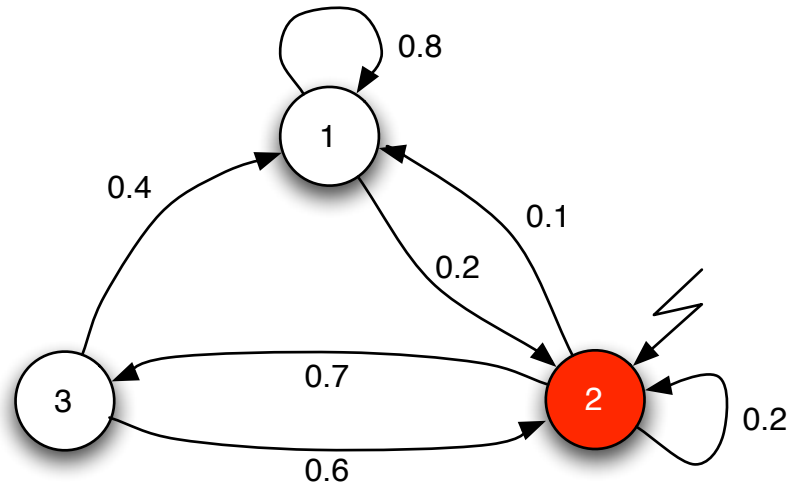
Our contributions

- ***New approach:*** Induce a HMM fitting the **dynamics** of the target model by extension of Markov chains properties.
- Fit distribution features badly modeled by existing methods
⇒ ***long-term dependencies***
- ***Induction algorithm*** for estimating model structure and parameters
⇒ Fit the **dynamics** observed in the learning sequence(s)
⇒ No restriction on the model topology

Dynamics of random walks in Markov chains

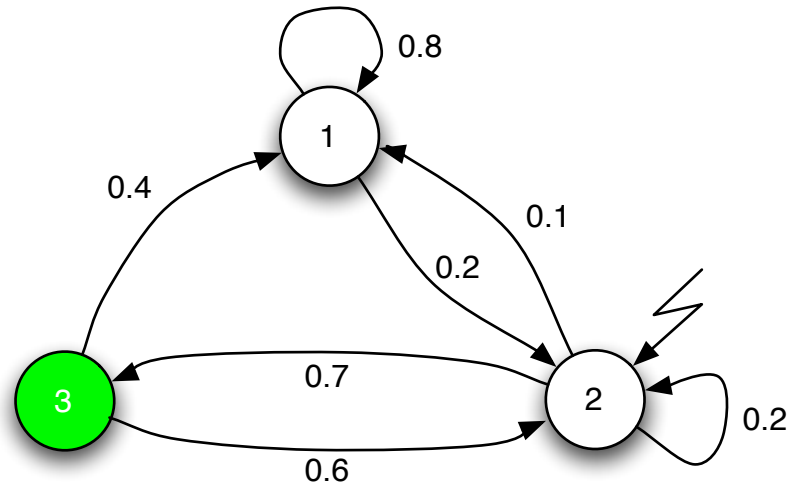


Dynamics of random walks in Markov chains



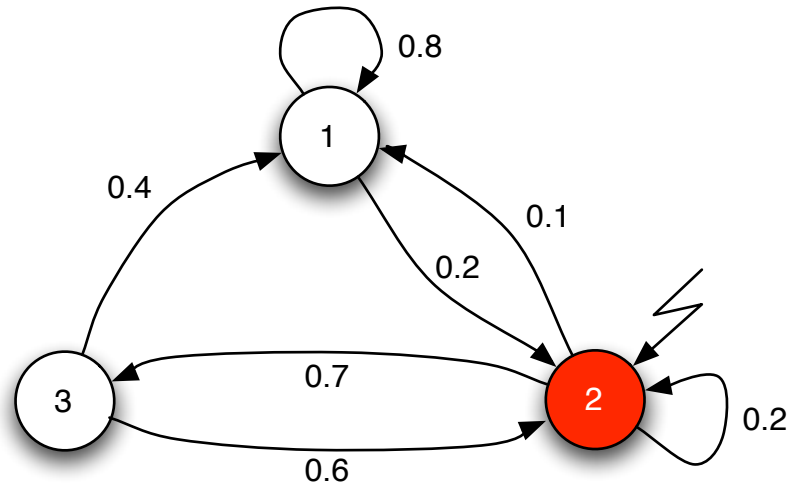
s = 2

Dynamics of random walks in Markov chains



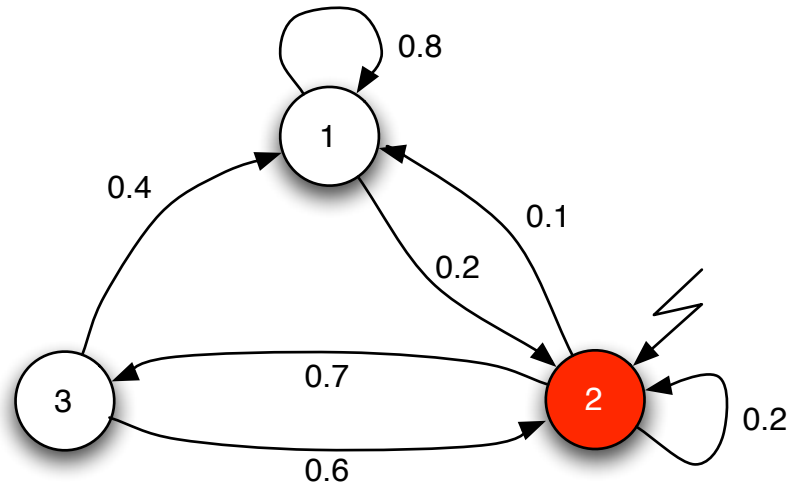
$s = 23$

Dynamics of random walks in Markov chains



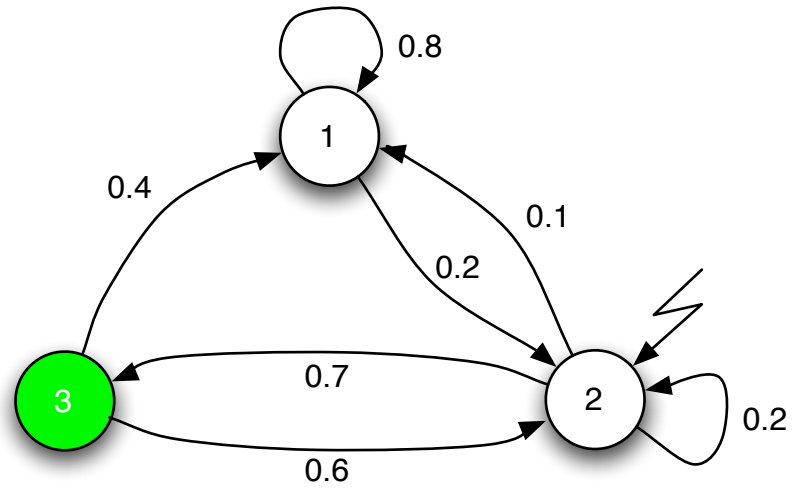
s = 2 3 2

Dynamics of random walks in Markov chains



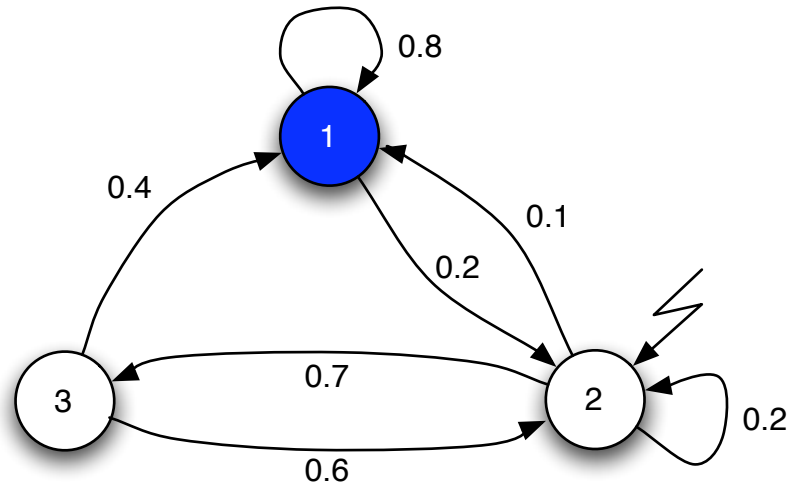
$s = 2322$

Dynamics of random walks in Markov chains



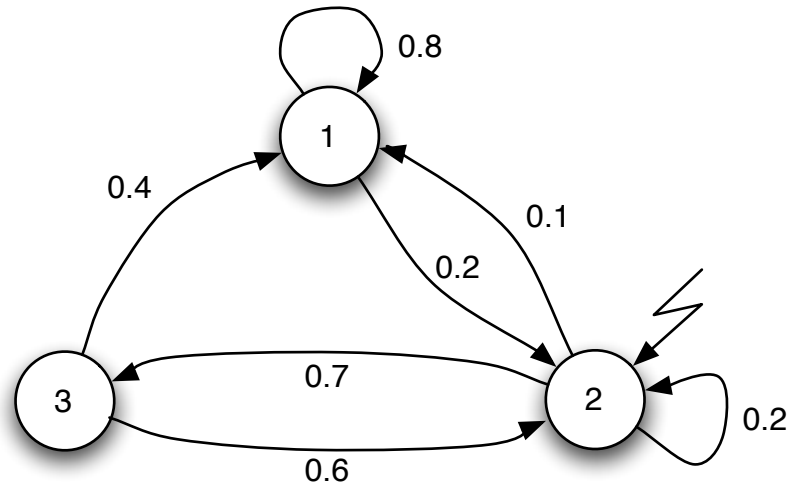
s = 2 3 2 2 3

Dynamics of random walks in Markov chains



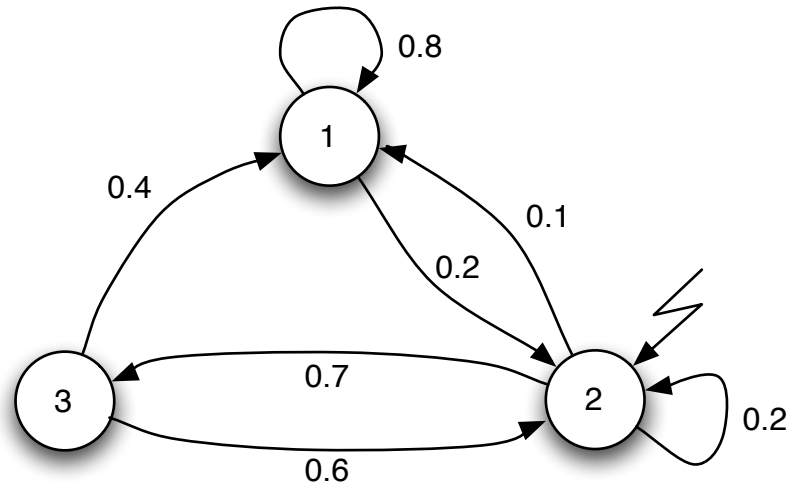
$\mathbf{s} = 2\ 3\ 2\ 2\ 3\ 1$

Dynamics of random walks in Markov chains



$\mathbf{s} = 2\ 3\ 2\ 2\ 3\ 1\ 1\ 1\ 2\ 3\ 2\ 3\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 3\ 1\ 1\ 1\ 1\ 1\ \dots$

Dynamics of random walks in Markov chains

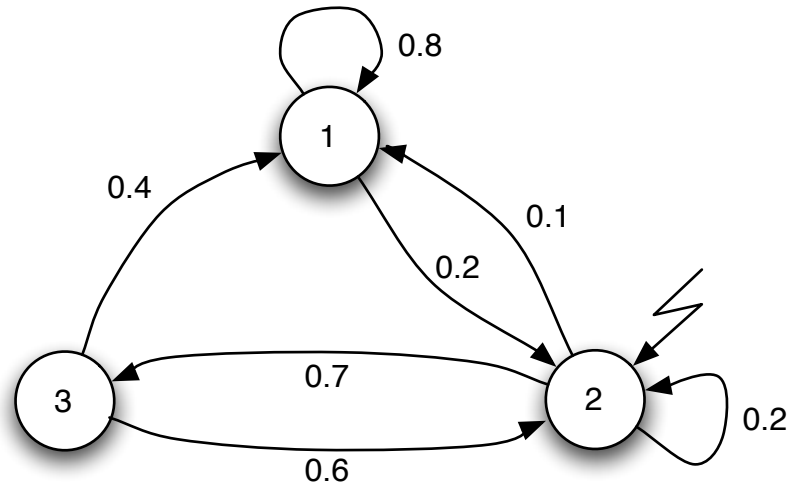


$\mathbf{s} = 2\ 3\ 2\ 2\ 3\ 1\ 1\ 1\ 2\ 3\ 2\ 3\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 2\ 3\ 1\ 1\ 1\ 1\ 1\ \dots$

$\hat{\pi} =$

1	2	3
---	---	---

Dynamics of random walks in Markov chains



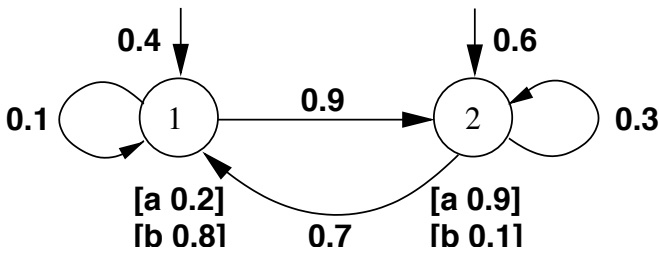
$\mathbf{s} = 2 \ 3 \ 2 \ 2 \ 3 \ 1 \ 1 \ 1 \ 2 \ 3 \ 2 \ 3 \ 2 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 2 \ 3 \ 1 \ 1 \ 1 \ 1 \ 1 \dots$

$\hat{\pi} =$ 1 2 3

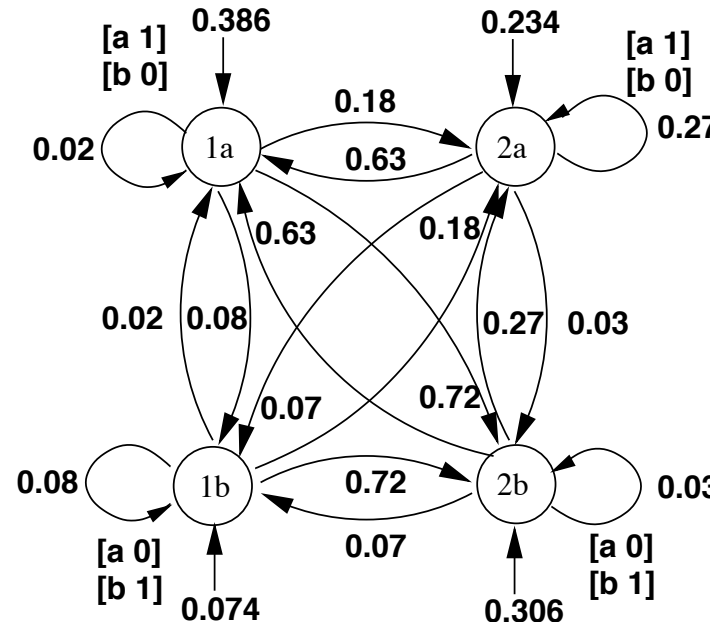
$\hat{M}_{3,1} = \text{Mean}(| \longleftrightarrow |, | \longleftrightarrow |, | \longleftrightarrow |, | \longleftrightarrow |, | \longleftrightarrow |, \dots)$

The **stationary distribution** π and the **Mean First Passage Times (MFPT)** M completely characterize the Markov chain [Kemeny, 82].

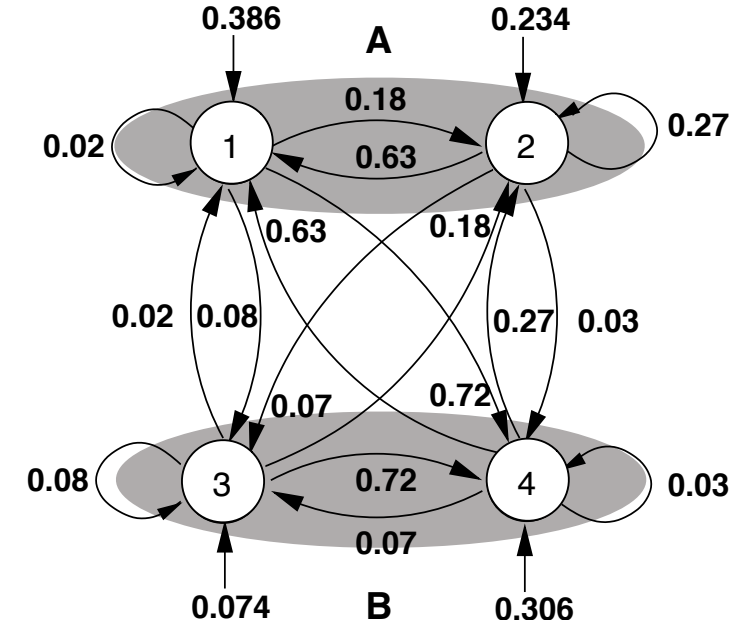
Back to HMMs: equivalent models



HMM



POMM



Lumped process

Random walks in HMMs \iff **Random walks in lumped processes**

We are interested in the **dynamics of random walks** :

- Stationary distribution
- Mean First Passage Times

Modeling long-term dependencies

A stochastic process contains *long-term dependencies* if

$$P(X_t | X_{t-1}, \dots, X_{t'}) \neq P(X_t | \underbrace{X_{t-1}, \dots, X_{t-p}}_{H:\text{process history}}) \text{ when } |H| < t - t'$$

The *relevant history size* $|H_{rel}|$ is the size of the history required to correctly model these dependencies.

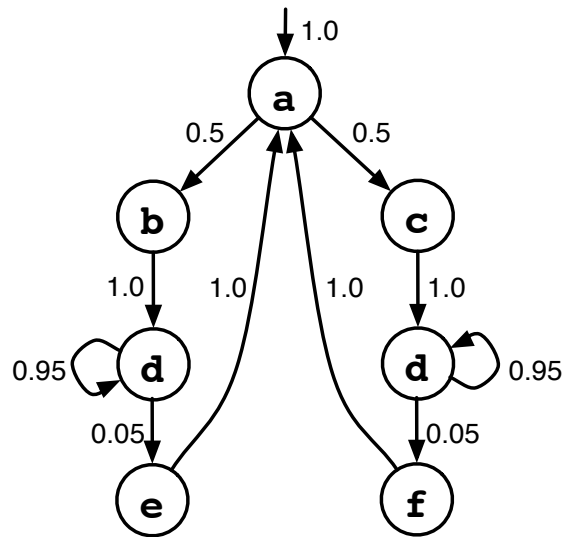
- if $|H_{rel}|$ is bounded : finite order MCs are appropriate models
- if $|H_{rel}|$ is unbounded: HMMs or POMMs are required

The **Perron theorem** [Senata, 81] applied on the HMM/POMM transition matrix:

⇒ *The use of a good model topology is necessary in order to store adequately the relevant history.*

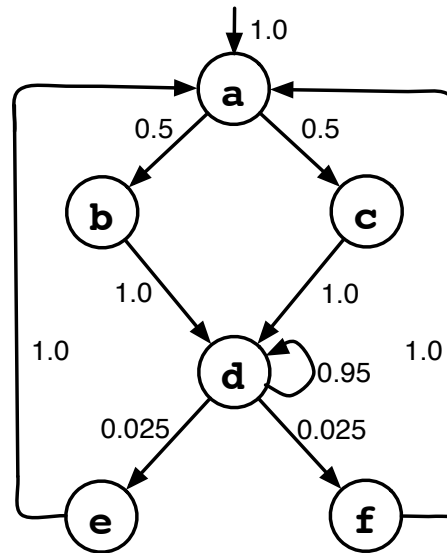
Long-term dependencies: finite order modeling and MFPT

Target model



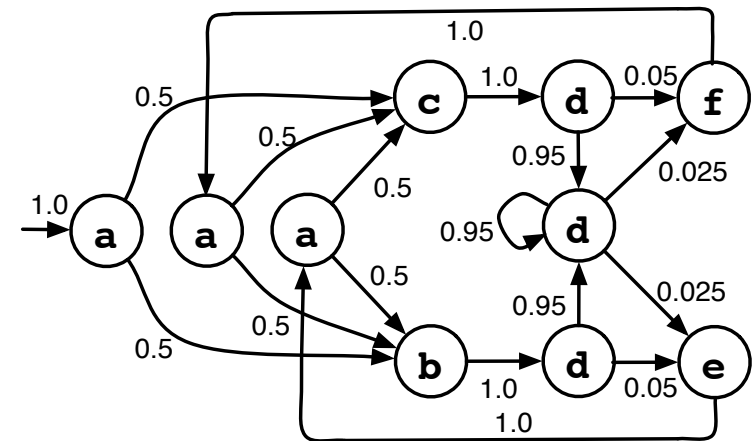
$$P(t_f < t_e | X_t = b) = 0$$

Order 1 MC



$$P(t_f < t_e | X_t = b) = 0.5$$

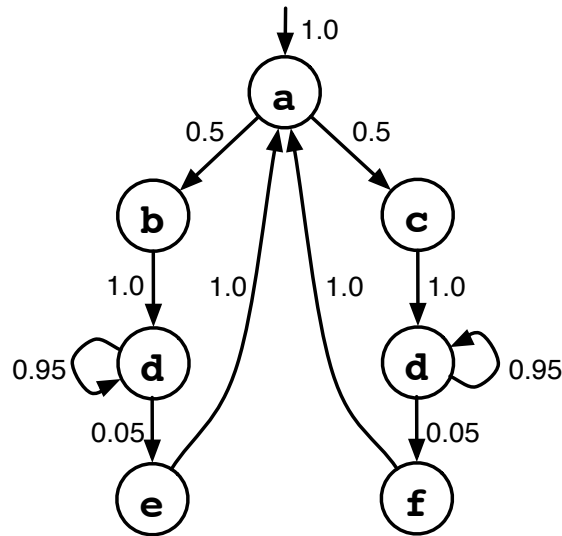
Order 2 MC



$$P(t_f < t_e | X_t = b) = 0.475$$

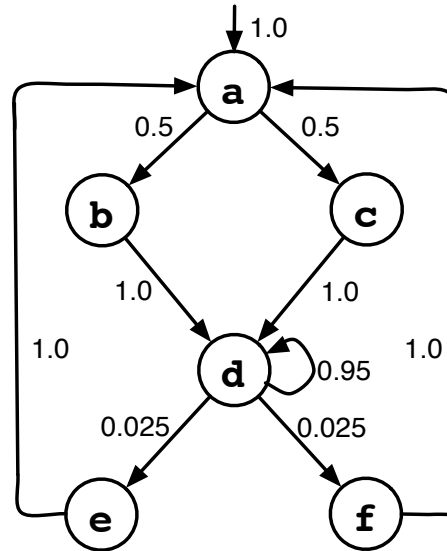
Long-term dependencies: finite order modeling and MFPT

Target model



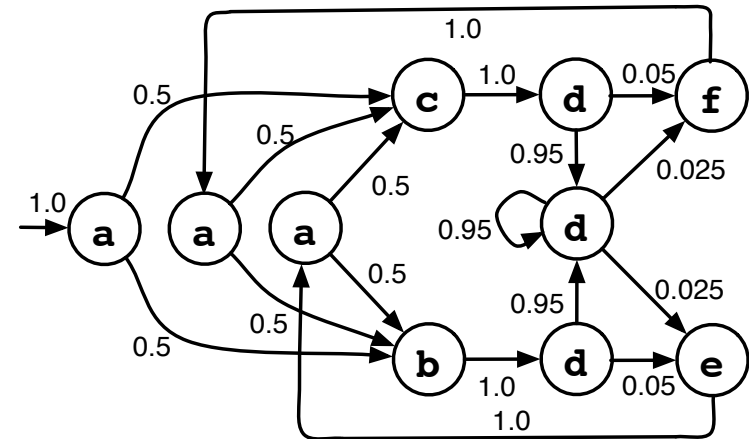
$$P(t_f < t_e | X_t = b) = 0$$

Order 1 MC



$$P(t_f < t_e | X_t = b) = 0.5$$

Order 2 MC



$$P(t_f < t_e | X_t = b) = 0.475$$

Related Mean First Passage Times

Target	e	f
b	21.0	67.0
c	67.0	21.0

MC_1	e	f
b	44.0	44.0
c	44.0	44.0

MC_2	e	f
b	42.85	45.15
c	45.15	42.85

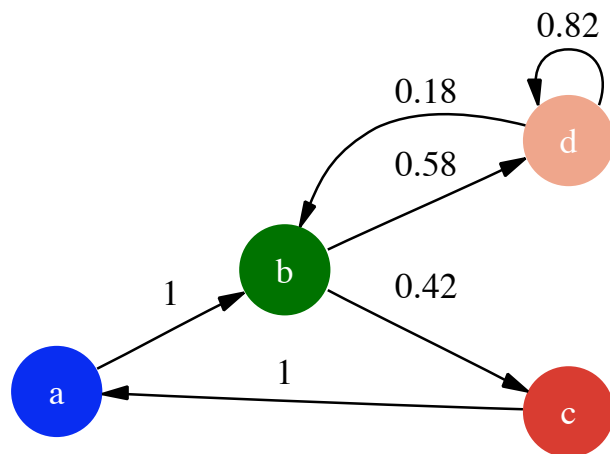
Induction algorithm

Input: -A string s from the target TP
-A precision parameter ϵ

Output: -A POMM EP

1. Initialization

- $\hat{M} \leftarrow$ Estimate MFPTs from the sample s
- $EP \leftarrow$ Estimate an **order 1 MC** from s



- $s = dbcabdddddcbcabddb \dots$

- $A_{a,b} = \frac{\text{count}(a,b)}{\text{count}(a)}$

- Model size = $|\Sigma|$

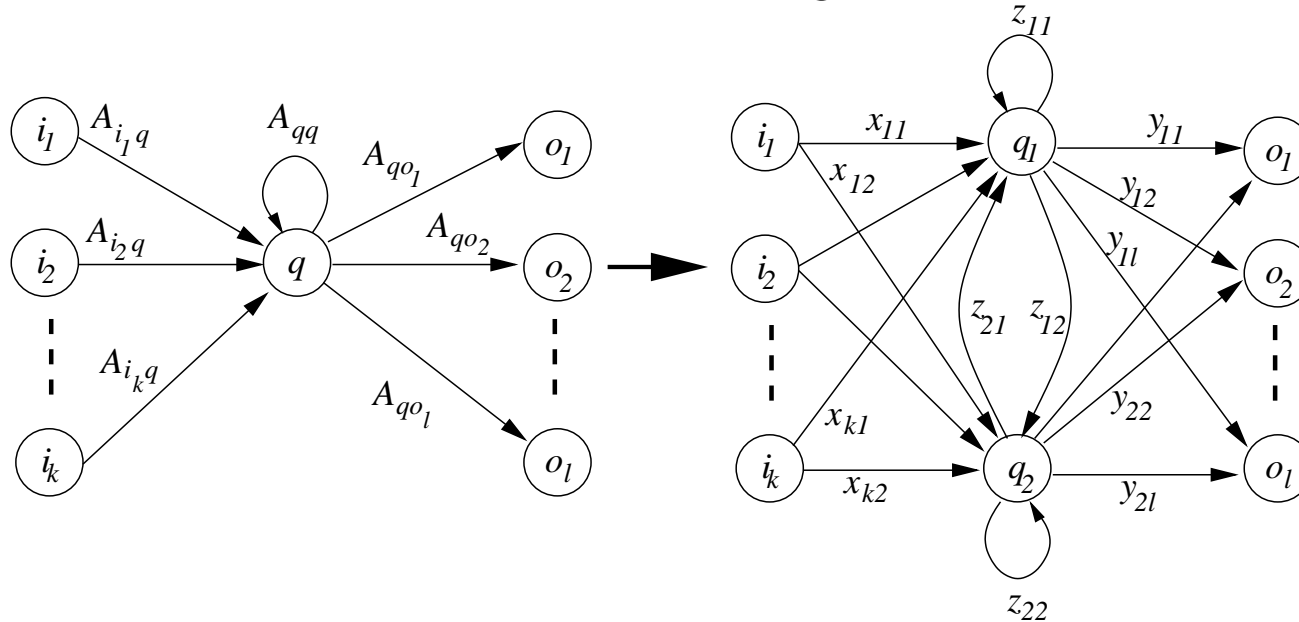
- Satisfies stationary distribution

- **Does not satisfy MFPT between letters**

Induction algorithm

2. Iterate

- Try every state q as a candidate for splitting



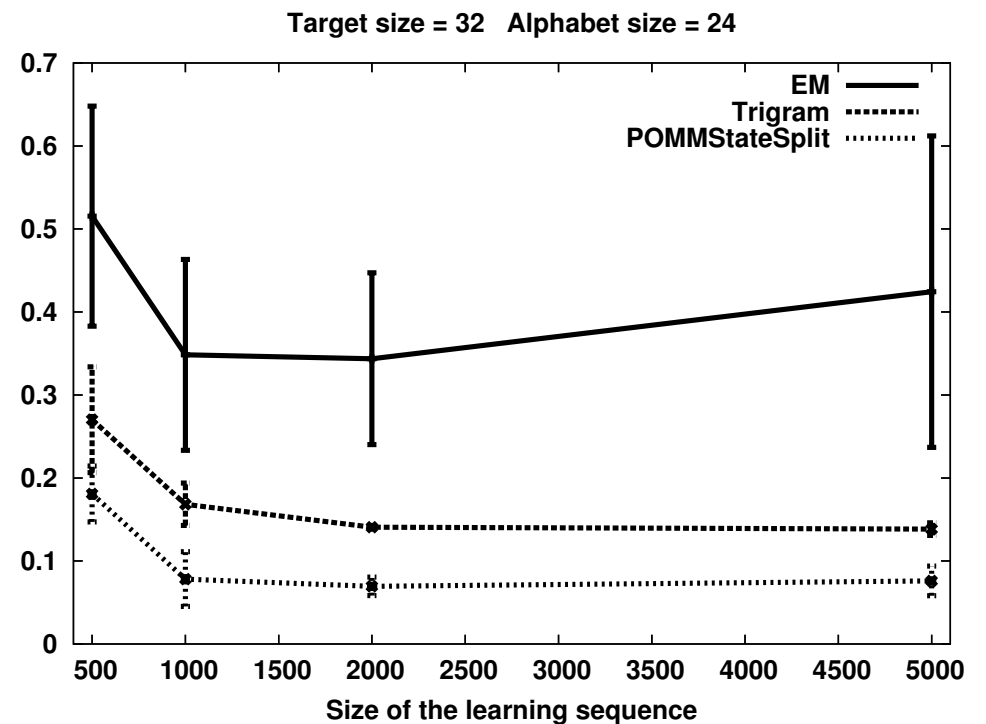
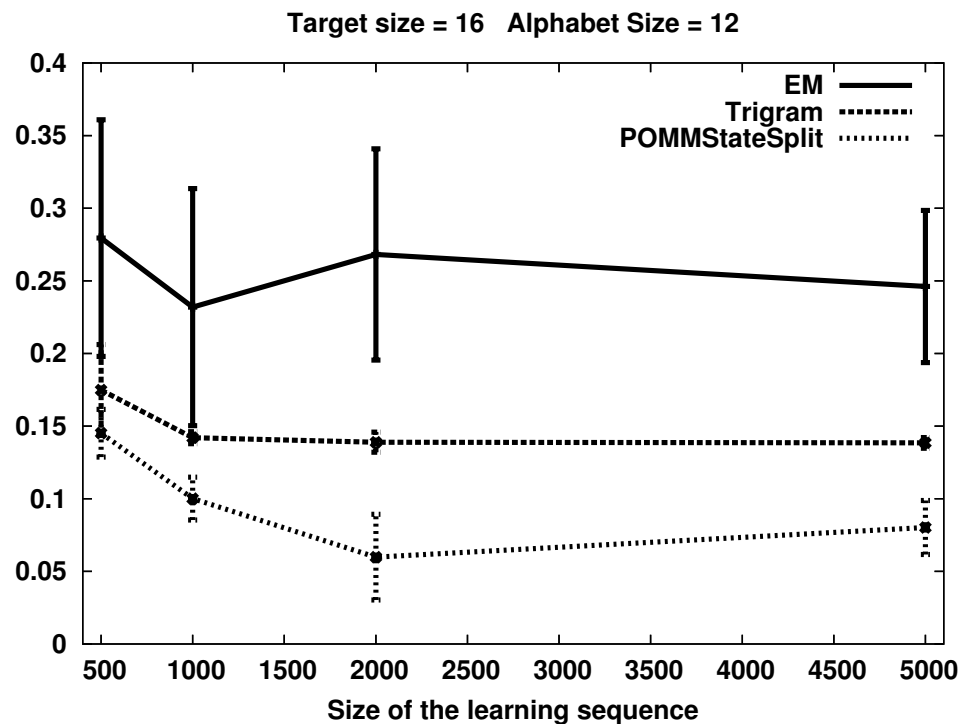
- Minimize $W(X, Y, Z) = \sum_{i,j=1}^{|\Sigma|} (\hat{M}_{ij} - M_{ij} // \kappa)^2$

such that $\begin{cases} \text{Blocks stationary distribution is unchanged} \\ \text{The model remains a proper POMM} \end{cases}$

- Choose the state splitting that maximize the log-likelihood

Experiments

- Random target models of increasing size are used
- 100 test sequences of length 1000 are generated for each target size
- Results are averaged on 10 training sequences



- **Performance measure:** relative likelihood loss with respect to the target model as function of the training size

Conclusion and future work

We proposed a novel approach to induce HMMs based on the observed dynamics in the sample. This technique is well-suited to model long-term dependencies.

Ongoing work

- Use of additional dynamics features in the sample
- Study the link between Petri nets and finite state machines
- Applying the algorithm on real-life data (business process logs)

References

- [1] J. Callut and P. Dupont. Inducing Hidden Markov Models to model long-term dependencies. In *16th European Conference on Machine Learning (ECML)*, number 3720 in Lecture Notes in Artificial Intelligence, pages 513–521, Porto, Portugal, October 2005. Springer-Verlag.
- [2] J. Callut and P. Dupont. Learning Hidden Markov Models to fit long-term dependencies. Technical Report 2005-9, Université catholique de Louvain, July 2005.
- [3] Y. Bengio and P. Frasconi. Diffusion of context and credit information in markovian models. *Journal of Artificial Intelligence Research*, 3:223–244, 1995.
- [4] M. Ostendorf and H. Singer. Hmm topology design using maximum likelihood successive state splitting. *Computer Speech and Language*, 11:17–41, 1997.

- [5] R. Carrasco and J. Oncina. Learning stochastic regular grammars by means of a state merging method. In *Proc. 2nd International Colloquium on Grammatical Inference - ICGI '94*, volume 862, pages 139–150. Springer-Verlag, 1994.

- [6] A. Stolcke and S.M. Omohundro. Hidden markov model induction by bayesian model merging. In C.L. Giles, S.J. Hanton, and J.D. Cowan, editors, *Advances in Neural Information Processing Systems*. Morgan Kaufman, 1993.

- [7] E. Senata. *Non-negative Matrices and Markov Chains*. Springer-Verlag, 1981.

- [8] Carl D. Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, 2000.