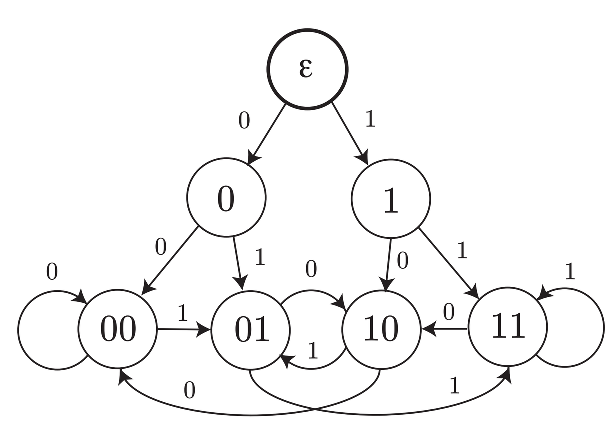
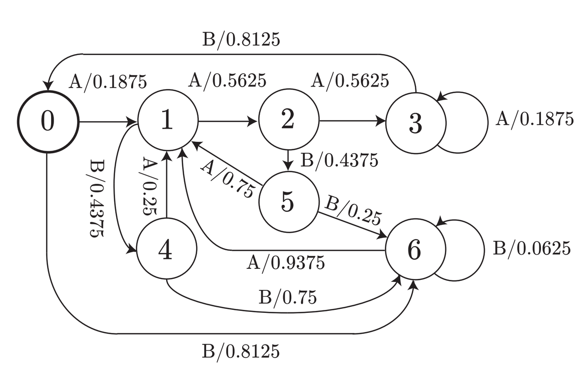
Bayesian Infinite Automata

David Pfau*, Nicholas Bartlett†, Frank Wood†

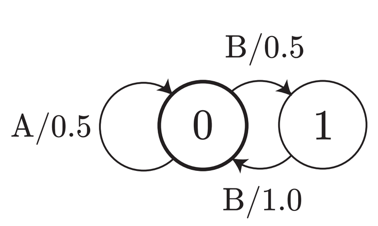
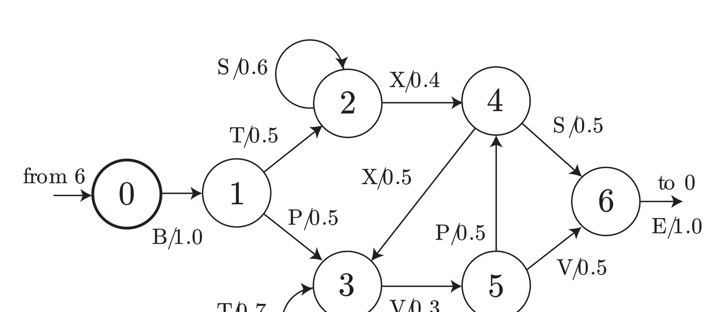
Overview

- *nth*-order Markov models, or m-gram models, are popular for learning sequences, but the size of the models blows up as n increases.
- We relax the problem by expanding the class of models to include all *probabilistic deterministic finite automata* (PDFA)[1], which includes m-gram models as a special case
- Inference is Bayesian - we define a prior over PDFAs of arbitrary size, using *hierarchical Pitman-Yor processes*[2]. We call the model the Probabilistic Deterministic *Infinite* Automata since there is no bound on the possible number of states of a sample
- Posterior inference via MCMC on natural language, DNA and synthetic grammars yield encouraging results

Finite Automata

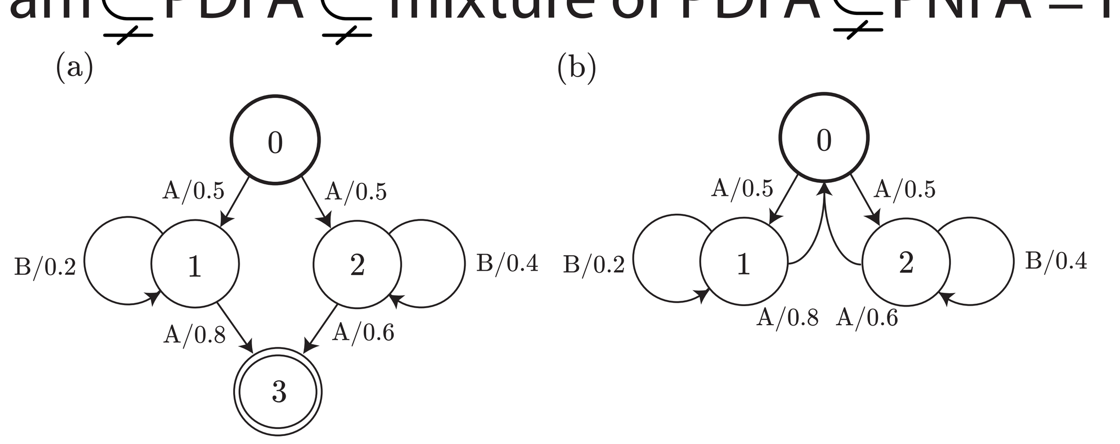



Trigram as DFA

The posterior of the PDIA is approximated with a mixture of PDFAs. From m-gram models to Hidden Markov Models, the model classes here form a simple hierarchy:

m-gram $\subsetneq$ PDFA $\subsetneq$ mixture of PDFA $\subsetneq$ PNFA = HMM*



(a) PNFA in mixture of PDFA (b) PNFA not in mixture of PDFA

$$\delta: \begin{array}{c|ccc} & \sigma_0 & \sigma_1 & \sigma_2 \\ \hline q_0 & q_3 & q_1 & q_{12} \\ q_1 & \bullet & q_3 & \bullet \\ q_2 & q_6 & q_4 & \bullet \\ q_3 & q_3 & \bullet & q_5 \\ q_4 & q_3 & q_{87} & q_5 \\ q_5 & \bullet & q_1 & q_2 \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

$$\pi: \begin{array}{c|ccc} & \sigma_0 & \sigma_1 & \sigma_2 \\ \hline q_0 & \multicolumn{3}{c}{\text{(iid probability vector)}} \\ q_1 & \multicolumn{3}{c}{\text{(iid probability vector)}} \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

$$\frac{q_0\ q_1\ q_3\ q_3\ q_3\ q_5\ q_2\ q_4\ q_5\ q_2 \dots}{1\ \ 1\ \ 0\ \ 0\ \ 2\ \ 2\ \ 1\ \ 2\ \ 1\ \ 0 \dots}$$

Generative Model

$$\mu \sim \mathrm{Dir}(\alpha_0/|Q|)$$
$$\phi_j \sim \mathrm{Dir}(\alpha\mu) \qquad j=0\dots|\Sigma|-1$$
$$\delta(q_i,\sigma_j) = \delta_{ij} \sim \phi_j \qquad i=0\dots|Q|-1$$
$$\pi_{q_i} \sim \mathrm{Dir}(\beta/|\Sigma|) \qquad i=0\dots|Q|-1$$
$$\xi_0 = q_0,\ \ \xi_t = \delta(\xi_{t-1},x_{t-1})$$
$$x_t \sim \mathrm{Mult}(\pi_{\xi_t})$$

where

- Q – finite set of states
- Σ – finite alphabet
- $\delta: Q\times\Sigma\to Q$ – transitions
- $\pi: Q\times\Sigma\to[0,1]$ – emissions
- $q_0\in Q$ – initial state
- $x_t\in\Sigma$ – data at time t
- $\xi_t\in Q$ – state at time t
- $\alpha,\alpha_0,\beta\geqslant 0$ – hyperparams

The limit as $|Q|\to\infty$ is well defined - a Hierarchical Dirichlet Process (HDP)[4]. Add discounts $d, d_0\in[0,1]$ to make it a Hierarchical Pitman-Yor process ($d, d_0=0 \Leftrightarrow$ HDP). Also, specify base distribution H (here geometric).
If μ and ϕ_j are marginalized out, then δ_{ij} are exchangeable.

Intuitively, δ_{ij} is likely similar to other $\delta_{i'j'}$, moreso if $j = j'$ (same symbol emitted from different states). Draws from a PYP cluster together, and rich clusters get richer.

Inference

- MCMC sampler for posterior - sample $\delta_{ij}|\delta_{-ij},x_{0:t},\alpha,\alpha_0,\beta$
- likelihood only depends on π through counts c_{ij}

$$p(x_{0:T}|\delta,c,\beta) = \prod_{i=0}^{|Q|-1} \frac{\Gamma(\beta)}{\Gamma(\frac{\beta}{|\Sigma|})^{|\Sigma|}} \frac{\prod_{j=1}^{|\Sigma|}\Gamma(\frac{\beta}{|\Sigma|}+c_{ij})}{\Gamma(\beta+\sum_{j=1}^{|\Sigma|}c_{ij})}$$

- δ_{ij} not encountered by the data can be ignored

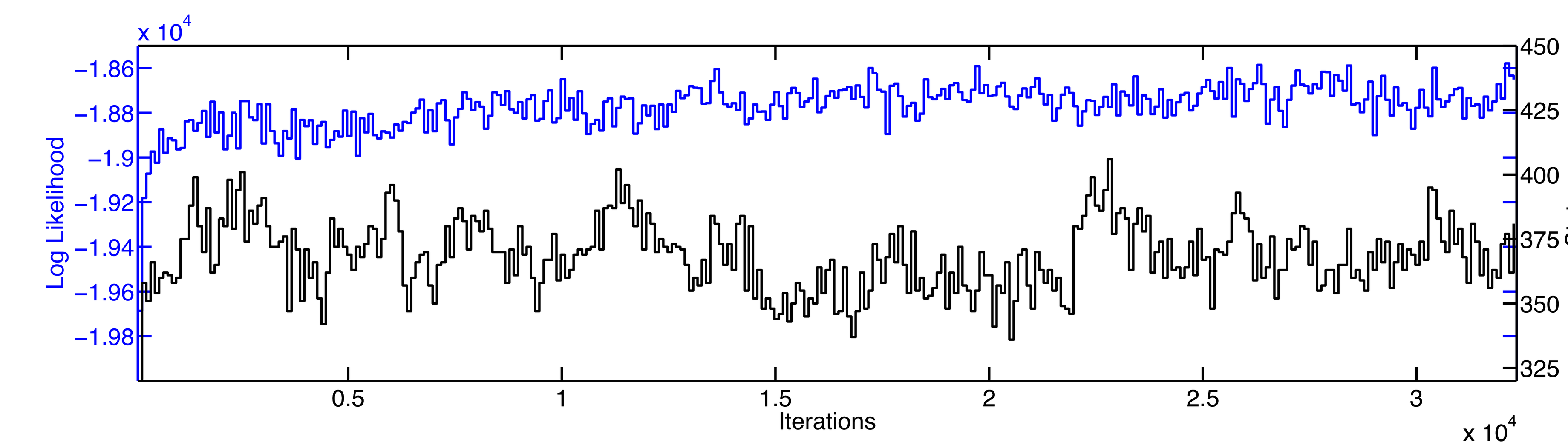
- If δ_{ij} is only transition to state $q_{i'}$, Gibbs sampling fails
- Instead use Metropolis-Hastings sampling for each δ_{ij}
- Propose from $\delta_{ij}|\delta_{-ij}$, accept from ratio of $p(x_{0:t}|\delta,\pi)$ for new and old δ_{ij}, sampling entries of δ from $\delta_{ij}|\delta_{-ij}$ as needed
- If proposal is accepted, remove entries from δ with 0 counts

Natural Language and DNA Prediction

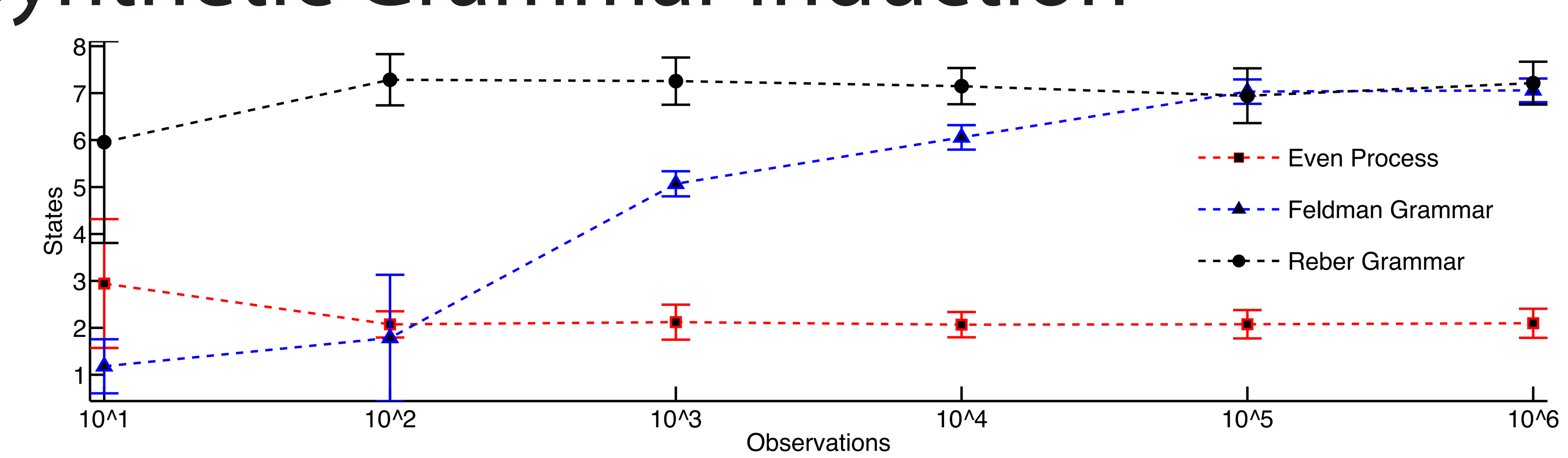
	PDIA	PDIA-MAP	HMM-EM	bigram	trigram	4-gram	5-gram	6-gram	SSM
AIW	5.13	5.46	7.89	9.71	6.45	5.13	4.80	4.69	4.78
	365.6	379	52	28	382	2,023	5,592	10,838	19,358
DNA	3.72	3.72	3.76	3.77	3.75	3.74	3.73	3.72	3.56
	64.7	54	19	5	21	85	341	1,365	314,166

Top rows: perplexity of held out data. Bottom: number of states

- Alice in Wonderland: 10k train, 4k test *"alice was beginning to..."*
- Mouse DNA: 150k train, 50k test *"CGTATATGCGCC..."*
- Controls: EM-trained HMM, HPYP smoothed n-gram[2], sequentially-trained sequence memoizer[5]
- Average predictions superior to predictions of "best" or MAP sample from PDIA posterior



Synthetic Grammar Induction



Future Directions

- Evaluation on larger data sets
- More efficient sampling - split-merge?
- How to tie together emission distributions between different states? (Like Kneser-Ney for m-grams)

References

[1] Rabin, M. Probabilistic automata. *Information and control*, Elsevier, 1963, 6, 230-245.
[2] Teh, Y. W. A Hierarchical Bayesian Language Model based on Pitman-Yor Processes. *Proceedings of the Association for Computational Linguistics*, 2006, 985-992.
[3] Dupont, P.; Denis, F. & Esposito, Y. Links between probabilistic automata and hidden Markov models: probability distributions, learning models and induction algorithms. *Pattern recognition*, Elsevier, 2005, 38, 1349-137.
[4] Teh, Y. W.; Jordan, M. I.; Beal, M. J. & Blei, D. M. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 2006, 101, 1566-1581.
[5] Wood, F.; Archambeau, C.; Gasthaus, J.; James, L. & Teh, Y. W. A Stochastic Memoizer for Sequence Data. *Proceedings of the 26th International Conference on Machine Learning*, 2009, 1129-1136.

* technically, PNFA without final state = HMM[3], but those are the only models we consider here