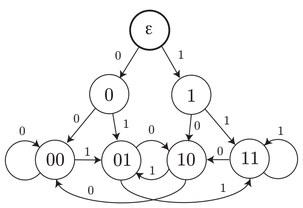
Bayesian Infinite Automata

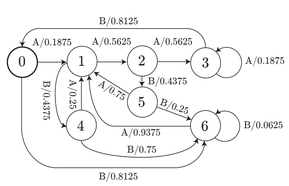
David Pfau*, Nicholas Bartlett†, Frank Wood†

Overview

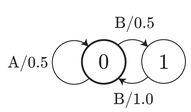
-n*th*-order Markov models, or m-gram models, are popular for learning sequences, but the size of the models blows up as n increases.
-We relax the problem by expanding the class of models to include all *probabilistic deterministic finite automata* (PDFA), which includes m-gram models as a special case
-Inference is Bayesian - we define a prior over PDFAs of arbitrary size, using *hierarchical Pitman-Yor processes*. 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

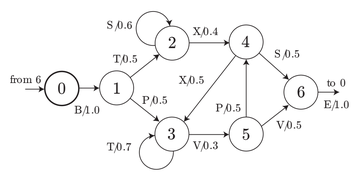
Finite Automata





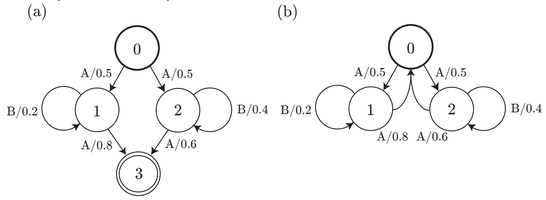
Trigram as DFA





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

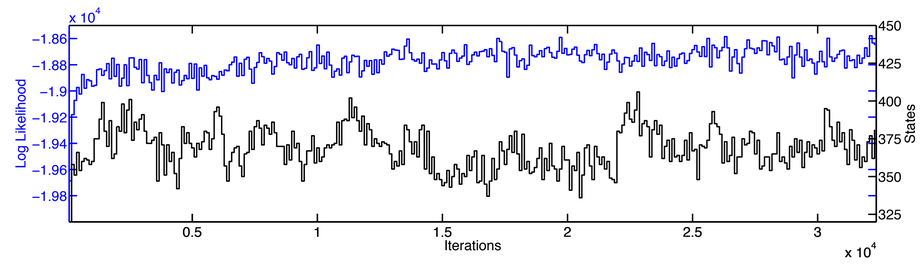
* technically, PNFA without final state = HMM, but those are the only models we consider here

Natural Language and DNA Prediction

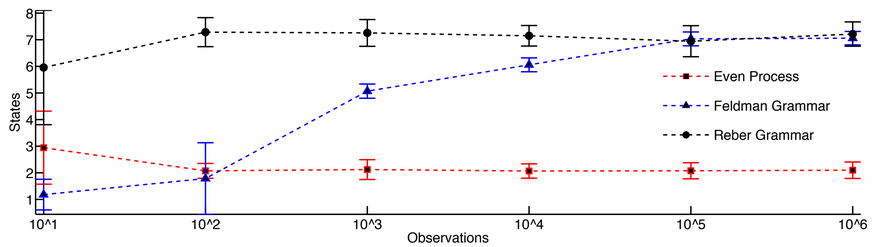
	PDIA	PDIA-MAP	HMM-EM	bigram	trigram	4-gram	5-gram	6-gram	SSM
AIW	5.13	5.46	7.89	9.71	6.45	5.13	4.80	4.69	4.78
	365.6	379	52	28	382	2,023	5,592	10,838	19,358
DNA	3.72	3.72	3.76	3.77	3.75	3.74	3.73	3.72	3.56
	64.7	54	19	5	21	85	341	1,365	314,166

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
- 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 n-grams)

{*Center for Theoretical Neuroscience, †Department of Statistics} Columbia University

pfau@neurotheory.columbia.edu