

# Bayesian Infinite Automata

David Pfau\*, Nicholas Bartlett†, Frank Wood†

## Overview

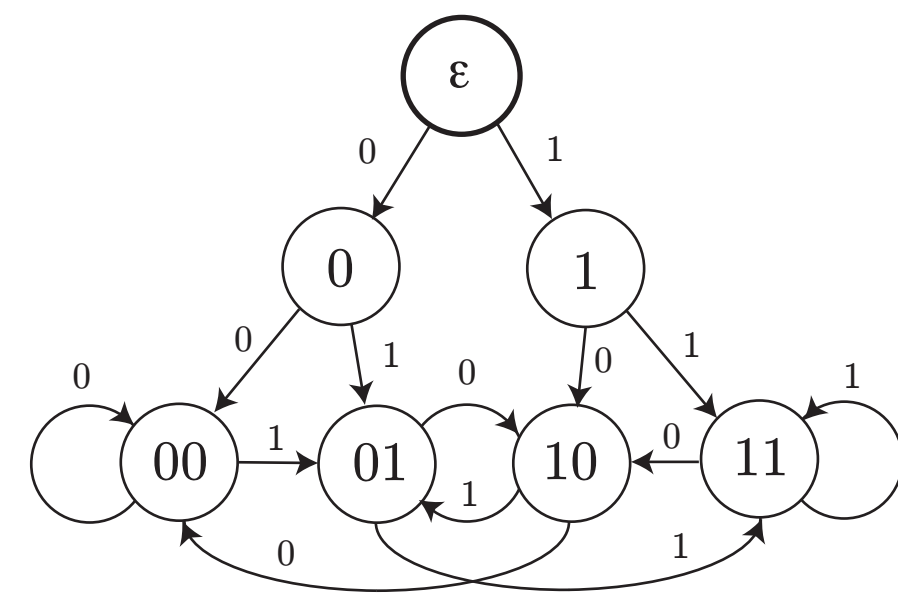
-*n*-th-order Markov models, or m-gram models, are popular for learning sequences, but the size of the models blows up as n increases.

-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

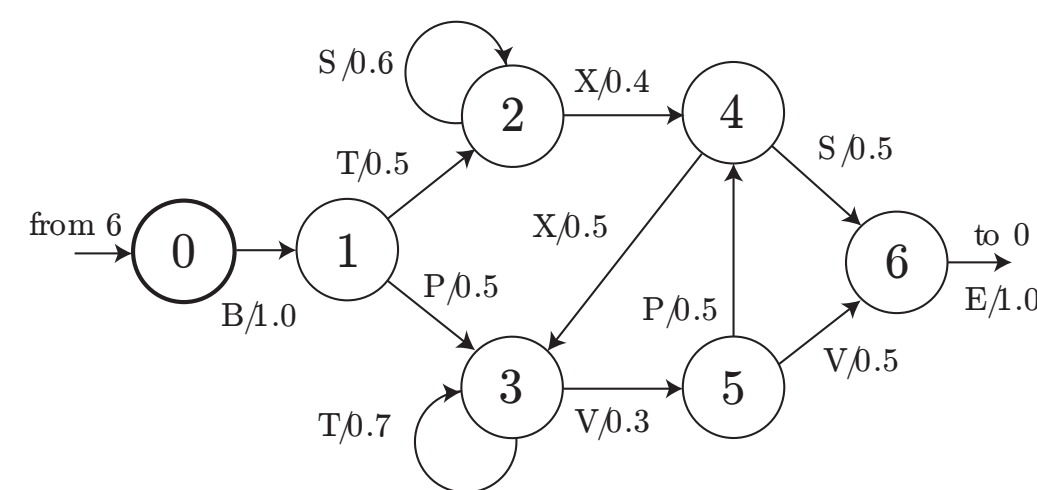
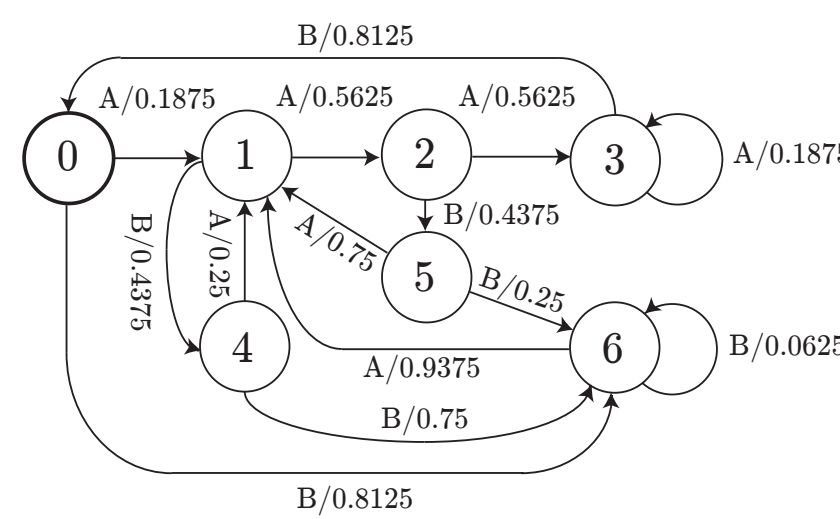
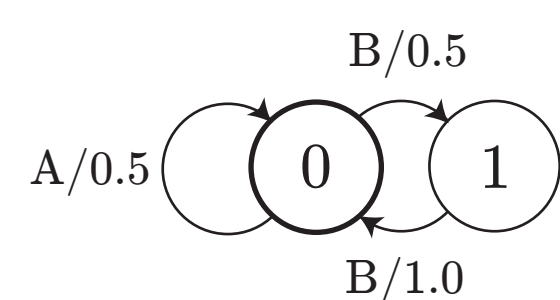
-Inference is Bayesian - we define a prior over PDFAs of arbitrary size, using *hierarchical Pitman-Yor processes*. We call the model the Probabilistic Deterministic *Infinite* Automata since there is no bound on the possible number of states of a sample

-Posterior inference via MCMC on natural language, DNA and synthetic grammars yield encouraging results

## Finite Automata

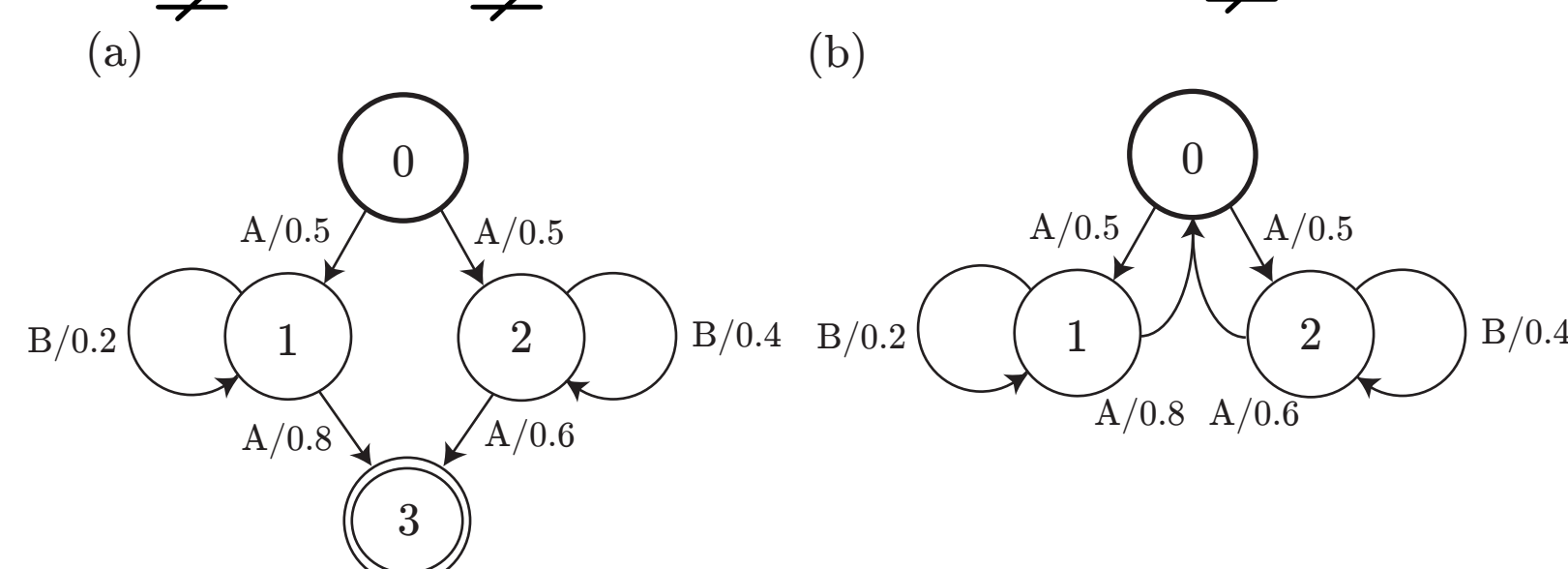


Trigram as DFA



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

## Notation

$M=(Q,\Sigma,\delta,\pi,q_0)$  – PDFA  
 $Q$  – finite set of states  
 $\Sigma$  – finite alphabet  
 $\delta:Q\times\Sigma\rightarrow Q$  – transitions  
 $\pi:Q\times\Sigma\rightarrow[0,1]$  – emissions  
 $q_0\in Q$  – initial state  
 $x_t\in\Sigma$  – data at time t  
 $\xi_t\in Q$  – state at time t  
 $\alpha,\alpha_0\geq 0$  – concentrations  
 $d,d_0\in[0,1]$  – discounts  
 $\beta\geq 0$  – emission param  
 $H$  – base distribution on  $\mathbb{N}$

PUT BIG  
TRANSITION  
MATRIX  
CARTOON  
HERE .

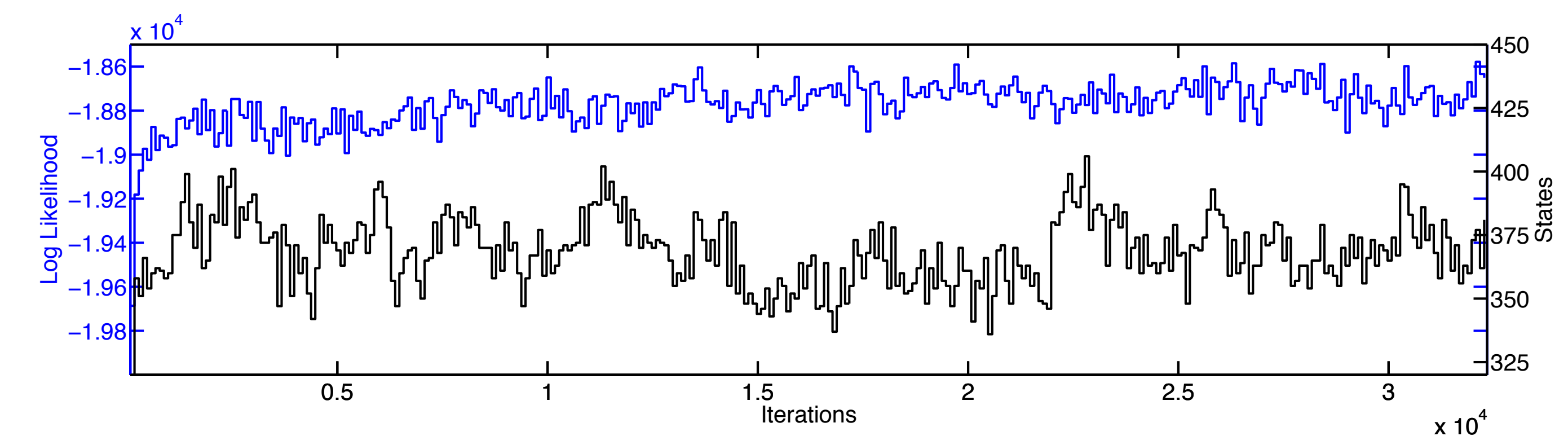
## Generative Model

## Natural Language and DNA Prediction

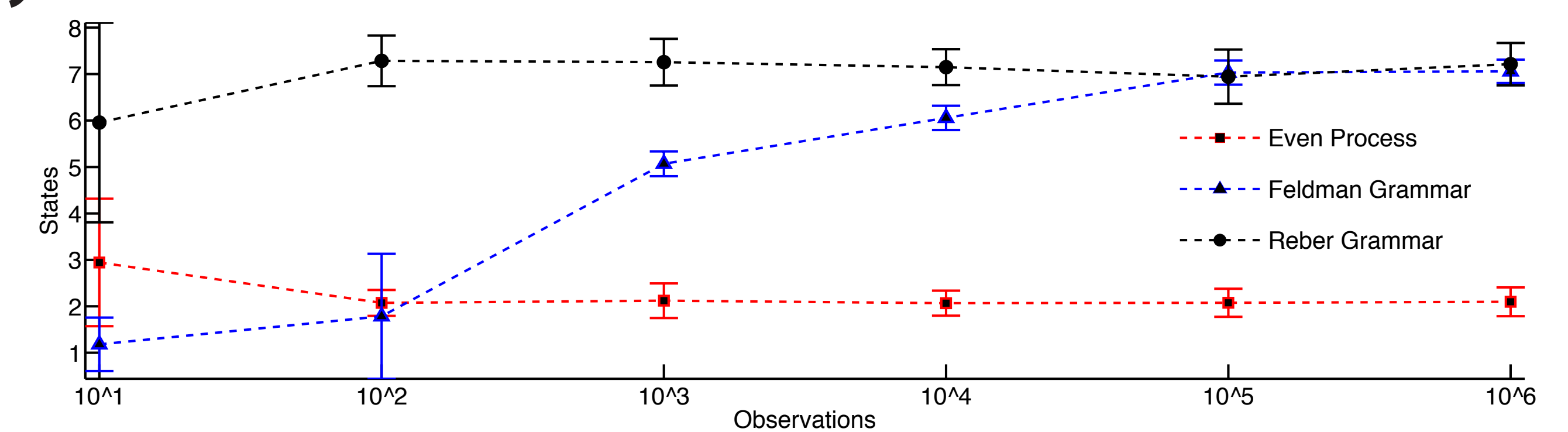
|     | PDIA          | PDIA-MAP    | HMM-EM     | bigram     | trigram     | 4-gram        | 5-gram        | 6-gram         | SSM             |
|-----|---------------|-------------|------------|------------|-------------|---------------|---------------|----------------|-----------------|
| AIW | 5.13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

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