The Sequence Memoizer

Frank Wood	Gatsby
Cedric Archambeau	UCL
Jan Gasthaus	Gatsby
Lancelot James	HKUST
Yee Whye Teh	Gatsby

Executive Summary

- Model
  - Smoothing Markov model of discrete sequences
  - Extension of hierarchical Pitman Yor process [Teh 2006]
    - Unbounded depth (context length)
- Algorithms and estimation
  - Linear time suffix-tree graphical model identification and construction
  - Standard Chinese restaurant franchise sampler
- Results
  - Maximum contextual information used during inference
  - Competitive language modelling results
    - Limit of n-gram language model as $n \rightarrow \infty$
  - Same computational cost as a Bayesian interpolating 5-gram language model

Executive Summary

- Uses
  - Any situation in which a low-order Markov model of discrete sequences is insufficient
  - Drop in replacement for smoothing Markov model
- Name?
  - "A Stochastic Memoizer for Sequence Data" $\rightarrow$ Sequence Memoizer (SM)
    - Describes posterior inference [Goodman et al '08]

Statistically Characterizing a Sequence

- Sequence Markov models are usually constructed by treating a sequence as a set of (exchangeable) observations in fixed-length contexts

$$oacac \rightarrow \begin{cases} o\|[] \\ a\|[] \\ c\|[] \\ a\|[] \\ c\|[] \end{cases} \quad oacac \rightarrow \begin{cases} a\|o \\ c\|a \\ a\|c \\ c\|a \end{cases} \quad oacac \rightarrow \begin{cases} c\|ao \\ a\|ca \\ c\|ac \end{cases} \quad oacac \rightarrow \begin{cases} a\|cao \\ c\|aca \end{cases}$$

unigram bigram trigram 4-gram

$$\dashrightarrow$$

Increasing context length / order of Markov model

Decreasing number of observations

Increasing number of conditional distributions to estimate (indexed by context)

Increasing power of model

Finite Order Markov Model

$$
\begin{aligned}
P(x_{1:N}) &= \prod_{i=1}^{N} P(x_i|x_1, \ldots x_{i-1}) \\
&\approx \prod_{i=1}^{N} P(x_i|x_{i-n+1}, \ldots x_{i-1}), \quad n = 2 \\
&= P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)\ldots
\end{aligned}
$$

- Example

$$
\begin{aligned}
P(\text{oacac}) &= P(\text{o})P(\text{a}|\text{o})P(\text{c}|\text{a})P(\text{a}|\text{c})P(\text{c}|\text{a}) \\
&= \mathcal{G}_{[]}(\text{o})\mathcal{G}_{[\text{o}]}(\text{a})\mathcal{G}_{[\text{c}]}(\text{a})\mathcal{G}_{[\text{a}]}(\text{c})\mathcal{G}_{[\text{c}]}(\text{a})
\end{aligned}
$$

Learning Discrete Conditional Distributions

- Discrete distribution $\leftrightarrow$ vector of parameters

$$\mathcal{G}_{[\mathbf{u}]} = [\pi_1, \ldots, \pi_K], K \in |\Sigma|$$

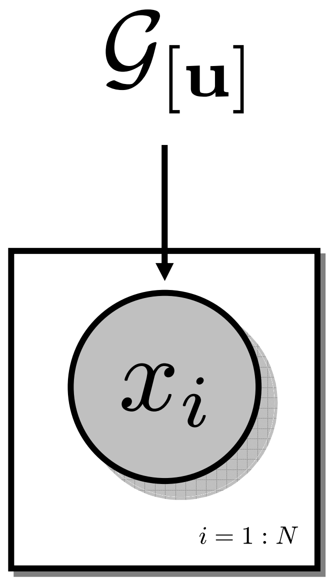
- Counting / Maximum likelihood estimation
  - Training sequence $x_{1:N}$

$$\hat{\mathcal{G}}_{[\mathbf{u}]}(X = k) = \hat{\pi}_k = \frac{\#\{\mathbf{u}k\}}{\#\{\mathbf{u}\}}$$

  - Predictive inference

$$P(X_{n+1}|x_1 \ldots x_N) = \hat{\mathcal{G}}_{[\mathbf{u}]}(X_{n+1})$$

- Example
  - Non-smoothed unigram model ($\mathbf{u} = \epsilon$)



Bayesian Smoothing

- Estimation

$$P(\mathcal{G}_{[\mathbf{u}]}|x_{1:n}) \propto P(x_{1:n}|\mathcal{G}_{[\mathbf{u}]})P(\mathcal{G}_{[\mathbf{u}]})$$

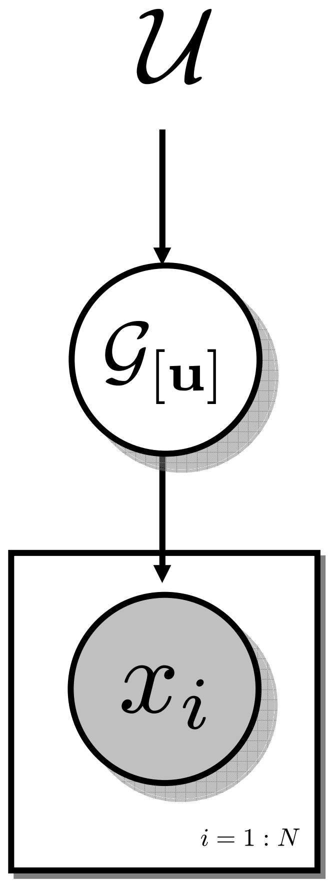
- Predictive inference

$$P(X_{n+1}|x_{1:n}) = \int P(X_{n+1}|\mathcal{G}_{[\mathbf{u}]})P(\mathcal{G}_{[\mathbf{u}]}|x_{1:n})d\mathcal{G}_{[\mathbf{u}]}$$

- Priors over distributions

$$\mathcal{G}_{[\mathbf{u}]} \sim \text{Dirichlet}(\mathcal{U}), \quad \mathcal{G}_{[\mathbf{u}]} \sim \text{PY}(d, c, \mathcal{U})$$

- Net effect
  - Inference is "smoothed" w.r.t. uncertainty about unknown *distribution*
- Example
  - Smoothed unigram ($\mathbf{u} = \epsilon$)



A Way To Tie Together Distributions

$$
\begin{aligned}
\mathcal{G}_{[\mathbf{u}]} &\sim \mathrm{PY}(\overset{\text{discount}}{\boxed{d}}, \overset{\text{concentration}}{\boxed{c}}, \underset{\text{base distribution}}{\boxed{G_{[\sigma(\mathbf{u})]}}}) \\
x_i &\sim \mathcal{G}_{[\mathbf{u}]}
\end{aligned}
$$

- Tool for tying together related distributions in hierarchical models
- Measure over measures
- Base measure is the "mean" measure

$$E[\mathcal{G}_{[\mathbf{u}]}(dx)] = \mathcal{G}_{[\sigma(\mathbf{u})]}(dx)$$

- A distribution drawn from a Pitman Yor process is related to its base distribution
  - (equal when $c = \infty$ or $d = 1$)

[Pitman and Yor '97]

Pitman-Yor Process Continued

- Generalization of the Dirichlet process ($d = 0$)
  - Different (power-law) properties
  - Better for text [Teh, 2006] and images [Sudderth and Jordan, 2009]
- Posterior predictive distribution

Can't actually do this integral this way

$$\begin{aligned} P(X_{N+1}|x_{1:N}; c, d) &\approx \int P(x_{N+1}|\mathcal{G}_{[\mathbf{u}]}) P(\mathcal{G}_{[\mathbf{u}]}|x_{1:N}; c, d) d\mathcal{G}_{[\mathbf{u}]} \\ &= \mathbb{E}\left[\frac{\sum_{k=1}^{K} (m_k - d)\mathbb{I}(\phi_k = X_{N+1})}{c + N} + \frac{c + dK}{c + N} \mathcal{G}_{[\sigma(\mathbf{u})]}(X_{N+1}) \right] \end{aligned}$$

- Forms the basis for straightforward, simple samplers
- Rule for stochastic memoization

Hierarchical Bayesian Smoothing

- Estimation

$$\Theta = \{\mathcal{G}_{[\mathbf{u}]}, \mathcal{G}_{[\mathbf{v}]}, \mathcal{G}_{[\mathbf{w}]}\}, \quad \mathbf{w} = \sigma(\mathbf{u}) = \sigma(\mathbf{v})$$
$$P(\Theta|x_{1:N}) \propto P(x_{1:N}|\Theta)P(\Theta)$$

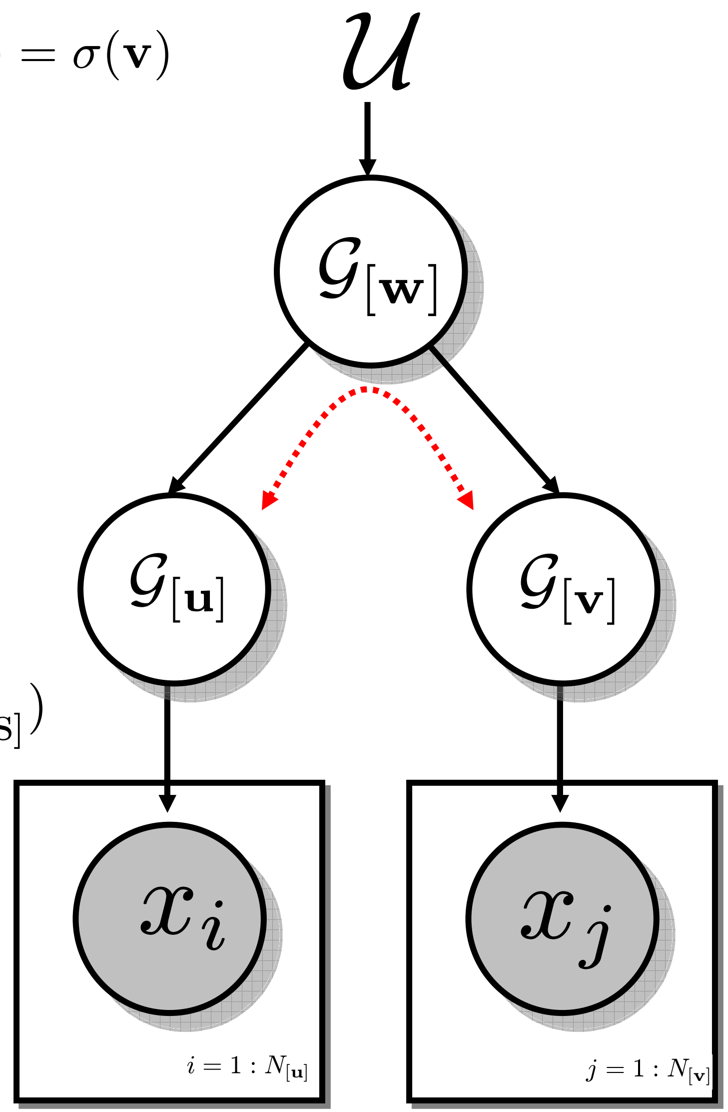
- Predictive inference

$$P(X_{N+1}|x_{1:N}) = \int P(X_{N+1}|\Theta)P(\Theta|x_{1:N})d\Theta$$

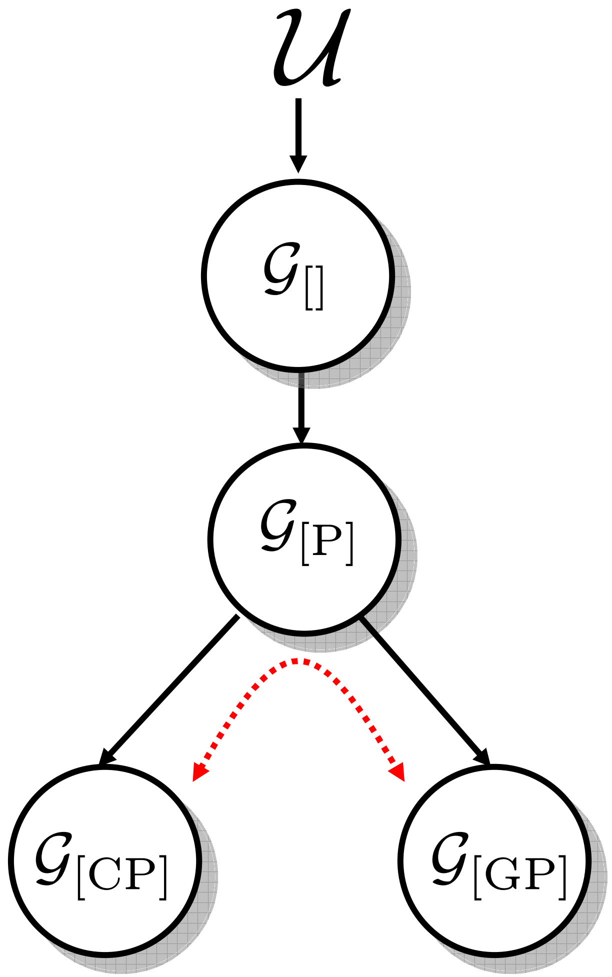
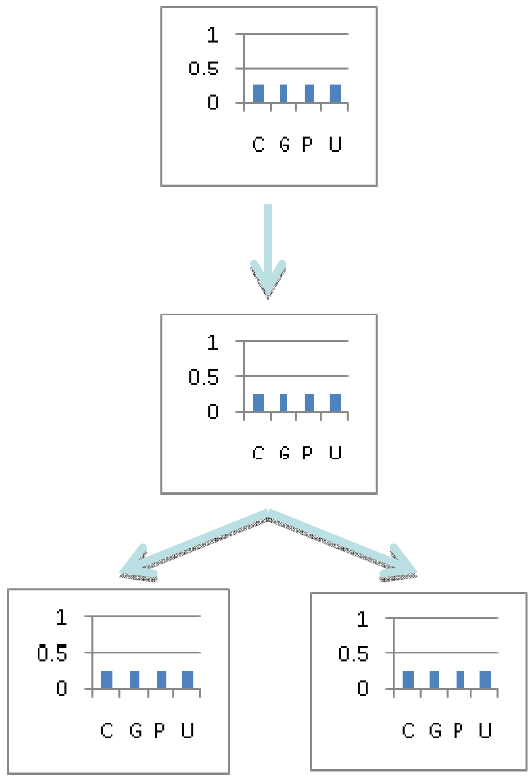
- Naturally related distributions tied together

$$\mathcal{G}_{[\text{the United States}]} \sim \text{PY}(d, c, \mathcal{G}_{[\text{United States}]})$$

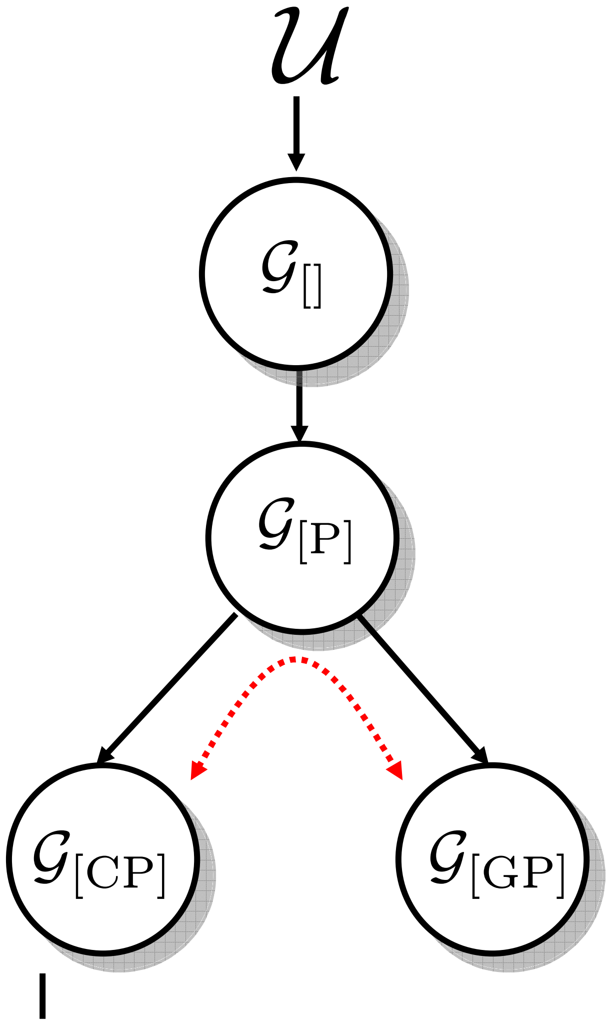
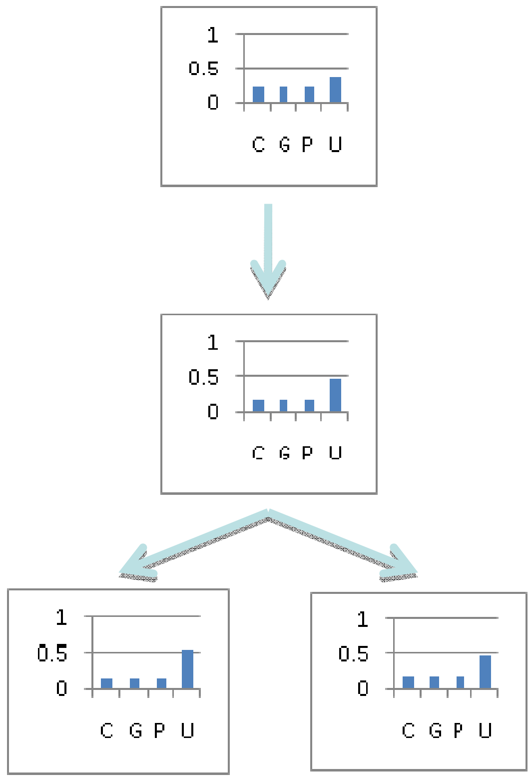
- Net effect
  - Observations in one context affect inference in other context.
  - Statistical strength is shared between similar contexts
- Example
  - Smoothing bi-gram ($\mathbf{w} = \epsilon$, $\mathbf{u}$,$\mathbf{v} \in \mathbf{\Sigma}$)



SM/HPYP Sharing in Action

Observations	Conditional Distributions	Posterior Predictive Probabilities
		

CRF Particle Filter Posterior Update

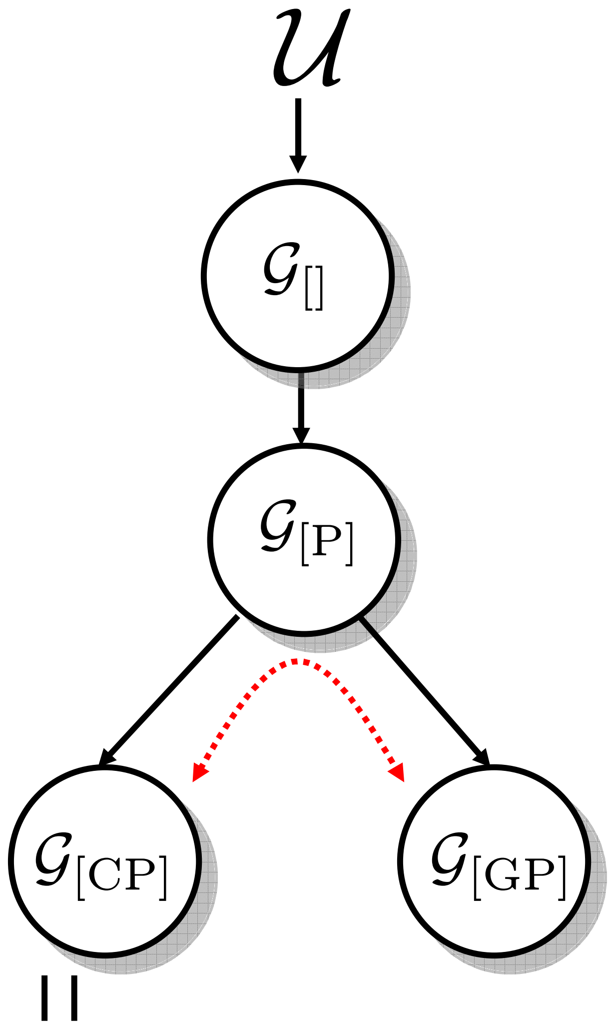
Observations	Conditional Distributions	Posterior Predictive Probabilities
CPU		

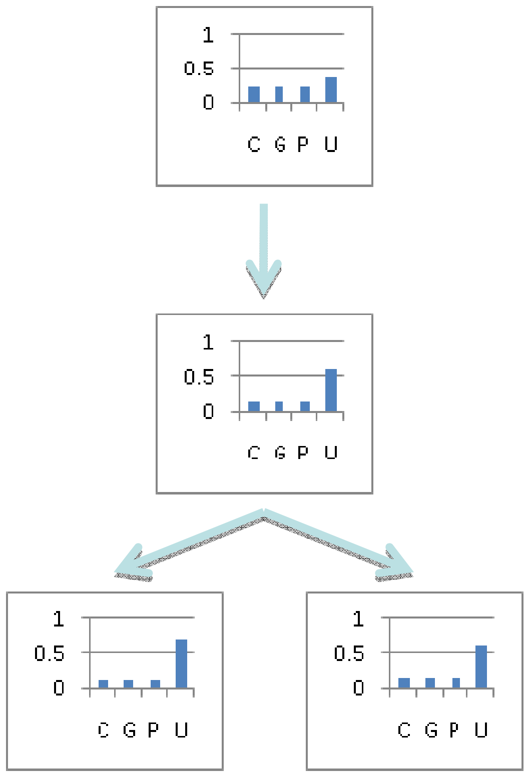
CRF Particle Filter Posterior Update

Observations	Conditional Distributions	Posterior Predictive Probabilities

CPU

CPU





HPYP LM Sharing Architecture

- Share statistical strength between sequentially related predictive conditional distributions
  - Estimates of highly specific conditional distributions

$$\mathcal{G}_{[\text{was on the}]}$$

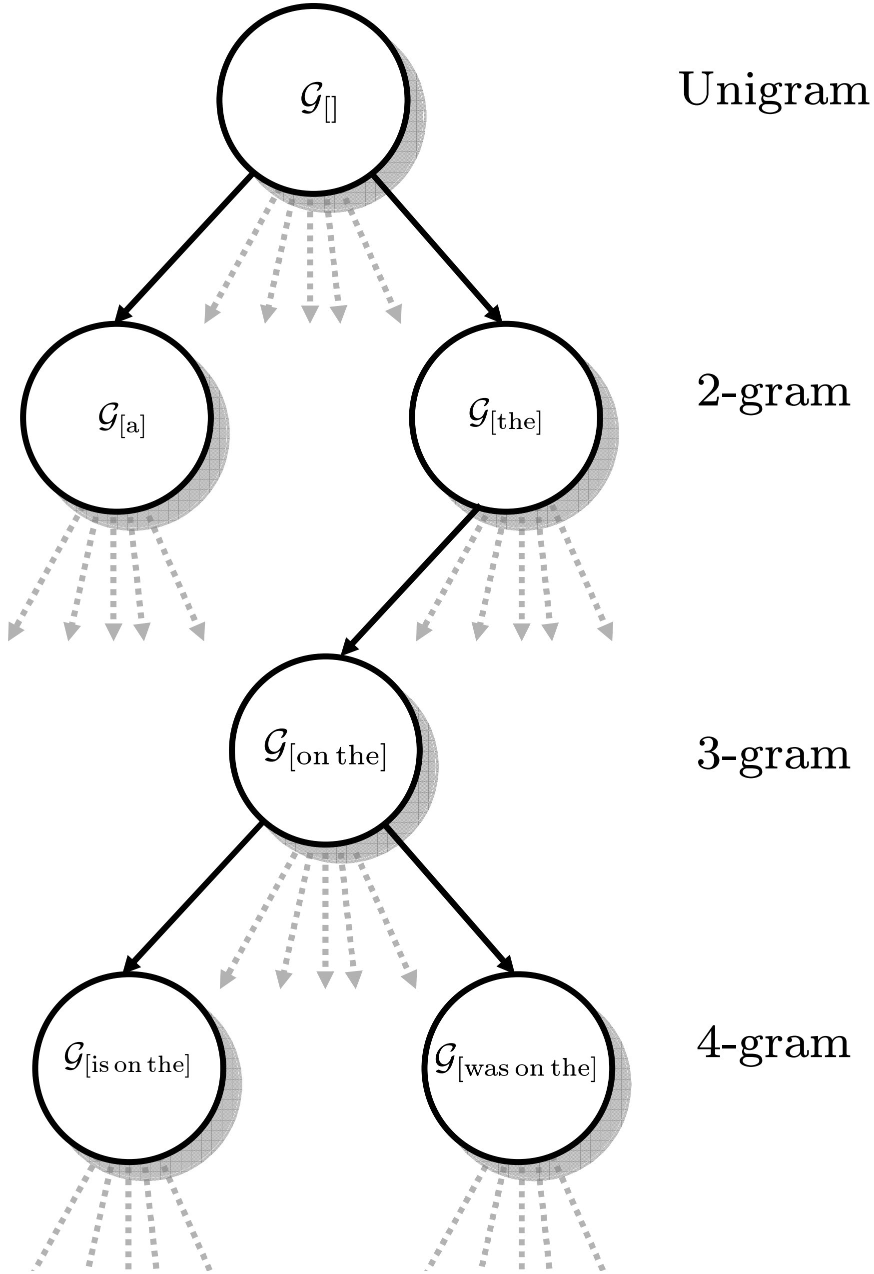
  - Are coupled with others that are related

$$\mathcal{G}_{[\text{is on the}]}$$

  - Through a single common, more-general shared ancestor

$$\mathcal{G}_{[\text{on the}]}$$

- Corresponds intuitively to back-off



Hierarchical Pitman Yor Process

$$\begin{aligned}
\mathcal{G}_{[]} \mid d_0, \mathcal{U} &\sim \mathrm{PY}(d_0, 0, \mathcal{U}) \\
\mathcal{G}_{[\mathbf{u}]} \mid d_{|\mathbf{u}|}, \mathcal{G}_{[\sigma(\mathbf{u})]} &\sim \mathrm{PY}(d_{|\mathbf{u}|}, 0, \mathcal{G}_{[\sigma(\mathbf{u})]}) \\
x_i \mid \mathbf{x}_{1:i-1} = \mathbf{u} &\sim \mathcal{G}_{[\mathbf{u}]} \\
i = 1, \ldots, T & \\
\forall \mathbf{u} \in \Sigma^{n-1} &
\end{aligned}$$

- Bayesian generalization of smoothing n-gram Markov model
- Language model : outperforms interpolated Kneser-Ney (KN) smoothing
- Efficient inference algorithms exist
  - [Goldwater et al ’05; Teh, ’06; Teh, Kurihara, Welling, ’08]
- Sharing between contexts that differ in most distant symbol only
- Finite depth

[Goldwater et al ’05, Teh ’06]

Alternative Sequence Characterization

- A sequence can be characterized by a set of *single* observations in unique contexts of growing length

$$oacac \to \begin{cases} o|[] \\ a|o \\ c|ao \\ a|cao \\ c|acao \end{cases}$$

Increasing context length

Always a single observation

Foreshadowing: all suffixes of the string “cacao”

“Non-Markov” Model

$$
\begin{aligned}
P(x_{1:N}) &= \prod_{i=1}^{N} P(x_i|x_1,\ldots x_{i-1}) \\
&= P(x_1)P(x_2|x_1)P(x_3|x_2,x_1)P(x_4|x_3,\ldots x_1)\ldots
\end{aligned}
$$

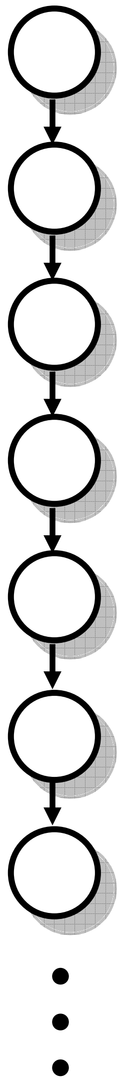
- Example

$$
P(\text{oacac}) = P(\text{o})P(\text{a}|\text{o})P(\text{c}|\text{oa})P(\text{a}|\text{oac})P(\text{c}|\text{oaca})
$$

- Smoothing essential
  - Only one observation in each context!
- Solution
  - Hierarchical sharing ala HPYP

Sequence Memoizer

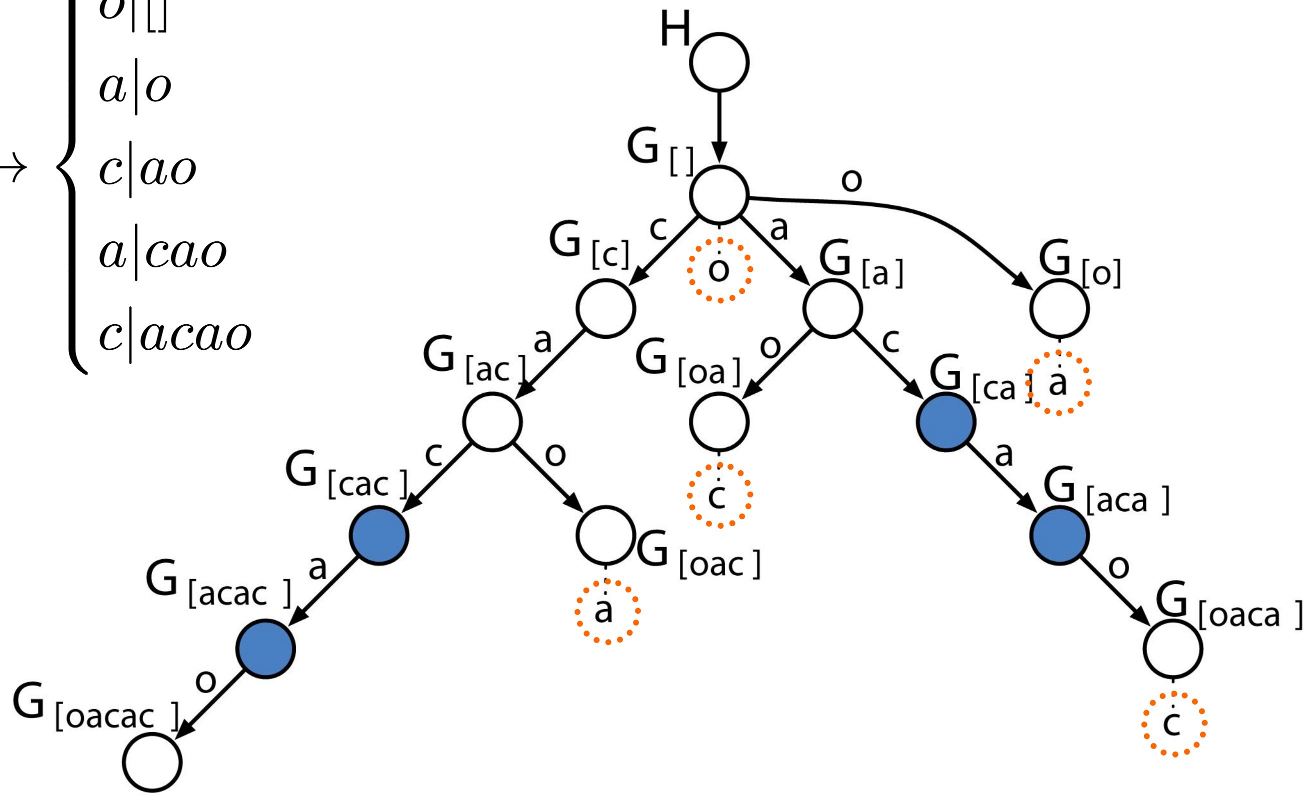
$$
\begin{aligned}
\mathcal{G}_{[]} \mid d_0, \mathcal{U} &\sim \mathrm{PY}(d_0, 0, \mathcal{U}) \\
\mathcal{G}_{[\mathbf{u}]} \mid d_{|\mathbf{u}|}, \mathcal{G}_{[\sigma(\mathbf{u})]} &\sim \mathrm{PY}(d_{|\mathbf{u}|}, 0, \mathcal{G}_{[\sigma(\mathbf{u})]}) \\
x_i \mid \mathbf{x}_{1:i-1} = \mathbf{u} &\sim \mathcal{G}_{[\mathbf{u}]} \\
i = 1, \ldots, T & \\
\forall \mathbf{u} \in \Sigma^{+} &
\end{aligned}
$$



- Eliminates Markov order selection
- Always uses full context when making predictions
- Linear time, linear space (in length of observation sequence) graphical model identification
- Performance is limit of n-gram as $n \to \infty$
- Same or less overall cost as 5-gram interpolating Kneser Ney

Graphical Model Trie

$$oacac \rightarrow \begin{cases} o \,|\, [] \\ a \,|\, o \\ c \,|\, ao \\ a \,|\, cao \\ c \,|\, acao \end{cases}$$



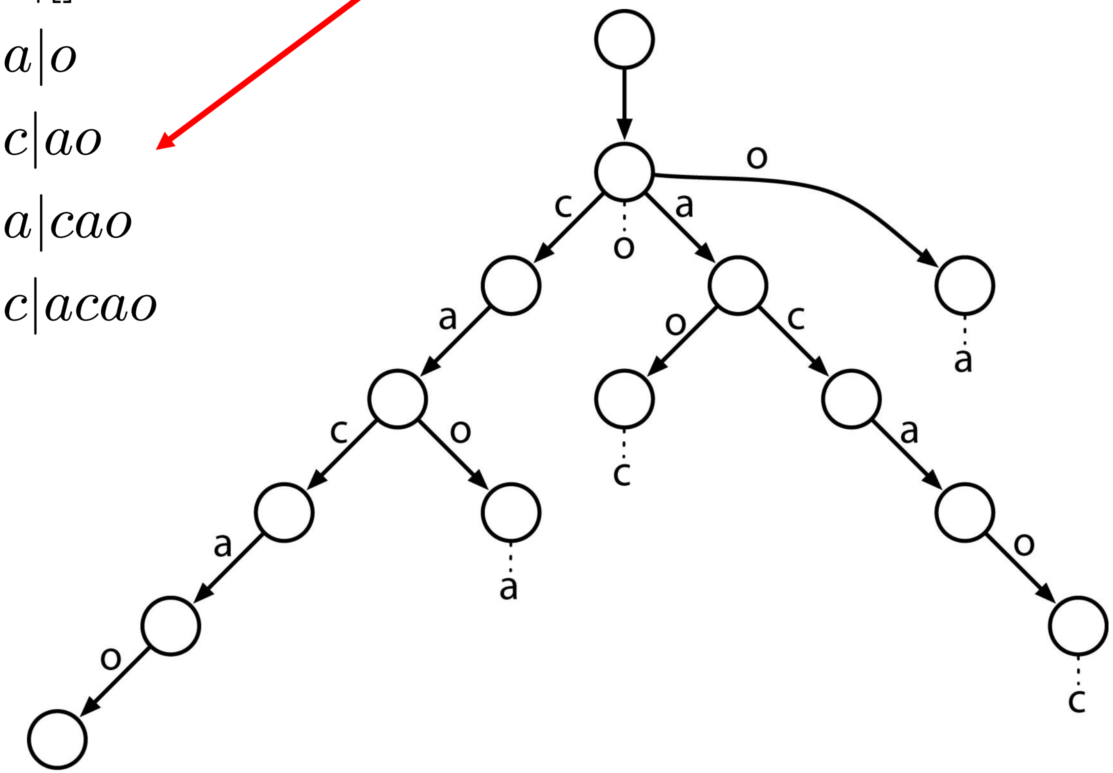
Observations

Latent conditional distributions with Pitman Yor priors / stochastic memoizers

Suffix Trie Datastructure

All suffixes of the string “cacao”

$$oacac \rightarrow \begin{cases} o \,|\, [\,] \\ a \,|\, o \\ c \,|\, ao \\ a \,|\, cao \\ c \,|\, acao \end{cases}$$



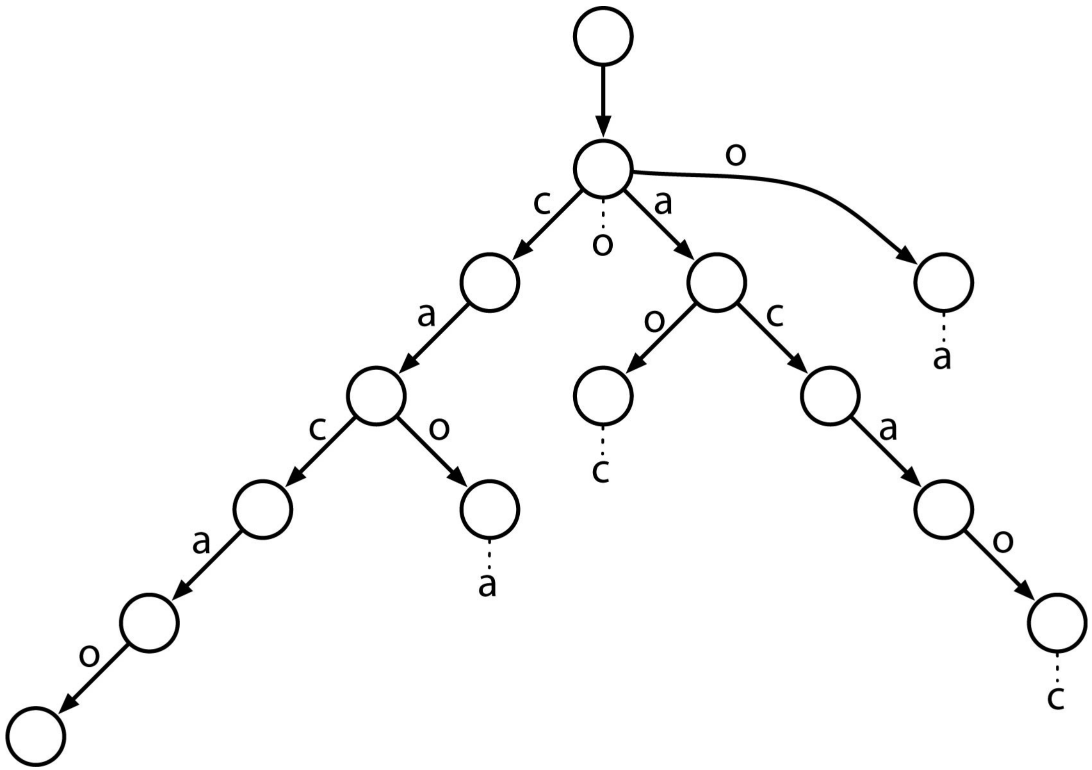
Suffix Trie Datastructure

- Deterministic finite automata that recognizes all suffixes of an input string.
- Requires O(N^2) time and space to build and store [Ukkonen, 95]
- Too intensive for any practical sequence modelling application.

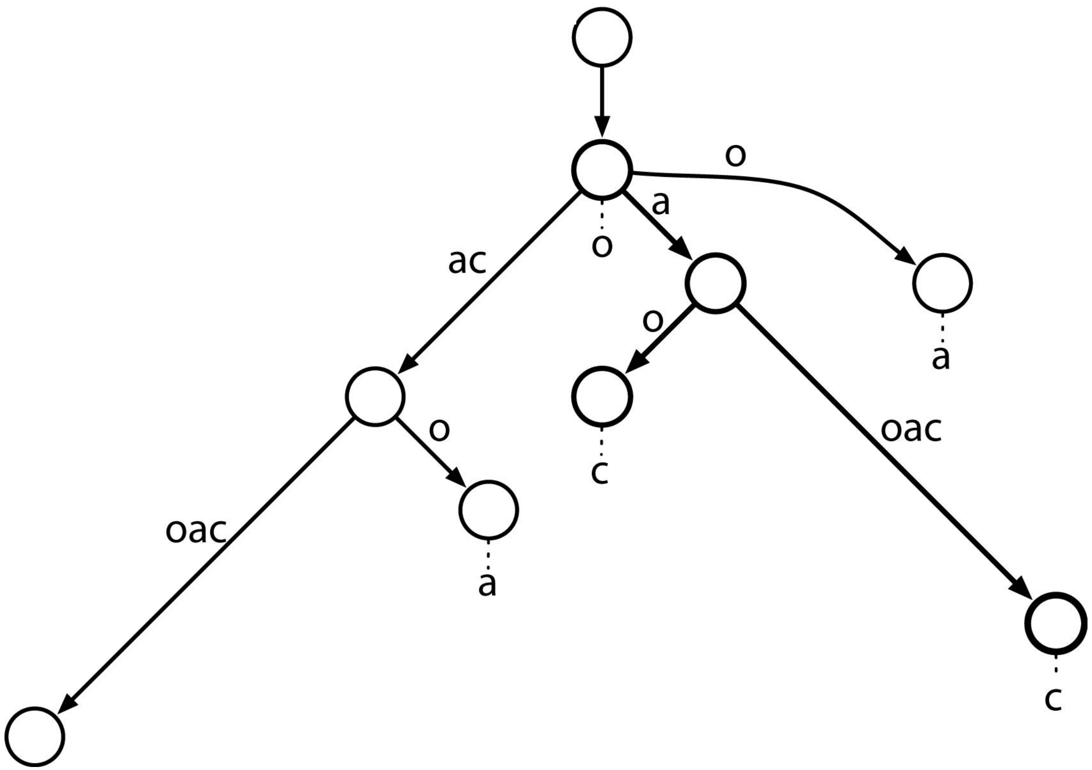
Suffix Tree

- Deterministic finite automata that recognizes all suffixes of an input string
- Uses path compression to reduce storage and construction computational complexity.
- Requires only O(N) time and space to build and store [Ukkonen, 95]
- Practical for large scale sequence modelling applications

Suffix Trie Datastructure



Suffix Tree Datastructure



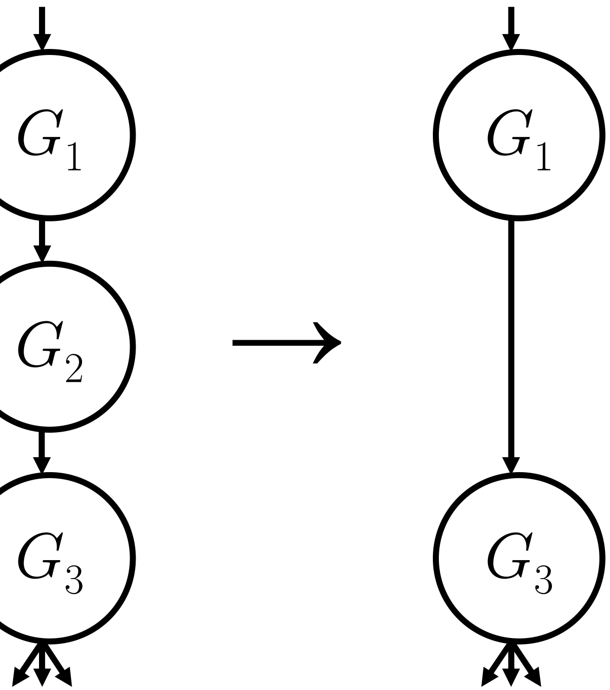
Graphical Model Identification

- This is a graphical model transformation under the covers.
- These compressed paths require being able to analytically marginalize out nodes from the graphical model
- The result of this marginalization can be thought of as providing a different set of caching rules to memoizers on the path-compressed edges

Marginalization

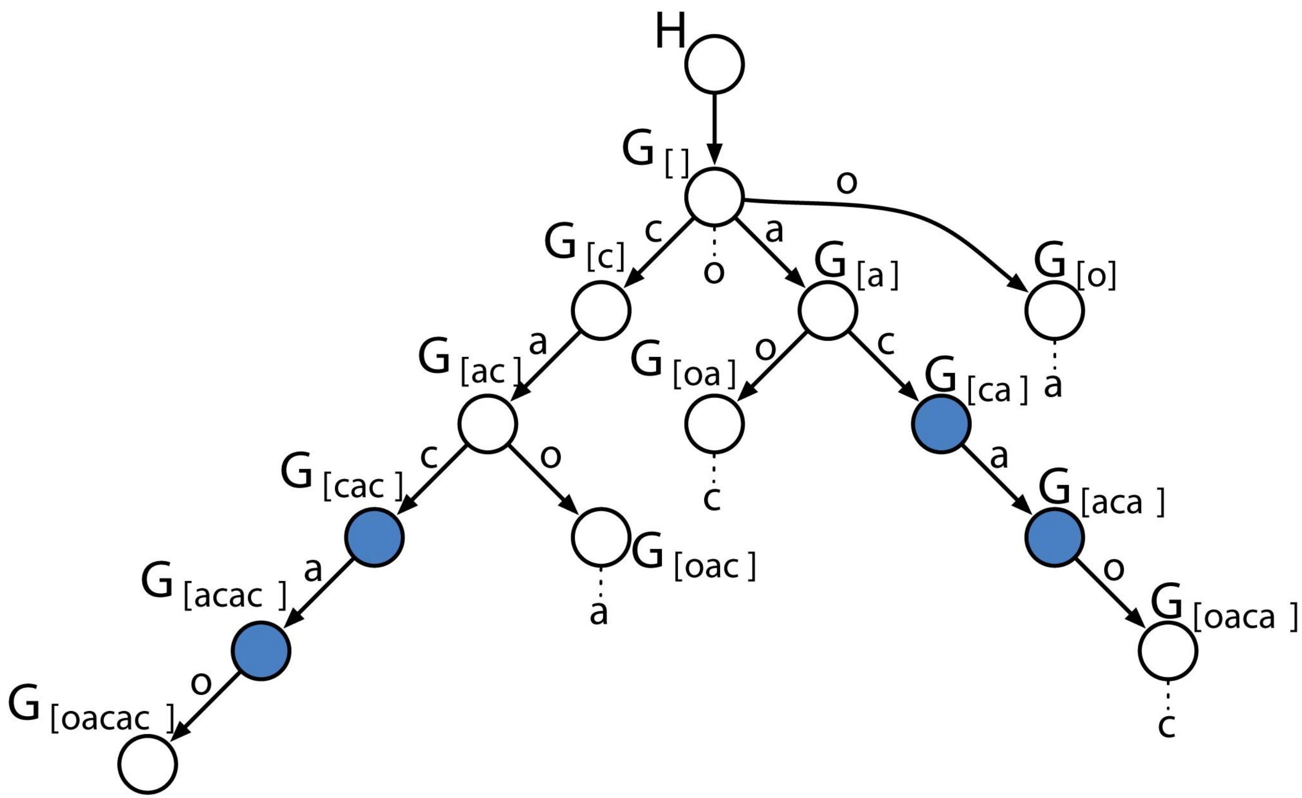
- Theorem 1: Coagulation

If $G_2|G_1 \sim \mathrm{PY}(d_1, 0, G_1)$ and $G_3|G_2 \sim \mathrm{PY}(d_2, 0, G_2)$
then $G_3|G_1 \sim \mathrm{PY}(d_1 d_2, 0, G_1)$ with G_2 marginalized out.

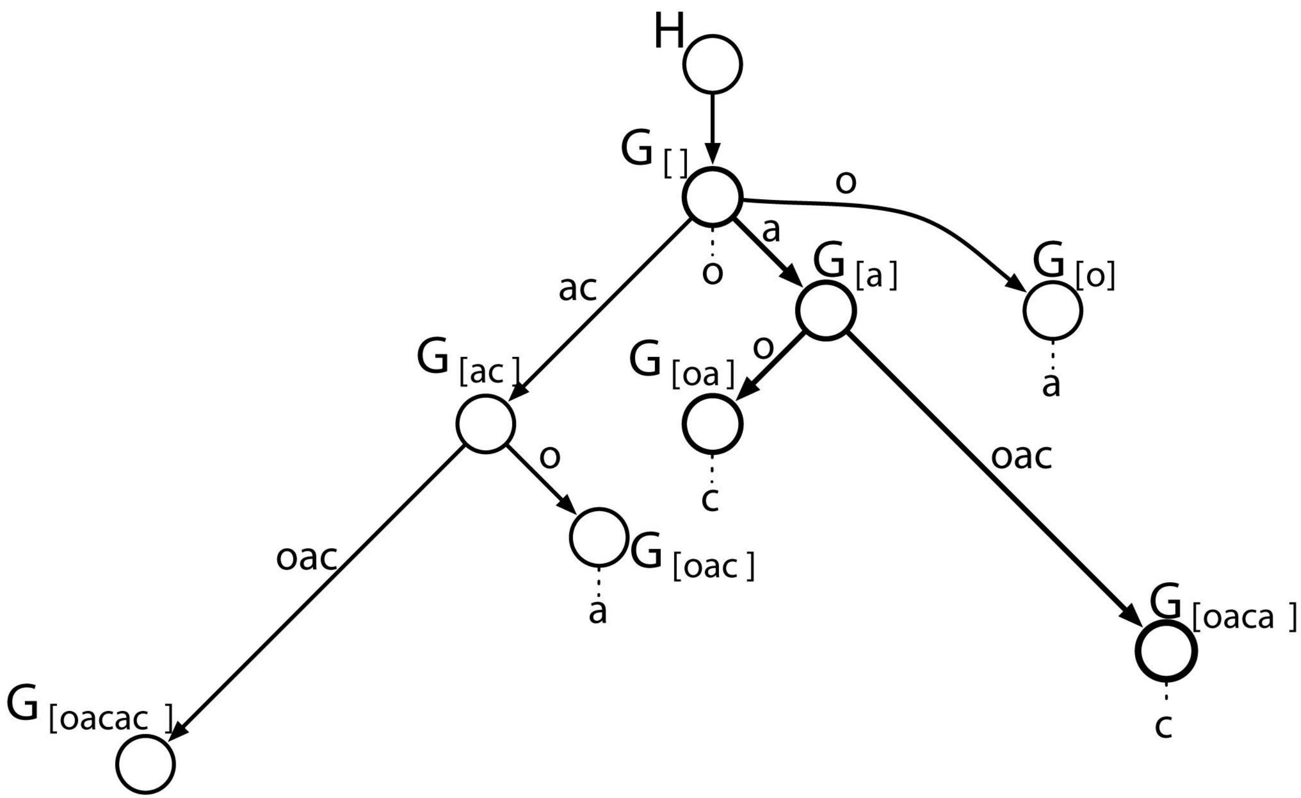


[Pitman '99; Ho, James, Lau '06; W., Archambeau, Gasthaus, James, Teh '09]

Graphical Model Trie



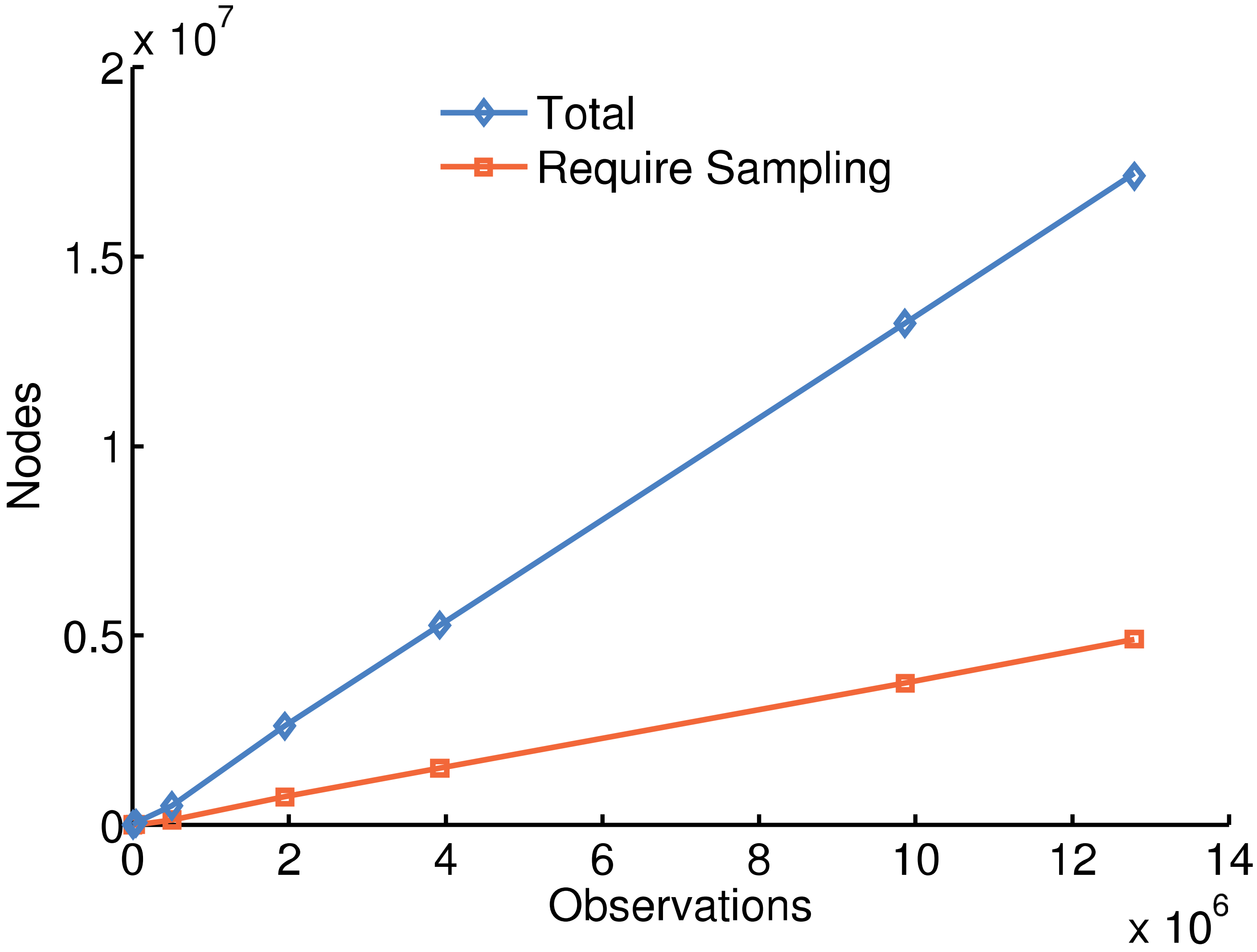
Graphical Model Tree



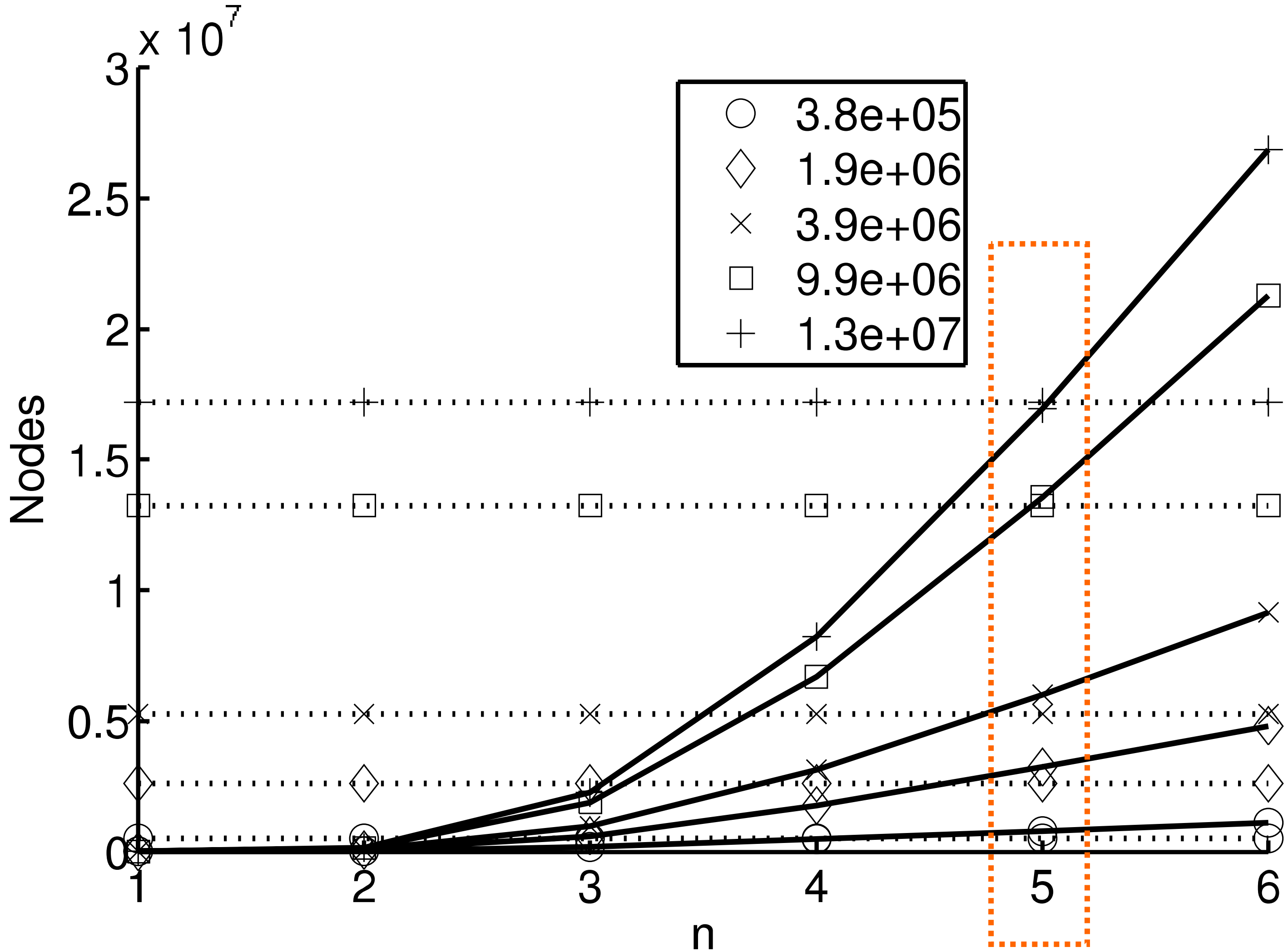
Graphical Model Initialization

- Given a single input sequence
  - Ukkonen's linear time suffix tree construction algorithm is run on its reverse to produce a prefix tree
  - This identifies the nodes in the graphical model we need to represent
  - The tree is traversed and path compressed parameters for the Pitman Yor processes are assigned to each remaining Pitman Yor process

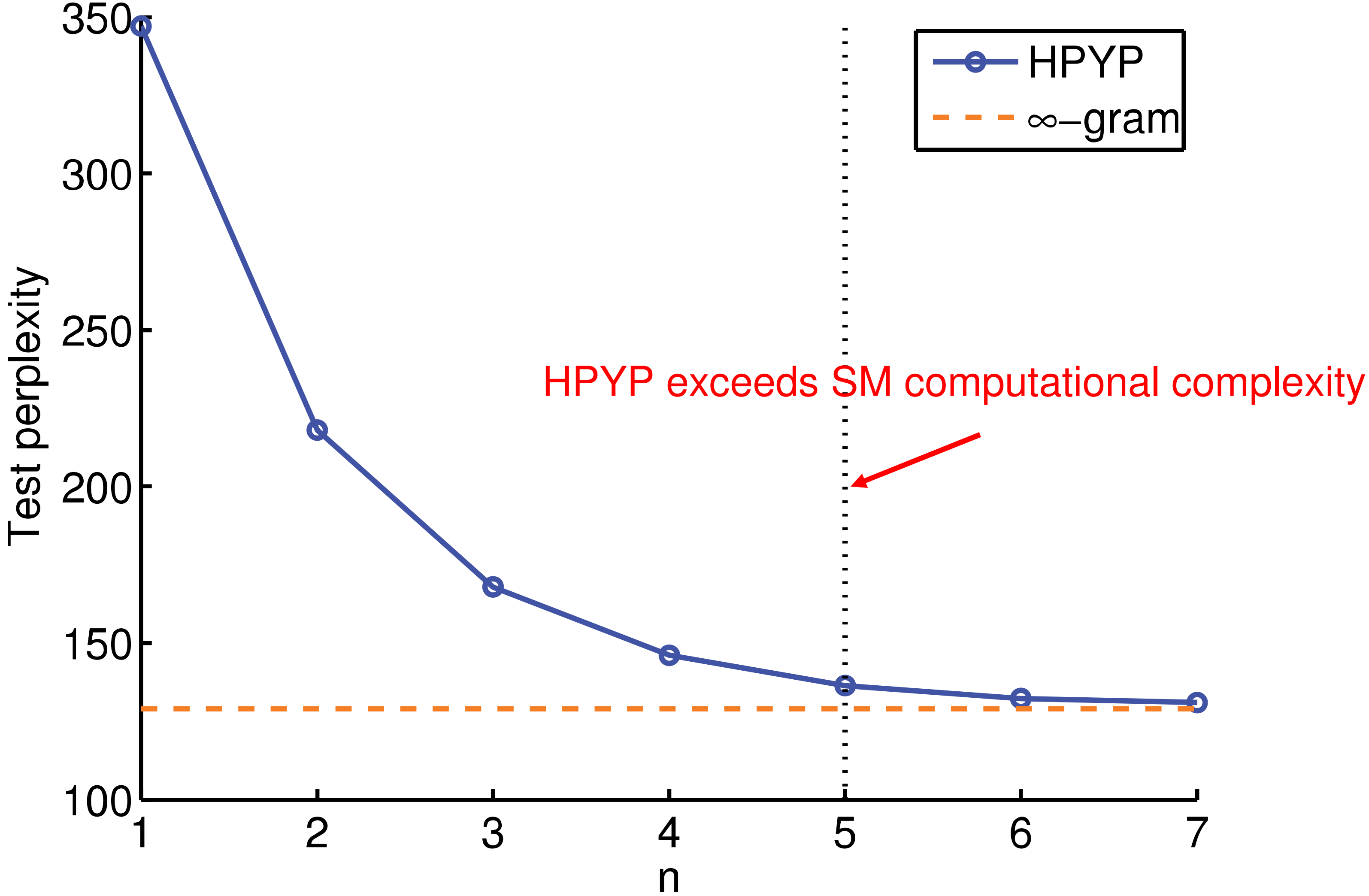
Nodes In The Graphical Model



Never build more than a 5-gram



Sequence Memoizer Bounds N-Gram Performance



Language Modelling Results

AP News Test Perplexity

[Mnih & Hinton, 2009]	112.1
[Bengio et al., 2003]	109.0
4-gram Modified Kneser-Ney [Teh, 2006]	102.4
4-gram HPYP [Teh, 2006]	101.9
Sequence Memoizer (SM)	96.9

The Sequence Memoizer

- The Sequence Memoizer is a deep (unbounded) smoothing Markov model
- It can be used to learn a joint distribution over discrete sequences in time and space linear in the length of a single observation sequence
- It is equivalent to a smoothing ∞-gram but costs no more to compute than a 5-gram