The Infinite Markov Model: A Nonparametric Bayesian approach

Daichi Mochihashi

NTT Communication Science Laboratories

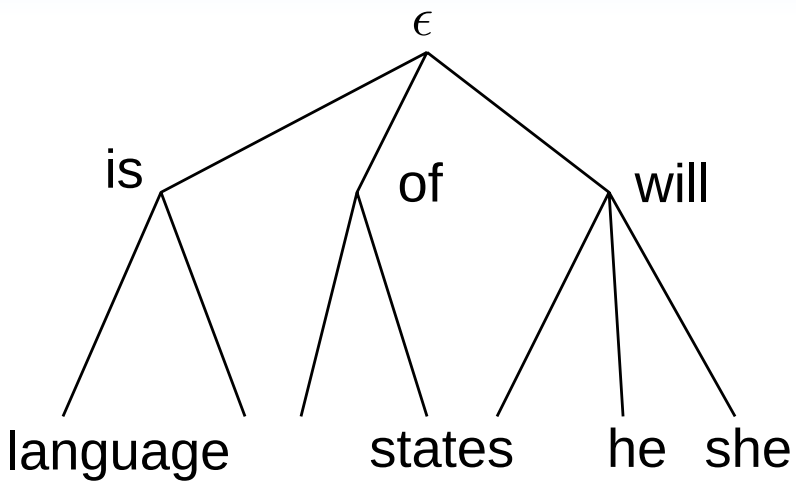
Postdoctoral Research Associate

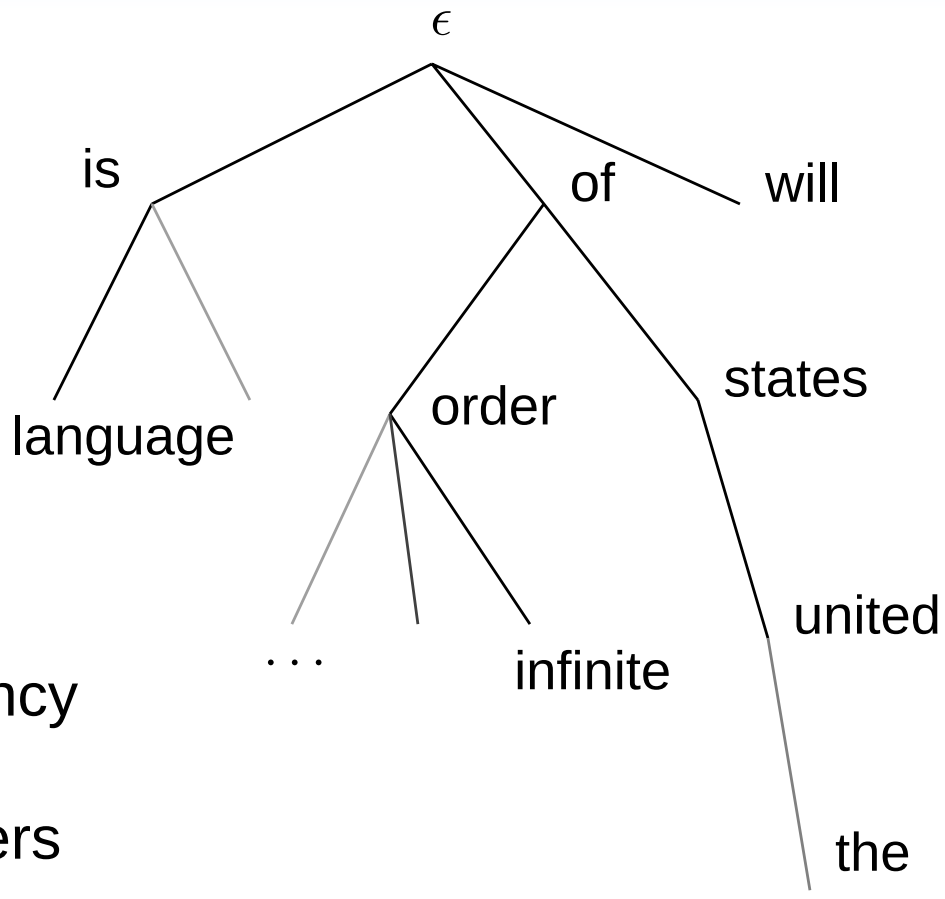
daichi@cslab.kecl.ntt.co.jp

ISM Bayesian Inference Workshop

2007-08-21

Overview





Fixed n-th order Markov model

- Fixed-order Markov dependency

  $\Downarrow$

  Infinitely variable Markov orders

- Prior $\rightarrow$ Posterior distribution for latent trees
  - How to draw an inference based on only the output sequences?

Infinite (Unbounded) Variable-order Markov model

Markov Models (1/3)

- Markov Models: first by Shannon (1948)
  - extremely simple model for discrete sequences
  - $p(s_t|s_{t-1}\cdots s_{t-n})$: n-th order Markov Model
- Widely used in sequence modeling:
  - Natural language processing & speech recognition
    "Could you show me the *ware* to Tokyo terminal?"
  - Music modeling: Transitions between musical notes
    "c<q6b-aq8>>c4r<dq6c<b-q8>>d4" (Vivaldi "L'estate")
  - Bioinformatics
    "...cctttccggtgatccgacaggttacg"
    - Enterobacteria phage Lambda: GeneBank J02459

    "...QYVTVFYGVPVWKEAKTHLICATDNS"
    - Amino acids: GP 120: Pfam

Markov Models (2/3)

- From Shannon: "*A mathematical theory of communication*" (1948)

3. THE SERIES OF APPROXIMATIONS TO ENGLISH

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

Markov Models (2/3)

- From Shannon: "*A mathematical theory of communication*" (1948)

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Markov Models (2/3)

- From Shannon: "*A mathematical theory of communication*" (1948)

  5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

     REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

  6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

     THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

- "n-gram" models $\cdots$ $(n-1)$ order Markov models over words
- Simple but accurate statistics for language
  - Widely used in speech recognition, statistical machine translation, and many others
  - We'll concentrate on this for example (but notice music and DNA)

Markov Models (3/3)

- Problem: What Markov order should we use?
  - *Exponentially large* number of states for higher-order models
  - Words $|V| = 10,000$: $|V|^2 = 100,000,000$ (3-grams), $|V|^3 = 1,000,000,000,000$ (4-grams), ...
- Fixed order model: *Full of noises*
  - Google 5-grams: compressed 24GB of word 5-tuples
  - many linguistically meaningless tuples $\rightarrow$ memory overflow
- Language will have non-stationary dependencies
  - "the united states of america"
  - "superior to"
- What about DNAs and musical transitions?

Variable Markov Models

- "Variable-order Markov Model"
  - Ron, Singer, Tishby (1994): machine learning
  - Buhlmann and Wyner (1999): statistics
- Application: Bioinformatics (Apostolico and Bejerano, 2000) (Leonardi, 2006), natural language processing (Pereira+, 1995) (Stolke, 1998) (Siu and Ostendorf 2000) ...
- But not so widely used (why?)

  ↓
- Previous approaches: Pruning a huge candidate model (of some maximum order)
  - Candidate model will be exponentially large

    $\cdots$ contradicts to our objective
  - Pruning criteria are inherently exogeneous

    $\cdots$ such as KL-divergences (why KL?)

    $\cdots$ how to interpret them?

Variable Markov Models (cont'd)

- Why only these post-hoc approaches?

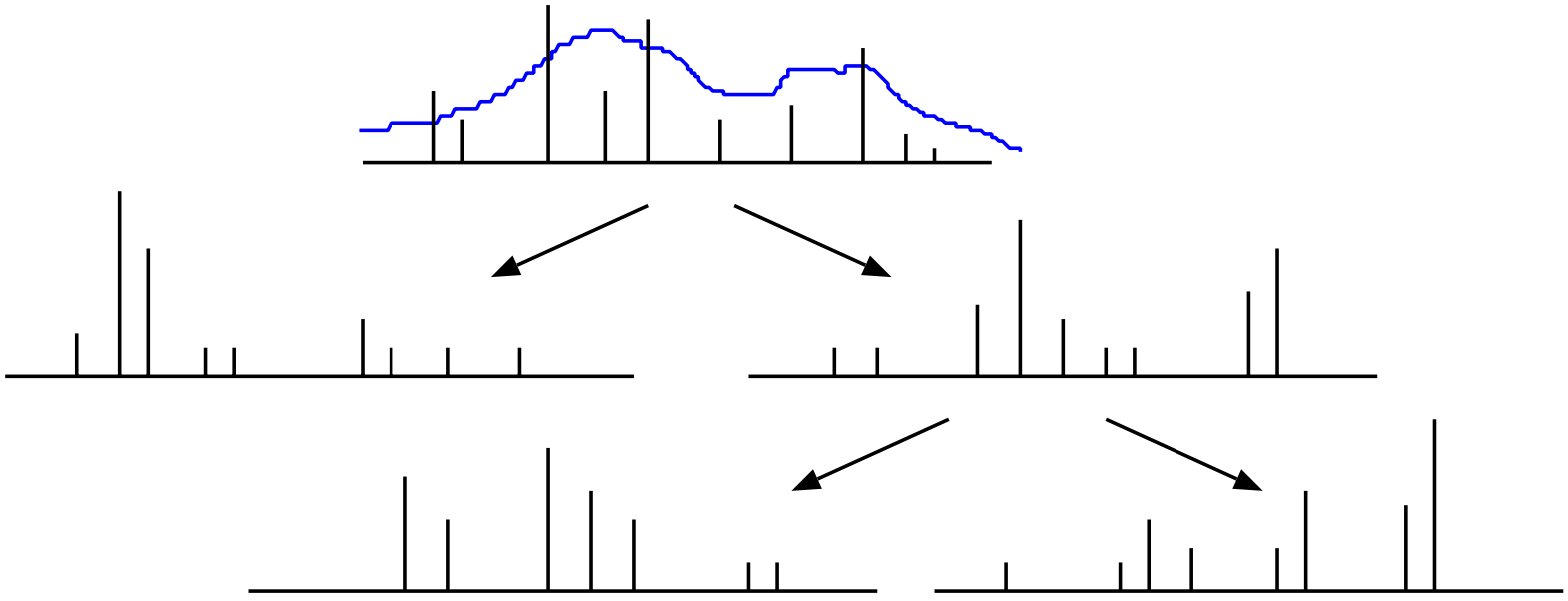
  $\Downarrow$

- Higher-order Markov Models get sparser and sparser
  - order-n Markov Model depends on order-(n-1) Markov Model
  - but, no complete generative models have existed
- However...

Bayesian Markov Model

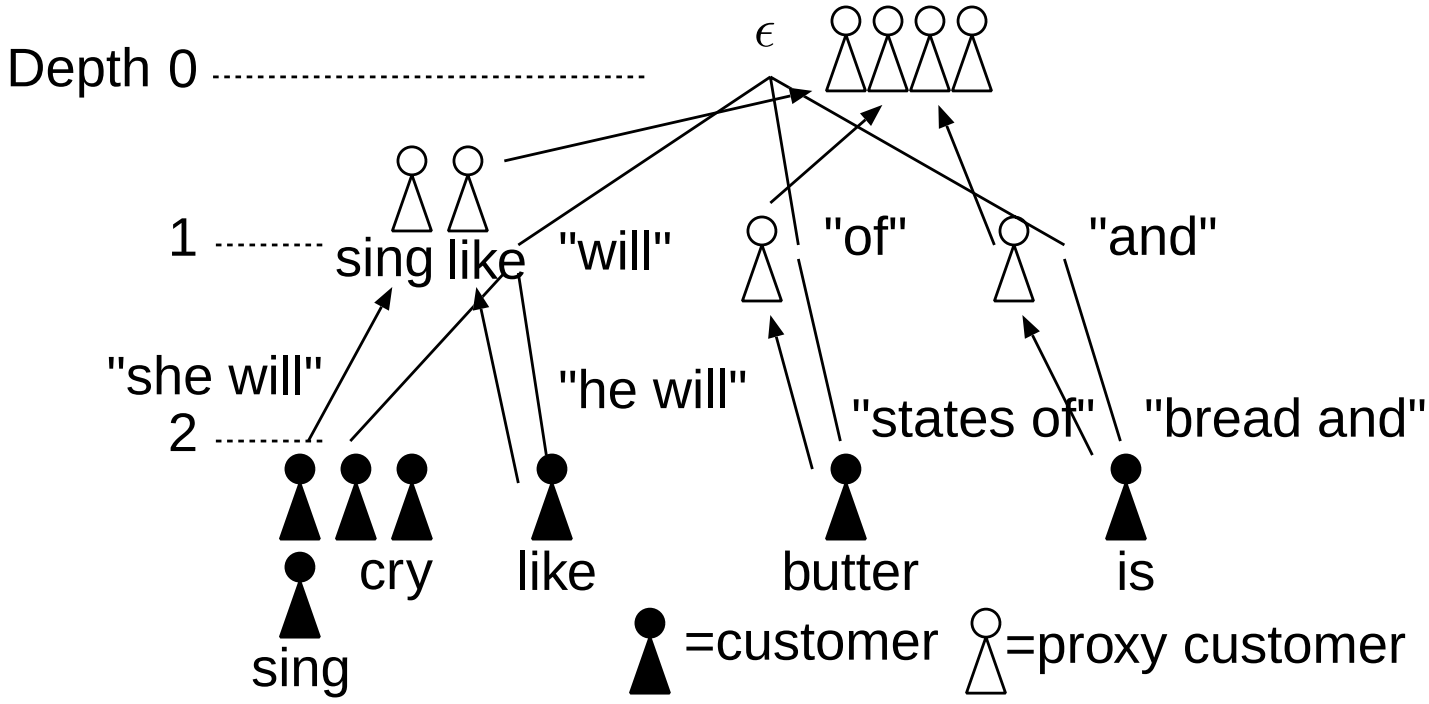
- Nonparametric Bayesian model of discrete distributions
  - Dirichlet processes: infinite dimension discrete distributions

    $\downarrow$
- Hierarchical Dirichlet Processes (Teh et al., 2004)
- Hierarchical Pitman-Yor Processes (Teh, 2006)
  - “Pitman-Yor” process = Two-parameter Poisson-Dirichlet process $\mathrm{PD}(\alpha, \theta)$ (Pitman and Yor 1997) in statistics



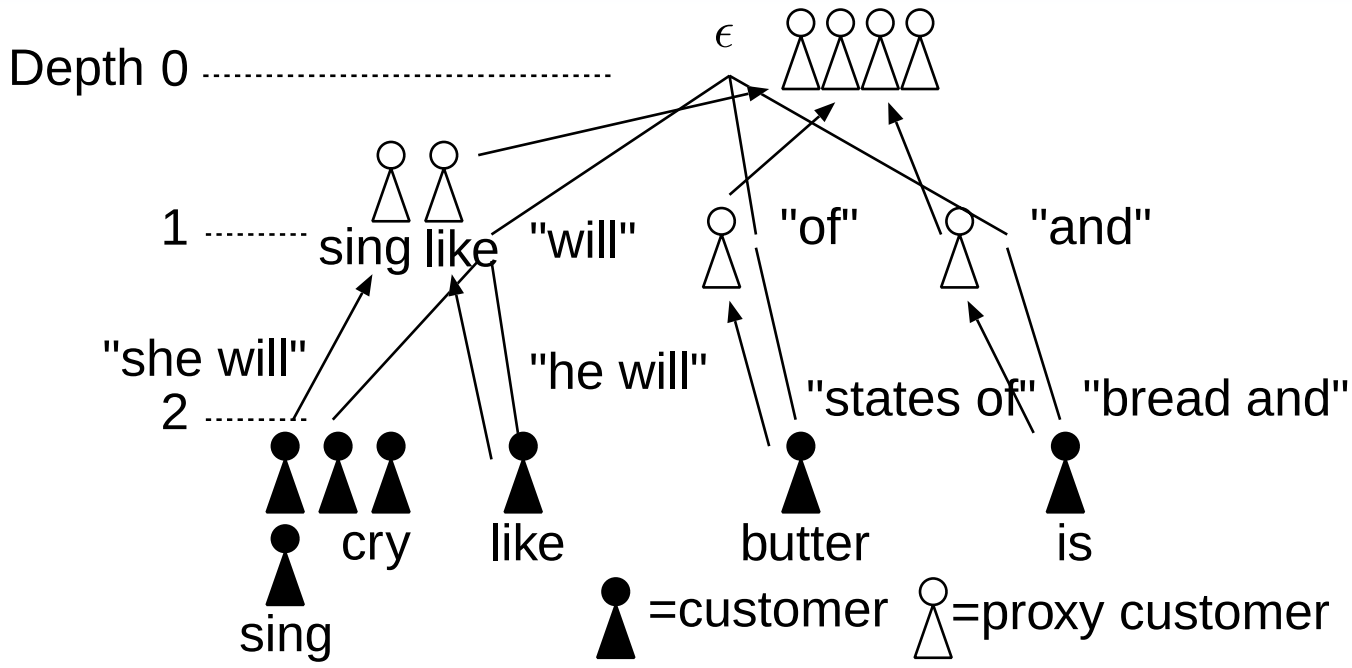
Hierarchical Pitman-Yor Language Model (Teh 2006)

- n-gram model = Suffix Tree of depth $(n-1)$
- For a 3-gram model,



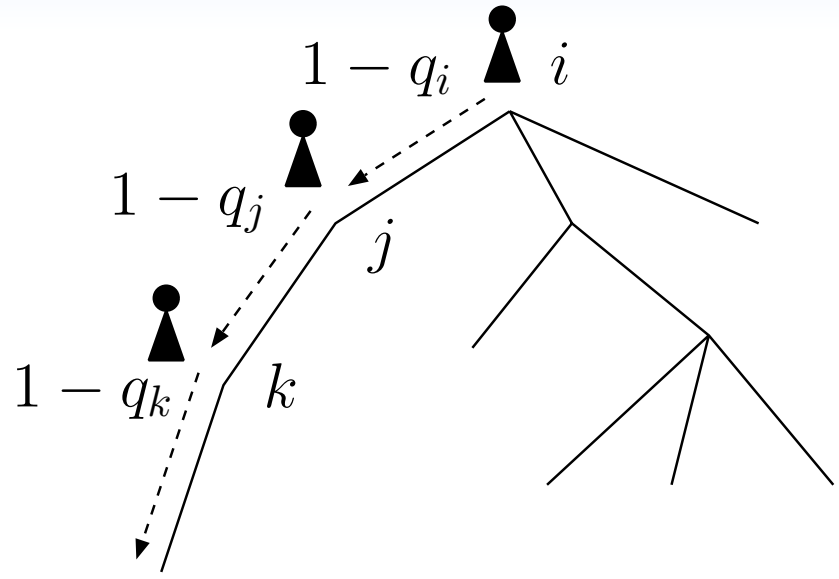
  - Descend "context node" backwards
  - Each context node is a Chinese restaurant
    - Send a "proxy customer" upwards and smoothing using the base measure in parent node

Problem with HPYLM



- Does it suffice that all customers reside in depth 2 in the Suffix Tree?
  - “will sing”: only one word dependency (2-gram)
  - “the united states of america”: 4 words dependency (5-gram)
- How to deploy customers in suitably different depths?

Variable-order Pitman-Yor Language Model (VPYLM)



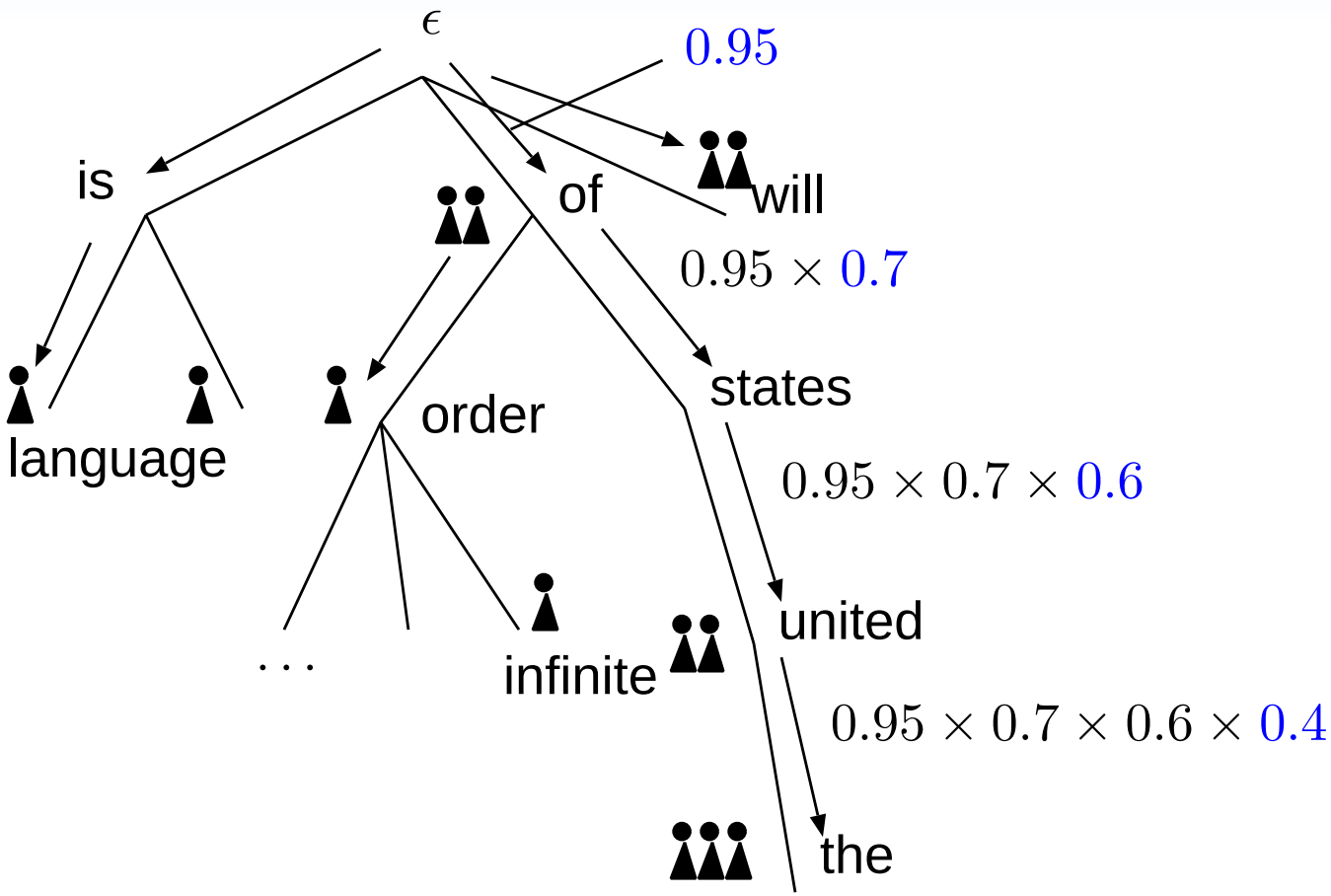
- Add a customer by stochastically decending a suffix tree from its root
- Each node i has a probability to stop at that node ($1-q_i$ equals the "penetration" probability)

$$q_i \sim \mathrm{Be}(\alpha, \beta)\,. \tag{0}$$

- Therefore, a customer will stop at depth n by the probability

$$p(n|h) = q_n \prod_{i=0}^{n-1} (1-q_i)\,. \tag{0}$$

Variable-order Pitman-Yor Language Model (2)



- "penetration" $1-q_i$'s are large $\cdots$ may reach deep nodes
  - Long Markov dependencies
- "penetration" are small $\cdots$ stop early for short Markov dependencies

Inference of VPYLM

- For the training data $\mathbf{w} = w_1 w_2 \cdots w_T$, latent Markov orders $\mathbf{n} = n_1 n_2 \cdots n_T$ exist:

$$p(\mathbf{w}) = \sum_{\mathbf{n}} \sum_{\mathbf{s}} p(\mathbf{w}, \mathbf{n}, \mathbf{s}) \tag{0}$$

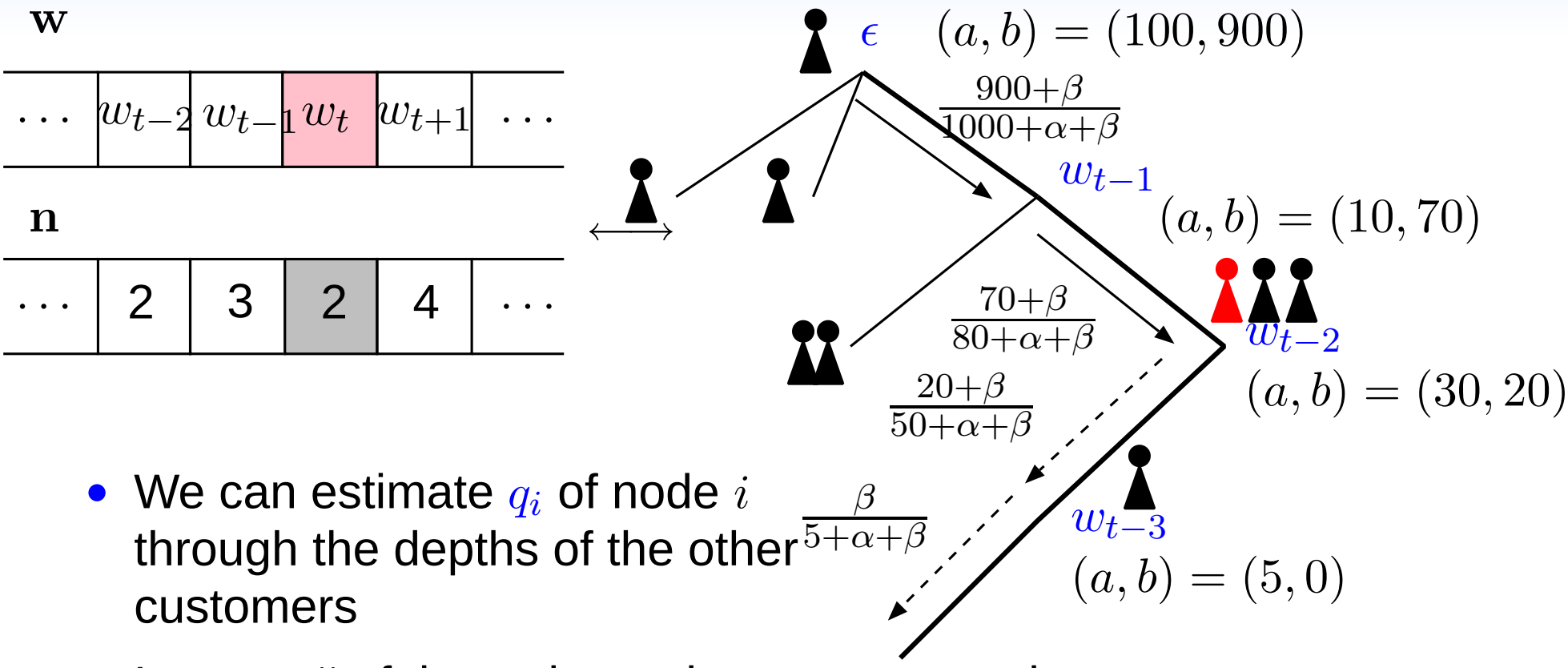
  - $\mathbf{s} = s_1 s_2 \cdots s_T$: seatings of proxy customers in parent nodes

- Gibbs sample $\mathbf{n}$ for inference:

$$p(n_t | \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \propto \underbrace{p(w_t | n_t, \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{n_t\text{-gram prediction}} \cdot \underbrace{p(n_t | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{\text{prob. to reach depth } n_t} \tag{0}$$

  - Trade-off between two terms (penalty for deep n_t)
  - How to compute the second term $p(n_t | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t})$?

Inference of VPYLM (2)



- We can estimate q_i of node i through the depths of the other customers
- Let a_i = # of times the node i was stopped at, b_i = # of times the node i was passed by:

$$p(n_t = n|\mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) = q_n \prod_{i=0}^{n-1}(1-q_i) \tag{0}$$

$$= \frac{a_n+\alpha}{a_n+b_n+\alpha+\beta} \prod_{i=0}^{n-1} \frac{b_i+\beta}{a_i+b_i+\alpha+\beta}\,. \tag{0}$$

Prediction

- Since we don't know the Markov order n beforehand, we integrate it out:

$$p(s|h) = \sum_n p(s,n|h) \tag{0}$$
$$= \sum_n p(s|h,n)p(n|h)\,. \tag{0}$$

- We can rewrite the above expression recursively:

$$p(s|h,n^+) = q_n \cdot \underbrace{p(s|h,n)}_{\text{Prediction at Depth } n} + (1-q_n) \cdot \underbrace{p(s|h,(n+1)^+)}_{\text{Prediction at Depths } n+}, \tag{0}$$

$$p(s|h) \equiv p(s|h,0^+)\,. \tag{0}$$

  - Stick-breaking process on an infinite tree, where breaking proportions will differ from branch to branch.

Perplexity results with # of nodes in the model

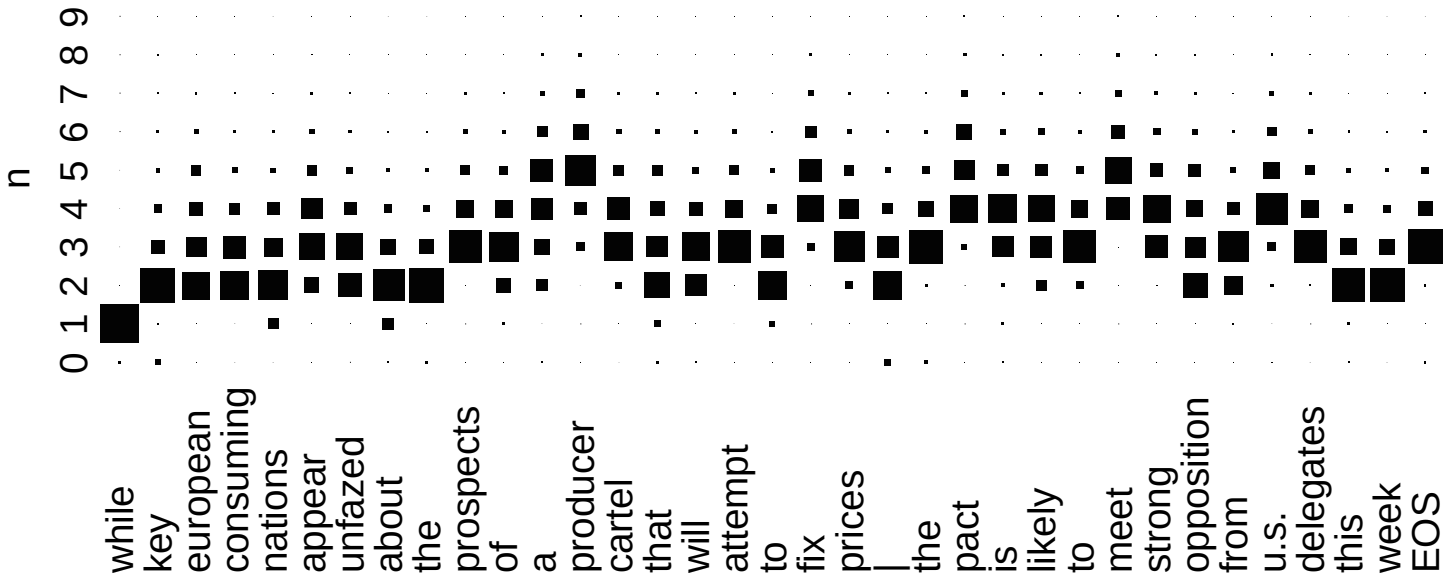
() NAB WSJ corpus (English)

n	SRILM	HPYLM	VPYLM	Nodes(H)	Nodes(V)
3	118.91	113.60	113.74	1,417K	1,344K
5	107.99	101.08	101.69	12,699K	7,466K
7	107.24	N/A	100.68	N/A	10,182K
8	107.21	N/A	100.58	N/A	10,434K
∞	—	—	117.65	—	10,392K

() Japanese Newspaper corpus

n	SRILM	HPYLM	VPYLM	Nodes(H)	Nodes(V)
3	84.74	78.06	78.22	1,341K	1,243K
5	77.88	68.36	69.35	12,140K	6,705K
7	77.51	N/A	68.63	N/A	9,134K
8	77.50	N/A	68.60	N/A	9,490K
∞	—	—	141.81	—	5,396K

Estimated Markov orders



- Estimated Markov orders from which each word has been generated.
- Hinton diagram of $p(n_t|\mathbf{w})$ used in Gibbs sampling for the training data.

“Stochastic phrase” from VMM (1/2)

- $p(s, n|h) = p(s|h, n)p(n|h)$
  - $\cdots$ Probability to generate s using the last n symbols of h as the context
    - For example, generate “Gaussians” from “mixture of”

      $\downarrow$

      “mixture of Gaussians”: *a phrase*
- $p(s, n|h)$ = cohesion strength of the stochastic phrase
  - Will not necessarily decay with length (like an empirical probability)
  - Enumerated by traversing the suffix tree in depth-first order

“Stochastic phrase” from VMM (2/2)

p	Stochastic phrase in the suffix tree
0.9784	primary new issues
0.9726	ˆ at the same time
0.9556	american telephone &
0.9512	is a unit of
0.9394	to # % from # %
0.8896	in a number of
0.8831	in new york stock exchange composite trading
0.8696	a merrill lynch & co.
0.7566	mechanism of the european monetary
0.7134	increase as a result of
0.6617	tiffany & co.
:	:

- “ˆ” = beginning-of-sentence, “#” = numbers

Random Walk generation from the language model

it was a singular man , fierce and quick-tempered , very foul-mouthed when he was angry , and of her muff and began to sob in a high treble key .
“ it seems to have made you , ” said he . ’what have i to his invariable success that the very possibility of something happening on the very morning of the wedding . ”
...

- Random walk generation from the 5-gram VPYLM trained on *“The Adventures of Sherlock Holmes.”*
  - We begin with an infinite number of “beginning-of-sentence” special symbols as the context.
- If we use vanilla 5-grams, overfitting will lead to a mere reproduction of the training data.

Infinite Character Markov Model

‘how queershaped little children drawling-desks, which would get through that dormouse!’ said alice; ‘let us all for anything the secondly, but it to have and another question, but i shalled out, ‘you are old,’ said the you’re trying to far out to sea.

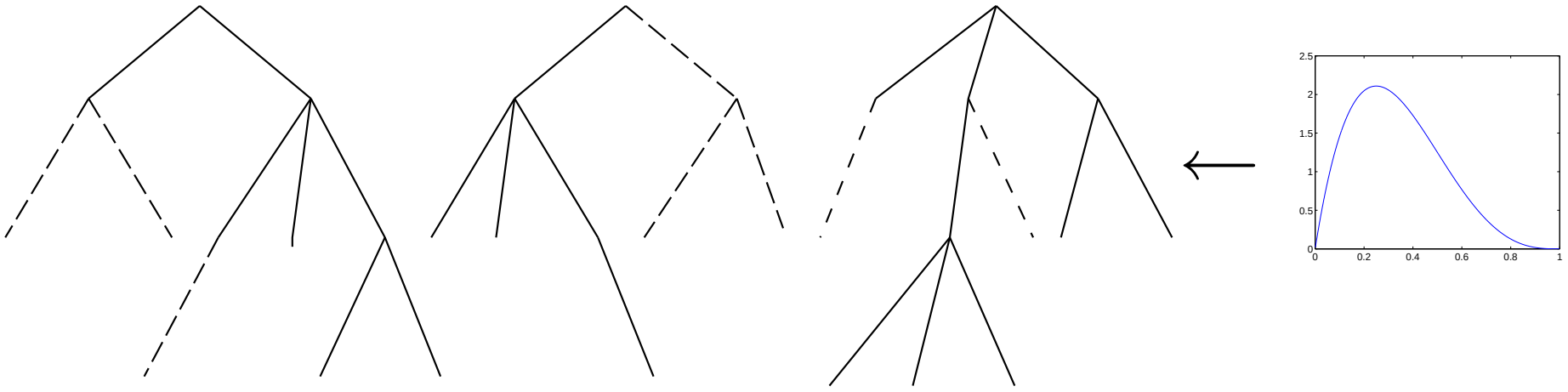
() Random walk generation from a character model.

Character	s a i d ␣ a l i c e ; ␣ ‘ l e t ␣ u s ␣ a l l ␣ f o r ␣ a n y t h i n g ␣ ···
Markov order	5 6 5 4 7 10 6 5 4 3 7 1 4 8 2 4 6 5 5 4 4 5 5 6 4 5 6 7 7 5 3 3 4 5 9 ···

() Markov orders used to generate each character above.

- Character-based Markov model trained on “Alice in Wonderland”.
  - Lowercased alphabets + space
  - OCR, compression, Morphology, . . .

Summary



- We defined a simple prior for *stochastic infinite trees*.
- We expect to use it for latent trees:
  - Variable resolution hierarchical clustering (cf. hLDA)
  - Deep semantic categories just when needed.
- Also for variable order HMM (pruning approach: Yi Wang et al. 2007, TPAMI)

Discussion

- Hierarchical stochastic process to generate q_i's.
  - Not a single Beta distribution
  - What is a good model for stochastic infinite trees?
- Mixture model on Trees
  - Mixing distributions are different from node to node
  - 1-gram word appearances are highly context dependent, but higher-order relationships aren't so much
- How to incorpolate different temporal scales?
  - Characters$\leftrightarrow$words, DNA$\leftrightarrow$RNA or larger units, ...
- Other nonparametric prior on discrete distributions
  - Better than Poisson-Dirichlet (Pitman-Yor).