



## *Present and Future of Text Modeling*

Daichi Mochihashi  
NTT Communication Science Laboratories  
*daichi@cslab.kecl.ntt.co.jp*

T-FaNT2,  
University of Tokyo

2008-2-13 (Wed)



# ”Text Modeling”

## ”Text Modeling”

”Bag-of-words”  
modeling

The Multinomial  
Simplex

Dirichlet Distribution

DCM (Polya)

distribution

DCM (Polya)

distribution (2)

Mixture of DCM

distributions

Performance of

Dirichlet Mixtures

Latent Dirichlet

Allocation (LDA)

Latent Dirichlet

Allocation (LDA) (2)

Limitation of LDA

LDA to cover the

whole simplex

LDA to cover the

whole simplex (2)

Exponential Family

DCM (Elkan 2006)

Dirichlet-Dirichlet

Allocation (DDA)

Work in Progress

Latent topics for

n-grams?

Chinese Restaurant

Process of HPYLM

Latent n-gram

...

## ■ What’s Text Modeling

- ◆ (Usually) bag-of-words style collections of words



- ◆ Idiosyncrasy modeling of **context**

## ■ “Context” =

- ◆ Document
- ◆ Sentence
- ◆ Utterances so far, ...

## ■ Word occurrences are *not homogeneous*.



# ”Bag-of-words” modeling

”Text Modeling”

”Bag-of-words”  
modeling

The Multinomial  
Simplex

Dirichlet Distribution  
DCM (Polya)  
distribution

DCM (Polya)  
distribution (2)

Mixture of DCM  
distributions

Performance of  
Dirichlet Mixtures

Latent Dirichlet  
Allocation (LDA)

Latent Dirichlet  
Allocation (LDA) (2)

Limitation of LDA  
LDA to cover the  
whole simplex

LDA to cover the  
whole simplex (2)

Exponential Family  
DCM (Elkan 2006)

Dirichlet-Dirichlet  
Allocation (DDA)

Work in Progress

Latent topics for  
n-grams?

Chinese Restaurant  
Process of HPYLM

Latent n-gram

$d_1 \rightarrow$  sea:2, habitat:1, ...

$d_2 \rightarrow$  economy:5, relation:1, international:2, ...

- Occurrences of words (features)
- Words are exchangeable (order doesn't matter)
- Counts are explicitly discrete ( $\Leftrightarrow$  Log-linear models)



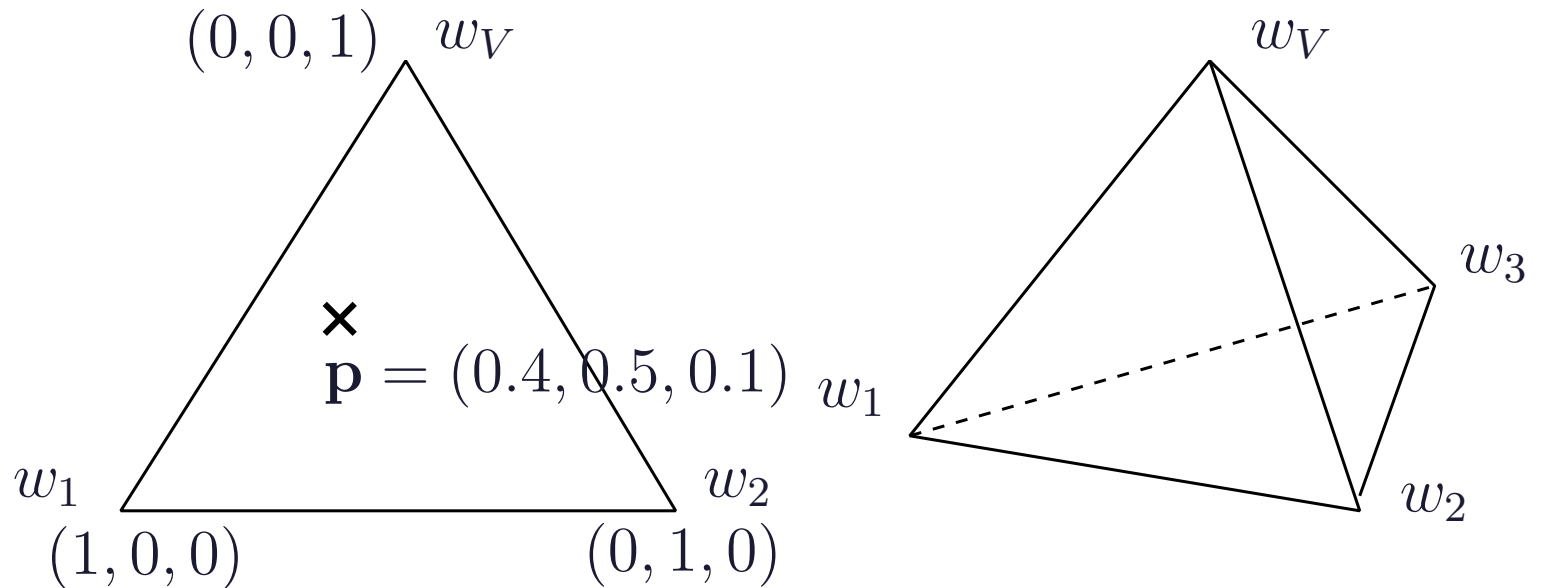
# The Multinomial Simplex

"Text Modeling"  
"Bag-of-words"  
modeling

## The Multinomial Simplex

Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions  
Performance of  
Dirichlet Mixtures  
Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
Dirichlet-Dirichlet  
Allocation (DDA)  
Work in Progress  
Latent topics for  
n-grams?

Chinese Restaurant  
Process of HPYLM  
Latent n-gram



- Each point  $\mathbf{p} \in \Delta(V-1)$  is a Multinomial parameter

$$\mathbf{p} = (p(w_1), p(w_2), \dots, p(w_V)) \quad (1)$$

- Words are generated i.i.d. from  $\mathbf{p}$ :

$$w_i \sim \mathbf{p}. \quad (i = 1 \dots N) \quad (2)$$



# Dirichlet Distribution

"Text Modeling"

"Bag-of-words"  
modeling

The Multinomial  
Simplex

**Dirichlet Distribution**

DCM (Polya)

distribution

DCM (Polya)

distribution (2)

Mixture of DCM

distributions

Performance of

Dirichlet Mixtures

Latent Dirichlet

Allocation (LDA)

Latent Dirichlet

Allocation (LDA) (2)

Limitation of LDA

LDA to cover the

whole simplex

LDA to cover the

whole simplex (2)

Exponential Family

DCM (Elkan 2006)

Dirichlet-Dirichlet

Allocation (DDA)

Work in Progress

Latent topics for

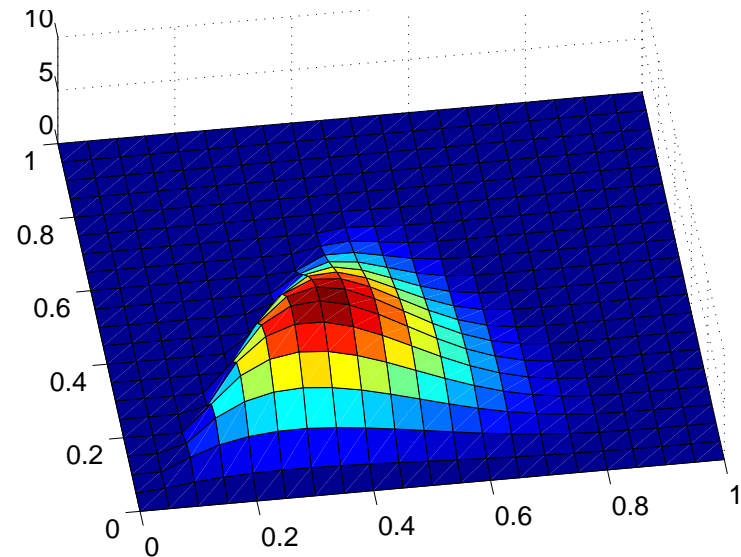
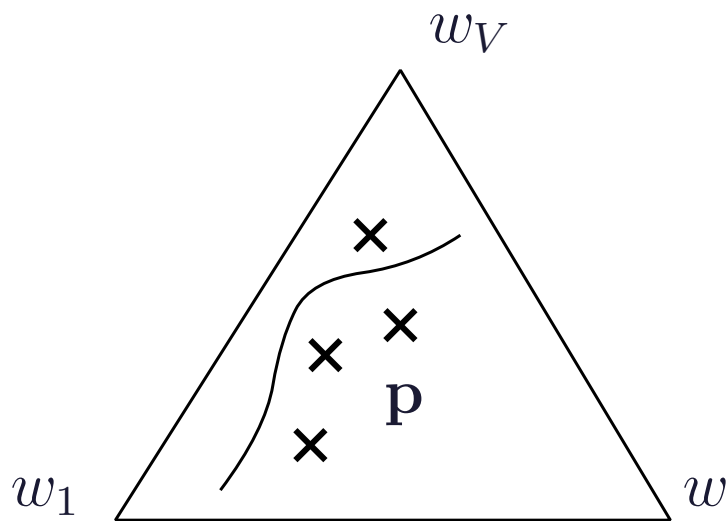
n-grams?

Chinese Restaurant

Process of HPYLM

Latent n-gram

...



- It is convenient (simplest) to put a Dirichlet prior on the distribution of  $\mathbf{p}$ 's:

$$p(\mathbf{p}|\boldsymbol{\alpha}) = \text{Dir}(\mathbf{p}|\boldsymbol{\alpha}) \tag{3}$$

$$\propto \prod_{i=1}^V p_i^{\alpha_i-1} \quad (\text{Normalization constant: } \frac{\prod_{i=1}^V \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^V \alpha_i)}) \tag{4}$$



# DCM (Polya) distribution

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
**DCM (Polya)  
 distribution**  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram

■ Generate  $\mathbf{w} = w_1 w_2 \cdots w_N$  in two steps:

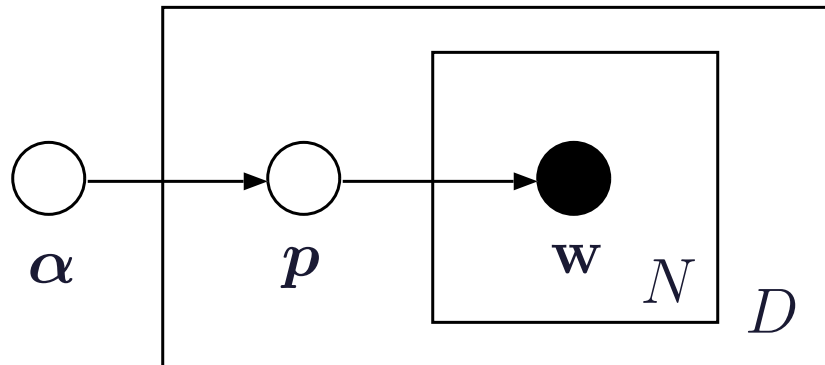
1. Draw  $\mathbf{p} \sim \text{Dir}(\mathbf{p}|\boldsymbol{\alpha})$ .
2. For  $i = 1 \cdots N$ ,
  - ◆ Draw  $w_i \sim \mathbf{p}$ .

■ Integrate out  $\mathbf{p}$  to get the DCM (Dirichlet Compound Multinomial) distribution:

$$p(\mathbf{w}) = \int_{\Delta} p(\mathbf{w}|\mathbf{p})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{p} \quad (5)$$

$$= \frac{\Gamma(\boldsymbol{\alpha})}{\Gamma(\boldsymbol{\alpha}+V)} \prod_{v \in \mathbf{w}} \frac{\Gamma(\alpha_v + n_v)}{\Gamma(\alpha_v)} \quad \left(\boldsymbol{\alpha} = \sum_{i=1}^V \alpha_i\right) \quad (6)$$

# DCM (Polya) distribution (2)



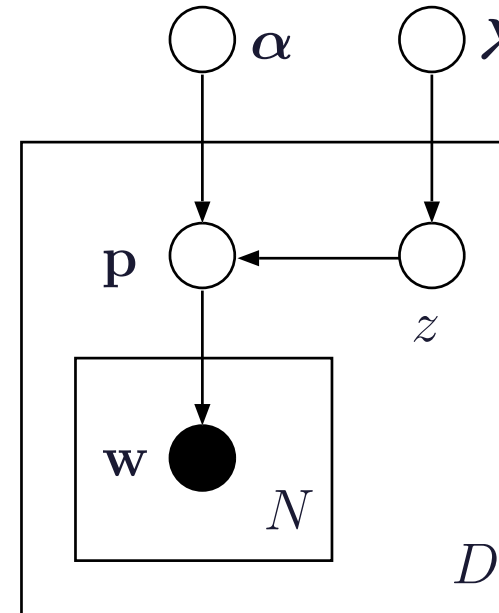
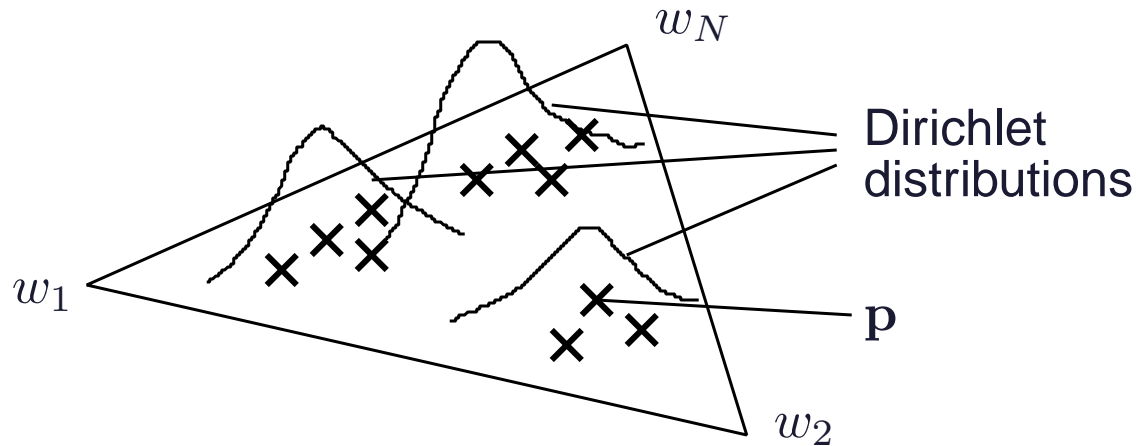
## ■ Characteristics of DCM distribution:

1. Feature occurrences are correlated (through  $p$ )
2. “Damping” of counts
  - (a) Counts  $n$  is effectively damped to  $\log n$
  - (b) “Chance of Two Noriegas is closer to  $p/2$  than  $p^2$ ”  
(Church 2000)



# Mixture of DCM distributions

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
**Mixture of DCM  
 distributions**  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram



- Dirichlet Mixtures (Sjölander+ 1996, Yamamoto+ 2003)
  - ◆ Parameter estimation: EM-Newton
  - ◆ Tool: <http://chasen.org/~daiti-m/dist/dm/>
- Unsupervised, complete Bayesian version of Naive Bayes
  - ◆ Perform better than NB



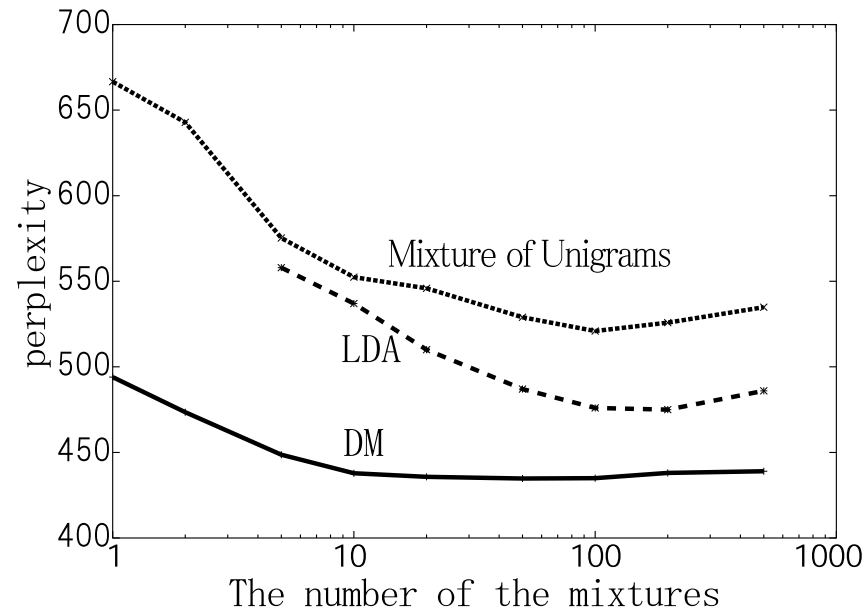
# Performance of Dirichlet Mixtures

"Text Modeling"  
"Bag-of-words"  
modeling  
The Multinomial  
Simplex  
Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions

## Performance of Dirichlet Mixtures

Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
Dirichlet-Dirichlet  
Allocation (DDA)  
Work in Progress  
Latent topics for  
n-grams?

Chinese Restaurant  
Process of HPYLM  
Latent n-gram



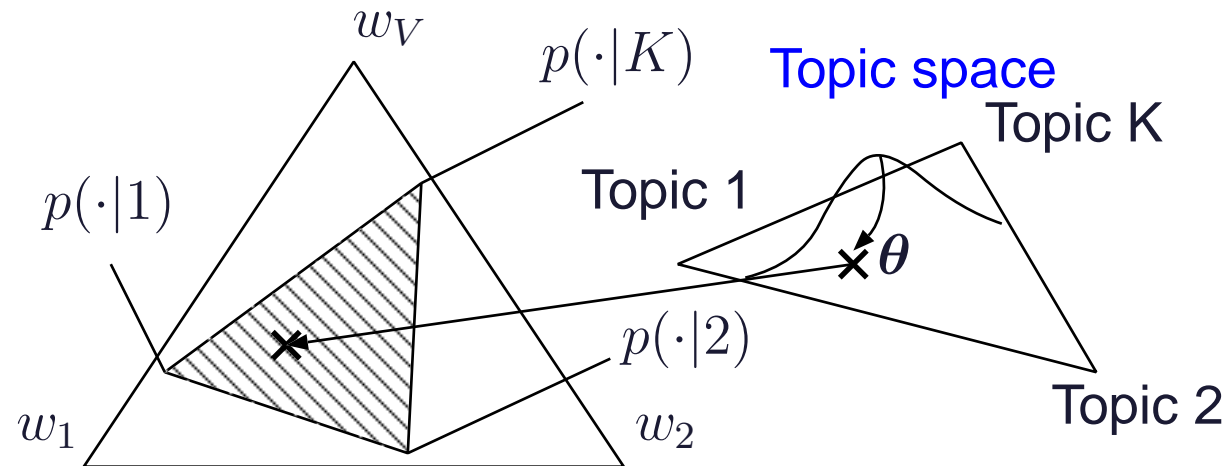
(Yamamoto+ 2005)

- Better than LDA (!)
- "Cache" property (damping multiple counts) is important in natural language.
- LDA?



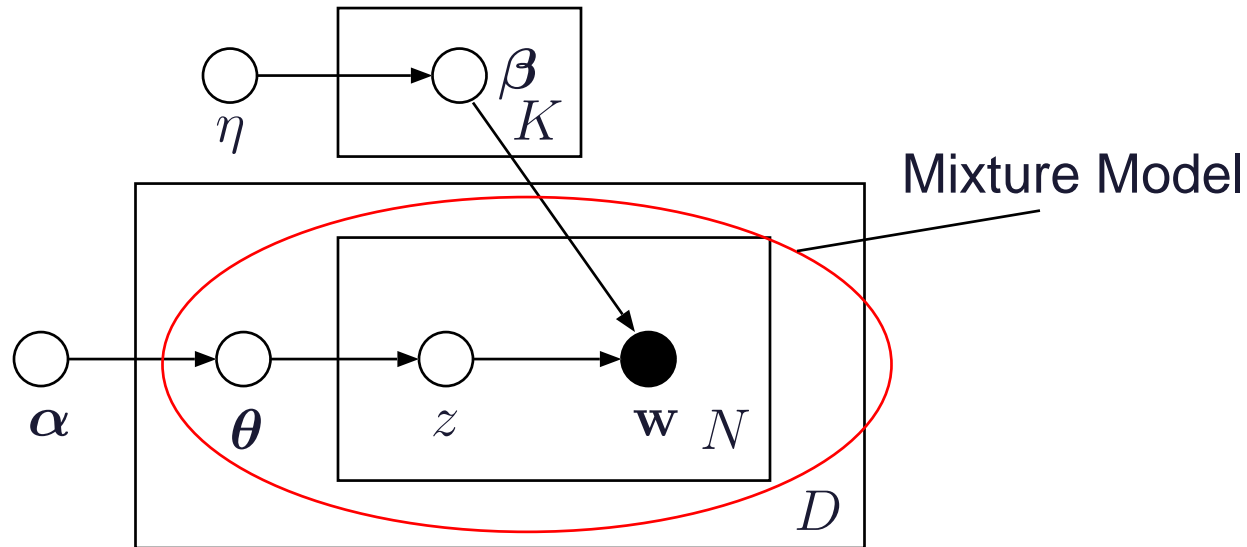
# Latent Dirichlet Allocation (LDA)

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
**Latent Dirichlet  
 Allocation (LDA)**  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram



- Each word will have different topic  $k$
- 3-stage generative process:
  1. Draw  $\theta \sim \text{Dir}(\alpha)$ . (topic mixture)
  2. For  $n = 1 \dots N$ ,
    - (a) Draw  $k \sim \theta$ .
    - (b) Draw  $w_n \sim p(w|k)$ .

# Latent Dirichlet Allocation (LDA) (2)



■ LDA is actually *a collection of mixture models*

◆ Mixture components are shared (HDP exactly does this)

■ Many applications

◆ Contextual SMT (Zhao and Xing 2007)

◆ Semi-supervised POS Tagging (Toutanova and Johnson 2007)

◆ Information retrieval, Computer vision, ...

"Text Modeling"

"Bag-of-words"  
modeling

The Multinomial  
Simplex

Dirichlet Distribution

DCM (Polya)

distribution

DCM (Polya)

distribution (2)

Mixture of DCM

distributions

Performance of

Dirichlet Mixtures

Latent Dirichlet

Allocation (LDA)

Latent Dirichlet  
Allocation (LDA) (2)

Limitation of LDA

LDA to cover the  
whole simplex

LDA to cover the  
whole simplex (2)

Exponential Family

DCM (Elkan 2006)

Dirichlet-Dirichlet

Allocation (DDA)

Work in Progress

Latent topics for

n-grams?

Chinese Restaurant

Process of HPYLM

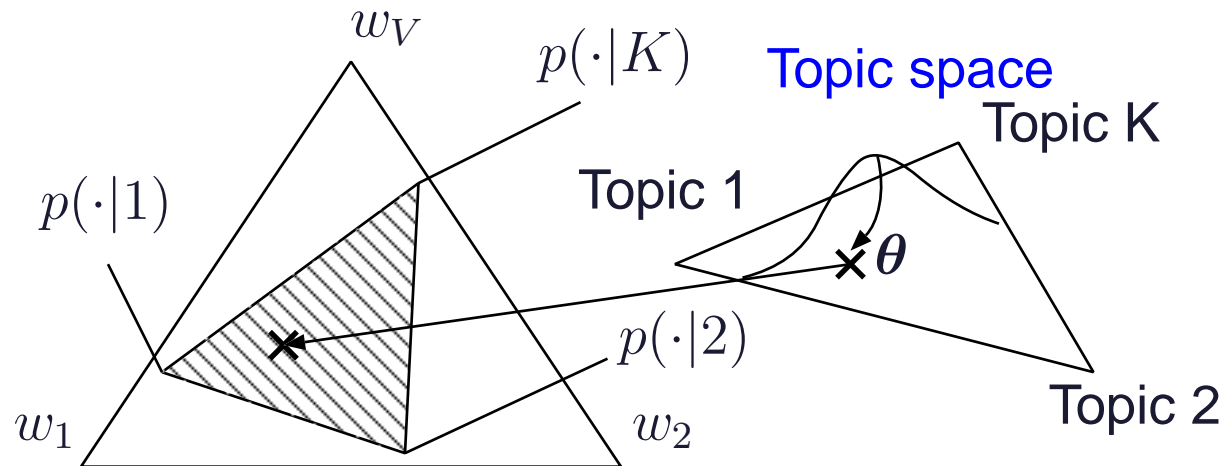
Latent n-gram

...



# Limitation of LDA

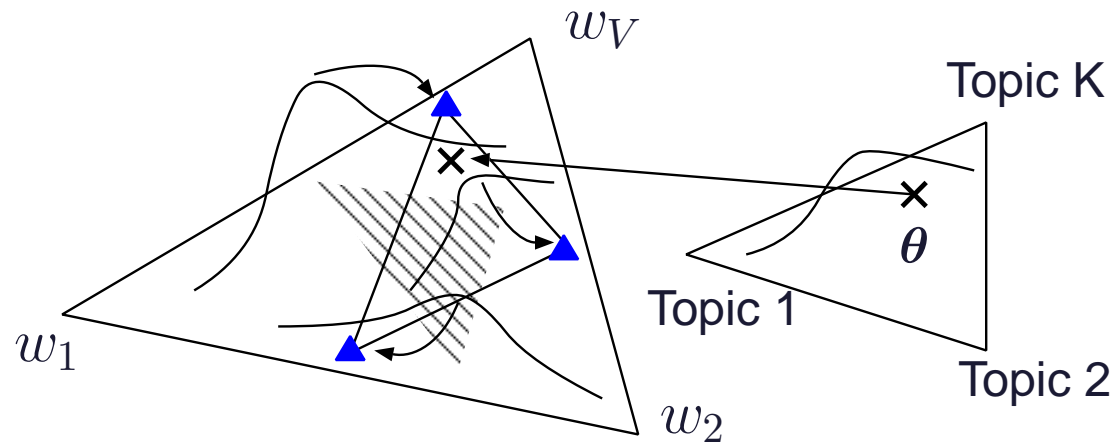
"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
**Limitation of LDA**  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram



- LDA models *only within the Topic subsimplex*
  - ◆ Each document is generated from a mixture of fixed coordinates
- Whole simplex modeling like DM? → DDA (this talk)
- Correlation between topics? → hLDA, CTM, PAM

# LDA to cover the whole simplex

- LDA cannot model outside topic subsimplex  
⇒ Plugging DCM into LDA?



1. Draw  $\theta \sim \text{Dir}(\theta | \alpha)$ .
2. For  $k = 1 \dots K$ ,
  - (a) Draw  $p(\cdot | k) \sim \text{Dir}(\beta_k)$ .
3. For  $n = 1 \dots N$ ,
  - (a) Draw  $k \sim \theta$ .
  - (b) Draw  $w_n \sim p(w | k)$ .

"Text Modeling"  
"Bag-of-words"  
modeling  
The Multinomial  
Simplex  
Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions  
Performance of  
Dirichlet Mixtures  
Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
Dirichlet-Dirichlet  
Allocation (DDA)  
Work in Progress  
Latent topics for  
n-grams?  
Chinese Restaurant  
Process of HPYLM  
Latent n-gram



# LDA to cover the whole simplex (2)

- Generally OK, but:

$$p(\mathbf{w}) =$$

$$\int_{\Delta} \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \sum_k \theta_k \frac{\Gamma(\beta_k)}{\Gamma(\beta_k + n_k)} \prod_{v \in \mathbf{w}} \frac{\Gamma(\beta_{kv} + n_{kv})}{\Gamma(\beta_{kv})} d\theta \quad (7)$$

- ◆ Too complex to optimize!
  - DCM: *not in the exponential family.*

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
**LDA to cover the  
 whole simplex (2)**  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram

# Exponential Family DCM (Elkan 2006)

## ■ DCM (Polya) distribution:

$$p(\mathbf{w}) = \frac{n!}{\prod_v n_v!} \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_v \frac{\Gamma(\alpha_v + n_v)}{\Gamma(\alpha_v)} \quad (8)$$

- ◆ For  $\alpha \ll 1$ ,  $\Gamma(\alpha+n)/\Gamma(\alpha) \simeq \alpha\Gamma(n)$ .
- ◆ Plugging this into (8) and using  $\Gamma(n) = (n-1)!$ , we get:

## ■ Exponential family DCM distribution:

$$q(\mathbf{w}) = \frac{n!}{\prod_v n_v!} \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_v \alpha_v (n_v - 1)! \quad (9)$$

$$= n! \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_v \mathbb{I}(n_v > 0) \frac{\alpha_v}{n_v} \quad (10)$$

- ◆ Exponential family and simpler form!

"Text Modeling"  
"Bag-of-words"  
modeling  
The Multinomial  
Simplex  
Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions  
Performance of  
Dirichlet Mixtures  
Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
Dirichlet-Dirichlet  
Allocation (DDA)  
Work in Progress  
Latent topics for  
n-grams?  
Chinese Restaurant  
Process of HPYLM  
Latent n-gram

# Dirichlet-Dirichlet Allocation (DDA)

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k + n_k - 1} \cdot n_k! \frac{\Gamma(\beta_k)}{\Gamma(\beta_k + n_k)} \prod_{v \in \mathbf{w}} \mathbb{I}(n_{kv} > 0)$$

$$\text{where } n_k = \sum_n \mathbb{I}(z_n = k) \quad \begin{matrix} (11) \\ (12) \end{matrix}$$

$$n_{kv} = \sum_n \sum_v \mathbb{I}(z_n = k) \mathbb{I}(w_n = v) \quad (13)$$

- Can model whole word simplex



- “Burst” phenomenon of natural language

Example:

- ◆ Multiple appearances of “Noriega” in a document
- ◆ ‘Toyota’ and ‘Nissan’ in “car” topic distribution

“Text Modeling”  
“Bag-of-words”  
modeling  
The Multinomial  
Simplex  
Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions  
Performance of  
Dirichlet Mixtures  
Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
**Dirichlet-Dirichlet  
Allocation (DDA)**  
Work in Progress  
Latent topics for  
n-grams?  
Chinese Restaurant  
Process of HPYLM  
Latent n-gram



# Work in Progress

"Text Modeling"  
"Bag-of-words"  
modeling  
The Multinomial  
Simplex  
Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions  
Performance of  
Dirichlet Mixtures  
Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
Dirichlet-Dirichlet  
Allocation (DDA)

## Work in Progress

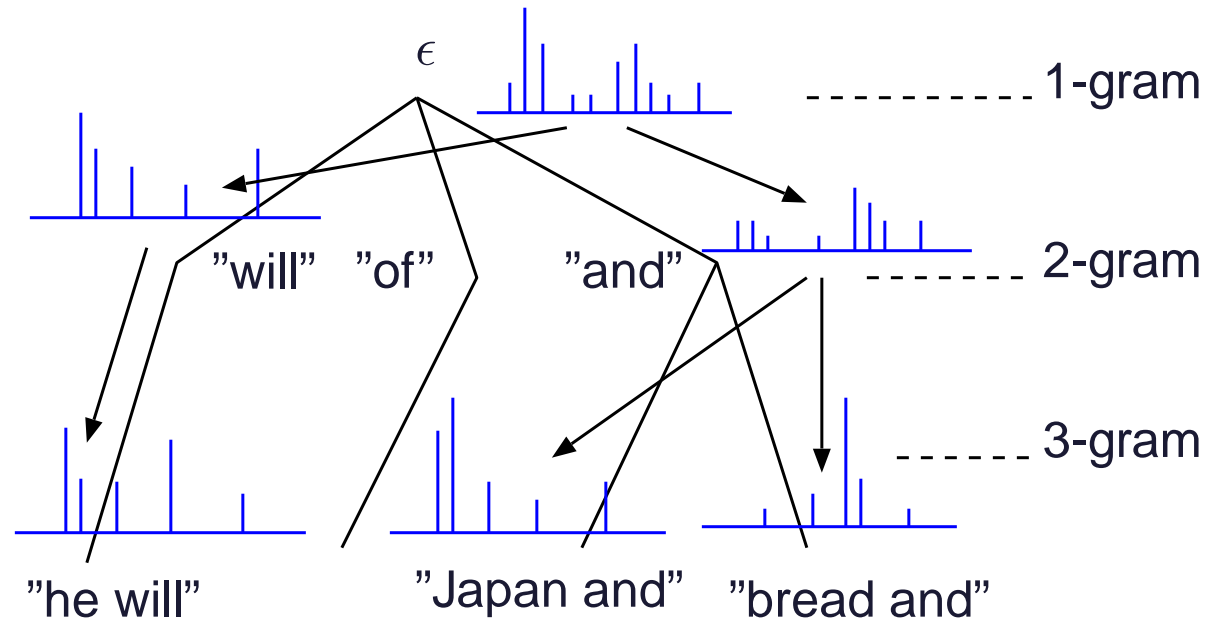
Latent topics for  
n-grams?  
Chinese Restaurant  
Process of HPYLM  
Latent n-gram

- A small piece of work, but can combine the benefits of both LDA and DM
  - ✓ Derived the variational lower bound
  - ◆ Experiments to compare with DM and vanilla LDA
- Lesson: EDCM is an useful exponential family distribution



# Latent topics for n-grams?

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress



- Bayesian n-gram model . . . Hierarchical Pitman-Yor Language Model (ACL 2006)
  - ◆ Hierarchical draw of  $n$ -gram distribution from  $(n - 1)$ -gram distribution
  - ◆ An extension of HDP

Latent topics for  
 n-grams?

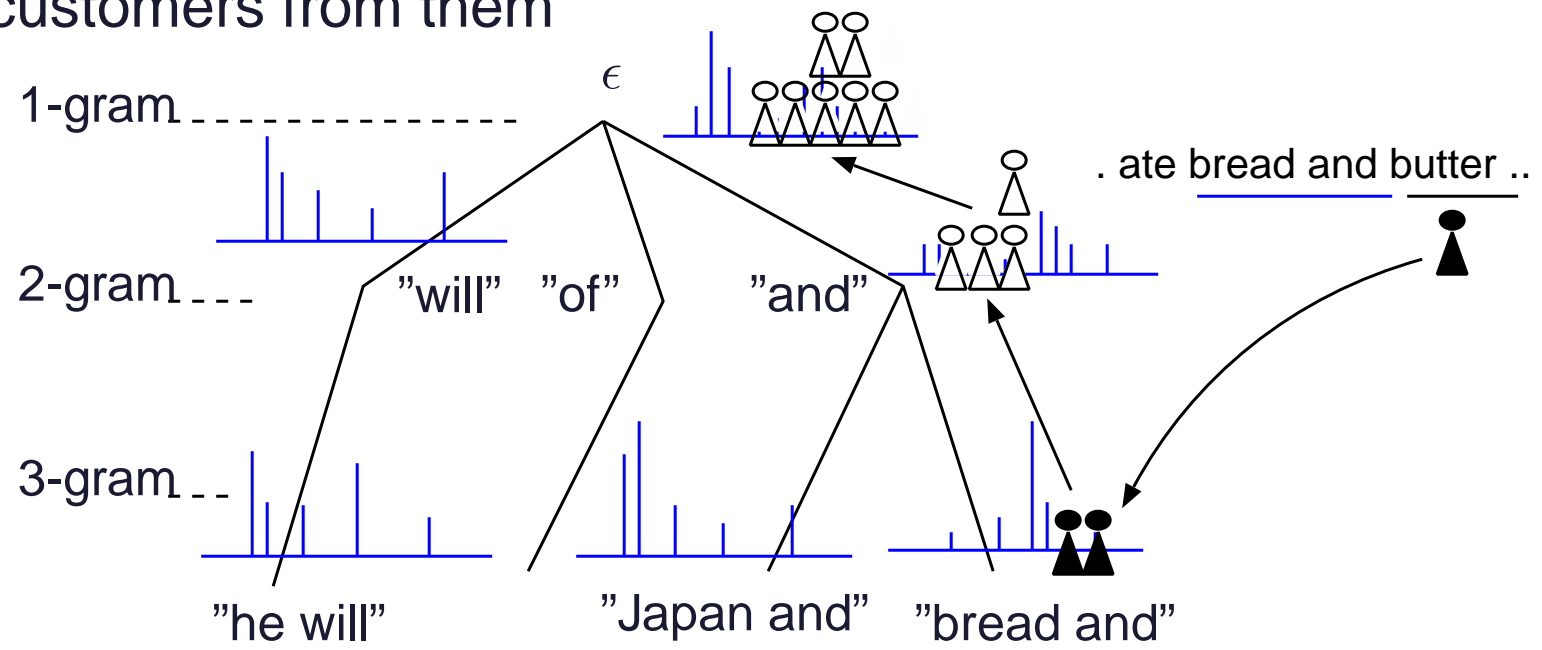
Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram



# Chinese Restaurant Process of HPYLM

"Text Modeling"  
 "Bag-of-words" modeling  
 The Multinomial Simplex  
 Dirichlet Distribution  
 DCM (Polya) distribution  
 DCM (Polya) distribution (2)  
 Mixture of DCM distributions  
 Performance of Dirichlet Mixtures  
 Latent Dirichlet Allocation (LDA)  
 Latent Dirichlet Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the whole simplex  
 LDA to cover the whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet Allocation (DDA)  
 Work in Progress  
 Latent topics for n-grams?

- Estimate n-gram distributions by finite draws = customers from them



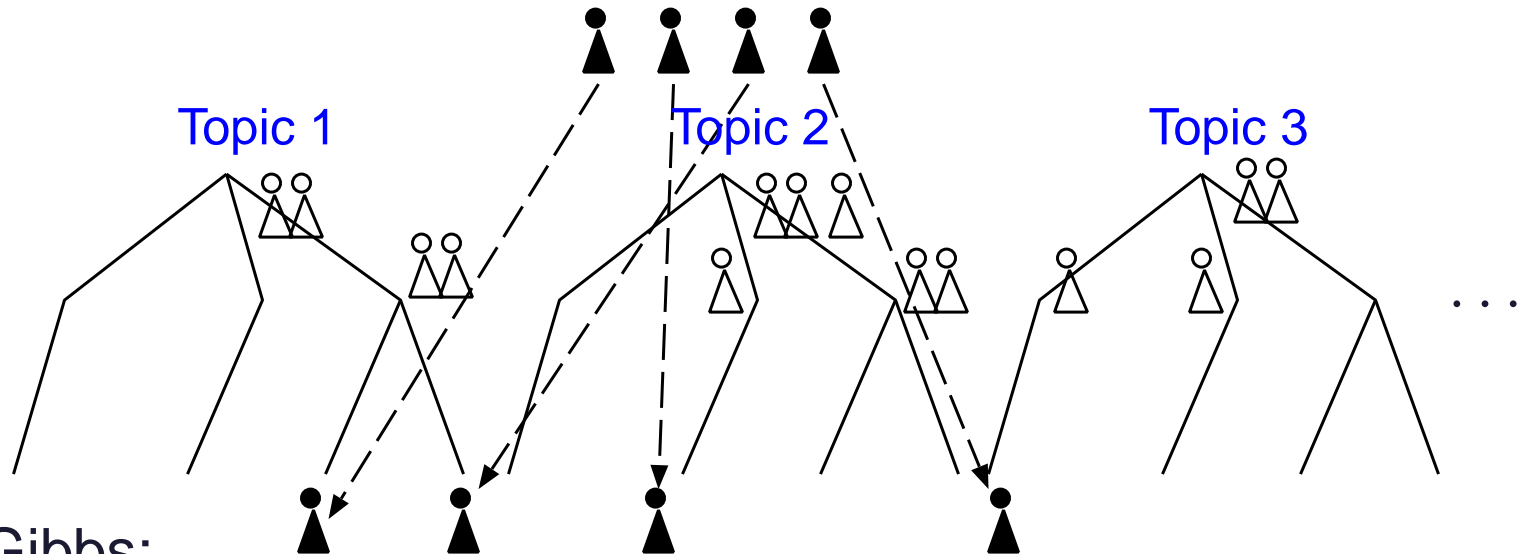
- "Real" customers only reside in the leaves
  - ◆  $< n$ -gram customers are "virtual" and sent by  $n$ -gram customers for smoothing purpose
- Gibbs sampler: remove a customer and add him again to stochastically optimize  $< n$ -gram customers for smoothing



# Latent n-gram Mixtures

- Straightforward approach: LDA of HPYLM (or VPYLM)

Text:  $\dots\dots w_1 w_2 w_3 w_4 \dots\dots$



■ Gibbs:

$$\text{Topic } k \mid w_t, d \propto p(w_t \mid \text{HPYLM}_k) \cdot (n_d^{-t}(k) + \alpha_k) \quad (14)$$

$n_d^{-t}(k)$  : # of customers in document  $d$ , assigned to topic  $k$  (excluding  $w_t$ )

"Text Modeling"  
 "Bag-of-words" modeling  
 The Multinomial Simplex  
 Dirichlet Distribution  
 DCM (Polya) distribution  
 DCM (Polya) distribution (2)  
 Mixture of DCM distributions  
 Performance of Dirichlet Mixtures  
 Latent Dirichlet Allocation (LDA)  
 Latent Dirichlet Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the whole simplex  
 LDA to cover the whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet Allocation (DDA)  
 Work in Progress  
 Latent topics for n-grams?

Chinese Restaurant Process of HPYLM

Latent n-gram



# Latent n-gram Mixtures (2)

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram

- NIPS papers dataset (1500 documents/3,261,224 words) with 5 mixtures

$p(n, s)$	Phrase	$p(n, s)$	Phrase	$p(n, s)$	Phrase
0.9904	in section #	0.9853	et al	0.9823	monte carlo
0.9900	the number of	0.9840	receptive field	0.9524	associative memory
0.9856	in order to	0.9630	excitatory and inhibitory	0.9081	as can be seen
0.9832	in table #	0.9266	in order to	0.8206	parzen windows
0.9752	dealing with	0.8939	primary visual cortex	0.8044	in the previous section
0.9693	with respect to	0.8756	corresponds to	0.7790	american institute of phy

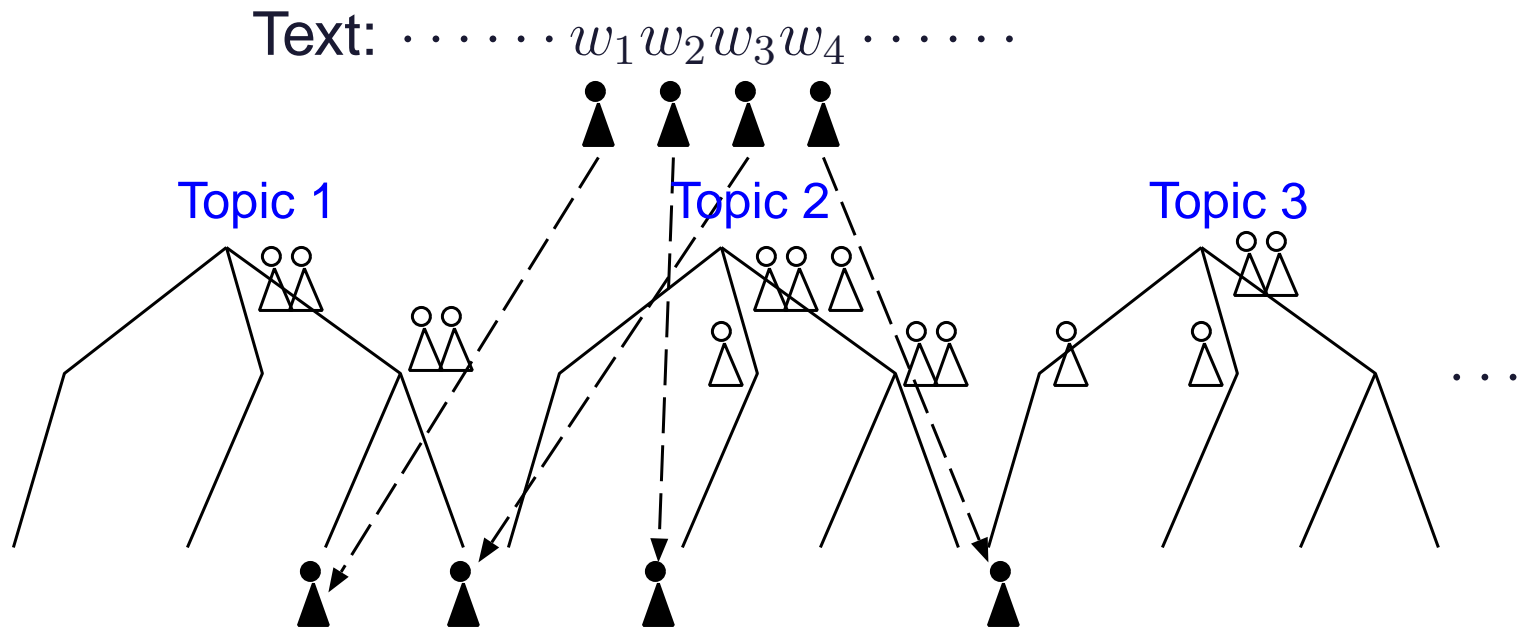
(a) Topic 0 (generic)                      (b) Topic 1                      (c) Topic 4

- Generally OK, but PPL is identical (117.28→116.62).
  - ◆ PPL increases on other dataset (such as AP)
  - ◆ Data sparseness problem!



# Data sparseness in n-gram mixtures

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram



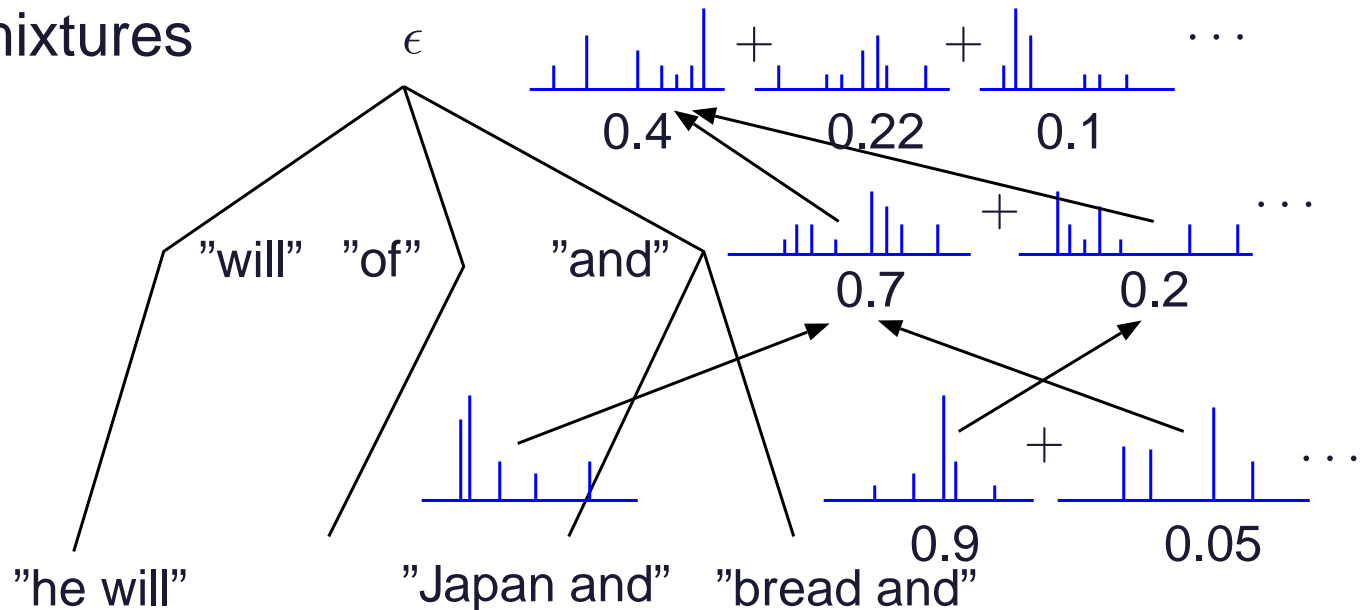
- LDA only mixes different trees
  - ⇓
  - Severe data sparseness problem
    - ◆ Many infrequent n-grams concentrate on specific trees
    - ◆ If topic weights for these trees are near zero, estimates are severely backed-off

Solution?



# Solution

- Possible solution: Nested (Poisson-)Dirichlet process (nDP) (Rodriguez+ 2006)
- Single tree, but measures on branches are infinite mixtures



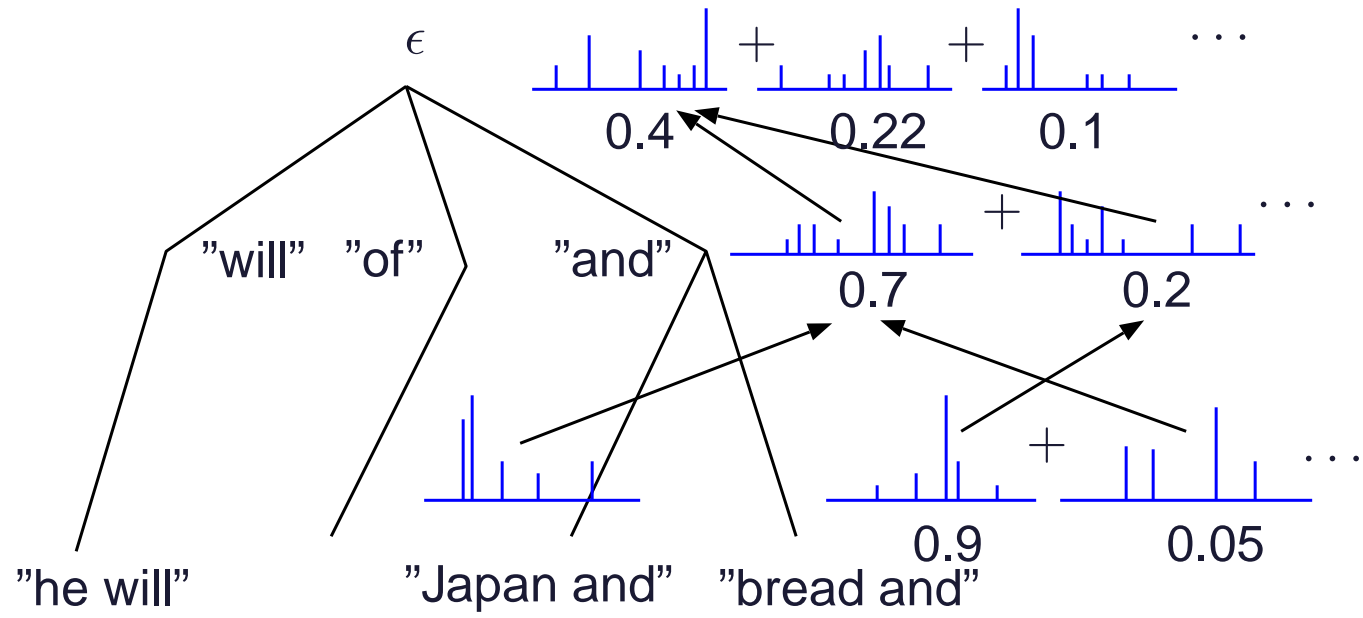
- ◆ Mixtures are governed by DP (thus many counts (eg. unigrams) induce large mixtures)
- ◆ Customers are grouped accordingly

"Text Modeling"  
"Bag-of-words"  
modeling  
The Multinomial  
Simplex  
Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions  
Performance of  
Dirichlet Mixtures  
Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
Dirichlet-Dirichlet  
Allocation (DDA)  
Work in Progress  
Latent topics for  
n-grams?  
Chinese Restaurant  
Process of HPYLM  
Latent n-gram



# Current problem

"Text Modeling"  
 "Bag-of-words"  
 modeling  
 The Multinomial  
 Simplex  
 Dirichlet Distribution  
 DCM (Polya)  
 distribution  
 DCM (Polya)  
 distribution (2)  
 Mixture of DCM  
 distributions  
 Performance of  
 Dirichlet Mixtures  
 Latent Dirichlet  
 Allocation (LDA)  
 Latent Dirichlet  
 Allocation (LDA) (2)  
 Limitation of LDA  
 LDA to cover the  
 whole simplex  
 LDA to cover the  
 whole simplex (2)  
 Exponential Family  
 DCM (Elkan 2006)  
 Dirichlet-Dirichlet  
 Allocation (DDA)  
 Work in Progress  
 Latent topics for  
 n-grams?  
 Chinese Restaurant  
 Process of HPYLM  
 Latent n-gram



- nDP (or nPD) is OK, but how can we introduce “document” (context) here?
- Context-dependent n-gram language models are important in NLP applications.



# Final Remark

- “Text Modeling” is a research for introducing context dependency in many NLP applications.
- LDA is useful, and can incorporate DM through EDCM distribution
- Latent topic-aware n-gram language models are important, and nDP may open the door

Thank you very much.

“Text Modeling”  
“Bag-of-words”  
modeling  
The Multinomial  
Simplex  
Dirichlet Distribution  
DCM (Polya)  
distribution  
DCM (Polya)  
distribution (2)  
Mixture of DCM  
distributions  
Performance of  
Dirichlet Mixtures  
Latent Dirichlet  
Allocation (LDA)  
Latent Dirichlet  
Allocation (LDA) (2)  
Limitation of LDA  
LDA to cover the  
whole simplex  
LDA to cover the  
whole simplex (2)  
Exponential Family  
DCM (Elkan 2006)  
Dirichlet-Dirichlet  
Allocation (DDA)  
Work in Progress  
Latent topics for  
n-grams?  
Chinese Restaurant  
Process of HPYLM  
Latent n-gram