Bayesian Nonparametric Information Processing

Jen-Tzung Chien

Department of Electrical and Computer Engineering

National Chiao Tung University, Taiwan

July 28, 2014

Table of contents

Why Information Processing?


copyright 2012 Paula M De Angelis

Challenges in information processing

- We are in an era of abundant data
- An enormous amount of multimedia data is available in internet which contains speech, text, image, music, video, social networks and any specialized technical data
- As more information becomes available, it becomes more difficult to find and discover what we need
- The collected data are prone to be noisy, non-labeled, non-aligned, mismatched, and ill-posed

Modeling tools

- We need new tools to help us organize, search, and understand these vast amounts of information

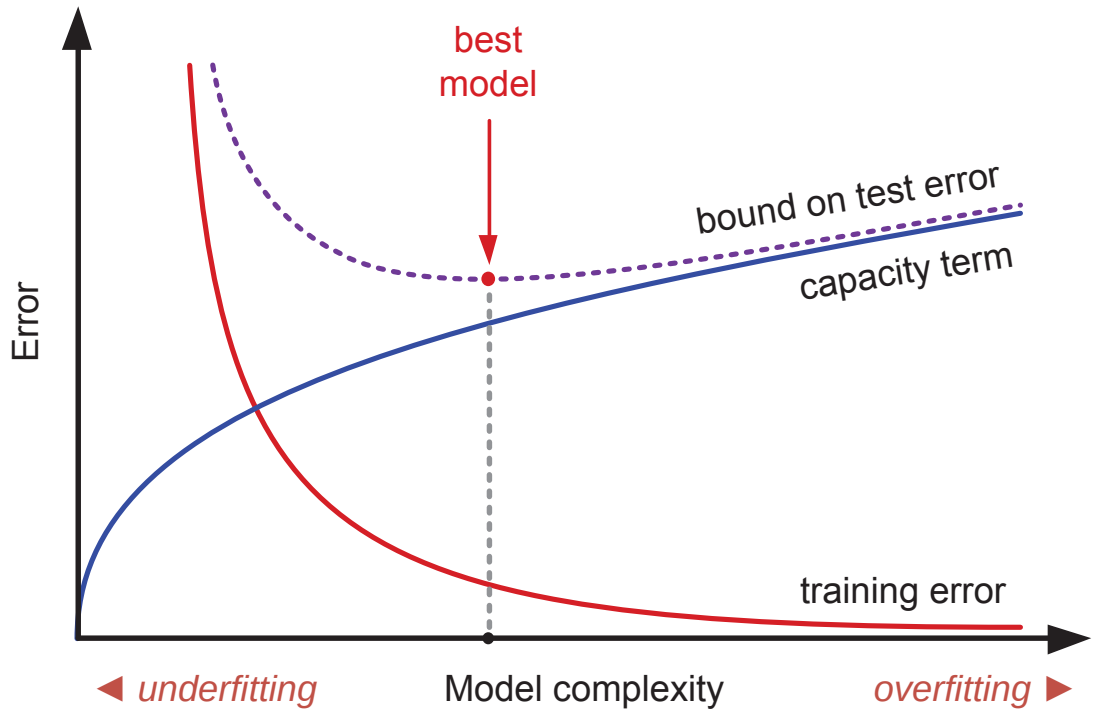
- Our modeling tools should
  - faithfully represent uncertainty in model structure and its parameters
  - reflect noise condition in observed data
  - be automated and adaptive
  - assure robustness
  - scalable for large data sets

- Uncertainty can be properly expressed by prior distribution and prior process

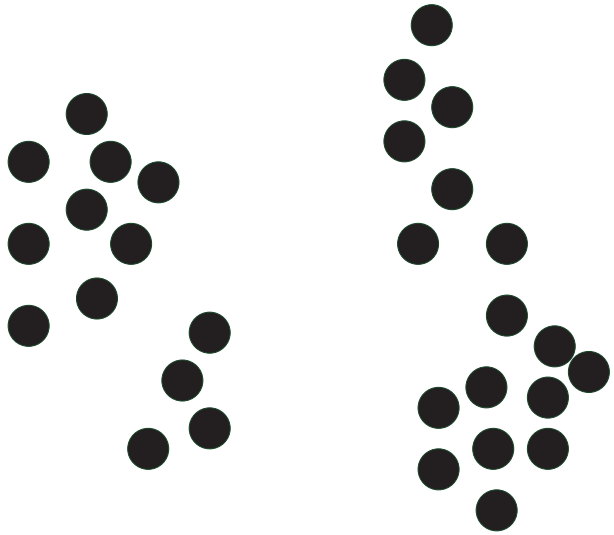
Why Bayesian Nonparametrics?

Model selection

- Model selection or averaging is often necessary to prevent overfitting and underfitting. We control model complexity to improve model generalization



Example: clustering



$$p(\mathcal{D}) = \sum_{k=1}^{K} p(\mathcal{D}|\lambda_k)p(\lambda_k)$$

$$p(\mathcal{M}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M})p(\mathcal{M})$$

$$\mathcal{M} \equiv \{\boldsymbol{\lambda}, K\}$$

Bayesian nonparametrics

- **Definition**: Bayesian model on an ∞-dimensional parameter space (Orbanz and Teh, 2010)

- Bayesian nonparametrics are used to characterize (Hjort et al., 2010)

  1. big parameter space
  2. construction of probability measure over these space

- **Idea**: construct a prior on probability distributions or density functions

- We work with the priors that are general stochastic processes (e.g. the measure on function space or on measure space) and the number of parameters could grow as more data are observed

Table of contents

Dirichlet distribution

- Dirichlet distribution is a distribution over the K-dimensional probability simplex (i.e. random partitions)

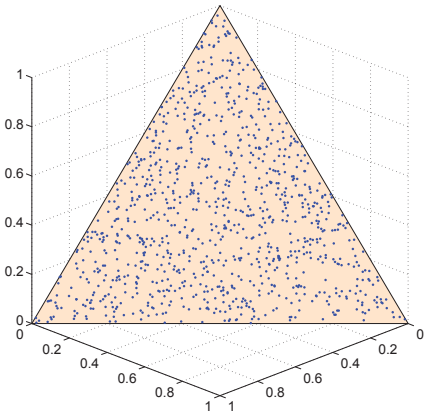
$$\{(\pi_1, \ldots, \pi_K) \mid 0 \leq \pi_k \leq 1,\ \textstyle\sum_{k=1}^{K} \pi_k = 1\}$$

- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is Dirichlet distributed with concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, if

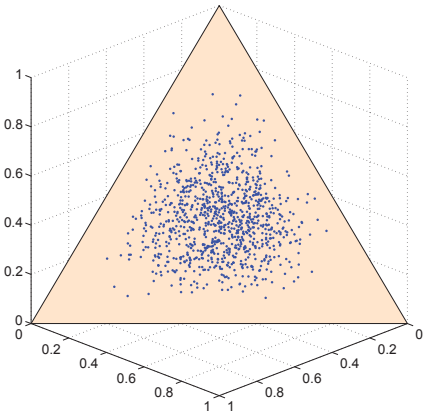
$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$$\text{where} \quad \alpha_k > 0, \ \text{for } k = 1, \ldots, K$$

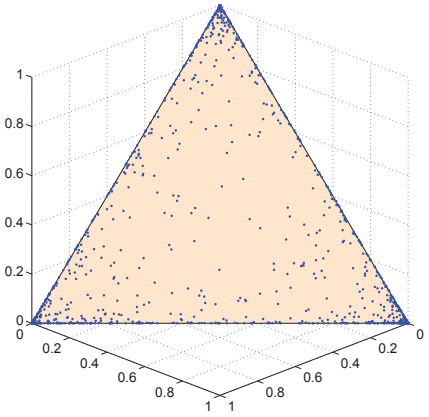
Probability simplex



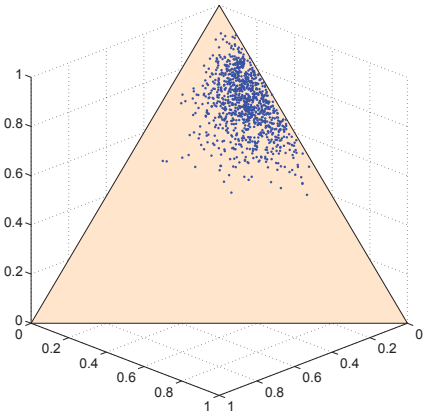
(a) $\boldsymbol{\alpha} = (1, 1, 1)$



(b) $\boldsymbol{\alpha} = (10, 10, 10)$



(c) $\boldsymbol{\alpha} = (0.2, 0.2, 0.2)$



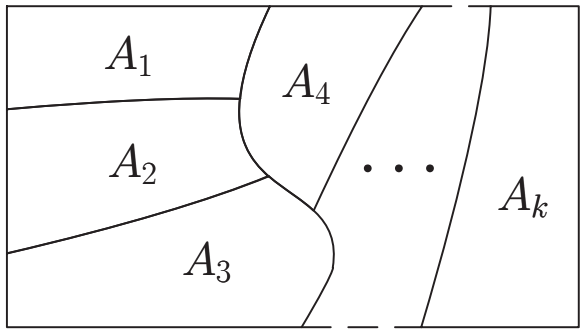
(d) $\boldsymbol{\alpha} = (2, 5, 15)$

Dirichlet process

- Dirichlet process (DP) $G \sim \mathrm{DP}(\alpha_0, G_0)$ is a distribution over probability measures

- Let G_0 be a distribution over probability space Θ, a random measure G is distributed according to a DP with the parameter $\alpha_0 > 0$, if for all natural numbers k and any partitions $A_1, \ldots, A_k \subset \Theta$

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_k))$$



- An ∞-dimensional generalization of the Dirichlet distribution

Properties

- Moment

$$\mathbb{E}[G(A)] = G_0(A) \qquad \mathrm{Var}(G(A)) = \frac{G_0(A)(1 - G_0(A))}{1 + \alpha_0}$$

- Posterior

$$(G(A_1), \ldots, G(A_k)) \mid \theta_{1:n} \sim \mathrm{Dir}(\alpha_0 G_0(A_1) + n_1, \ldots, \alpha_0 G_0(A_k) + n_k)$$

$$G \mid \theta_{1:n} \sim \mathrm{DP}\left(\alpha_0 + n, \sum_{i=1}^{n} \frac{1}{\alpha_0 + n} \delta_{\theta_i} + \frac{\alpha_0}{\alpha_0 + n} G_0\right)$$

where $n_k = |\{i : \theta_i \in A_k\}|$

- Posterior predictive distribution

$$p(\theta_{n+1} \in A \mid \theta_{1:n}) = \mathbb{E}[G(A) \mid \theta_{1:n}] = \sum_{i=1}^{n} \frac{1}{\alpha_0 + n} \delta_{\theta_i}(A) + \frac{\alpha_0}{\alpha_0 + n} G_0(A)$$

Stick-breaking process

- Image a stick with unit length
  - we can break the stick infinitely
  - we obtain the k-dimensional vector $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$ whose components are subject to $\sum_k \pi_k = 1$

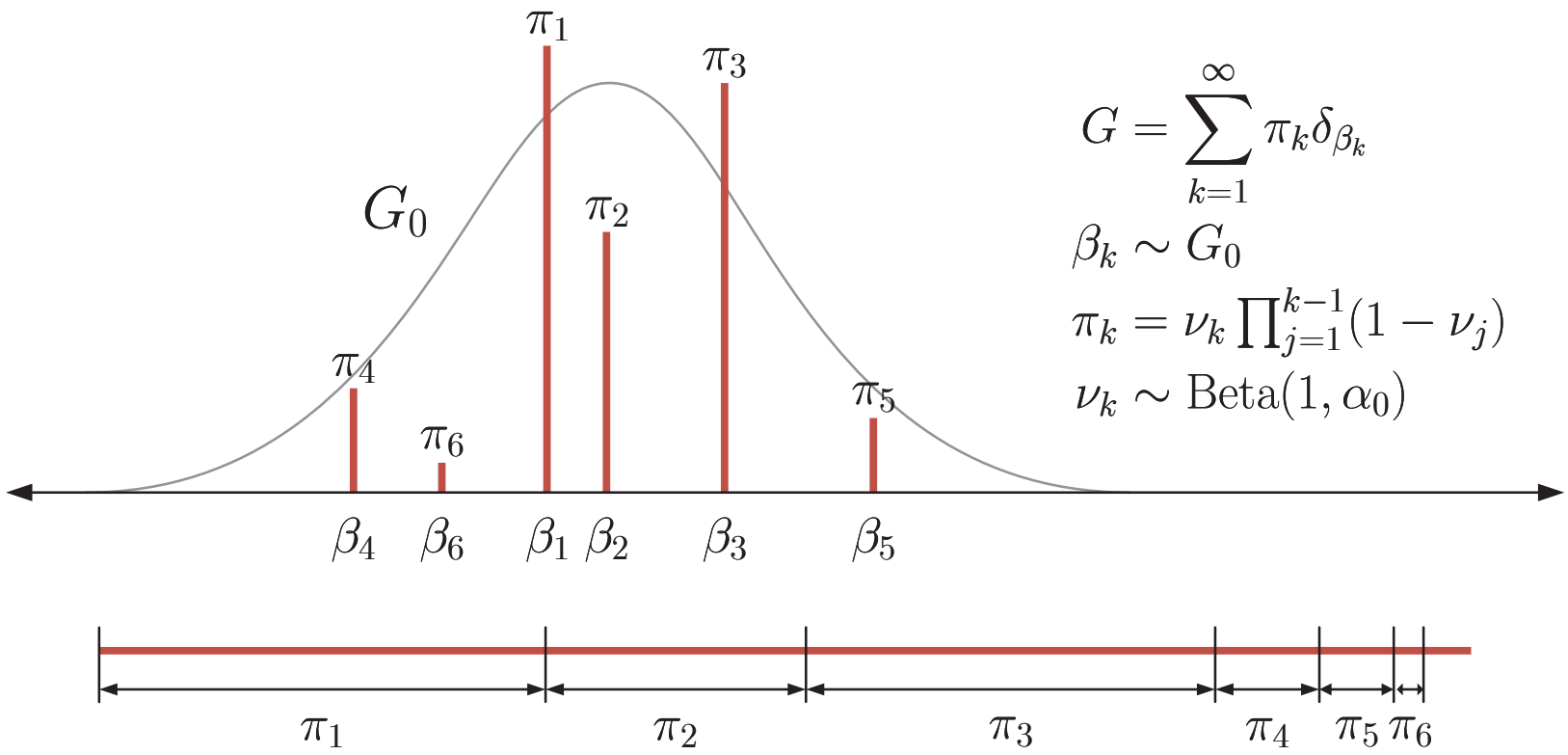
- A mixture model with infinite components is built by

$$\beta_k \sim G_0, \quad \nu_k | \alpha_0 \sim \text{Beta}(1, \alpha_0)$$

$$\pi_k = \nu_k \prod_{j=1}^{k-1} (1 - \nu_j), \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\beta_k}$$

- δ_{β_k} denotes an atom at location β_k

Illustration



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\beta_k}$$

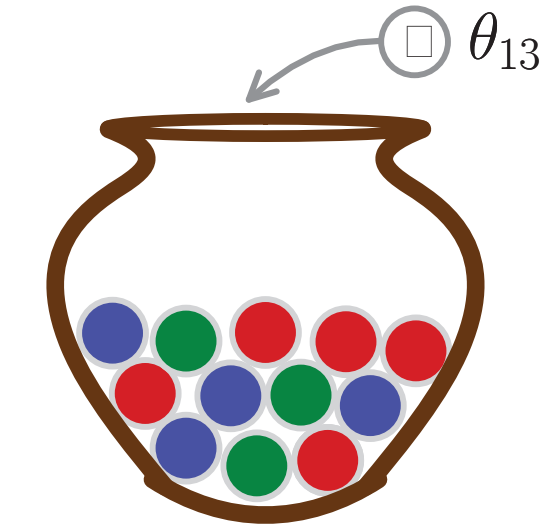
$$\beta_k \sim G_0$$

$$\pi_k = \nu_k \prod_{j=1}^{k-1} (1 - \nu_j)$$

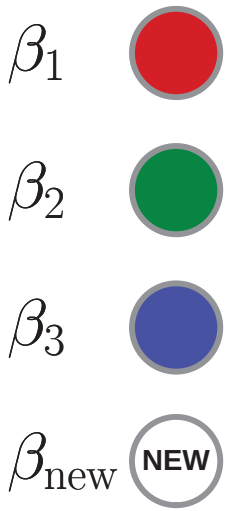
$$\nu_k \sim \text{Beta}(1, \alpha_0)$$

(Sethuraman, 1994)

Polya Urn process



$G_0 \triangleq$ distr over colorspace



$\beta_1 \quad \frac{5}{12+\alpha_0}$

$\beta_2 \quad \frac{3}{12+\alpha_0}$

$\beta_3 \quad \frac{4}{12+\alpha_0}$

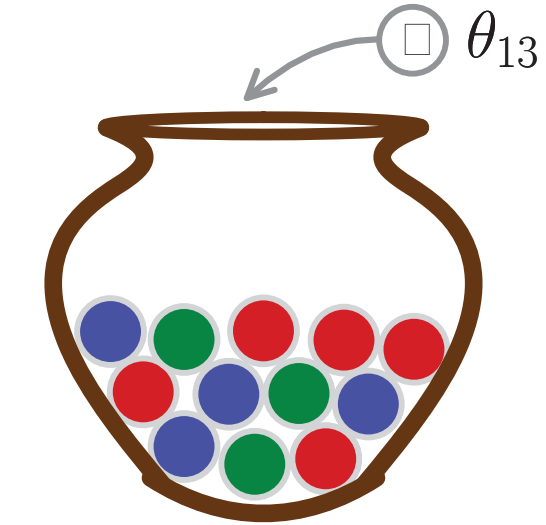
$\beta_{\text{new}} \quad \frac{\alpha_0}{12+\alpha_0}$

$$p(\text{color } k \mid \text{balls in urn}) = \frac{n_k}{i-1+\alpha_0}$$

$$p(\text{new color} \mid \text{balls in urn}) = \frac{\alpha_0}{i-1+\alpha_0}$$

(Blackwell and MacQueen, 1973)

Polya Urn process



$G_0 \triangleq$ distr over colorspace

β_1 $\frac{5}{12+\alpha_0}$

β_2 $\frac{3}{12+\alpha_0}$

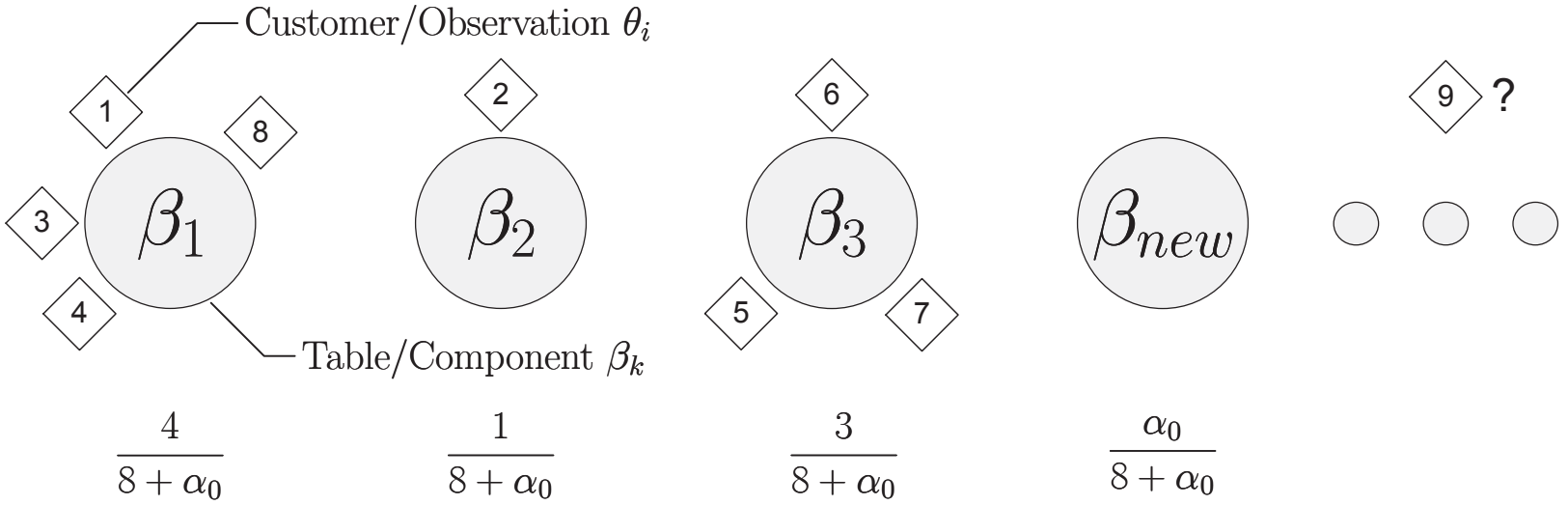
β_3 $\frac{4}{12+\alpha_0}$

β_{new} NEW $\frac{\alpha_0}{12+\alpha_0}$

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{n_k}{i-1+\alpha_0} \delta_{\beta_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

(Blackwell and MacQueen, 1973)

Chinese restaurant process

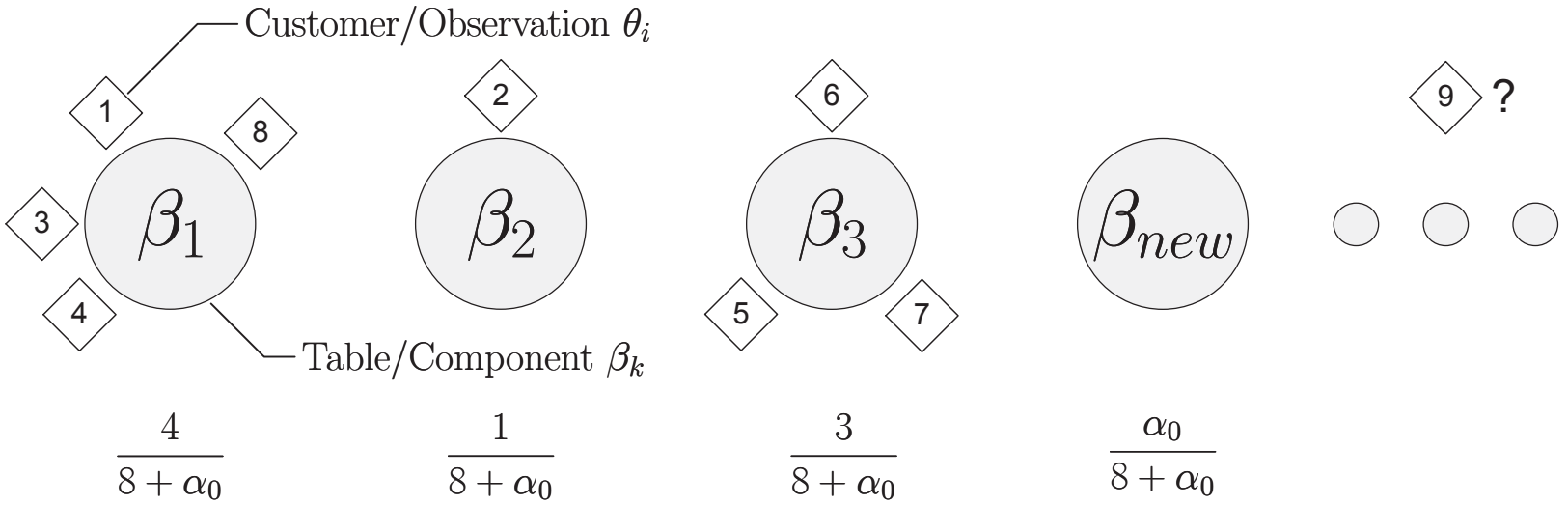


$$p(\text{occupied table } k \mid \text{previous customers}) = \frac{n_k}{i-1+\alpha_0}$$

$$p(\text{new table} \mid \text{previous customers}) = \frac{\alpha_0}{i-1+\alpha_0}$$

(Aldous, 1985)

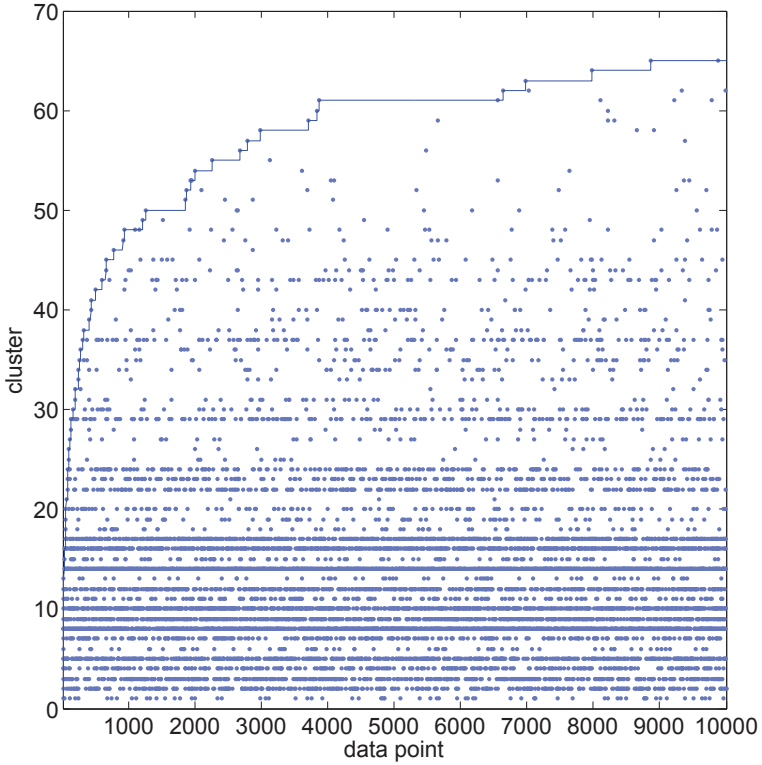
Chinese restaurant process

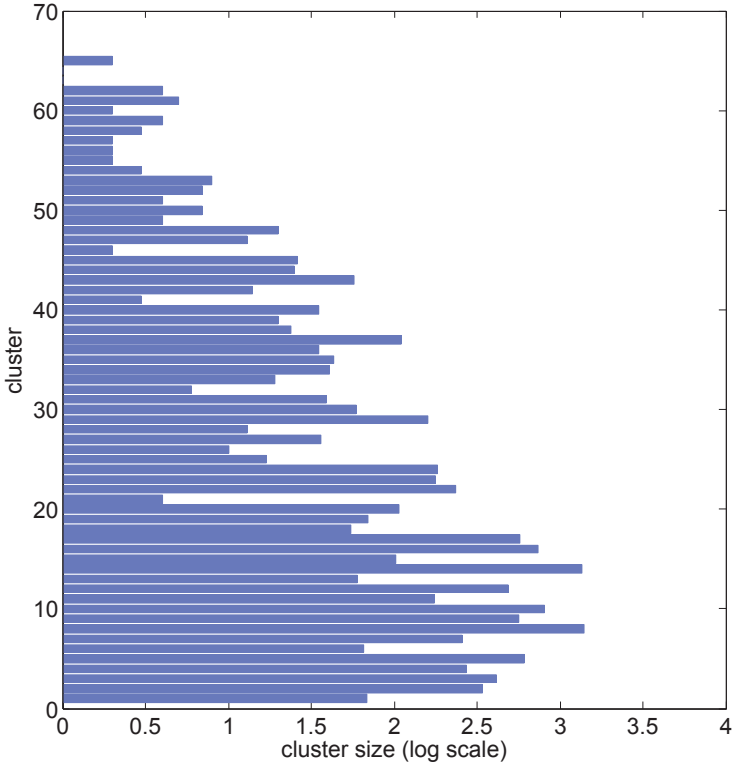


$$\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{n_k}{i-1+\alpha_0} \delta_{\beta_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

(Aldous, 1985)

Asymptotics





Dirichlet process with $\alpha_0 = 20$

number of clusters $K \propto O(\alpha_0 \log N)$

Hierarchical and nested models

- Hierarchical DP
  - analyses the groups of data items by introducing dependencies through G_0

$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H) \\
G_j \mid G_0 &\sim \mathrm{DP}(\alpha_0, G_0) && \text{for each } j \\
\theta_{ji} \mid G_j &\sim G_j && \text{for each } j, i
\end{aligned}
$$

- Nested DP
  - partitions one set of data items into different groups, and analyses each group separately

$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\alpha_0, \mathrm{DP}(\gamma, H)) \\
G_i &\sim G_0 && \text{for each } i \\
\theta_i \mid G_i &\sim G_i && \text{for each } i
\end{aligned}
$$

(Teh, 2010, Rodrìguez et al., 2008)

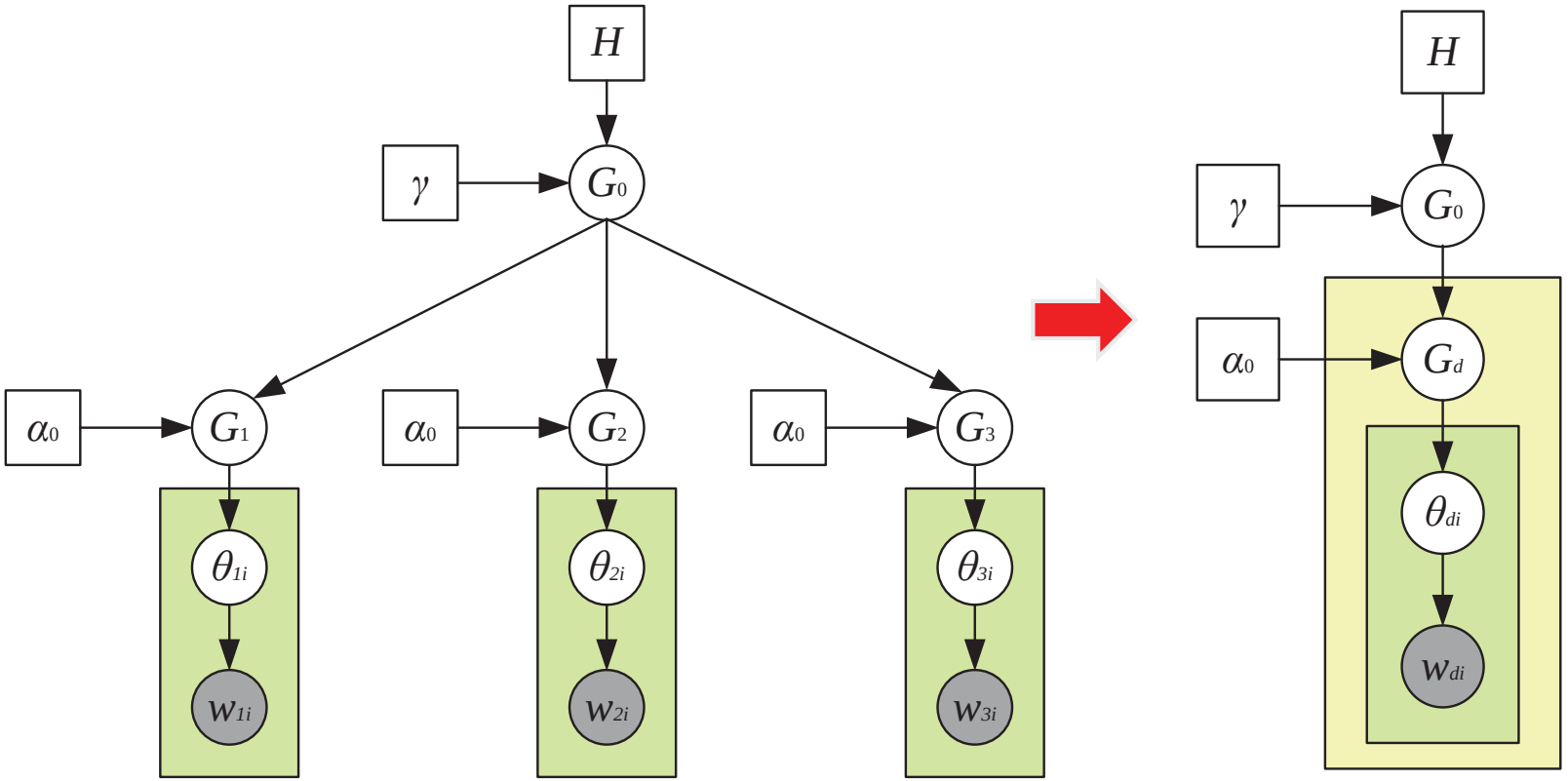
Hierarchical Dirichlet process

- Teh et al. (2006) conducts Bayesian nonparametric learning from the groups of data or more specifically a set of documents
- Two layers of Dirichlet process (DP) are used

$$G_0 \sim \text{DP}(\gamma, H), \qquad G_d \sim \text{DP}(\alpha_0, G_0)$$

- Each document d is associated with a draw from individual G_d
- Base measure of G_d comes from the other DP G_0 which is a nonparametric prior and is shared among different documents
- HDP deals with the problem of mixed membership in representation of grouped data

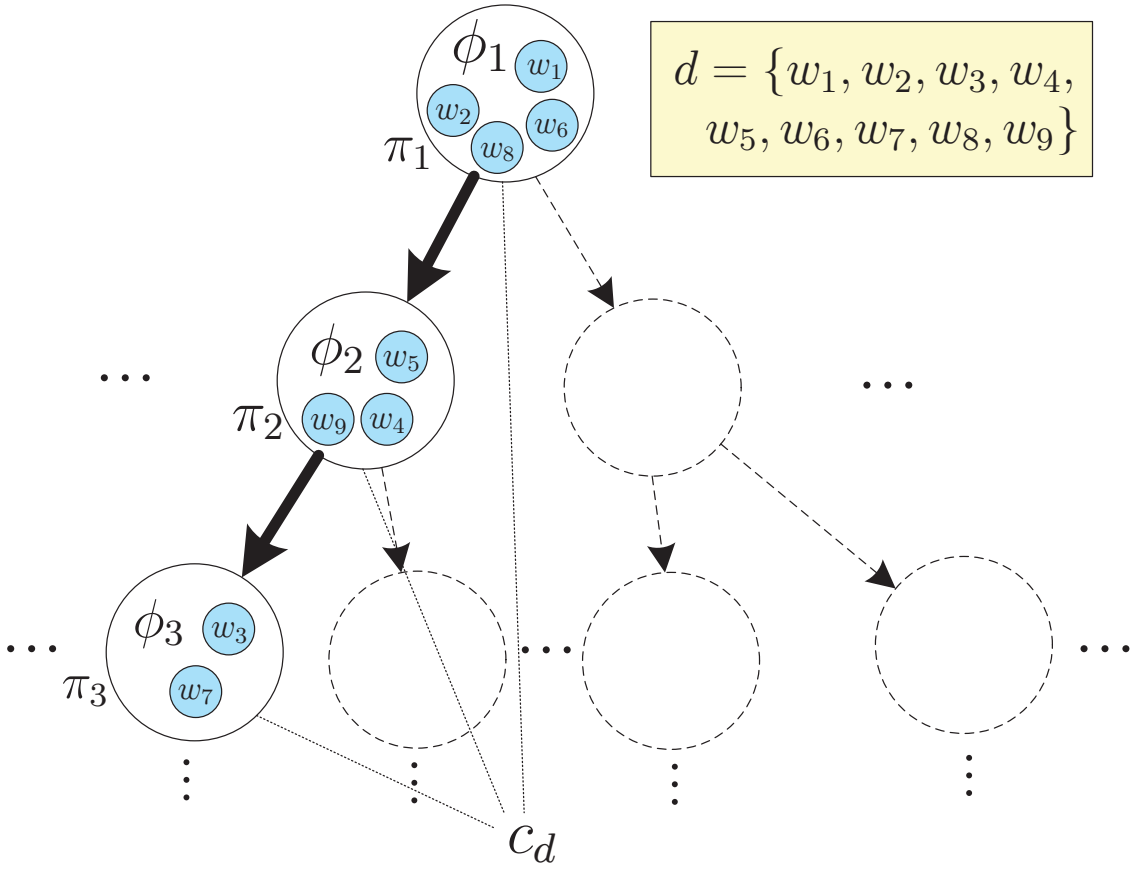
Graphical representation



The nested Chinese restaurant process

- Hierarchies of the topics are explored (Blei et al., 2010)
- Tree model with infinite tree branches and layers is built according to the Bayesian nonparametrics
- Each customer or document d visits the restaurants and selects an associated tree path consisting of topics with different levels of sharing
- Words in a document are drawn from a mixture model of topics given a tree path c_d

Illustration



Model construction

1. For each node k in the infinite tree
   (a) Draw a topic with parameter $\phi_k|H \sim H$.

2. For each document $\mathbf{w}_d = \{w_{di}|i = 1, \ldots, N_d\}$
   (a) Draw a tree path by $c_d \sim \text{nCRP}(\alpha_0)$.
   (b) Draw topic proportions over layers of the tree path c_d by stick-breaking process $\boldsymbol{\pi}_d|\gamma \sim \text{GEM}(\gamma)$.
   (c) For each word w_{di}
      i. Choose a layer or a topic by $z_{di} = k|\boldsymbol{\pi}_d \sim \boldsymbol{\pi}_d$.
      ii. Choose a word based on topic $z_{di} = k$ by $w_{di}|z_{di}, c_d, \{\phi_k\}_{k=1}^{\infty} \sim \text{Mult}(\theta_{di} = \phi_{c_d(z_{di})})$.

Pitman-Yor process

- Pitman-Yor process draws a longer tail probability measure than DP by $G \sim \mathrm{PY}(d, \alpha_0, G_0)$
- Pitman-Yor process produces power law behavior with K in an order of $O(\alpha_0 N^{\rho})$
- When $d = 0$, Pitman-Yor process degenerates into Dirichlet process

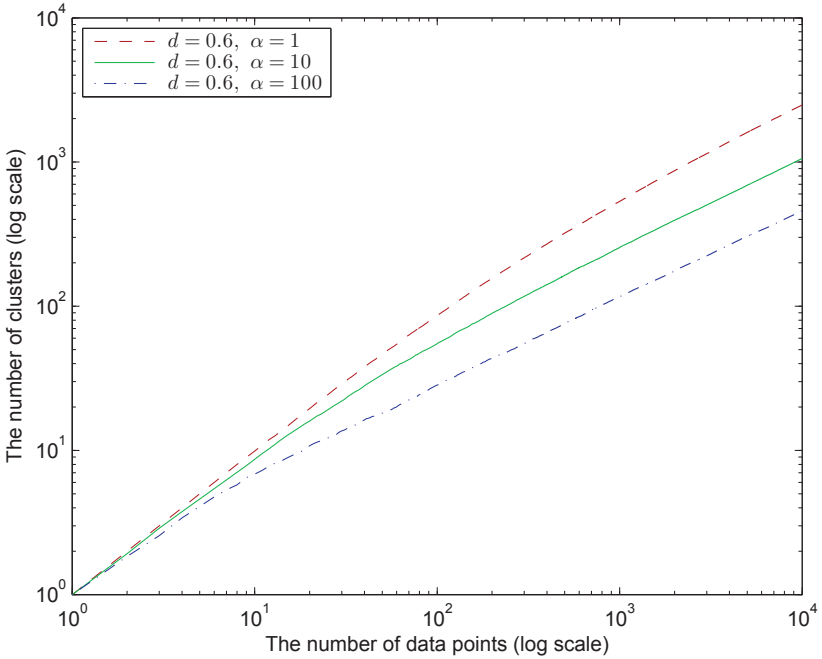
$$p(\text{occupied table } k \mid \text{previous customers}) = \frac{n_k - d}{i - 1 + \alpha_0}$$

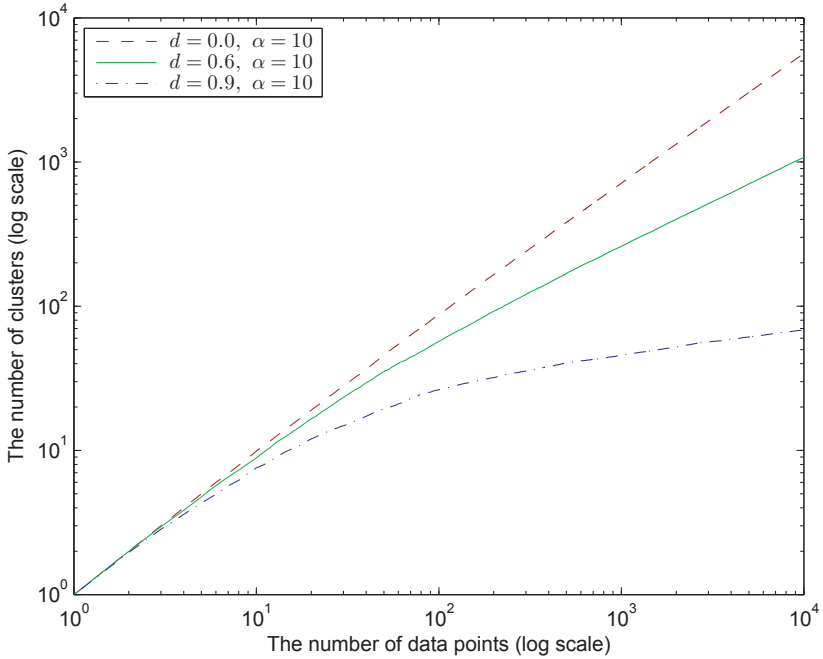
$$p(\text{next new table} \mid \text{previous customers}) = \frac{\alpha_0 + dt}{i - 1 + \alpha_0}$$

(Teh, 2006)

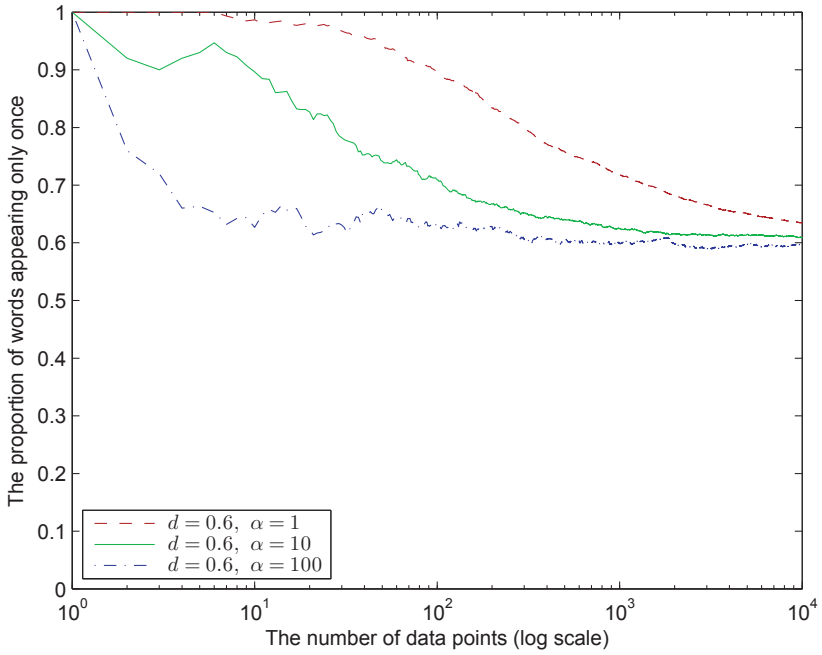
Power-law property

- Pitman-Yor process (Pitman & Yor 1997) produces power-law distributions which resemble those seen in natural language
  - rich-gets-richer property





Power-law property



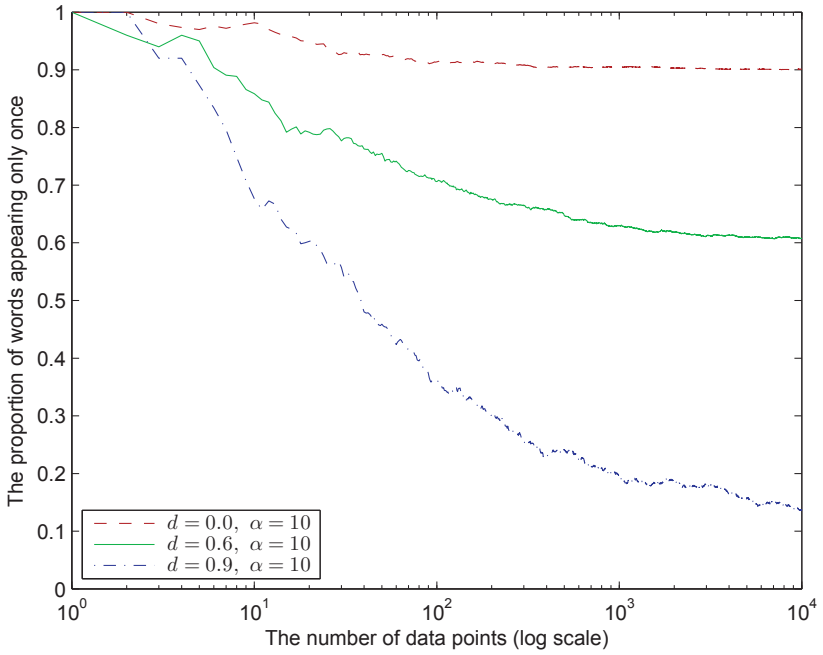


Table of contents

Why Document Summarization?

Introduction

- We extend the topic model for document representation based on the hierarchical Dirichlet process (Teh et al., 2006)
- A hierarchical tree model is constructed and applied for document summarization
- We aim to select the thematic sentences from multiple documents for a summary
- We usually observe heterogeneous documents where the topics are ambiguous, inconsistent and diverse
- A good summary system should reflect diverse topics of documents and keep redundancy to a minimum.

Some issues

- How to conduct deep unsupervised learning
- How to group the sentences into themes and group the words into topics
- How to determine the number of themes and topics from data
- How to discover the tree structure of themes and topics
- How to deal with the heterogeneous documents

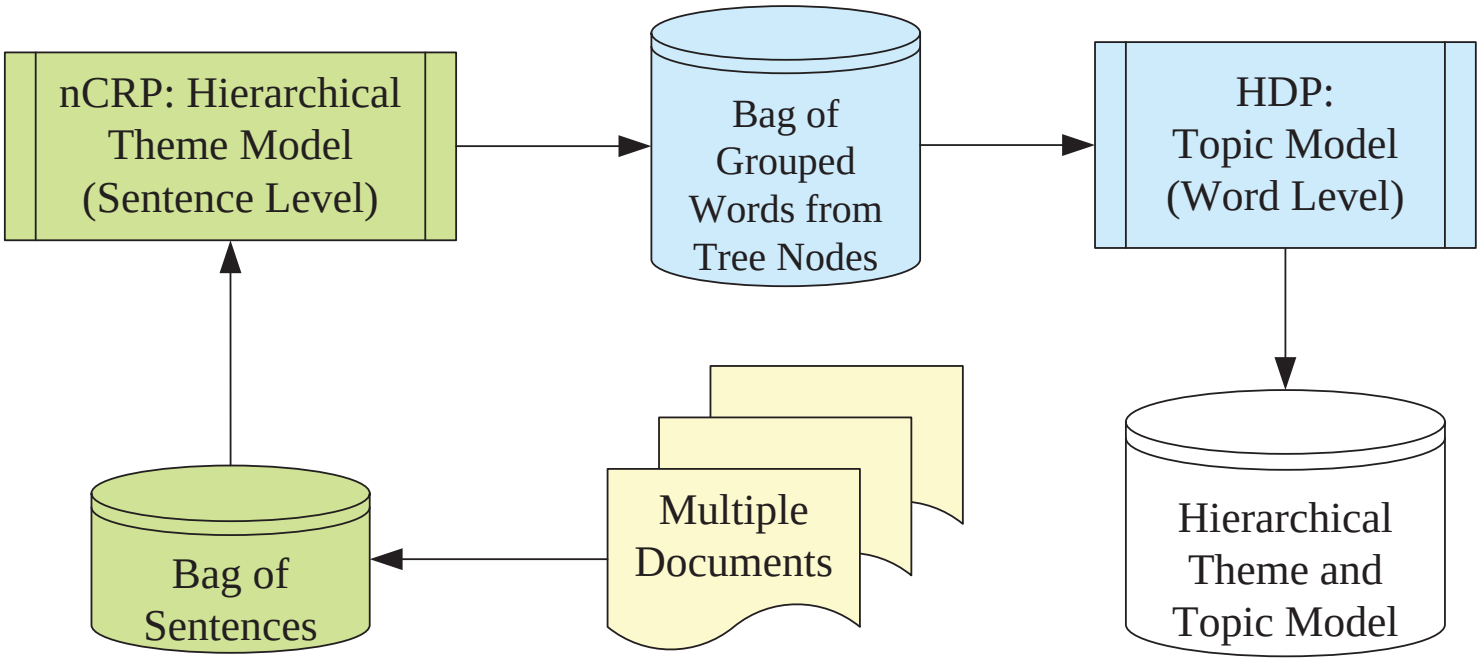
Motivation

- We construct a tree model for document representation where
  - each node is formed by the thematically-related sentences
  - sentences are driven by the sentence-based nCRP
  - each path reflects the hierarchical themes for a document
  - each word is drawn from a topic model given a theme
  - the words in different nodes are seen as the grouped data which are sampled by a HDP

- Heterogeneous documents are ubiquitous
  - long articles contain diverse and confusing themes
  - a subtree with multiple paths is selected to compensate the ambiguous themes for the sentences in a document

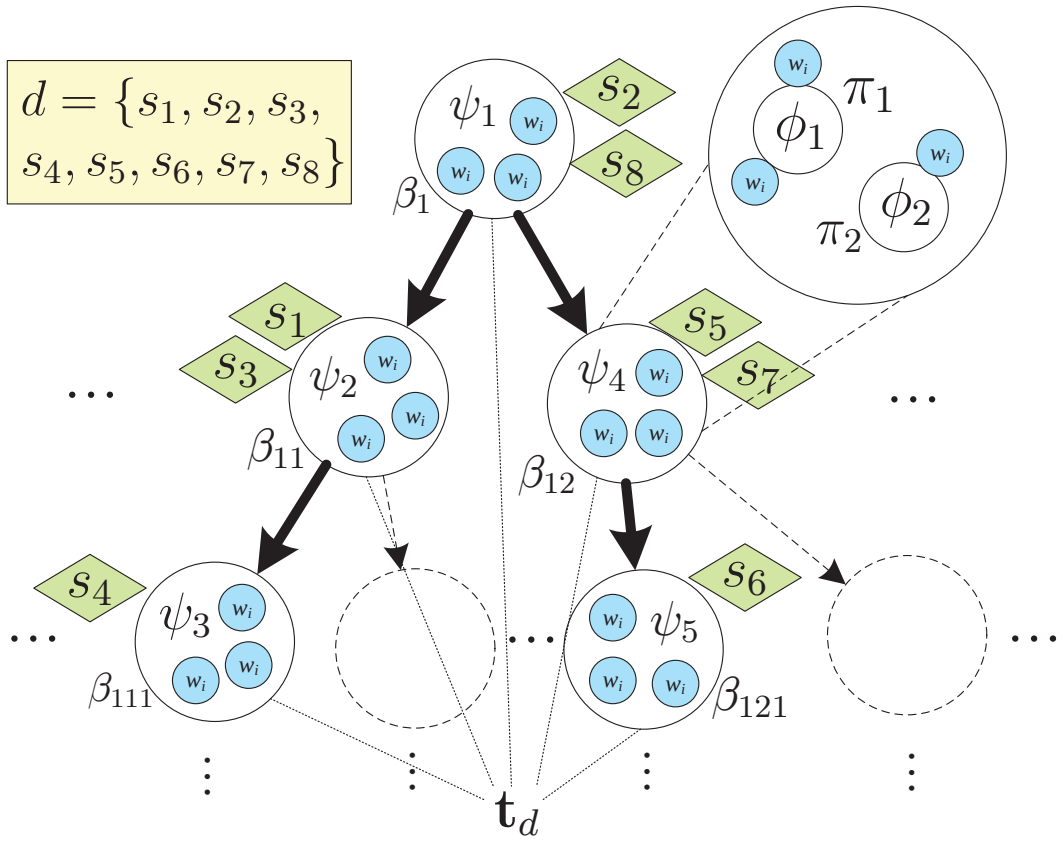
(Chien and Chang, 2013)

Systematic diagram



Hierarchical theme and topic model

- Each document consists of a "bag of sentences" while each sentence consists of a "bag of words"



Two-stage procedure

- Each sentence s_j of a document d is drawn from a mixture model of themes $\{\psi_l\}_{l=1}^{\infty}$ along its corresponding tree paths $\mathbf{c}_d$

$$\sum_{l=1}^{\infty} \beta_{dl} \cdot \delta_{\psi_l}$$

- Each word w_i of the sentences in a node is drawn by a mixture model of topics $\{\phi_k\}_{k=1}^{\infty}$

$$\sum_{k=1}^{\infty} \pi_{lk} \cdot \delta_{\phi_k}$$

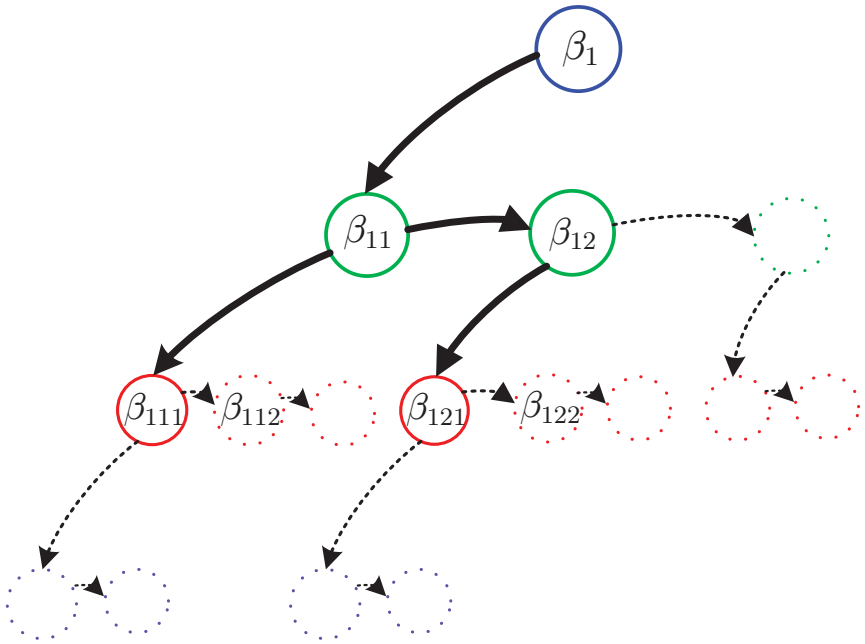
- The hierarchy of themes and topics is built by $\psi_l \sim \sum_k \pi_{lk} \cdot \phi_k$

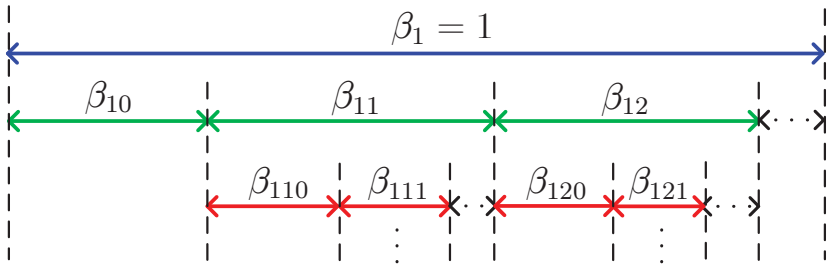
Model construction

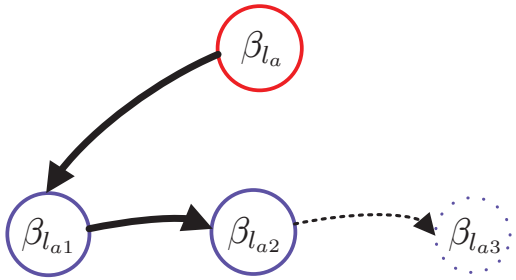
1. For each node or theme l in the infinite tree
   (a) For each topic k in a tree node, draw a topic with parameter $\phi_k|H \sim H$
   (b) Draw topic proportions by stick-breaking process $\boldsymbol{\pi}_l|\gamma_w \sim \text{GEM}(\gamma_w)$.
   (c) Theme model is constructed by $\psi_l \sim \sum_k \pi_{lk}\phi_k$.

2. For each document $\mathbf{w}_d = \{\mathbf{w}_{dj}\}$
   (a) Draw subtree paths $\mathbf{t}_d = \{t_{dj}\} \sim \text{snCRP}(\alpha_0)$.
   (b) Draw theme proportions over subtree paths $\mathbf{t}_d$ in different layers by tree stick-breaking process $\boldsymbol{\beta}_d|\gamma_s \sim tree\text{GEM}(\gamma_s)$.
   (c) For each sentence $\mathbf{w}_{dj} = \{w_{dji}\}$ with a tree path t_{dj}
      i. Choose a layer or a theme $y_{dj} = l|\boldsymbol{\beta}_d \sim \boldsymbol{\beta}_d$.
      ii. For each word w_{dji} or simply w_{di}
         A. Choose a topic by $z_{di} = k|\boldsymbol{\pi}_l \sim \boldsymbol{\pi}_l$.
         B. Choose a word based on topic $z_{di} = k$ by $w_{di}|z_{di}, t_{dj}, \{\phi_k\}_{k=1}^{\infty} \sim \text{Mult}(\theta_{dji} = \phi_{t_{dj}(z_{di})})$.

Tree stick-breaking process







Distribution of theme proportions

- A subtree of document d is sampled by $\mathbf{t}_d \sim \text{snCRP}(\alpha_0)$. TSBP is then performed to determine the theme proportions

$$\boldsymbol{\beta}_d | \gamma_s \sim tree\text{GEM}(\gamma_s)$$

- $Tree$GEM distribution is obtained through the procedure
  - beta variable of a child node l_{ac} is drawn by $\beta'_{l_{ac}} \sim \text{Beta}(1, \gamma_s)$
  - probability of this draw is multiplied by the theme proportion θ_{l_a} of an ancestor node l_a to find theme proportion for its chide nodes l_{ac}

$$\beta_{l_{ac}} = \beta_{l_a} \beta'_{l_{ac}} \prod_{j=1}^{c-1} (1 - \beta'_{l_{aj}})$$

HDP for groups of words

- Each word w_i of a sentence s_j from document d is sampled from a mixture model of topics according to an HDP
- Stick-breaking process is conducted to realize an DP where the topic proportions $\boldsymbol{\pi_l} = \{\pi_{lk}\}$ are sampled for individual node or theme l
- Each word w_i in tree node l is generated by topics ϕ_k and topic proportions π_{lk}

$$\phi_k|H \sim H, \quad \boldsymbol{\pi}_l|\gamma_w \sim \text{GEM}(\gamma_w)$$

$$z_{di} = k|\boldsymbol{\pi}_l \sim \boldsymbol{\pi}_l, \quad w_{di}|z_{di}, t_{dj}, \{\phi_k\}_{k=1}^{\infty} \sim \text{Mult}(\phi_{t_{dj}(z_{di})})$$

- We realize the snCRP compound HDP
- Clustering of sentences is obtained for document summarization

Sampling of a subtree

- Gibbs sampling is developed to infer posterior parameters
- A subtree with multiple paths $\mathbf{t}_d = \{t_{dj}\}$ is sampled according to

$$p(t_{dj} = t|\mathbf{t}_{d(-j)}, \mathbf{w}, \mathbf{y}_d, \alpha_0, \eta) \propto p(t_{dj} = t|\mathbf{t}_{d(-j)}, \alpha_0) \times p(\mathbf{w}_{dj}|\mathbf{w}_{d(-j)}, \mathbf{t}_d, \mathbf{y}_d, \eta)$$

where

$$p(t_{dj} = t|\mathbf{t}_{d(-j)}, \alpha_0) = \begin{cases} \frac{n_{d,t}}{n_d - 1 + \alpha_0} & \text{if an occupied path } t \text{ is chosen} \\ \frac{\alpha_0}{n_d - 1 + \alpha_0} & \text{if a new path is chosen} \end{cases}$$

- Here, $n_{d,t}$ denotes the number of sentences in document $\mathbf{w}_d$ that are allocated along tree path t and c_d denotes total number of sentences in this document

Sampling of themes and topics

- Sampling of themes
  - given the current paths selected by snCRP, we sample a tree node at level or theme l for each sentence s_j according to

$$p(y_{dj} = l|\mathbf{w}_d, \mathbf{y}_{d(-j)}, \mathbf{t}_d, \gamma_s, \eta) \propto p(y_{dj} = l|\mathbf{y}_{d(-j)}, \mathbf{t}_d, \gamma_s) \times p(\mathbf{w}_{dj}|\mathbf{w}_{d(-j)}, \mathbf{y}_d, \mathbf{t}_d, \eta)$$

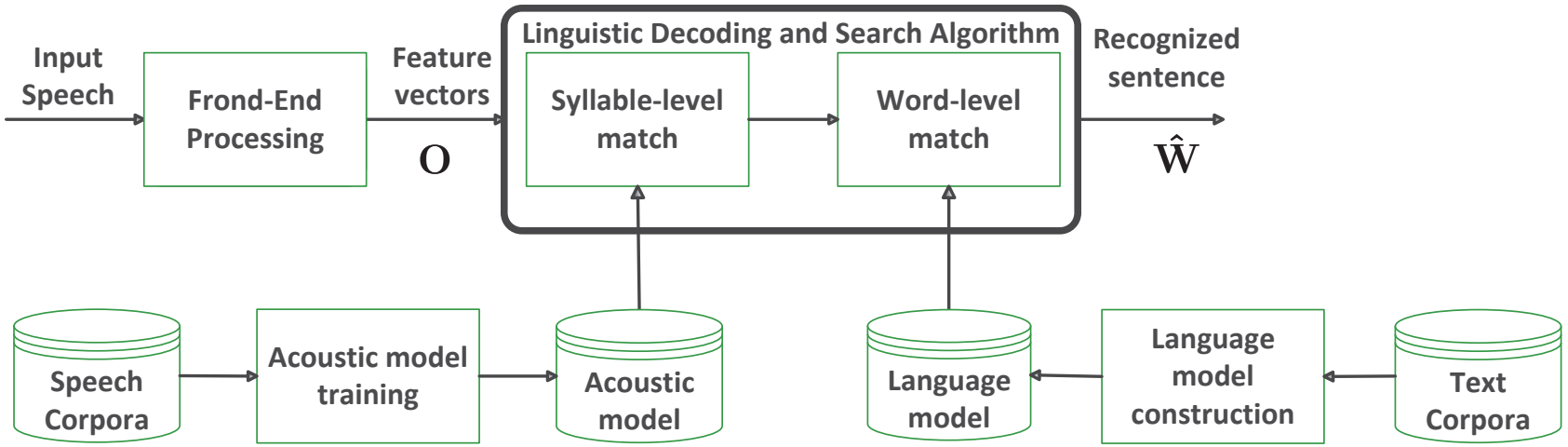
- Sampling of topics
  - stick-breaking process draws the topics for words w_{dji} in different tree nodes according to

$$p(z_{di} = k|\mathbf{w}, \mathbf{z}_{-(di)}, y_{dj} = l, \gamma_w, \eta) \propto p(z_{di} = k|\mathbf{z}_{-(di)}, y_{dj} = l, \gamma_w) \times p(w_{di}|\mathbf{w}_{-(di)}, \mathbf{z}, y_{dj} = l, \eta)$$

Table of contents

Why Language Modeling?

Automatic speech recognition



$$\hat{W} = \arg\max_{W} p(W|O) = \arg\max_{W} P_{\Theta}(O|W) P_{\Gamma}(W)$$

Language modeling

- LM plays an important role in speech recognition for finding the best word sequence $\hat{W}$ according to the Bayes decision rule.

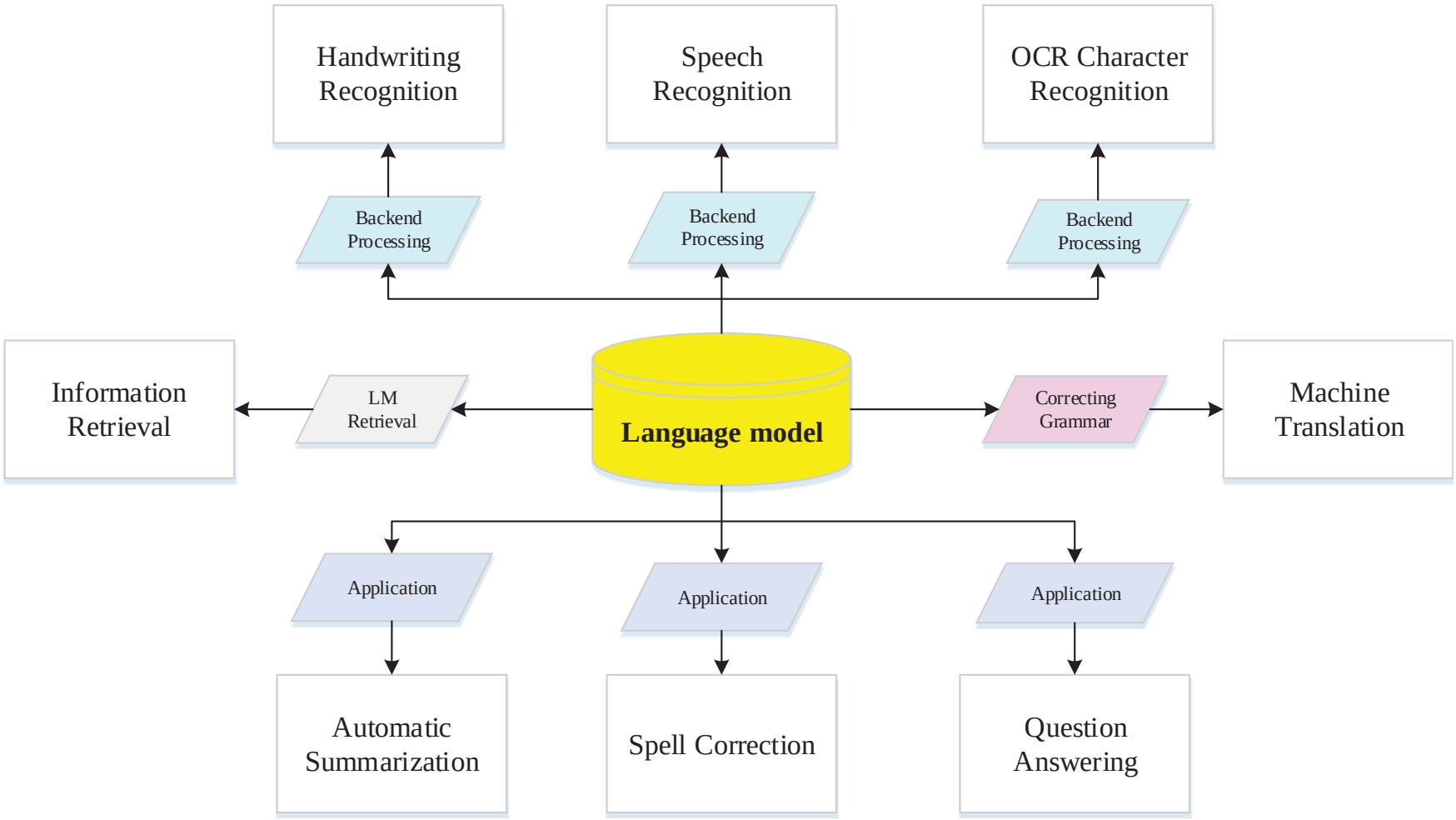
- LM based on n-gram

$$p(W) = p(w_1, \cdots, w_T) = \prod_{i=1}^{T} p(w_i|w_1, \cdots, w_{i-1}) \approx \prod_{i=1}^{T} p(w_i|w_{i-n+1}^{i-1})$$

- N-gram probability based on maximum likelihood is calculated by

$$p_{\mathrm{ML}}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i})}{c(w_{i-n+1}^{i-1}\blacksquare)} \quad \text{where} \quad c(w_{i-n+1}^{i-1}\blacksquare) = \sum_{w_i} c(w_{i-n+1}^{i})$$

Applications of language model



Issues in language modeling

- Data sparseness problem - model smoothing
  - class-based LM
  - backoff method
  - continuous space LM, neural network LM
- Insufficient long-distance regularity - topic/class information
  - latent Dirichlet allocation (LDA) (Blei et al., 2003)
  - cache information
- Model regularization - mismatch between training and test data

Class-based n-gram

- A simple approach to tackle data sparseness problem is to consider the transition probabilities between classes rather than words (P. Brown and Mercer, 1992)

$$p(w_i|w_{i-n+1}^{i-1}) \approx p(w_i|c_i)p(c_i|c_{i-n+1}^{i-1})$$

- The class of a word is assigned according to the word clustering based on the metric of mutual information

- Hard clustering is considered to find the word class

Interpolation smoothing

- Chen and Goodman (1999) surveyed a series of smoothing algorithms which cope with zero probability estimates for n-grams not observed in the training corpus

- Interpolation smoothing is performed by linearly combining higher-order n-grams with lower-order n-grams

$$p_{\text{INT}}(w_i|w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}} p_{\text{ML}}(w_i|w_{i-n+1}^{i-1})$$

$$+(1 - \lambda_{w_{i-n+1}^{i-1}}) p_{\text{INT}}(w_i|w_{i-n+2}^{i-1})$$

- Parameter $\lambda_{w_{i-n+1}^{i-1}}$ is estimated for each w_{i-n+1}^{i-1} by maximum likelihood method

Backoff smoothing

- Katz smoothing (Katz, 1987) combines higher-order models with lower-order models by correcting the counts in ML model

- For a bigram, the count $r = c(w_{i-1}^i)$ is corrected by

$$c_{\mathsf{KZ}}(w_{i-1}^i) = \begin{cases} d_r r & \text{if } r > 0 \\ \alpha(w_{i-1}) p_{\mathsf{ML}}(w_i) & \text{if } r = 0 \end{cases}$$

  where $\alpha(w_{i-1})$ is chosen to meet $\sum_{w_i} c_{\mathsf{KZ}}(w_{i-1}^i) = \sum_{w_i} c(w_{i-1}^i)$

- LM with Katz smoothing is calculated by

$$p_{\mathsf{KZ}}(w_i|w_{i-1}) = \frac{c_{\mathsf{KZ}}(w_{i-1}^i)}{\sum_{w_i} c_{\mathsf{KZ}}(w_{i-1}^i)}$$

Kneser-Ney LM

- Kneser-Ney (KN) smoothing (Kneser and Ney, 1995) calculates the interpolated model by

$$p_{\mathsf{KN}}(w_i|w_{i-n+1}^{i-1}) = \frac{\max\{c(w_{i-n+1}^{i}) - d, 0\}}{c(w_{i-n+1}^{i-1}\blacksquare)} + \frac{d \cdot N_{1+}(w_{i-n+1}^{i-1}\blacksquare)}{c(w_{i-n+1}^{i-1}\blacksquare)} p_{\mathsf{KN}}(w_i|w_{i-n+2}^{i-1})$$

Backoff model is given by

$$p_{\mathsf{KN}}(w_i|w_{i-n+1}^{i-1}) = \begin{cases} \frac{\max\{c(w_{i-n+1}^{i})-d,0\}}{c(w_{i-n+1}^{i-1}\blacksquare)} & \text{if } c(w_{i-n+1}^{i}) > 0 \\ \alpha(w_{i-n+1}^{i-1}) p_{\mathsf{KN}}(w_i|w_{i-n+2}^{i-1}) & \text{if } c(w_{i-n+1}^{i}) = 0 \end{cases}$$

Bayesian nonparametric LM

- Hierarchical Pitman-Yor (HPY)-LM is an Bayesian extension of KN-LM where nonparametric prior is considered
- HPY-LM is estimated from the HPY process (Teh 2006)
- Gibbs sampling was applied for model inference based on the Chinese restaurant metaphor
- HPY-LM improved the performance over KN-LM for speech recognition (Huang & Renals 2010)
- Power law property in natural language is reflected by HPY-LM

HPY Process

- HPY process (Teh, 2006) is formed by

$$G_\phi \sim \mathrm{PY}(d_0, \alpha_0, G_b) \quad \text{and} \quad G_U \sim \mathrm{PY}\left(d_{|U|}, \alpha_{|U|}, G_{\pi(U)}\right)$$

$$p(w_i|w_{i-1}, \cdots\cdots, w_{i-n+2}, w_{i-n+1})$$

$$G_\emptyset \sim \mathrm{PY}(d_0, \alpha_0, G_0)$$

$$G_1 \sim \mathrm{PY}(d_1, \alpha_1, G_\emptyset)$$

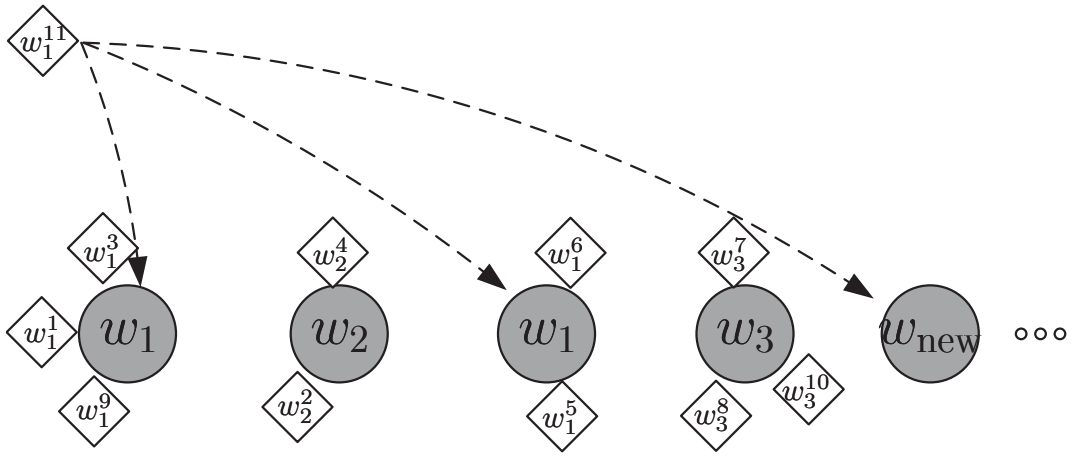
$$\cdots \qquad \cdots$$

$$G_{\pi(U)} \sim \mathrm{PY}\left(d_{|\pi(U)|}, \alpha_{|\pi(U)|}, G_{\pi(\pi(U))}\right)$$

$$G_U \sim \mathrm{PY}\left(d_{|U|}, \alpha_{|U|}, G_{\pi(U)}\right)$$

Chinese restaurant metaphor

- Pitman-Yor process PY(d, α, G_0) is realized by



$$p(\text{occupied table } k|\text{previous customers}) = \frac{c_k - d}{\alpha + c_\cdot}$$

$$p(\text{new table } k|\text{previous customers}) = \frac{\alpha + dt_\cdot}{\alpha + c_\cdot}$$

$$p(z_i = k|\mathbf{z}_{-i}) = \begin{cases} \frac{c_k - d}{\alpha + c_\cdot}, & 1 \le k \le t_\cdot \\ \frac{\alpha + dt_\cdot}{\alpha + c_\cdot}, & k = t_\cdot + 1 \end{cases}$$

Hierarchical Pitman-Yor LM

- Given the seating arrangement S, HPY-LM is calculated by

$$p(w|S, d_0, \alpha_0) = \sum_{k=1}^{t_\cdot} \frac{c_k - d_0}{\alpha_0 + c_\cdot} \delta_{kw} + \frac{\alpha_0 + d_0 t_\cdot}{\alpha_0 + c_\cdot} G_b(w)$$

$$= \frac{c_k - d_0 t_w}{\alpha_0 + c_\cdot} + \frac{\alpha_0 + d_0 t_\cdot}{\alpha_0 + c_\cdot} G_b(w)$$

More specifically, we have

$$p_{\mathsf{HPY}}(w_i|w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^{i}) - d_{|w_{i-n+1}^{i-1}|} N_{1+}(w_{i-n+1}^{i})}{\alpha_{|w_{i-n+1}^{i-1}|} + c(w_{i-n+1}^{i-1}\cdot)}$$

$$+ \frac{\alpha_{|w_{i-n+1}^{i-1}|} + d_{|w_{i-n+1}^{i-1}|} N_{1+}(w_{i-n+1}^{i-1}\cdot)}{\alpha_{|w_{i-n+1}^{i-1}|} + c(w_{i-n+1}^{i-1}\cdot)} p_{\mathsf{HPY}}(w_i|w_{i-n+2}^{i-1})$$

Hierarchical Pitman-Yor-Dirichlet LM

- We propose the hierarchical Pitman-Yor-Dirichlet (HPYD) LM
- Backoff smoothing and topic clustering are performed via the Bayesian nonparametric learning
- A hybrid probability measure is drawn from HPYD process to sample the smoothed topic-based LM
- A new Chinese restaurant scenario is implemented for HPYD-LM via Gibbs sampling
- HPYD-LM extracts the semantic topics and reflects the power-law property for natural language

(Chien and Chang, 2013)

Topic-based LM

- Topic-based LM (Gildea and Hofmann, 1999) captures the long-range word dependencies through latent topics

$$p(w_i|w_{i-n+1}^{i-1}) = \sum_{z_i} p(w_i|w_{i-n+1}^{i-1}, z_i)p(z_i|w_{i-n+1}^{i-1})$$

- Latent Dirichlet allocation (LDA) (Blei et al., 2003) was proposed for topic modeling and applied to build LDA-LM (Tam and Schultz, 2005) through LM adaptation

- Parametric mixture models with fixed number of topics

Hierarchical Pitman-Yor-Dirichlet Process

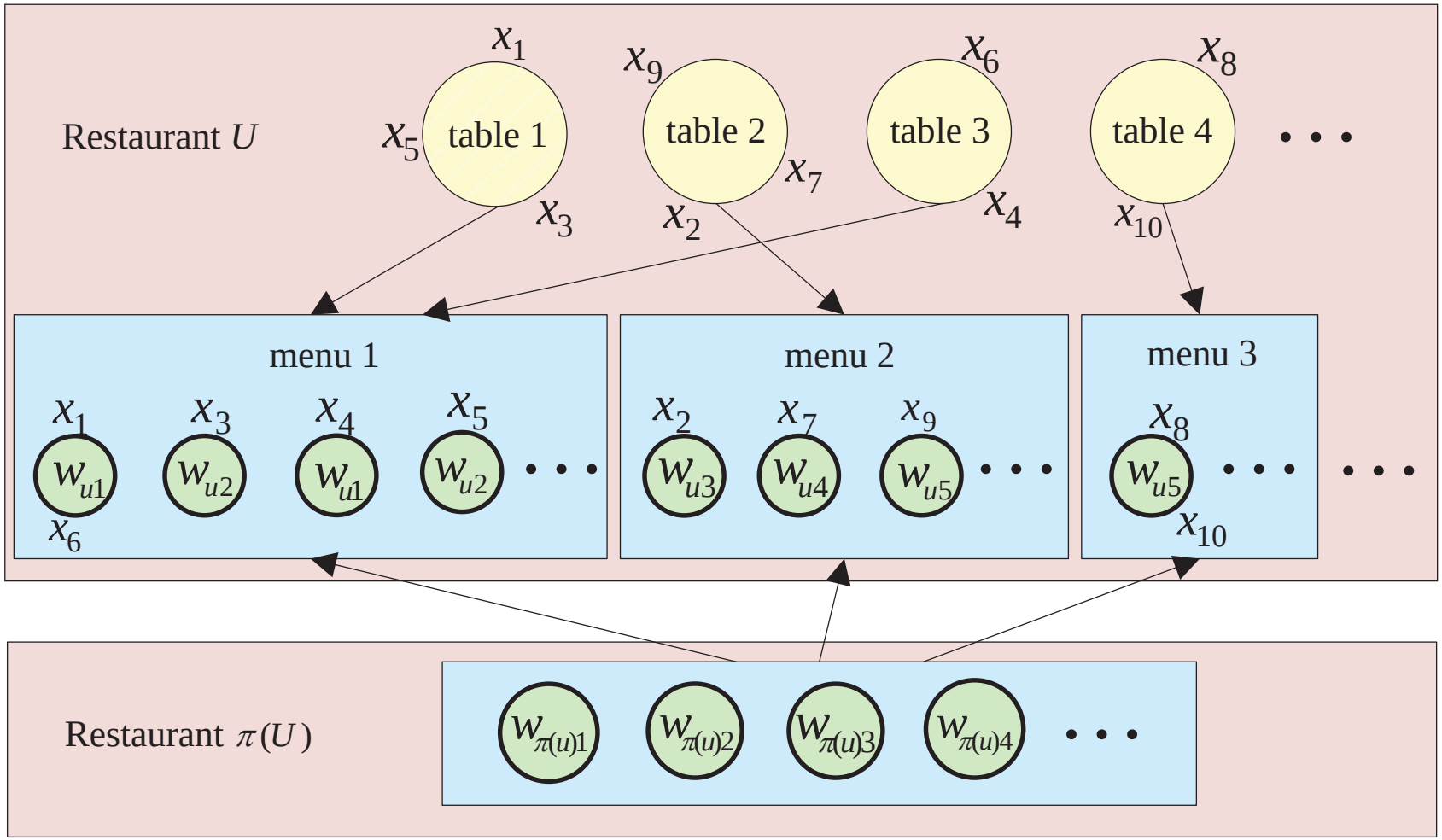
- HPYD process is realized by recursively sampling the topic-dependent n-gram $p(w_i|w_{i-n+1}^{i-1}, z_i) \triangleq H_{w_{i-n+1}^{i-1} z_i}$ and then sampling the n-gram $p(w_i|w_{i-n+1}^{i-1}) \triangleq G_{w_{i-n+1}^{i}}$

$$H_{w_{i-n+1}^{i-1} z_i} \sim \mathrm{PY}(\alpha_n, d_n, G_{w_{i-n+1}^{i-1}})$$
$$G_{w_{i-n+1}^{i}} \sim \mathrm{DP}(\gamma_n, H_{w_{i-n+1}^{i-1} z_i})$$

- HPYD probability measure is produced by

$$H_{w_{i-n+1}^{i-1}(z_i=k)} \sim \frac{c_{ukw_i\cdot} - d_n t_{ukw_i}}{c_{uk\cdot\cdot} + \alpha_n} + \frac{\alpha_n + d_n t_{u\cdot w_i}}{c_{u\cdots} + \alpha_n} G_{w_{i-n+1}^{i-1}}$$
$$G_{w_{i-n+1}^{i}} \sim \sum_{t=1}^{m_{u\cdot}} \frac{n_{ut}}{n_{u\cdot} + \gamma_n} H_{w_{i-n+1}^{i-1}(z_i=t)} + \frac{\gamma_n}{n_{u\cdot} + \gamma_n} H_{w_{i-n+1}^{i-1} z_i}$$

Hierarchical Chinese restaurant process



Bayesian Inference

- We sample tables, menus and dishes according to the posterior probabilities

$$p(t_i = t|\mathbf{t}_{-i}, \mathbf{z}, \mathbf{w}, U)$$

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{t}, \mathbf{w}, U)$$

$$p(l_i = w|\mathbf{l}_{-i}, z_i = k, \mathbf{z}_{-i}, \mathbf{w}_{-i}, U)$$

  respectively, where $\mathbf{w} = \{w_i, \mathbf{w}_{-i}\}$ and "$-$" is self-exception

- Generating new table $t = t_{\text{new}}$, new menu $k = k_{new}$ and new dish $l = l_{\text{new}}$ is driven from these posterior probabilities

HPYD language model

- HPYD processs is developed to produce the HPYD LM

$$p(w_i = w|\mathbf{w}^{-i}, \mathbf{z}, \boldsymbol{\lambda}, U) \propto \sum_{k=1}^{K} \frac{\sum_t n_{u(t=k)(-i)}}{n_{u\cdot(-i)} + \gamma_n}$$
$$\times \left[\frac{c_{ukw\cdot(-i)} - d_n t_{ukw}}{c_{uk\cdot\cdot(-i)} + \alpha_n} + \frac{\alpha_n + d_n t_{uk\cdot}}{c_{uk\cdot\cdot(-i)} + \alpha_n} p(w|\pi(U)) \right]$$
$$+ \frac{\gamma_n}{n_{u\cdot(-i)} + \gamma_n} p(w|\mathbf{w}_{-i}, z_i = k_{\mathsf{new}}, \mathbf{z}_{-i}, \boldsymbol{\lambda}, U)$$

Noations

n_{ut} number of customers or words sitting at table t

c_{ukwl} number of customers ordering dish l which is labelled by a distinct word w from menu k given context U

t_{uk} number of tables that choose menu k

t_{ukw} number of dishes in menu k which are ordered by a distinct word w

we have $n_{u\cdot} = \sum_t n_{ut}$ and $c_{ukw\cdot} = \sum_l c_{ukwl}$

Gibbs sampling algorithm

function AddCustomer(U, w)
  draw $d_{|U|} \sim \text{Beta}(a_{|U|}, b_{|U|})$
  draw $\alpha_{|U|} \sim \text{Gamma}(e_{|U|}, f_{|U|})$
  set γ and $\gamma_{|U|}$
  for customer $i \leftarrow 1$ **to** N_u **do**
    sit in a table in restaurant U
      select an occupied table t with probabiltiy $\frac{n_{ut}}{n_{u\cdot}+\gamma_{|U|}}$, or
      select a new table with probabiltiy $\frac{\gamma_{|U|}}{n_{u\cdot}+\gamma_{|U|}}$, then
      draw an existing menu k in proportion to $t_{\cdot k}$, or
      draw a new menu in proportion to γ_0
      order a dish in a distinct menu
      select an ordered dish l labelled by w with $\frac{\max\{0, c_{ukwl} - d_{|U|}\}}{c_{uk\cdot\cdot}+\alpha_{|U|}}$
      select a new dish with $\frac{\alpha_{|U|}+d_{|U|}t_{u\cdot\cdot}}{c_{u\cdot\cdot}+\alpha_{|U|}} G_{\pi(|U|)}(w)$
  end for
end function

Table of contents

Conclusions

- Bayesian nonparametric learning methods were surveyed for document modeling and language modeling.
- Application for document summarization and speech recognition
- A hierarchical theme model was constructed according to a sentence-level nCRP while the topic model was established through a word-level HDP
- A tree stick-breaking procedure was implemented to draw a subtree for a heterogeneous document
- Unsupervised structural learning from different levels of grouped data could be done
- A sentence-based nCRP compound HDP was presented

- We presented a compound process which combined HPY for constructing the topic-dependent backoff LMs and HDP for integrating these LMs into the topic mixture model
- A hierarchical Pitman-Yor-Dirichlet process was developed
- Gibbs sampling was implemented for these compound processes
- A hierarchical Chinese restaurant process was developed for HPYD language model
- We are working on parallel processing algorithm for Bayesian nonparametric learning
- Rapid inference algorithm, variational inference for Bayesian nonparametrics, ...
- More applications including image modeling, music signal processing, ...