Machine Learning for Speech and Language Processing

Jen-Tzung Chien

Hong Kong Polytechnic University, July 24, 2012

TABLE OF CONTENTS

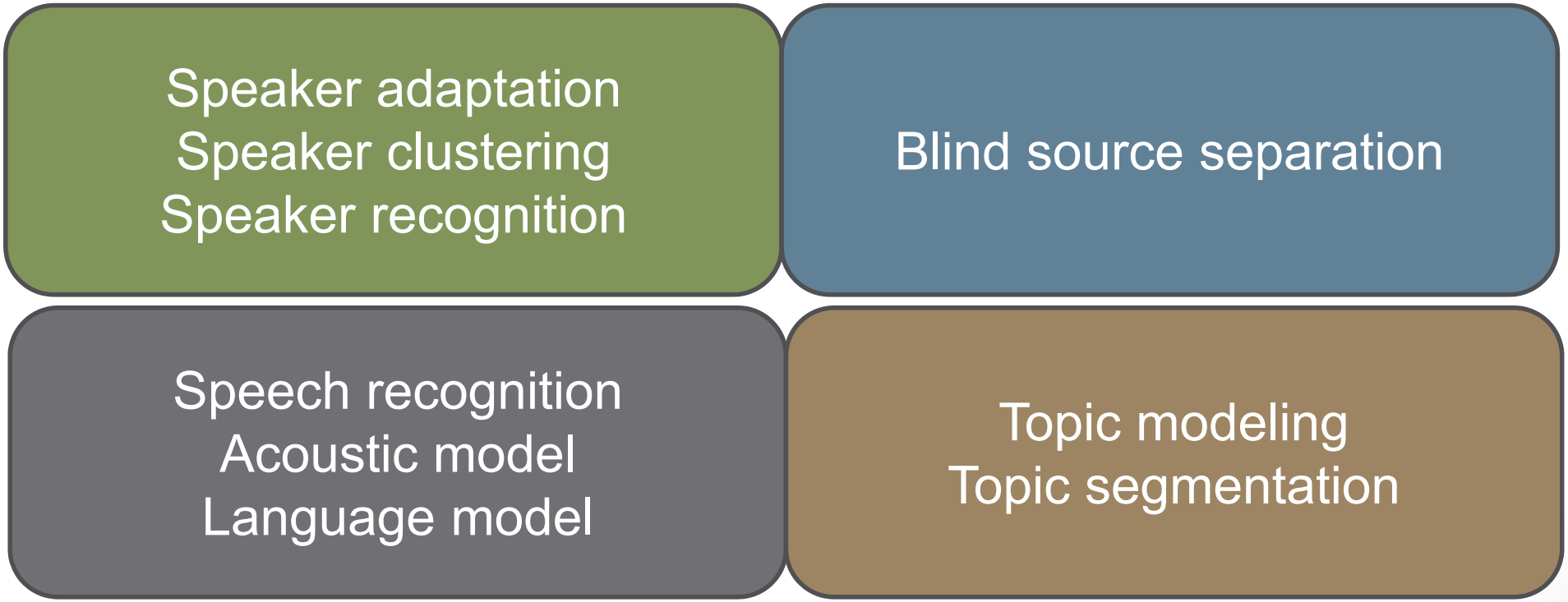
Why machine learning for speech and language processing?

- *Speech and language processing is a large scale problem*
  - *Big data*
  - *High dimensionality and complicated models*

  .....

This lecture provides some case studies of Bayesian learning applications to speech and language processing

Application Areas



We emphasize on presenting our real experience of why and how we apply Bayesian approaches.

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.
- The collected data are prone to be *noisy*, *mislabeled*, *misaligned*, *mismatched*, and *ill-posed*.
- Probabilistic models may be improperly-assumed, over-estimated, or under-estimated.

Modeling Tools

- We need tools for *modeling*, *analyzing*, *searching*, *recognizing* and *understanding* real-world data.
- 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
- *Bayesian theory* provides desirable tools. Uncertainty can be properly expressed by *prior distribution* or *prior process*.

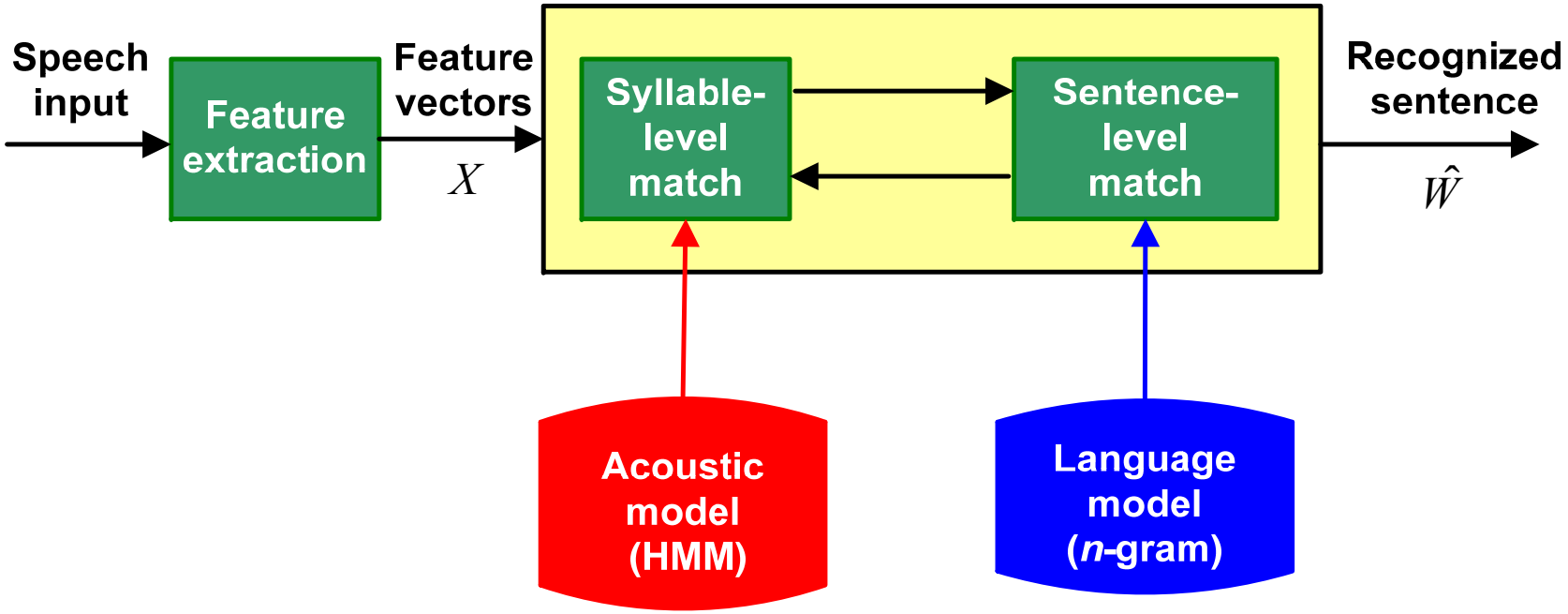
Bayes Rule

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

- Strength of *beliefs* (degrees of plausibility) should be numerically represented.
- Prior beliefs and expert knowledge can be incorporated into the analysis along with the data.
- Bayes rule tells us how to calculate *inverse probability* and do *inference* about hypothesis from data.
- Bayes rule allows us to estimate unknown quantities, adapt models, make predictions and learn from data.

TABLE OF CONTENTS

Speech Recognition System



$$\hat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} p(X|W)\,p(W)$$

Continuous-Density HMMs

- Given $X = \{\mathbf{x}_t\}_{t=1}^{T}$, we accumulate the likelihood function of individual frames based on *Gaussian mixture model* (GMM).

$$p(X \mid \lambda) = \sum_{S=\{s_t\}} \left[\pi_{s_1} p(\mathbf{x}_1 \mid \lambda_{s_1}) \prod_{t=2}^{T} [a_{s_{t-1}s_t} p(\mathbf{x}_t \mid \lambda_{s_t})] \right]$$

$$p(\mathbf{x}_t \mid \lambda_i) = \sum_{j=1}^{J} \omega_{ij} N(\mathbf{x}_t \mid \boldsymbol{\mu}_{ij}, R_{ij})$$

$$\propto \sum_{j=1}^{J} \omega_{ij} \mid R_{ij} \mid^{1/2} \exp\left[-\frac{1}{2} (\mathbf{x}_t - \boldsymbol{\mu}_{ij})^T R_{ij} (\mathbf{x}_t - \boldsymbol{\mu}_{ij}) \right]$$

Why Bayesian Acoustic Model?

- How do we estimate acoustic model from heterogeneous speech data? Is *ML*-based *GMM* a right model?
- Are Gaussians over-trained? Too many Gaussians? Are all Gaussians relevant to represent a new speech frame?
- Can we minimize the *model assumption error*? Can we change model structure?
- How model regularization is assured for unknown test conditions? How model uncertainty is considered?
- Should we collect *unlimited* speech data for LVCSR?

Basis Representation

- Sparse representation of data $\mathbf{x} \in R^D$

$$\mathbf{x} = \Phi \mathbf{w}$$

  - basis vectors $\Phi = [\boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_N]$
  - sensing weights $\mathbf{w} \in R^N$
  - reconstruction errors $\|\mathbf{x} - \Phi \mathbf{w}\|_2^2$

- Sensing weights are prone to be *sparse* especially in representation of *ill-posed* data.

Model Construction

- We characterize the *reconstruction error* of an observation $\mathbf{x}_t$ by a Gaussian density with zero mean and *state-dependent* precision matrix R_i. State-dependent *basis vectors* $\Phi_i = [\boldsymbol{\varphi}_{i1}, \cdots, \boldsymbol{\varphi}_{iN}]$ are adopted.

$$p(\mathbf{x}_t \mid \lambda_i) \propto |R_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \Phi_i \mathbf{w}_t)^T R_i (\mathbf{x}_t - \Phi_i \mathbf{w}_t)\right]$$

$$= |R_i|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \sum_{n=1}^{N} \boldsymbol{\varphi}_{in} w_{tn})^T R_i (\mathbf{x}_t - \sum_{n=1}^{N} \boldsymbol{\varphi}_{in} w_{tn})\right]$$

- Point estimates of weight parameters are *unreliable*.

Sparse Bayesian Sensing

- Bayesian sensing aims to yield the "*error bars*" or "distribution estimates" of the true signals.

- *Prior* density of sensing weights is incorporated

$$p(\mathbf{w}_t \mid A_i) = N(\mathbf{w}_t \mid 0, \mathrm{diag}\{\alpha_{in}^{-1}\}) = \prod_{n=1}^{N} N(w_{tn} \mid 0, \alpha_{in}^{-1})$$

  where A_i is a *state-dependent* precision matrix.

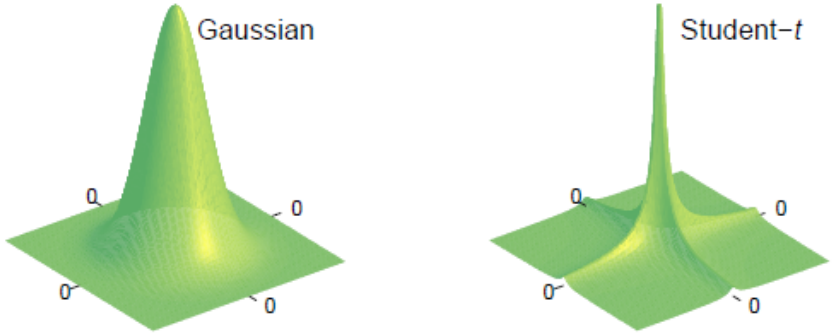
- Precision parameter α_{in} is called *automatic relevance determination* (ARD) which reflects how an observation is relevant to a basis vector.

Automatic Relevance Determination

- If ARD is modeled by a gamma density, the marginal distribution of weights turns out to be a *Student's t distribution* which is a *sparse* prior.

$$p(\mathbf{w}_t \mid a_i, b_i) = \prod_{n=1}^{N} \int_0^{\infty} N(w_{tn} \mid 0, \alpha_{in}^{-1}) \mathrm{Gam}(\alpha_{in} \mid a_i, b_i) d\alpha_{in}$$

$$\propto \prod_{n=1}^{N} (b_i + w_{tn}^2/2)^{-(a_i+1/2)}$$



[Tipping, JMLR 2001]

BS-HMM Parameters

- BS-HMM parameters include
  - basis vectors $\Phi_i = [\varphi_{i1}, \cdots, \varphi_{iN}]$
  - precision matrix of sensing weights A_i
  - precision matrix of reconstruction errors R_i

- *Maximum likelihood (ML) type II* estimation is performed by

$$\lambda^{\mathrm{ML}} = \{\pi_i^{\mathrm{ML}}, a_{ki}^{\mathrm{ML}}, A_i^{\mathrm{ML}}, \Phi_i^{\mathrm{ML}}, R_i^{\mathrm{ML}}\} = \arg \max_{\{\pi_i, a_{ki}, A_i, \Phi_i, R_i\}} p(X \mid \{\pi_i, a_{ki}, A_i, \Phi_i, R_i\})$$

- EM algorithm is applied for parameter estimation.

Solutions to Precision Matrices

- *Precision matrix of sensing weights*

$$\hat{A}_i^{-1} = \Sigma_i + \frac{\sum_t \gamma_t(i)\mathbf{m}_{ti}\mathbf{m}_{ti}^T}{\sum_t \gamma_t(i)} \equiv \Theta(A_i)$$

where $\mathbf{m}_{ti} = \mathbf{w}_t^{\text{MAP}} = \Sigma_i \Phi_i^T R_i \mathbf{x}_t$

- *Precision matrix of reconstruction errors*

$$\hat{R}_i^{-1} = \frac{\sum_t \gamma_t(i)[\Phi_i \Sigma_i \Phi_i^T + (\mathbf{x}_t - \Phi_i \mathbf{m}_{ti})(\mathbf{x}_t - \Phi_i \mathbf{m}_{ti})^T]}{\sum_t \gamma_t(i)} \equiv \Psi(R_i)$$

Solutions to Basis Vectors

- *Basis vectors*

$$\hat{\Phi}_i = \left[\sum_t \gamma_t(i)\mathbf{x}_t\mathbf{m}_{ti}^T\right]\left[\sum_t \gamma_t(i)(\Sigma_i + \mathbf{m}_{ti}\mathbf{m}_{ti}^T)\right]^{-1}$$
$$= \frac{\sum_t \gamma_t(i)\mathbf{x}_t\mathbf{m}_{ti}^T\hat{A}_i}{\sum_t \gamma_t(i)}$$
$$\equiv \Xi(\Phi_i).$$

- A hybrid *dictionary learning* and *basis representation* is performed.
- *Model compression* can be done by discarding the basis vectors corresponding to the largest ARDs.

Experimental Setup

- 1800 hours of Arabic broadcast news training data
- VTL-warped PLP cepstra with LDA and STC
- Speaker adaptation with VTLN, FMLLR and multiple MLLR
- Feature and model space discriminative training with boosted MMI [Povey'08]
- Acoustic models have 5000 states and
  - 800K Gaussians for the baseline
  - 417K Gaussians for the BSHMMs (initialized from 2.8M Gaussians)
- Recognition vocabulary: 795K words
- Language model: 4-gram with 884M n-grams

Model Compression Using ARD

- Acoustic models built with discriminative feature-space transforms [Povey'05]
- Discard 50% of basis vectors corresponding to the largest precision values after training
- Results before and after discriminative training of the parameters:

Model	Training	DEV07	DEV08	DEV09
original	ML type II	12.0%	13.9%	17.4%
compressed	ML type II	12.4%	14.2%	17.6%
original	boosted MMI	10.7%	11.9%	15.0%
compressed	boosted MMI	10.4%	11.7%	14.8%

TABLE OF CONTENTS

Why Bayesian Language Model?

$$p(W) = p(w_1, \cdots, w_T) = \prod_{i=1}^{T} p(w_i \mid w_1, w_2, \cdots, w_{i-1}) \cong \prod_{i-1}^{T} p\left(w_i \mid w_{i-n+1}^{i-1}\right)$$

Domain Mismatch

- Bayesian model adaptation [Masataki et al. 1997]
- Prior is determined from general domain data
- MAP adaptation

Data Sparseness

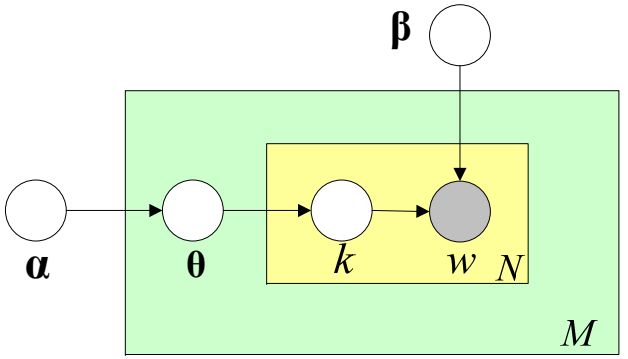
- model smoothing
- backoff method: *n*-grams are estimated by interpolating with (*n*-1)-grams
- hierarchical Pitman-Yor language model [Teh 2006]
- Bayesian nonparametrics

Insufficient Long Distance Information

- topic or class information
- latent Dirichlet allocation
- VB estimation

Latent Dirichlet Allocation [Blei et al. 2003]

- To improve the generalization to unseen documents, a *Dirichlet prior* is used to model the topic distribution.



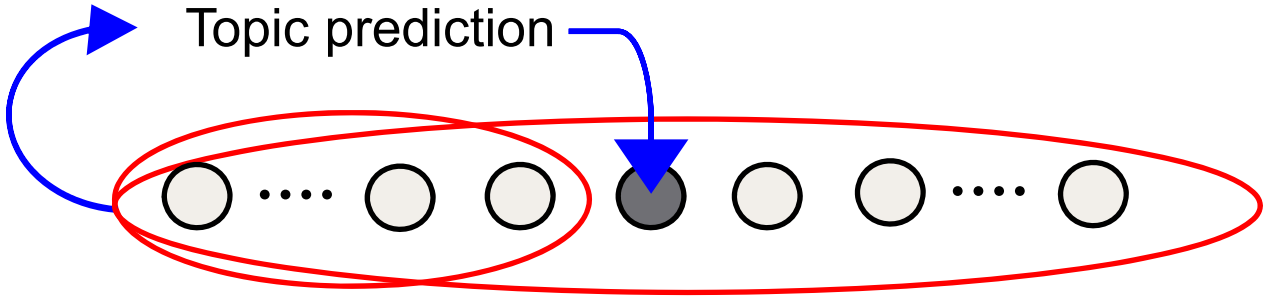
- Document probability

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} \sum_{k_n=1}^{K} p(k_n \mid \boldsymbol{\theta})\, p(w_n \mid k_n, \boldsymbol{\beta})\, d\boldsymbol{\theta}$$

- *Variational Bayesian EM* (*VB-EM*) algorithm is applied for parameter estimation.

LDA LM Adaptation [Tam and Schultz 2005, 2006]

- Estimation of topic probability using VB-EM
  - from *historical words*
  - from transcription of a *whole sentence*



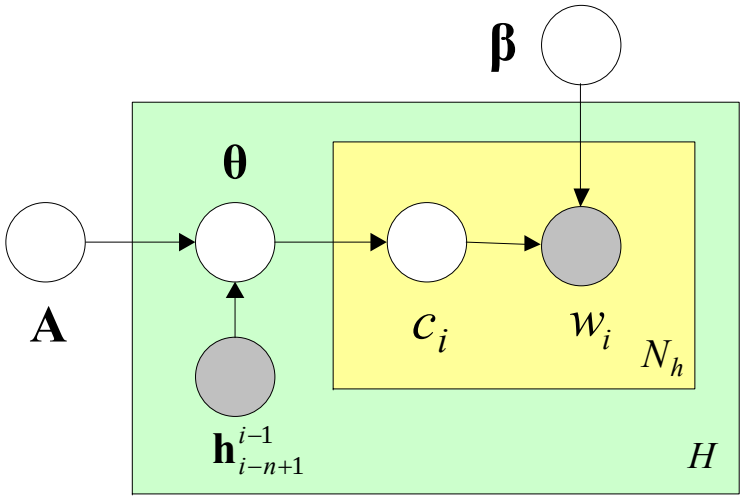
- Interpolation or unigram scaling method were applied for language model adaptation.

$$p(w|h) = \lambda p_{n\text{-gram}}(w|h) + (1-\lambda) p_{\text{LDA}}(w)$$

$$p(w|h) = \frac{p_{\text{LDA}}(w)}{p_{n-\text{gram}}(w)} p_{n\text{-gram}}(w|h)$$

Dirichlet Class Language Model

- Class model is directly built from *n*-gram events.
- DCLM acts as a new *Bayesian class language model* where prior density of the topic variable is involved.

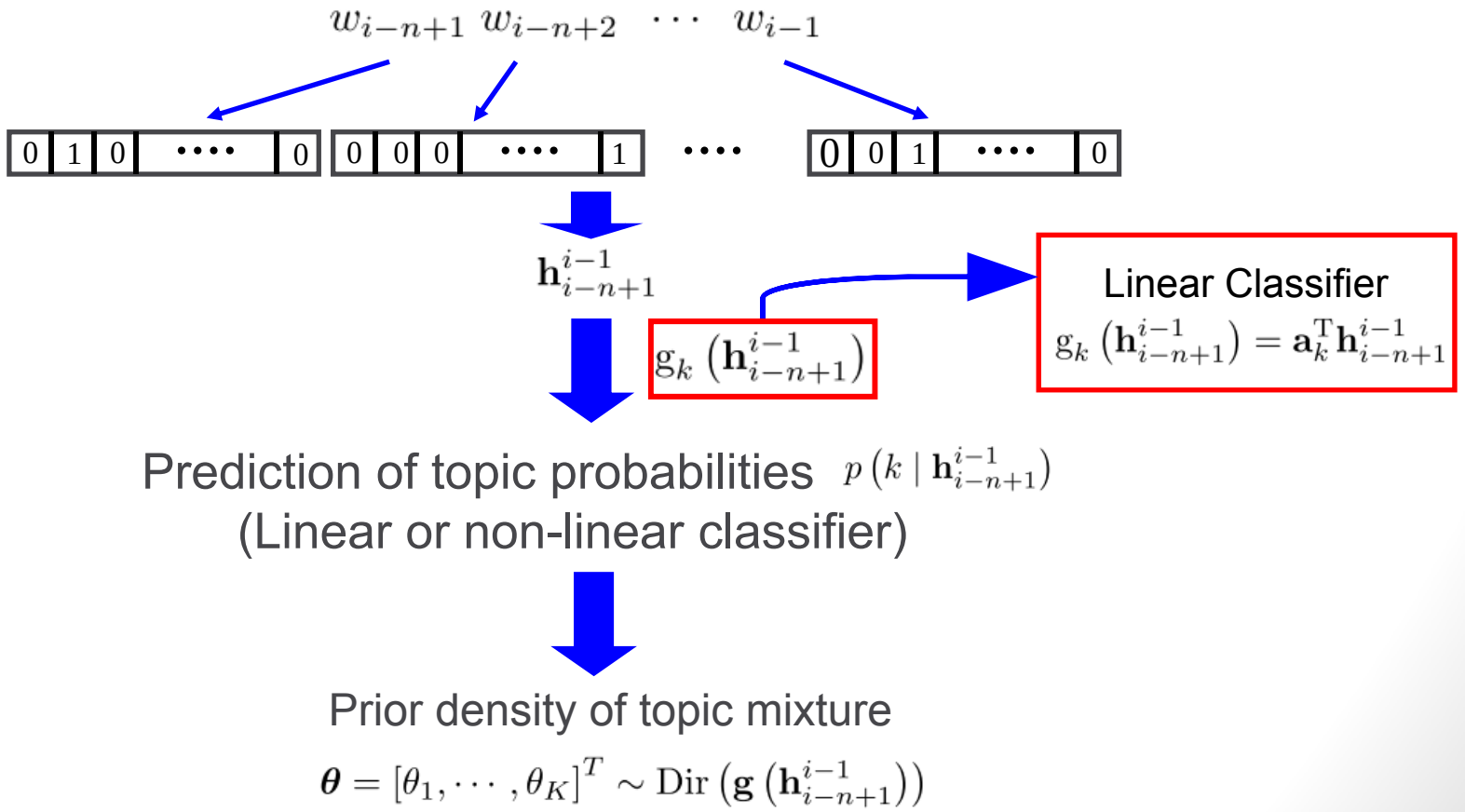


H: no. of histories in the training data

N_h: no. of words following the history

History Representation

- The $n-1$ historical words w_{i-n+1}^{i-1} are represented by an $(n-1)V \times 1$ vector.



Dirichlet Class Language Model

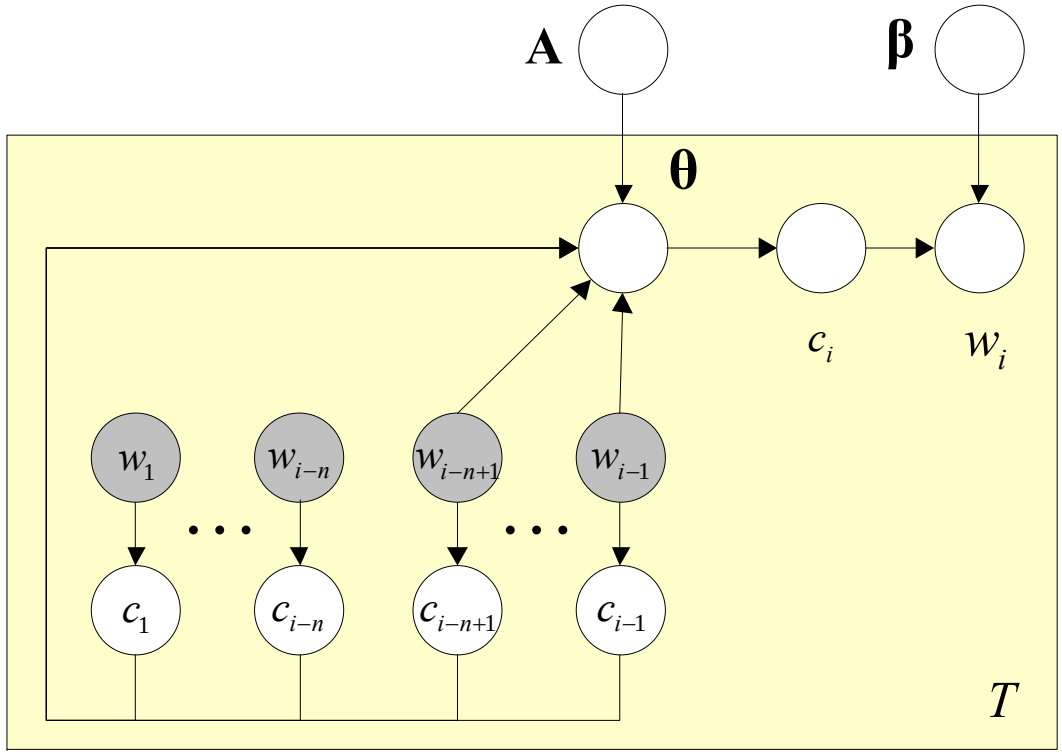
$$p\left(w_i \mid \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta}\right) = \sum_{c_i=1}^{C} p\left(w_i \mid c_i, \boldsymbol{\beta}\right) \int p\left(\boldsymbol{\theta} \mid \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}\right) p\left(c_i \mid \boldsymbol{\theta}\right) d\boldsymbol{\theta}$$

$$= \sum_{c=1}^{C} \beta_{ic} \frac{\mathbf{a}_c^{\mathrm{T}} \mathbf{h}_{i-n+1}^{i-1}}{\sum_{j=1}^{C} \mathbf{a}_j^{\mathrm{T}} \mathbf{h}_{i-n+1}^{i-1}}$$

- DCLM conducts *unsupervised learning* and determined the classes through VB-EM procedure.

Cache DCLM

- *Long-distance topic information* is detected and incorporated into calculation of cache DCLM.



Evaluation Corpus

- Wall Street Journal (WSJ0) was used.
- Nov92 ARPA CSR benchmark test was followed.
  - Acoustic HMMs:
    - 7240 training utterances
    - 4004 development utterances
    - 333 test utterances
  - Language models:
    - 87~89 WSJ text corpus with 38M words
    - 20K non-verbalized punctuation, closed vocabulary
  - HTK was used for model training and LVCSR decoding.

Frequent Words of Latent Topics/Classes

	Topic /Class	Frequent Words in Latent Topic
LDA	*Family*	*toy, kids, theater, event, season, shoe, teen, children's, plays, films, sport, magazines, Christmas, bowling, husband, anniversary, girls, festival, couple, parents, wife, friends*
	Election	*candidates, race, voters, challenger, democrat, state's, selection, county, front, delegates, elections, reverend, republicans, polls, conventions, label, politician, ballots*
	War	*troops, killed, Iraqi, attack, ship, violence, fighting, soldiers, mines, Iranian, independence, marines, revolution, died, nation, protect, armed, democracy, violent, commander*
DCLM	*Quantity*	*five, seven, two, eight, cents, six, one, nine, four, three, zero, million, point, percent, years, megabyte, minutes, milligrams, bushels, miles, marks, pounds, yen, dollars*
	Business	*exchange, prices, futures, index, market, sales, revenue, earnings, trading, plans, development, business, funds, organization, traders, ownership, holdings, investment*
	In+	*addition, the, fact, American, October, recent, contrast, Europe, June, Tokyo, July, March, turn, other, Washington, order, Chicago, case, China, general, my, which*

WER (%) vs Training Data Size

	Size of Training Data			
	6M	**12M**	**18M**	**38M**
Baseline LM	**39.19 (-)**	**21.25 (-)**	**15.79 (-)**	**12.89 (-)**
Cache LM	**38.13 (*2.7*)**	**20.92 (*1.6*)**	**15.56 (*1.5*)**	**12.74 (*1.2*)**
NNLM	**35.51 (*9.4*)**	**19.55 (*8.0*)**	**14.99 (*5.1*)**	**12.40 (*3.8*)**
Class-based LM	**35.49 (*9.4*)**	**19.7 (*7.3*)**	**15.03 (*4.8*)**	**12.42 (*3.6*)**
LDA LM	**35.86 (*8.5*)**	**19.67 (*7.4*)**	**14.73 (*6.7*)**	**12.16 (*5.7*)**
DCLM (*C*=200)	**35.91 (*8.4*)**	**19.59 (*7.8*)**	**14.61 (*7.5*)**	**11.96 (*7.2*)**
Cache DCLM (*C*=200)	**34.15 (*12.9*)**	**19.32 (*9.1*)**	**14.47 (*8.4*)**	**11.91 (*7.6*)**
DCLM (*C*=500)	**35.21 (*10.2*)**	**19.21 (*9.6*)**	**14.29 (9.5)**	**11.71 (*9.2*)**
Cache DCLM (*C*=500)	**33.92 (*13.4*)**	**19.02 (*10.5*)**	**14.17 (10.3)**	**11.63 (*9.8*)**

WER (%) vs Vocabulary Size

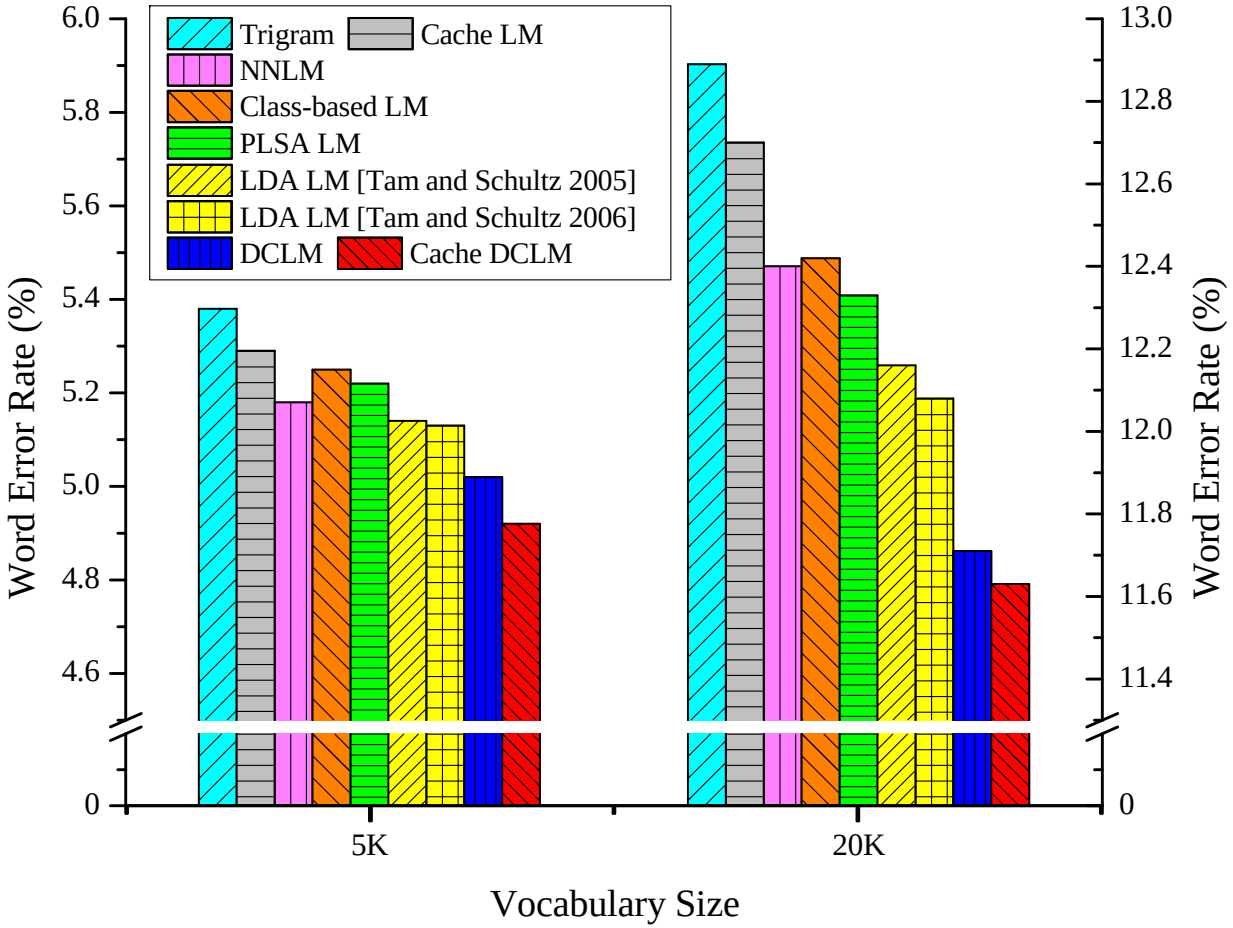


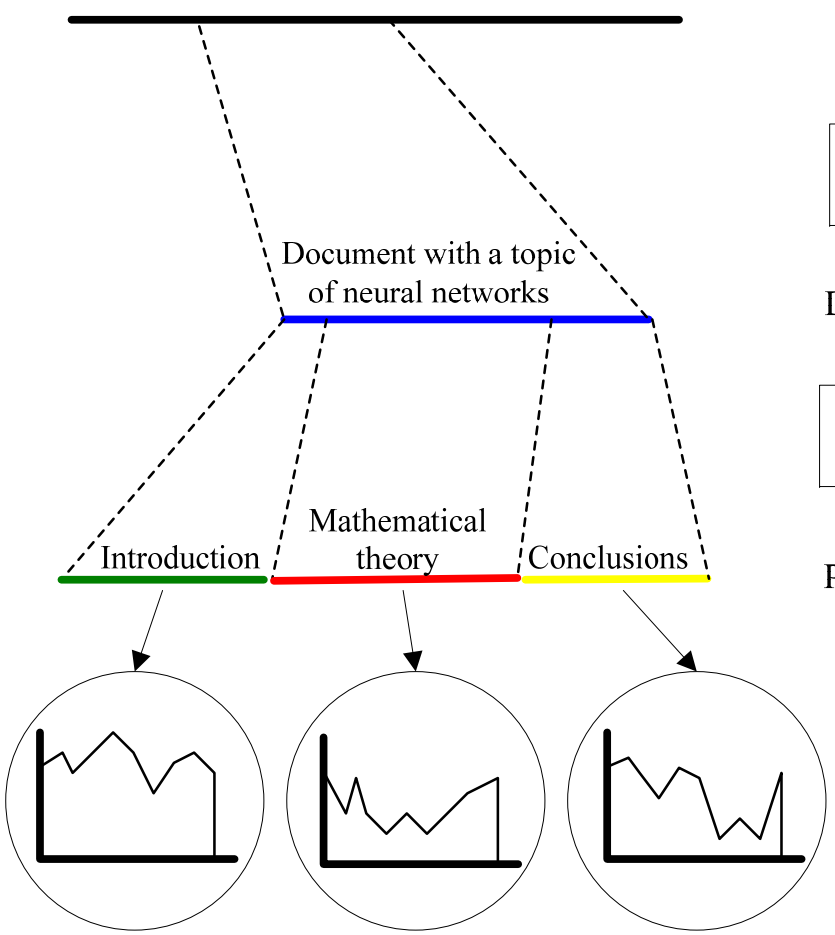
TABLE OF CONTENTS

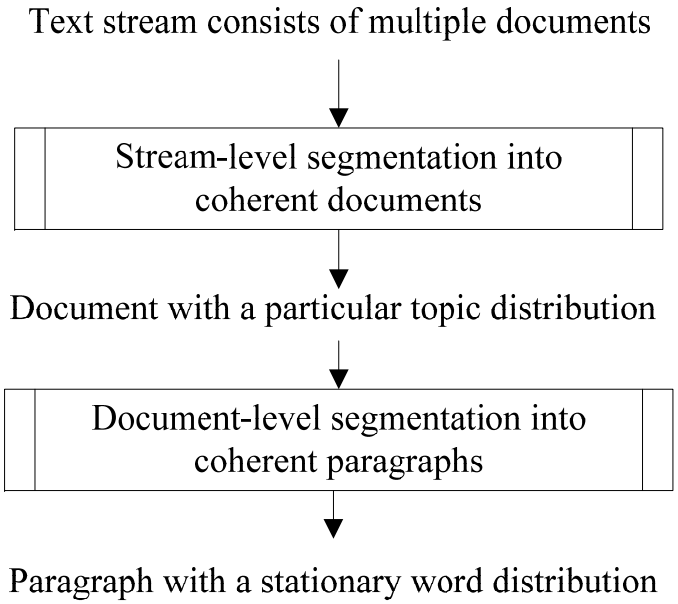
Why Bayesian Text Segmentation?

- Why *text segmentation*?
  - information retrieval
    - readers are only interested in specific parts of a document
  - document/story summarization
    - document may contain multiple topics in different segments
  - speech recognition
    - variations of word distributions exist in a spoken document

- Why *Bayesian*?
  - segment boundaries are ambiguous
  - number of segments is unknown
  - segmentation with beliefs
  - topic model provides beliefs for topic tracking and segmentation
  - VB estimation

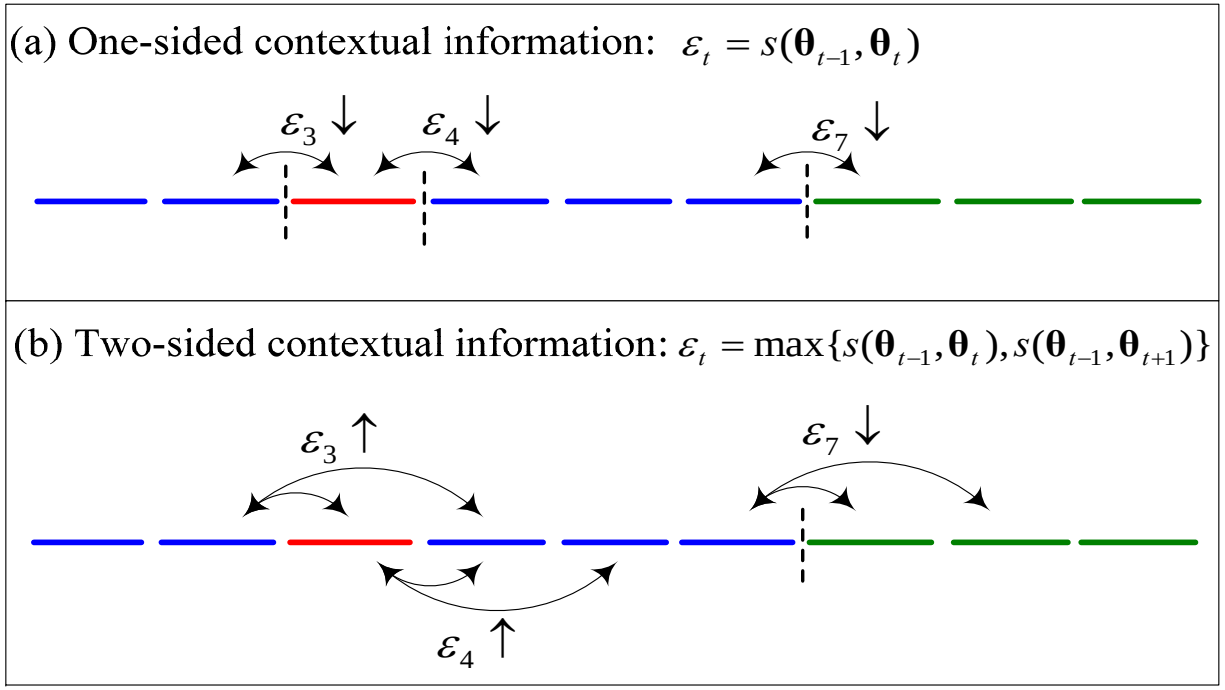
Topic-Based Hierarchical Segmentation



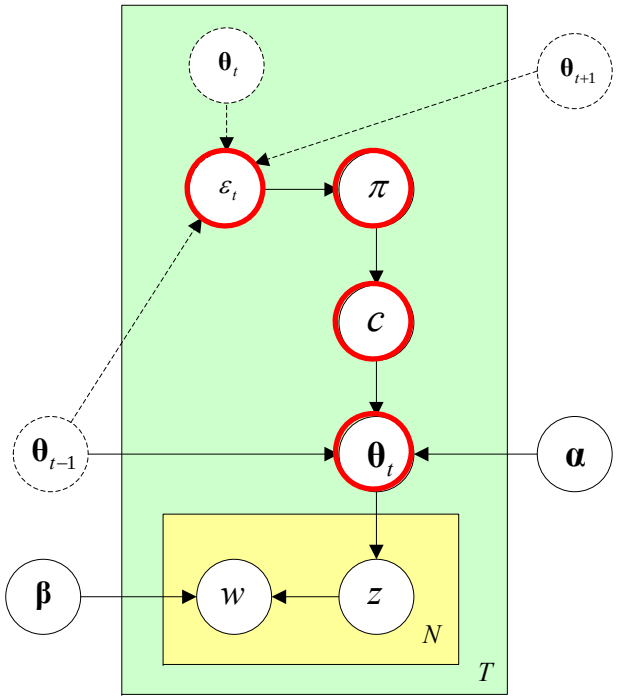


Different Contextual Information

- *Two-sided contextual information* is used to alleviate segmentation errors of non-topic sentences.



Model Construction



$$s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t) = \frac{\boldsymbol{\theta}_{t-1} \cdot \boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_{t-1}\| \|\boldsymbol{\theta}_t\|}$$

$$\varepsilon_t = \max\{s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t), s(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1})\}.$$

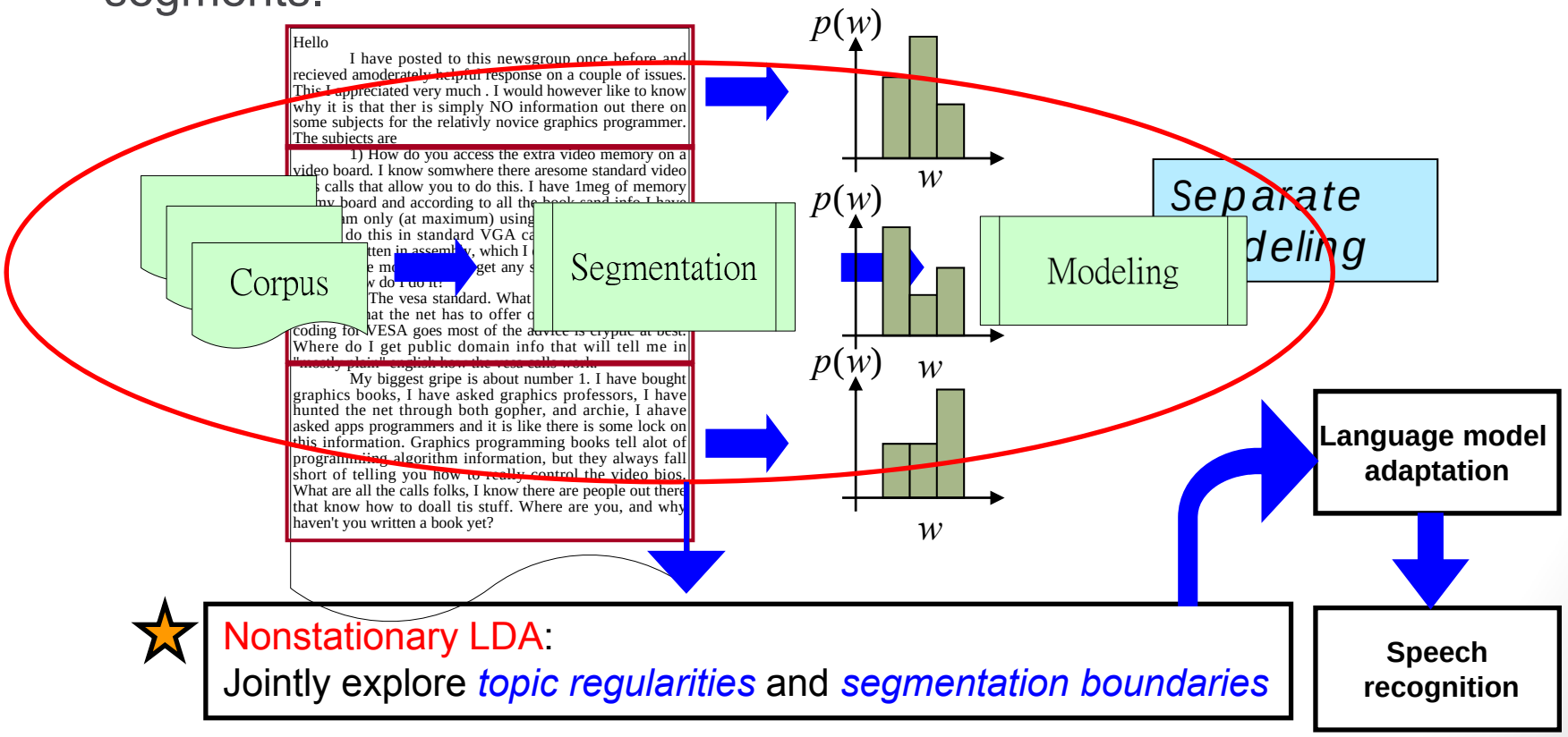
$$\pi \sim p(\pi \mid \varepsilon_t) = \frac{1}{\Gamma(\varepsilon_t)\Gamma(1-\varepsilon_t)} \pi^{-\varepsilon_t} (1-\pi)^{\varepsilon_t - 1}$$

$$c \sim p(c \mid \pi) = \begin{cases} \pi & \text{if } c = 1 \\ 1-\pi & \text{if } c = 0 \end{cases}$$

$$\boldsymbol{\theta}_t \sim p(\boldsymbol{\theta}_t \mid c, \boldsymbol{\alpha}, \boldsymbol{\theta}_{t-1}) = \begin{cases} \dfrac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1} & \text{if } c = 1 \\ \delta(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}) & \text{if } c = 0 \end{cases}$$

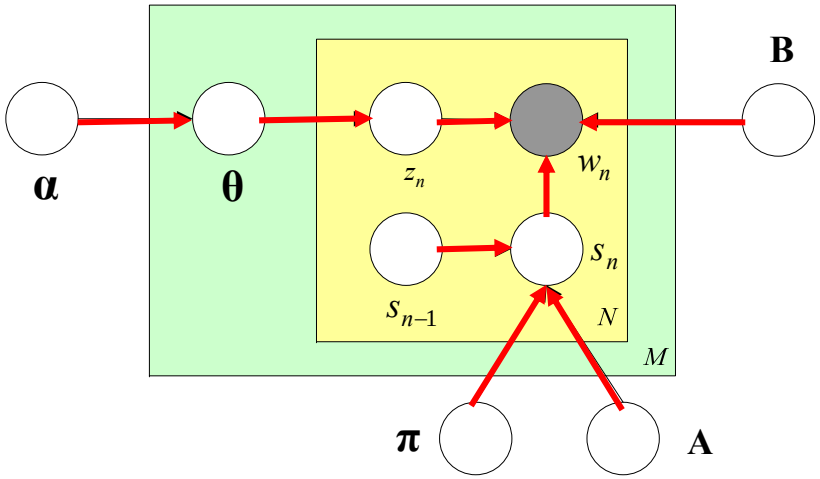
Topic-Based Segmentation

- *Words* in a document are *non-stationary*.
  - style or distribution of the same words is varied in different segments.



Model Construction

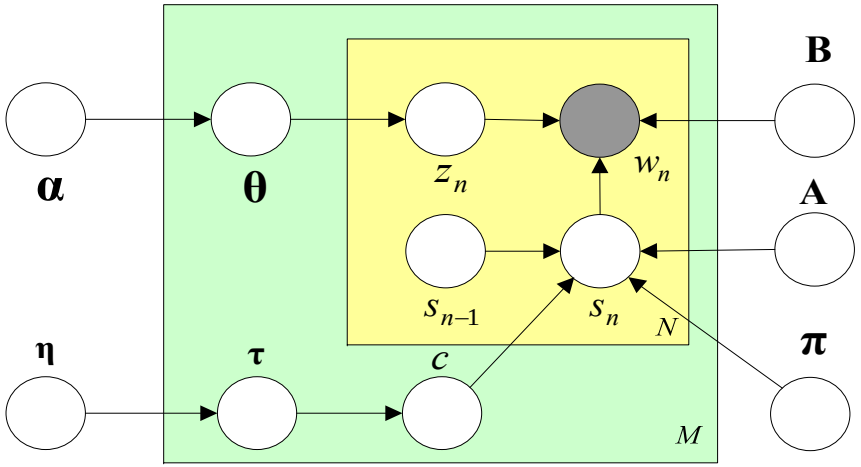
- Generation process of a document



$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\pi}, \mathbf{A})$$
$$= \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \sum_{s} \prod_{n=1}^{N} \sum_{z_n=1}^{K} p(w_n \mid z_n, s_n, \mathbf{B})\, p(z_n \mid \boldsymbol{\theta})\, p(s_n \mid s_{n-1}, \boldsymbol{\pi}, \mathbf{A})\, d\boldsymbol{\theta}$$

Adaptive Segmentation Model

- Different documents have different numbers of segments.
- A *style variable* c is introduced to indicate the number of stylistic changes in a document.



τ : style mixture

η: Dirichlet prior

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\pi}, \mathbf{A}, \mathbf{B}) = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\tau}} \sum_{c,\mathrm{s},\mathrm{z}} p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})\, p(\mathbf{z} \mid \boldsymbol{\theta})$$
$$\times p(\boldsymbol{\tau} \mid \boldsymbol{\eta})\, p(c \mid \boldsymbol{\tau})\, p(\mathbf{s} \mid \boldsymbol{\pi}_c, \mathbf{A}_c)\, p(\mathbf{w} \mid \mathbf{z}, \mathbf{s}, \mathbf{B})\, d\boldsymbol{\tau} d\boldsymbol{\theta}$$

TABLE OF CONTENTS

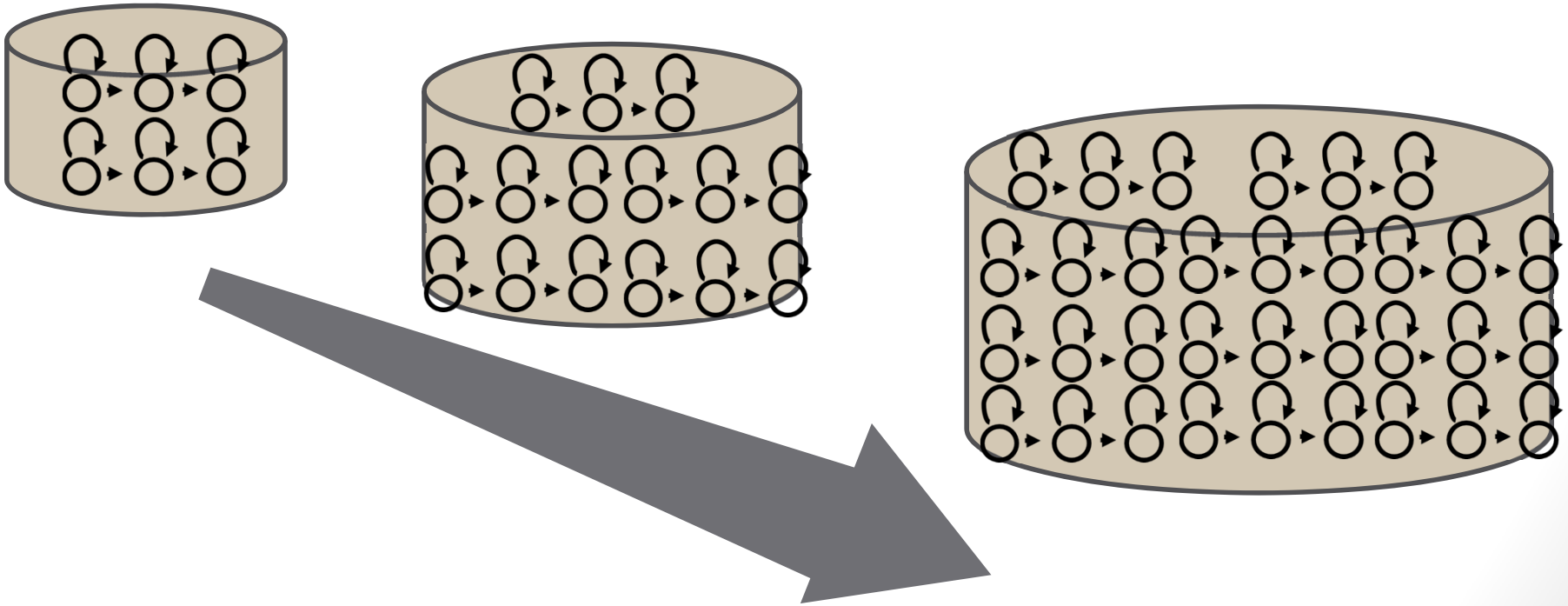
Why Bayesian Nonparametrics?

- Why *structural learning*?
  - model selection issue is tackled
  - speech and text data are complicated in high-dimensional space
  - latent information is rich and exists in a hierarchical way
  - structure of latent variables provides crucial information

- Why *Bayesian nonparametrics*?
  - flexible, scalable & realistic
  - unbounded number of topics
  - Gibbs sampling is performed
  - sentence selection for document summarization
  - hierarchical topic model for sentences
  - relations between word topics and sentence topics are explored

Scalable Modeling

- ***Bayesian nonparametrics for big speech data***



  - Incrementally updating model parameters and sizes as data increase

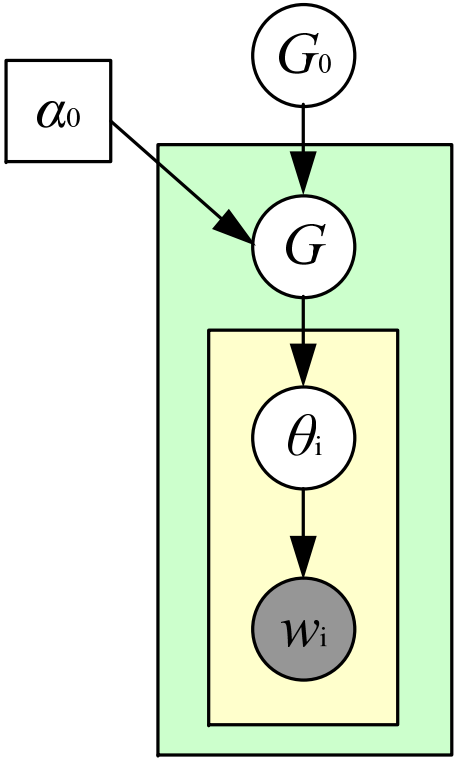
Dirichlet Process [Ferguson 1973]

- *DP* is an extension of Dirichlet distribution which defines a *distribution over distributions*.

$$G \sim \mathrm{DP}(\alpha_0, G_0)$$

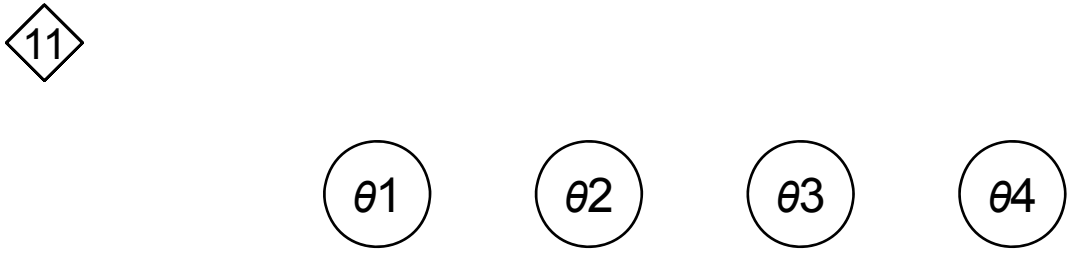
G_0 : base measure

α_0 : concentration parameter

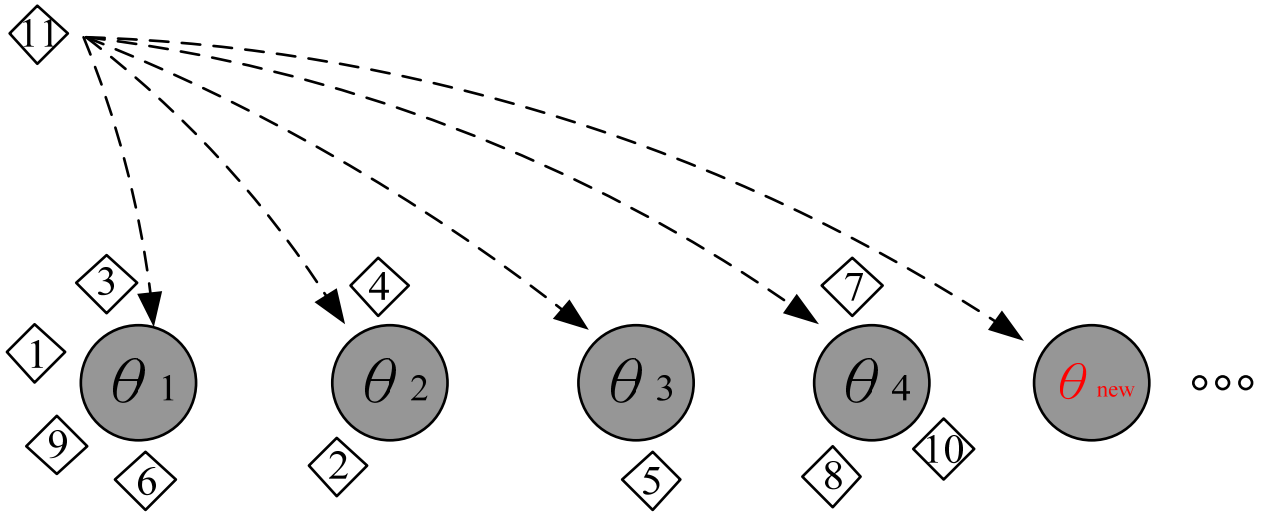


Chinese Restaurant Process [Aldous 1985]

- *CRP* is used to realize DP.
- Imagine a Chinese restaurant
  - infinite tables and infinite capacity
  - a sequence of *N* customers arrives
  - each customer chooses one table



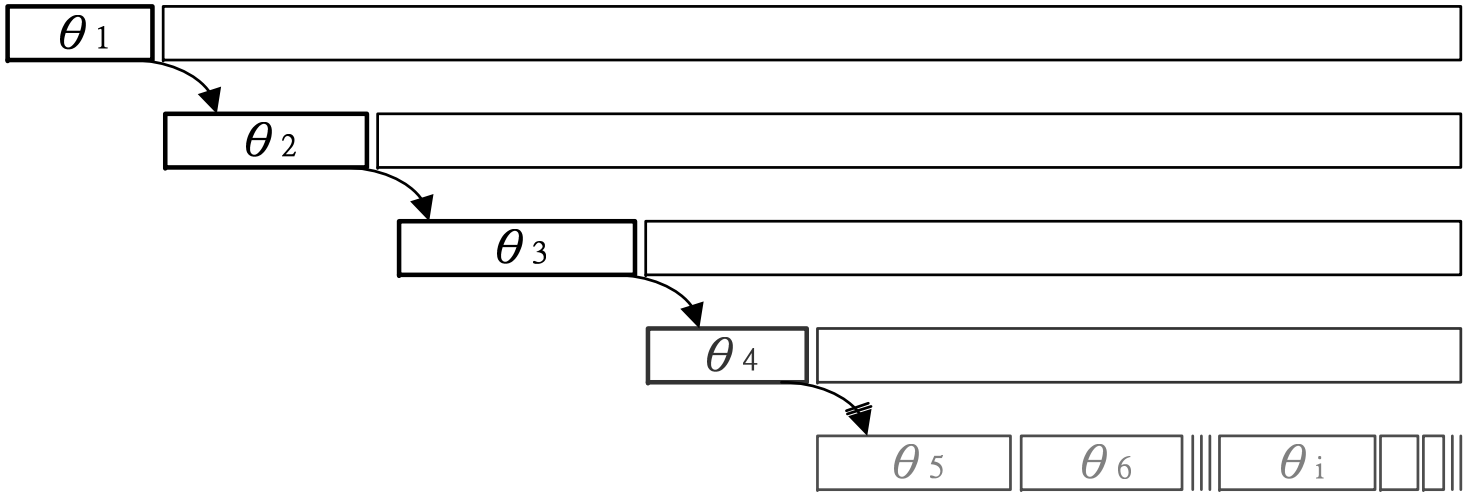
CRP Probability Model



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

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

Stick-Breaking Process [Pitman 2002]



SBP probabilistic model:

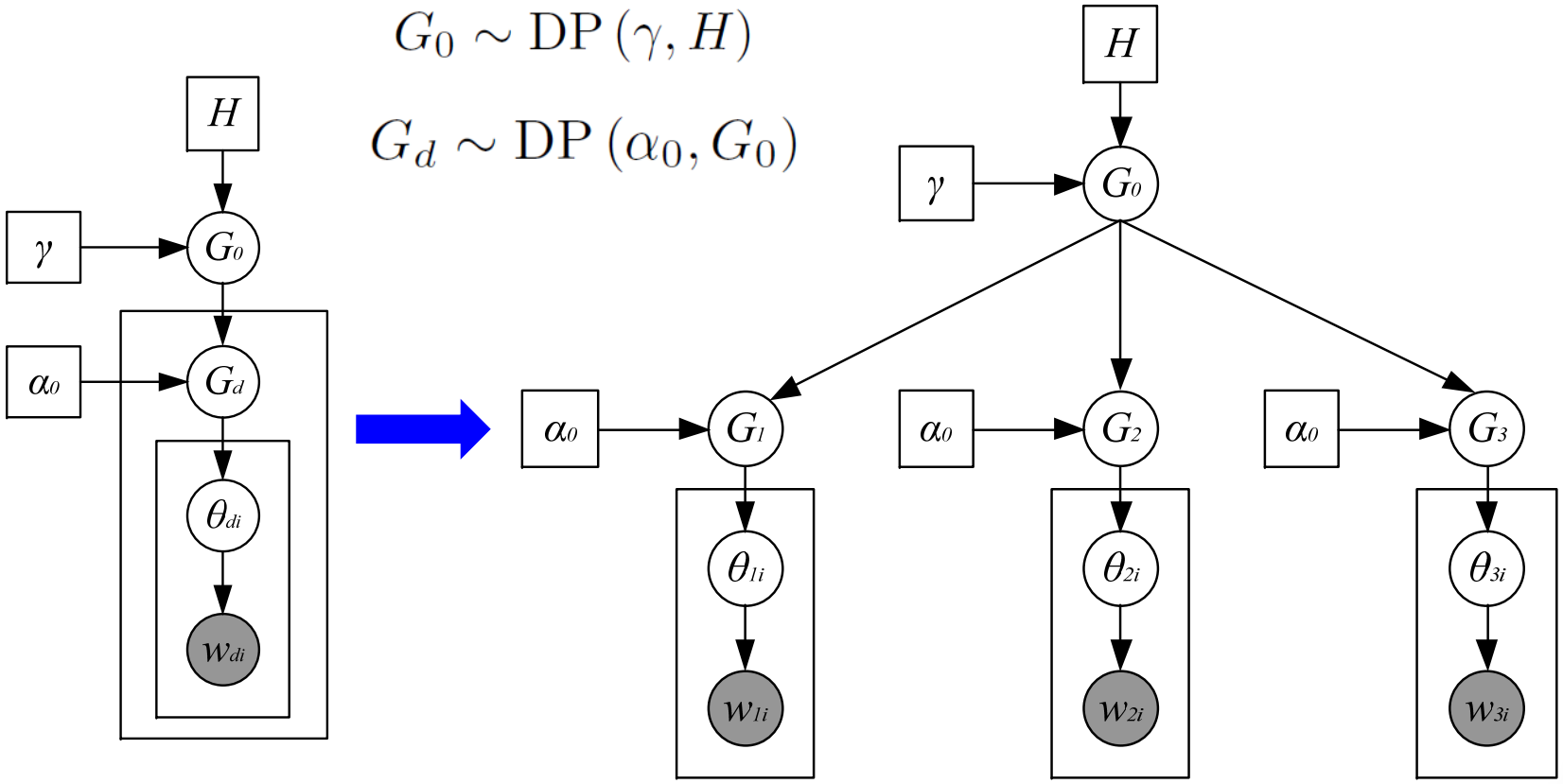
$$\beta_k \mid \alpha_0, G_0 \sim G_0 \qquad V_k \mid \alpha_0, G_0 \sim \mathrm{Beta}\,(1, \alpha_0)$$

$$\theta_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \qquad G = \sum_{k=1}^{\infty} \theta_k \delta_{\beta_k}$$

Hierarchical Dirichlet Process [Teh 2006]

$$G_0 \sim \mathrm{DP}(\gamma, H)$$

$$G_d \sim \mathrm{DP}(\alpha_0, G_0)$$

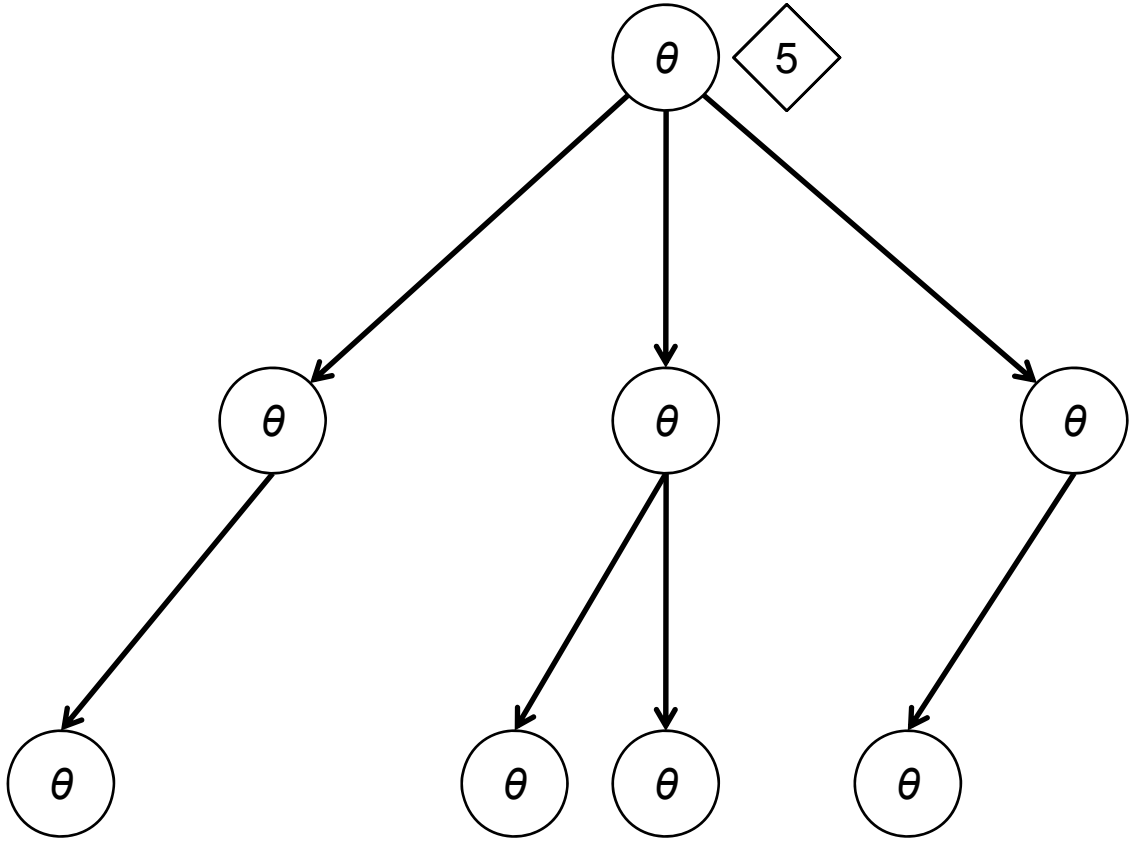


Hierarchical LDA & nCRP

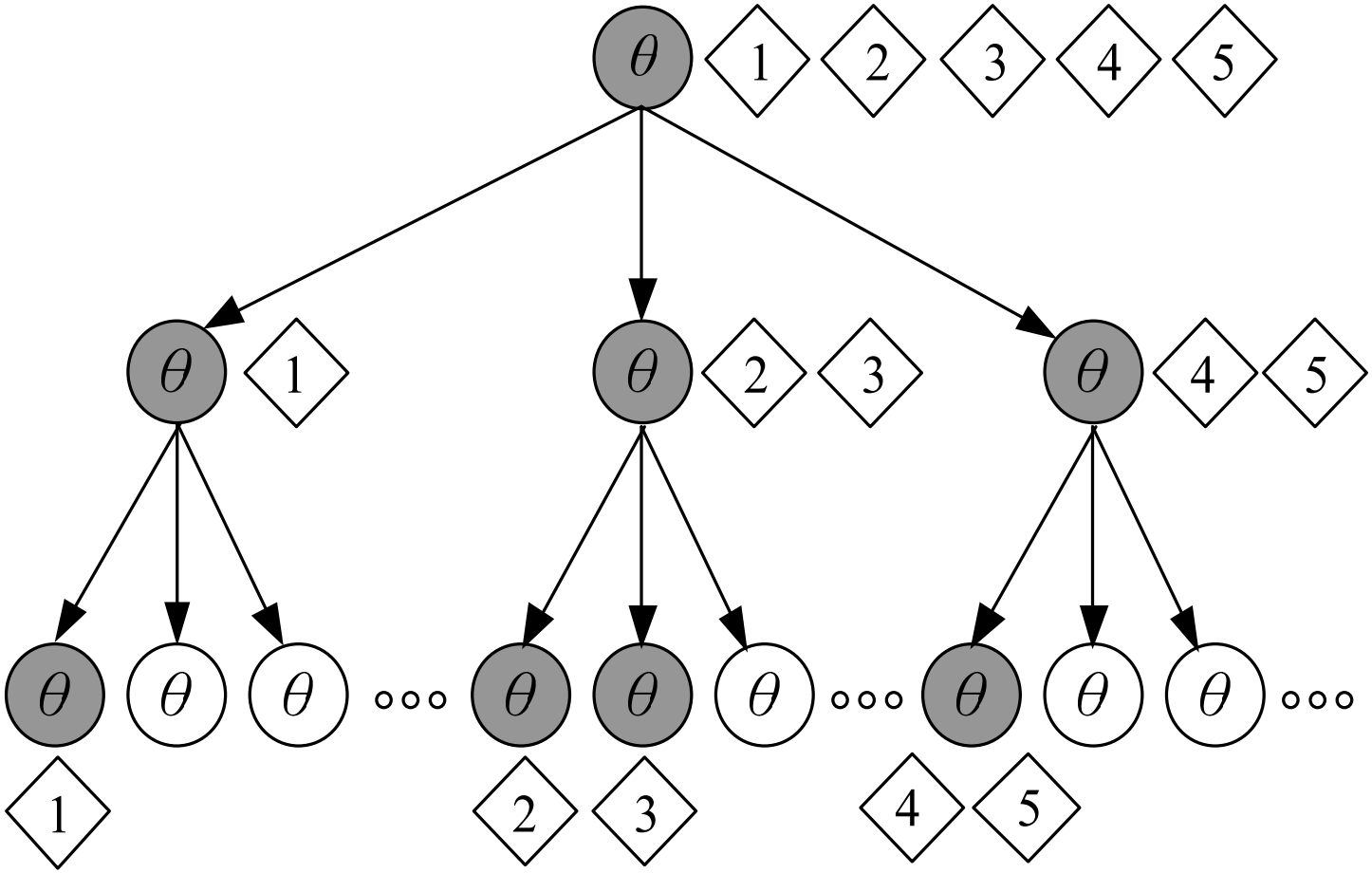
- The hLDA is a hierarchical topic model which uses *nCRP* to select paths and *SBP* to model the infinite deep path for LDA.
- Scenario of nCRP is introduced for document model based on *single paths*.
  - a tourist arranges restaurants for his culinary vacation
  - he enjoys dinner at the restaurant which offers infinite tables at night
  - after dinner he got a pass to the next restaurant and he went
  - he got a pass again and he repeats this procedure forever

Nested Chinese Restaurant Process

[Blei et al. 2004, 2010]



Tree Model of nCRP

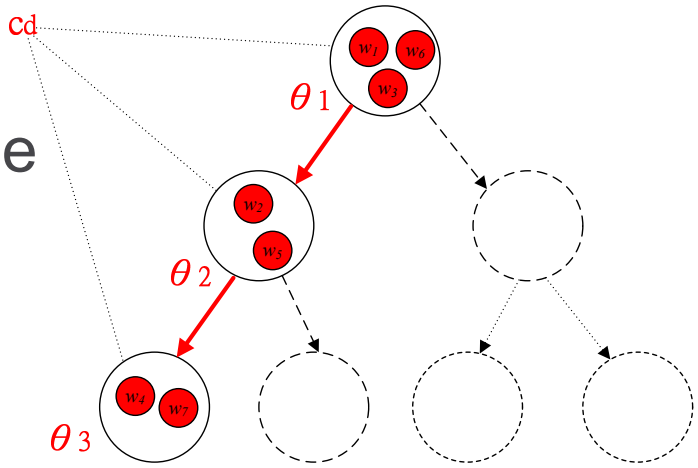


Model Construction for hLDA

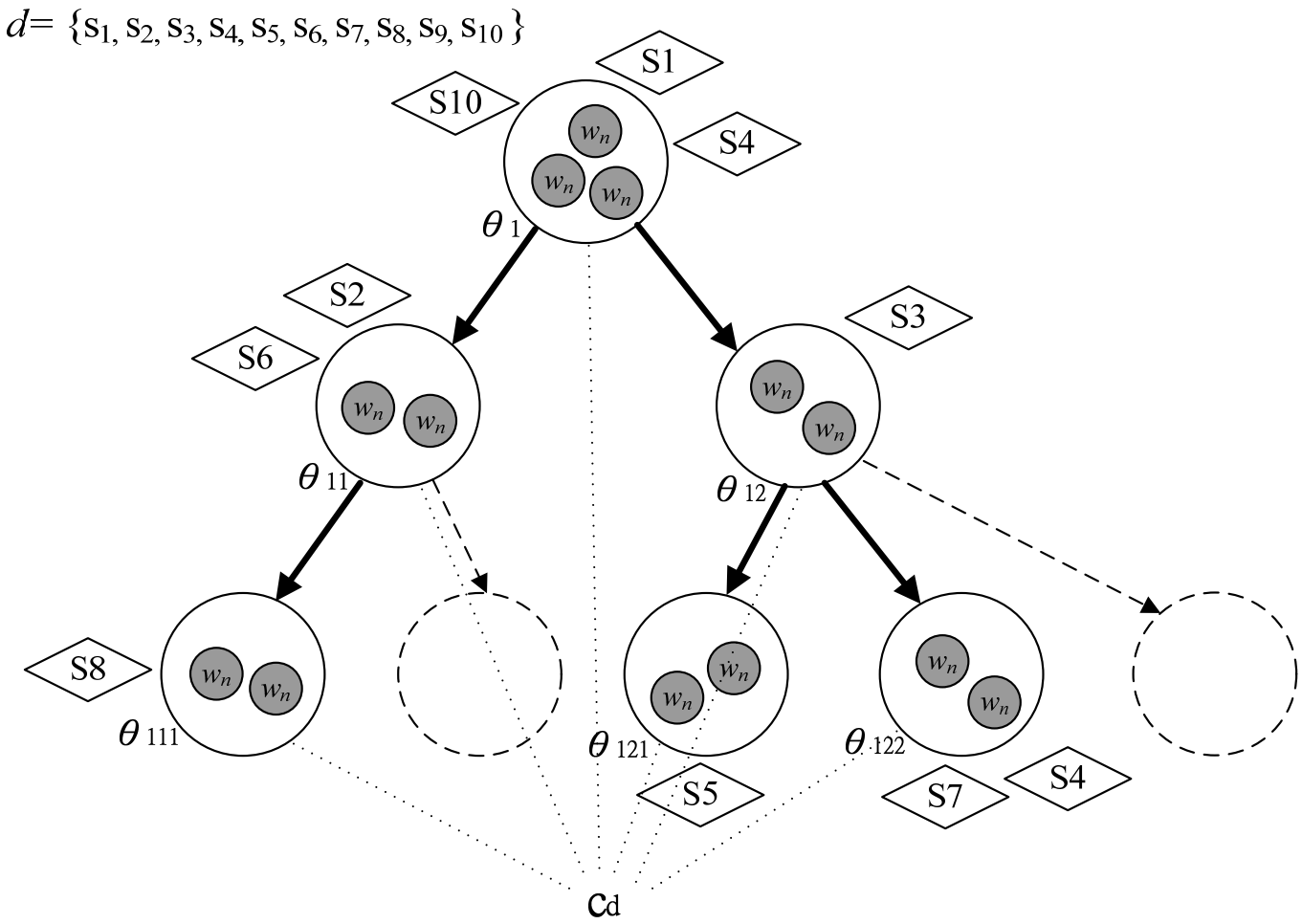
- For each table $k \in T$ in the infinite tree
  - Draw a topic $\beta_k \sim \text{Dirichlet}(\eta)$
- For each document $d \in \{1, 2, \cdots, D\}$
  - Draw $\mathbf{c}_d \sim \text{nCRP}(\gamma)$
  - Draw a distribution over levels in the tree,

$$\theta_d \mid \{m, \pi\} \sim \text{GEM}(m, \pi)$$

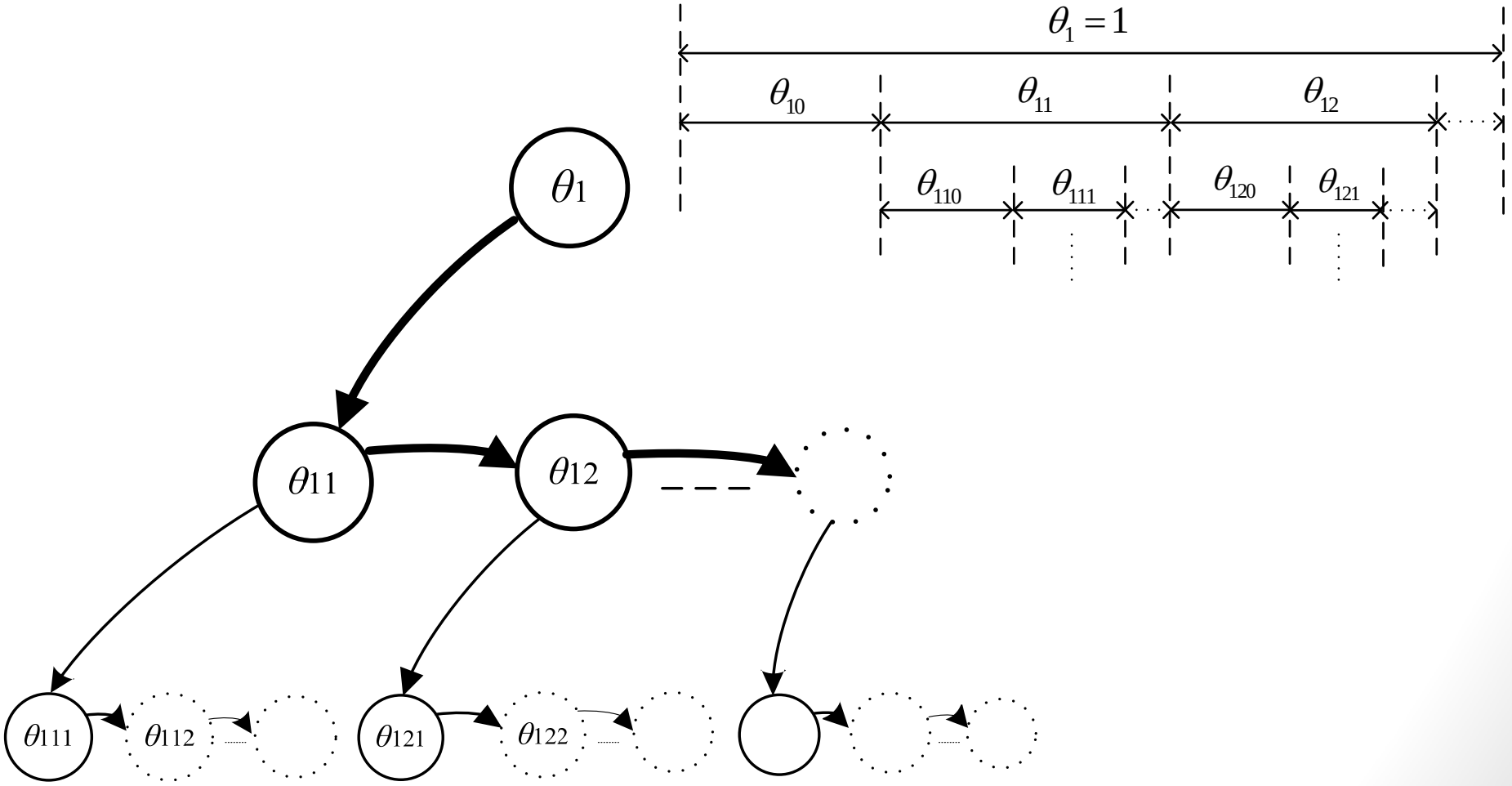
  - For each word
    - Choose level $z_{d,n} \mid \theta_d \sim \text{Discrete}(\theta_d)$
    - Choose word $w_{d,n} \mid \{z_{d,n}, \mathbf{c}_d, \beta\} \sim \text{Discrete}(\beta_{\mathbf{c}_d}[z_{d,n}])$



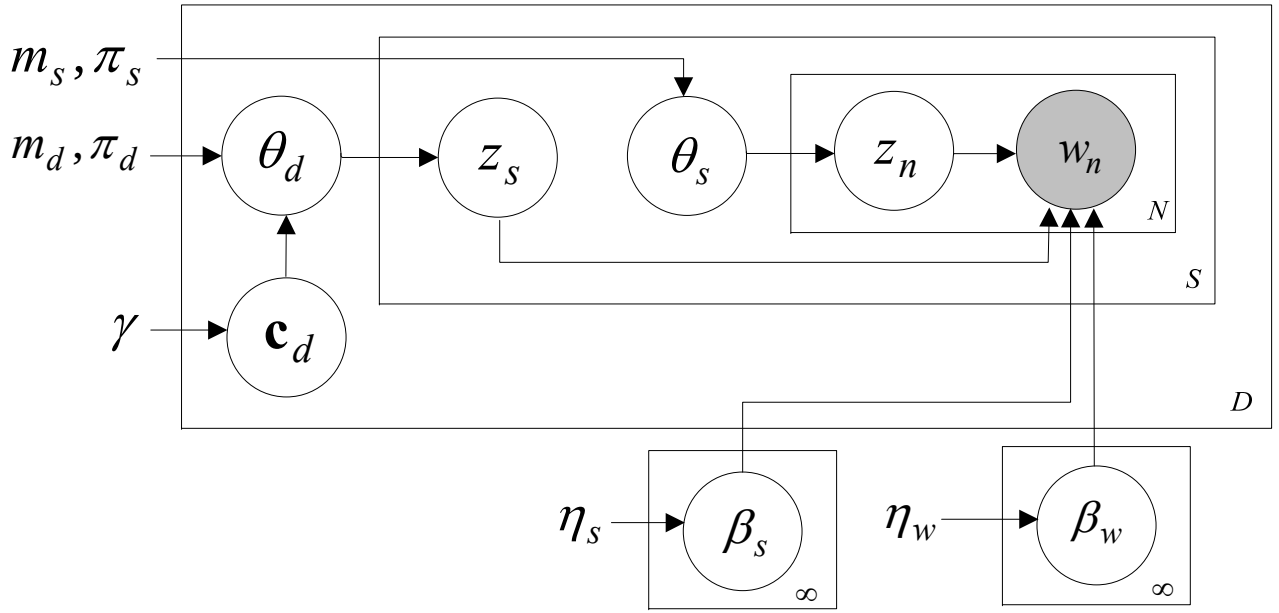
Hierarchical Topic and Sentence Model (HTSM)



Hierarchical Stick-Breaking Process (hSBP)



Graphical Model



- Approximate inference using *Gibbs sampling* is developed to infer model parameters.

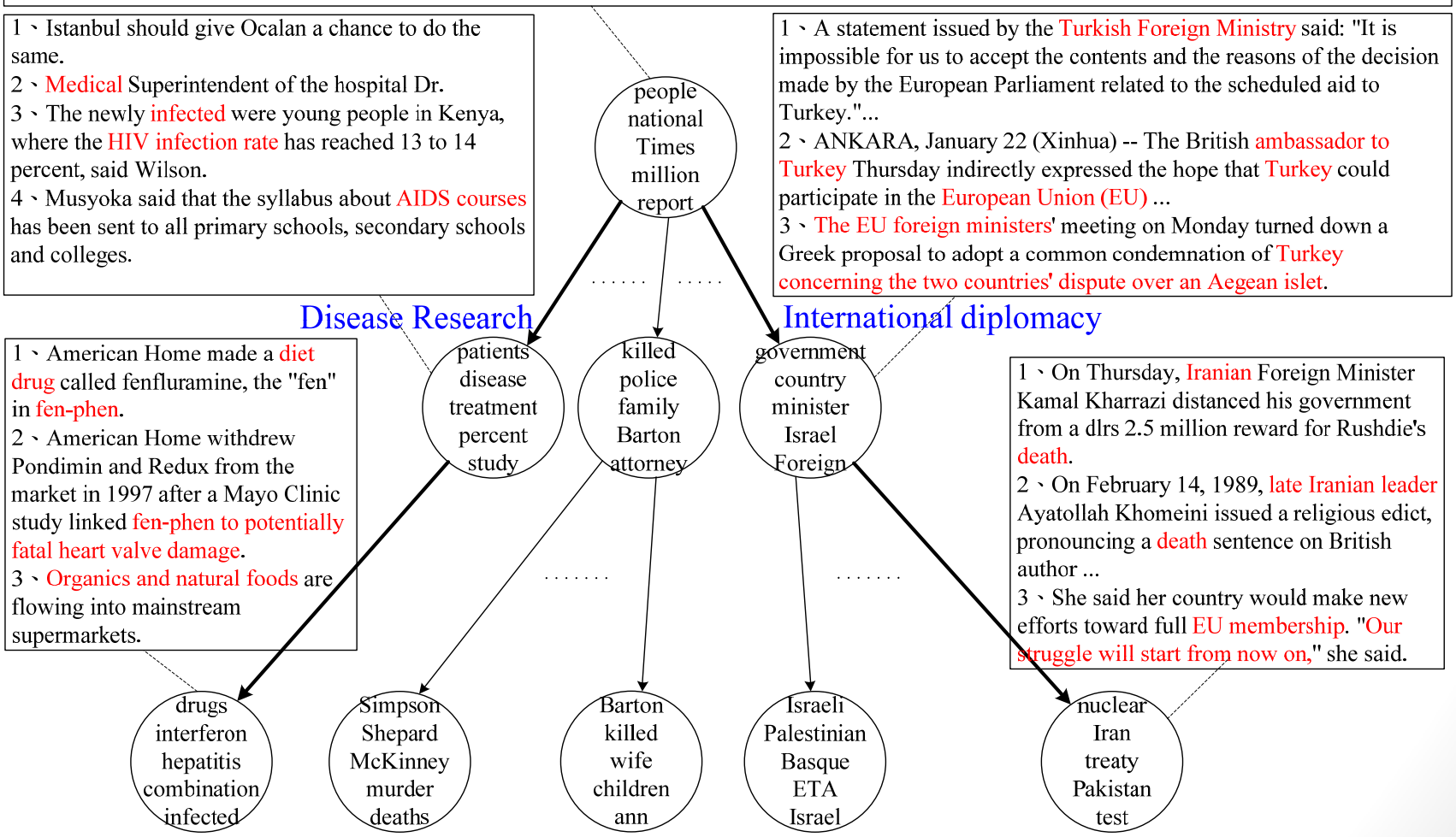
Experimental Setup

- *DUC 2007 corpus* (http://duc.nist.gov/): 45 documents. Each document contained news articles from 45 topics. There were 25-50 news articles in a topic.
- Randomly choose 40 topics as training set and 5 topics as test set.
- Vocabulary size :18395.
- Initial values were specified for three-layer parameters

$$\eta_w = \eta_s = [0.05, 0.025, 0.0125]^T$$

$$m_d = m_s = 0.035, \pi_d = \pi_s = 100, \gamma = 0.5$$

1、HANOI, October 15 (Xinhua) -- Drought in Vietnam has caused a serious water shortage affecting about 3 million people in the recent months, Vietnam's English newspaper The Saigon Times Daily reported Thursday.
2、BANGKOK, November 10 (Xinhua) -- Thailand is considering using the European single currency, the euro, in the country's foreign reserves, the Nation reported Tuesday.
3、Turkey is a European country," State Department spokesman Nicholas Burns told reporters. "We strongly believe that the European Union should allow the possibility of Turkish membership in the future." The 15-nation EU rejected a membership of Turkey last week.



Results on Document Summarization

- *Recall*, *precision*, and *F-measure* were compared.
- *ROUGE-N* measures the matched *n*-gram between reference and automatic summaries

	ROUGE-1			ROUGE-2			ROUGE-L		
	P	R	F	P	R	F	P	R	F
VSM	0.3289	0.2961	0.3118	0.0496	0.0435	0.0465	0.2991	0.2737	0.2857
LDA	0.3839	0.3387	0.3598	0.0767	0.0578	0.0670	0.3387	0.2993	0.3106
HTSM	**0.4100**	**0.3869**	**0.3976**	**0.0936**	**0.0888**	**0.0911**	**0.3695**	**0.3489**	**0.3585**

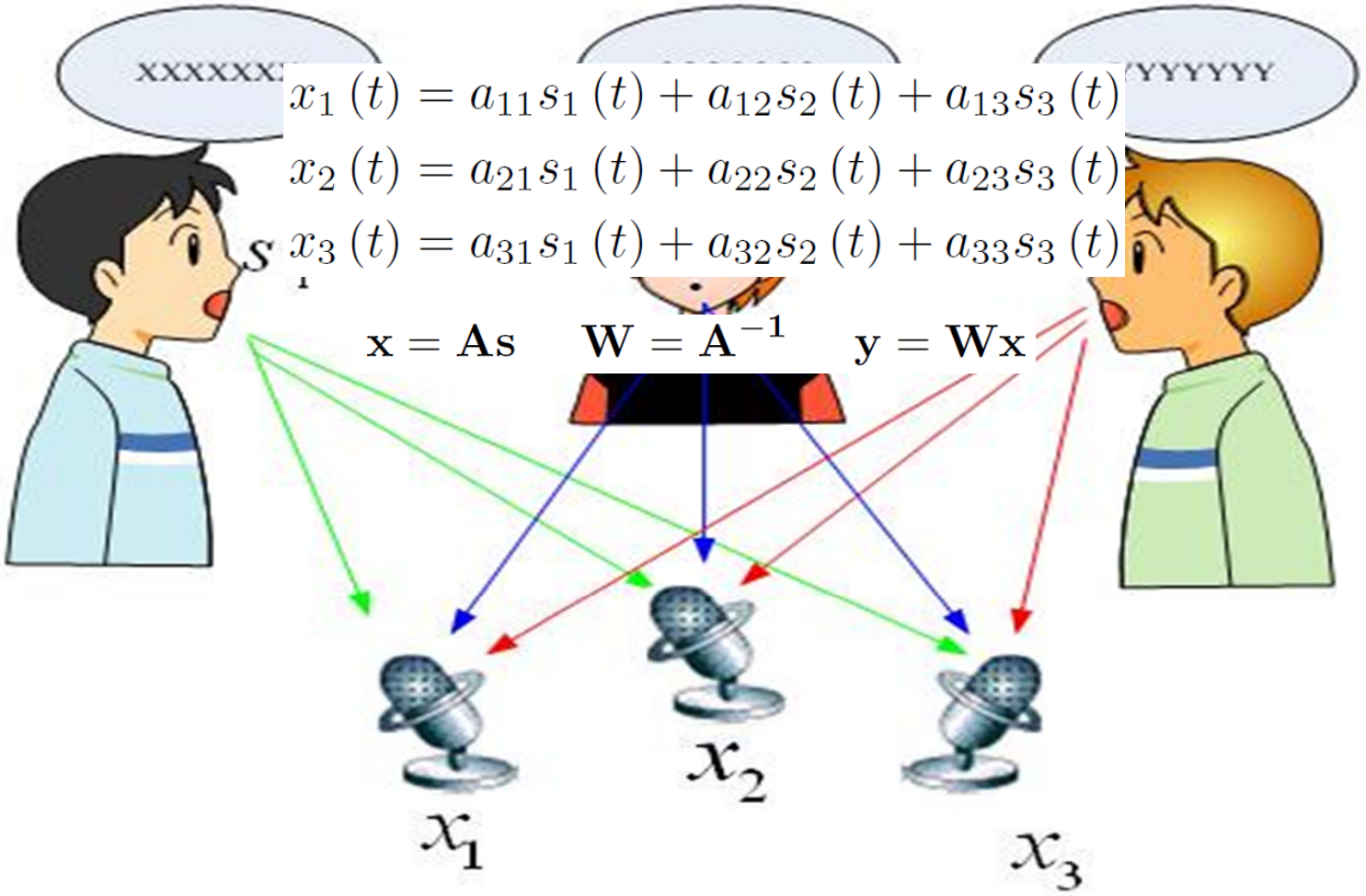
TABLE OF CONTENTS

Why Bayesian Source Separation?

- Real-world *blind source separation*
  - separation of a set of signals from a set of mixed signals without the aid of source signals and mixing process
  - number of sources is unknown
  - general solution to unsupervised structural learning
  - BSS is a dynamic time-varying system
  - mixing process is nonstationary

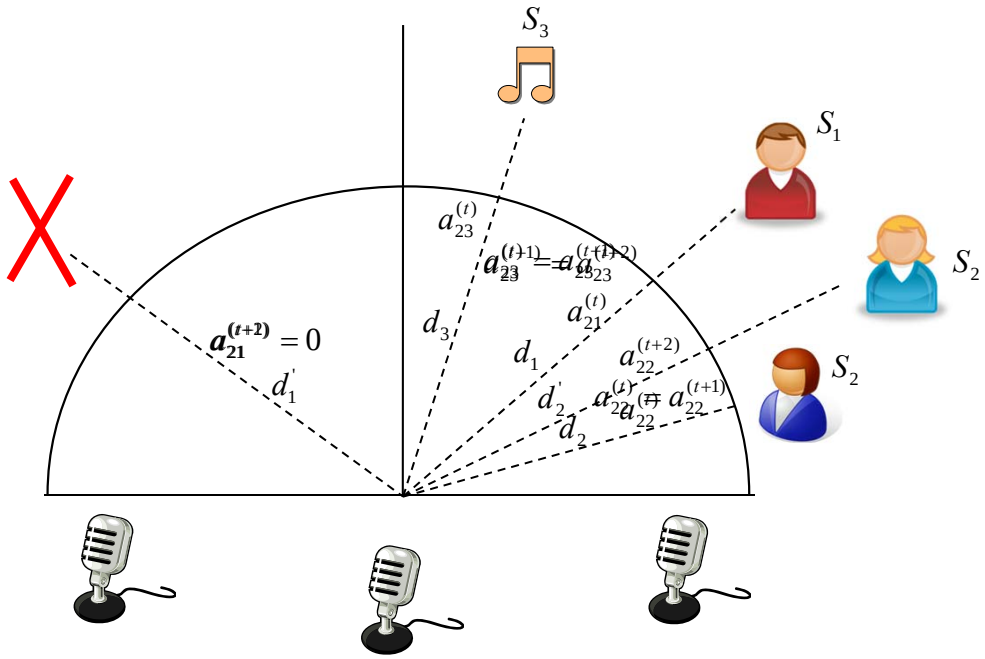
- Why *Bayesian*?
  - Bayesian method using ARD can determine the number of sources
  - recursive Bayesian for online tracking of nonstationary conditions
  - Gaussian process provides a nonparametric solution to temporal structure of time-varying mixing system. GP is a Bayesian method.
  - VB estimation is performed

Blind Source Separation



Nonstationary Mixing Systems

- Time-varying mixing matrix
- Source signals may abruptly appear or disappear

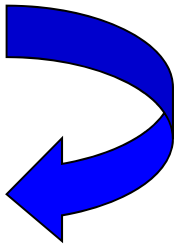


Nonstationary Bayesian (NB) Learning

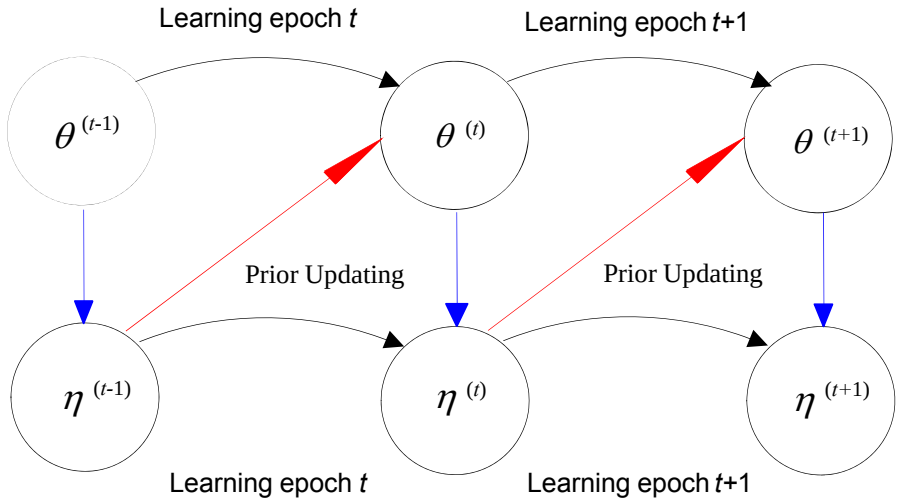
- *Maximum a posteriori* estimation of NB-ICA parameters and compensation parameters

$$\theta^{(t)} = \arg\max_{\theta} \; p(X_t|\theta^{(t)}, \eta^{(t-1)})p(\theta)$$

$$\eta^{(t)} = \arg\max_{\eta} \; p(X_t|\theta^{(t)}, \eta)p(\eta|\phi^{t-1})$$



updating



Model Construction

- *Noisy ICA* model $\mathbf{x}_t = A\mathbf{s}_t + \boldsymbol{\varepsilon}_t$
- Likelihood function of an observation $\mathbf{x}_t$

$$p(\mathbf{x_t}|\mathbf{A}^{(t)}, \mathbf{s}^{(t)}, \beta^{(t)}) = N(\mathbf{x}_t|\mathbf{A}^{(t)}\mathbf{s}^{(t)}, \beta^{(t)^{-1}} I_N)$$

- Distribution of model parameters
  - source $p(\mathbf{s}^{(t)}|\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\gamma}^{(t)}) = \prod_{m=1}^{M} \left[\sum_{k=1}^{K} \pi_k^{(t)} N(s_m^{(t)}|\mu_k^{(t)}, \gamma_k^{(t)^{-1}}) \right]$
  - mixing matrix $p(A^{(t)}|\boldsymbol{\alpha}^{(t)}) = \prod_{m=1}^{M} \left[\prod_{n=1}^{N} N(a_{nm}^{(t)}|0, \alpha_m^{(t)^{-1}}) \right]$
  - noise $p(\varepsilon_t|\beta^{(t)}) = N(\varepsilon_t|0, \beta^{(t)^{-1}} I_N)$

Prior & Marginal Distributions

- Prior distributions
  - precision of noise $p(\beta^{(t)}|u_\beta, w_\beta) = \mathrm{Gam}(\beta^{(t)}|u_\beta, w_\beta)$
  - precision of mixing matrix $p(\boldsymbol{\alpha}^{(t)}|u_\alpha, w_\alpha) = \prod_{m=1}^{M} \mathrm{Gam}(\alpha_m^{(t)}|u_\alpha, w_\alpha)$

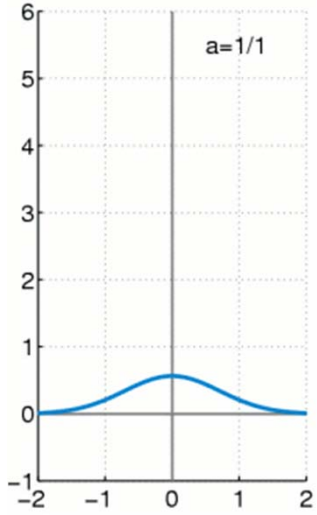
- *Marginal likelihood* of NB-ICA model

$$p(X) = \prod_{t=1}^{T} \int p(\mathbf{x}_t|A^{(t)}, \mathbf{s}^{(t)}, \boldsymbol{\alpha}^{(t)}, \beta^{(t)}) p(A^{(t)}|\boldsymbol{\alpha}^{(t)}) p(\boldsymbol{\alpha}^{(t)}|u_\alpha^{(t)}, w_\alpha^{(t)})$$
$$\times p(\mathbf{s}^{(t)}|\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\gamma}^{(t)}) p(\boldsymbol{\beta}^{(t)}|\boldsymbol{u}_\beta^{(t)}, \mathbf{w}_\beta^{(t)}) dA^{(t)} d\mathbf{s}^{(t)} d\boldsymbol{\alpha}^{(t)} d\boldsymbol{\beta}^{(t)}$$

Automatic Relevance Determination

- Detection of source signals

$$\alpha_m^{(t)} = \begin{cases} \infty & , \quad a_m^{(t)} = \{a_{nm}^{(t)}\} \to 0 \\ < \infty & , \quad a_m^{(t)} = \{a_{nm}^{(t)}\} \neq 0 \end{cases}$$



  - number of sources can be determined

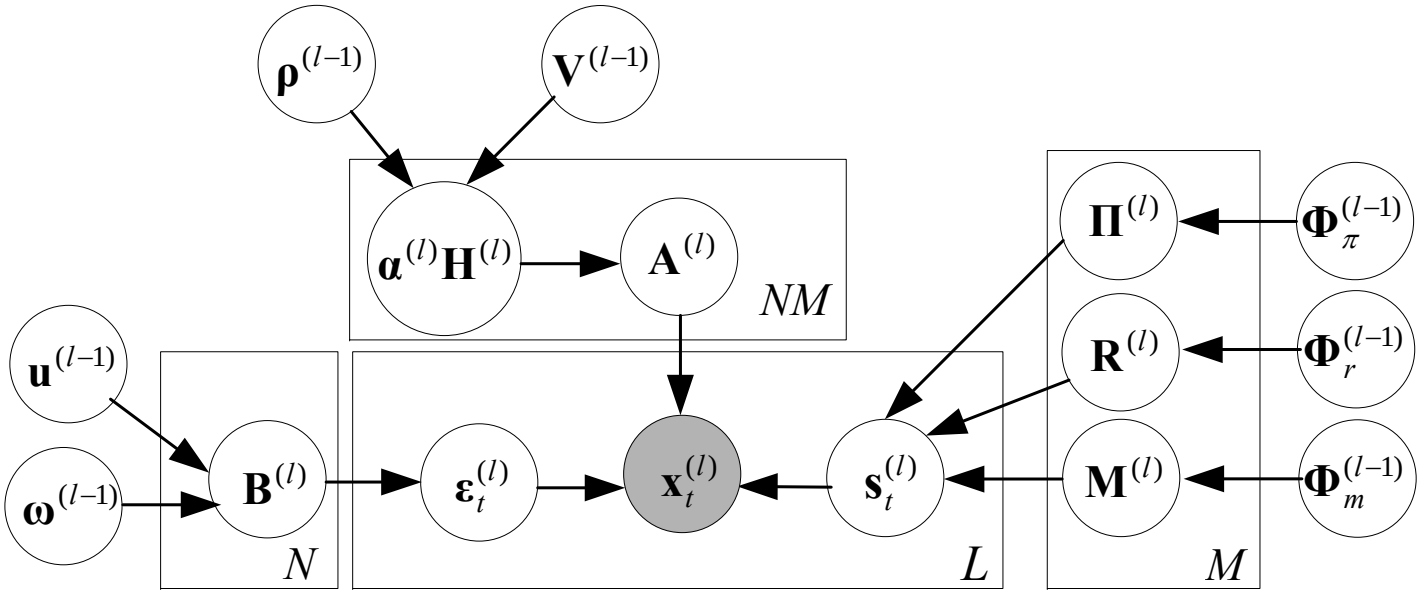
Compensation for Nonstationary ICA

$$G_{\eta^{(t)}}(\alpha^{(t)}) = \eta^{(t)}\alpha^{(t)}$$

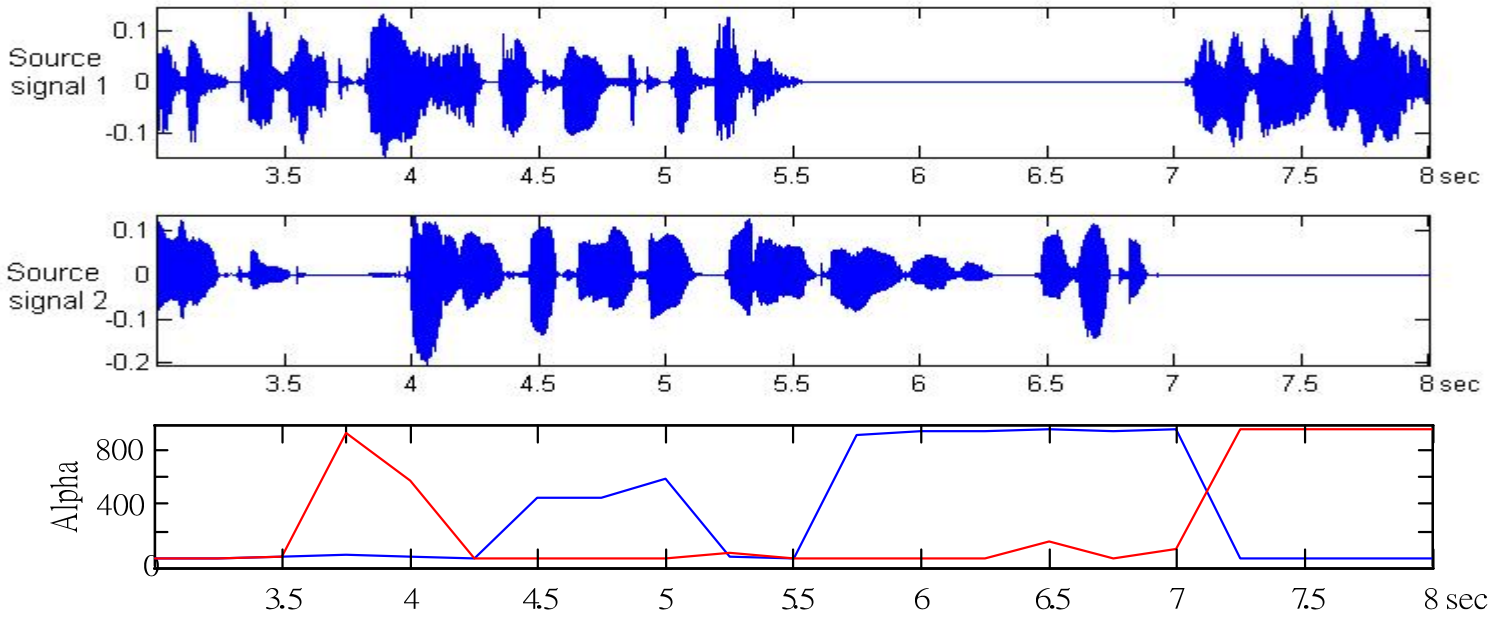
- Prior density of compensation parameter
  - conjugate prior (Wishart distribution)

$$p(\eta^{(t)}|\varphi^{t-1} = \{\nu^{t-1}, \Lambda^{t-1}\})c\left(M, \nu^{t-1}\right)\left|\frac{1}{2}\nu^{t-1}\Lambda^{t-1}\right|^{\left(\nu^{t-1}-1\right)/2}$$
$$\times\left|\eta^{(t)}\right|^{\left(\nu^{t-1}-M-2\right)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\nu^{t-1}\Lambda^{t-1}\eta^{(t)}\right)\right]$$
$$c\left(M, \nu^{t-1}\right) = \left(\pi^{M(M-1)/4}\prod_{m=1}^{M}\Gamma\left(\left(\nu^{t-1}-m\right)/2\right)\right)^{-1}$$

Graphical Model for NB-ICA



Source Signals and ARD Curves



Blue: first source signal

Red: second source signal

Online Gaussian Process (OLGP)

- Basic ideas
  - incrementally detect the status of source signals and estimate the corresponding distributions from online observation data $\chi^t = \left\{ X^{(1)}, X^{(2)}, \cdots, X^{(t)} \right\}$.
  - *temporal structure* of time-varying mixing coefficients $A^{(t)}$ are characterized by *Gaussian process*.
  - Gaussian process is a nonparametric model which defines the prior distribution over functions for Bayesian inference.

Model Construction

- *Noisy ICA* model $\mathbf{x}^{(t)} = A^{(t)}\mathbf{s}^{(t)} + \boldsymbol{\varepsilon}^{(t)}$
- Likelihood function

$$p(\mathbf{x}^{(t,i)} \mid \mathbf{A}^{(t,i)}, \mathbf{s}^{(t,i)}, \beta^{(t,i)}) = N(\mathbf{x}^{(t,i)} \mid \mathbf{A}^{(t,i)}\mathbf{s}^{(t,i)}, \beta^{(t,i)^{-1}} I_N)$$

- Distribution of model parameters
  - source

$$p(\mathbf{s}^{(t,i)} \mid \boldsymbol{\pi}^{(t,i)}, \boldsymbol{\mu}^{(t,i)}, \boldsymbol{\gamma}^{(t,i)}) = \prod_{m=1}^{M} \left[\sum_{k=1}^{K} \pi_{m,k}^{(t,i)} N(s_m^{(t,i)} \mid \mu_{m,k}^{(t,i)}, \gamma_{m,k}^{(t,i)^{-1}}) \right]$$

  - noise

$$p(\varepsilon^{(t,i)} \mid \beta) = N(\varepsilon^{(t,i)} \mid 0, \beta_m^{(t,i)^{-1}} I_N)$$

  - $p(\beta^{(t,i)}) = \mathrm{Gam}(u_\beta^{(t,i)}, w_\beta^{(t,i)})$

Gaussian Process

- Mixing matrix
  - $A^{(t)}$ is generated by the latent function $f_{nm}(\cdot)$

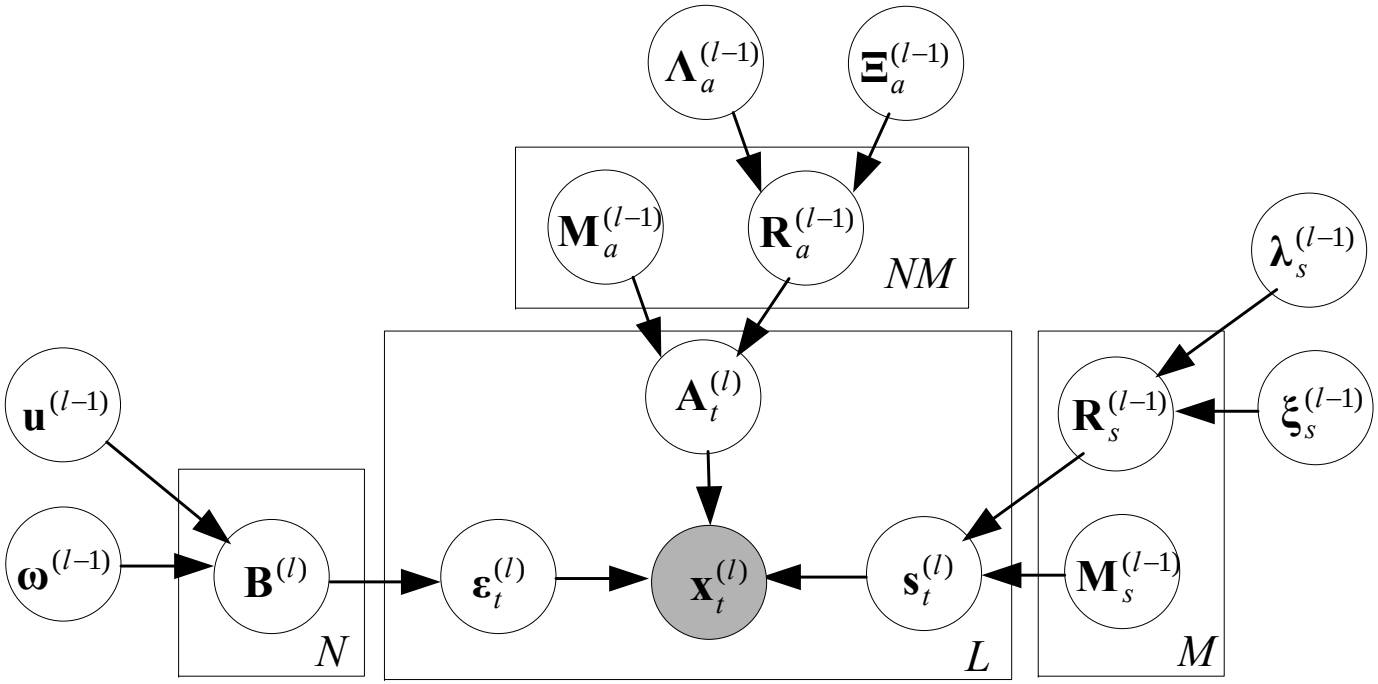
$$a_{nm}^{(t,i)} = f_{nm}\left(\mathbf{a}_{nm}^{(t,i-1,i-p)}\right)$$
$$\mathbf{a}_{nm}^{(t,i-1,i-p)} = \left[a_{nm}^{(t,i-1p)} a_{nm}^{(t,i-2)} \cdots a_{nm}^{(t,i-p)}\right]^T$$

  - GP is adopted to describe the distribution of $f_{nm}(\cdot)$

$$f_{nm}(\mathbf{a}_{nm}^{(t,i-1,i-p)}) \sim N(0, \kappa(\mathbf{a}_{nm}^{(t,i-1,i-p)}, \mathbf{a}_{nm}^{(t,\tau-1,\tau-p)}))$$
$$\kappa(\mathbf{a}_{nm}^{(t,i-1,i-p)}, \mathbf{a}_{nm}^{(t,\tau-1,\tau-p)}) = \rho_{nm}^{(t,i)} \exp\left\{-\frac{\lambda_{nm}^{(t,i)}}{2}\left\|\mathbf{a}_{nm}^{(t,i-1,i-p)} - \mathbf{a}_{nm}^{(t,\tau-1,\tau-p)}\right\|^2\right\}$$

  - $\left\{\lambda_{nm}^{(t,i)}, \rho_{nm}^{(t,i)}\right\}$ are hyperparameters of kernel function

Graphical Model for OLGP-ICA



Experimental Setup

- Nonstationary source separation using source signals from
  - http://www.kecl.ntt.co.jp/icl/signal/

- Nonstationary scenarios
  - status of source signals: active or inactive
  - source signals or sensors are moving: nonstationary mixing matrix

$$A^{(t)} = \begin{bmatrix} \cos(2\pi f_1 t) & \sin(2\pi f_2 t) \\ -\sin(2\pi f_1 t) & \cos(2\pi f_2 t) \end{bmatrix}$$

$$f_1 = 1/20 \text{ Hz} \qquad f_2 = 1/10 \text{ Hz}$$

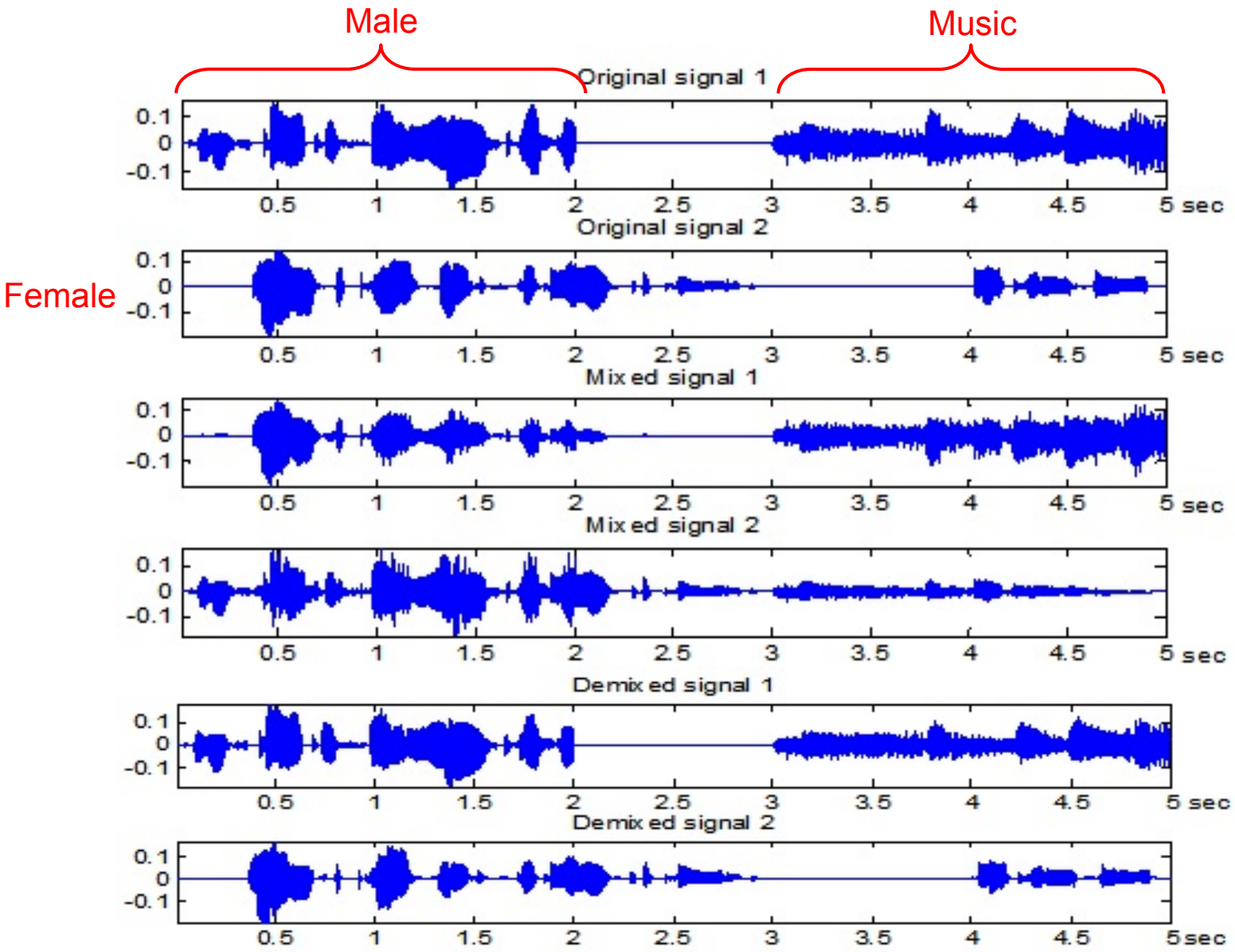

Male
Music
Original signal 1
Female
Original signal 2
Mixed signal 1
Mixed signal 2
Demixed signal 1
Demixed signal 2
0.1
0
-0.1
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5 sec

TABLE OF CONTENTS

Summary

- A *sparse Bayesian learning* was developed for HMM-based acoustic modeling.
- A Dirichlet class language model was developed for Bayesian modeling of *n*-grams.
- A Markov chain was used to characterize *temporal word variations* in a document. Boundaries in text streams were detected by *Bayesian topic* model.
- A scalable tree model of sentence topics was built using Bayesian nonparametrics for document summarization.
- An *online Gaussian process* ICA was presented for nonstationary and temporally correlated source separation.

Future Direction

- ***Model regularization issue is ubiquitous in speech and language processing***
  - think more seriously about our problems at hand
  - systematically extract latent information
  - carefully represent model uncertainty
  - Bayesian nonparametrics & Markov switching process.
- ***Transfer learning for domain and environment adaptation***
  - covariate-shift method
  - recursive Bayes & online learning
  - active learning and semi-supervised learning for reducing labeling costs of speech and text data

Thanks to

G. Saon

C.-H. Chueh

H.-L. Hsieh

Y.-L. Chang

Thank you for your attention!

References

- George Saon and Jen-Tzung Chien, “Bayesian sensing hidden Markov models”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 43-54, 2012.
- Jen-Tzung Chien and Chuang-Hua Chueh, “Topic-based hierarchical segmentation”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 55-66, 2012.
- Jen-Tzung Chien and Chuang-Hua Chueh, “Dirichlet class language models for speech recognition”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 482-495, 2011.
- Jen-Tzung Chien and Jung-Chun Chen, “Recursive Bayesian linear regression for adaptive classification”, *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 565-575, 2009.
- Hsin-Lung Hsieh and Jen-Tzung Chien, “Nonstationary and temporally-correlated source separation using Gaussian process”, *ICASSP*, pp. 2120-2123, 2011.
- Hsin-Lung Hsieh and Jen-Tzung Chien, “Online Gaussian process for nonstationary speech separation”, *INTERSPEECH*, pp. 394-397, 2010.
- Hsin-Lung Hsieh and Jen-Tzung Chien, “Online Bayesian learning for dynamic source separation”, *ICASSP*, pp. 1950-1953, 2010.
- Ying-Lan Chang, Jui-Jung Hung and Jen-Tzung Chien, “Bayesian nonparametric modeling of hierarchical topics and sentences”, *MLSP*, 2011.