Bayesian Learning Approaches for Speech Recognition

Jen-Tzung Chien

National Cheng Kung University, Taiwan

Outline

- Introduction
- Bayesian Adaptation and Predictive Classification
- Bayesian Model Comparison
- Bayesian Large Margin HMMs
- Bayesian Topic Language Model
- Conclusions

Introduction

Why Bayesian?

- *Certainty* knowledge
  - Explicit information to learn
  - We can define proper data structure or rule for the certainty knowledge
- Different people may have different opinions for the same problem
  - We may not have a perfect rule for a problem
- *Uncertainty* knowledge
  - Implicit information
  - Hard to learn
- Useful information is often uncertain
- We cannot build a complete knowledge in many cases

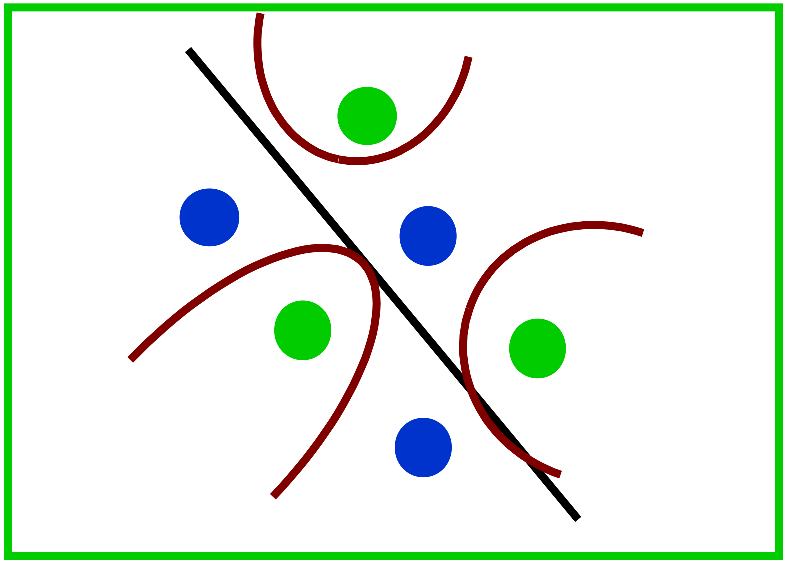
Generalization

- How much can we trust isolated data points?

- Optimal decision surface is a line

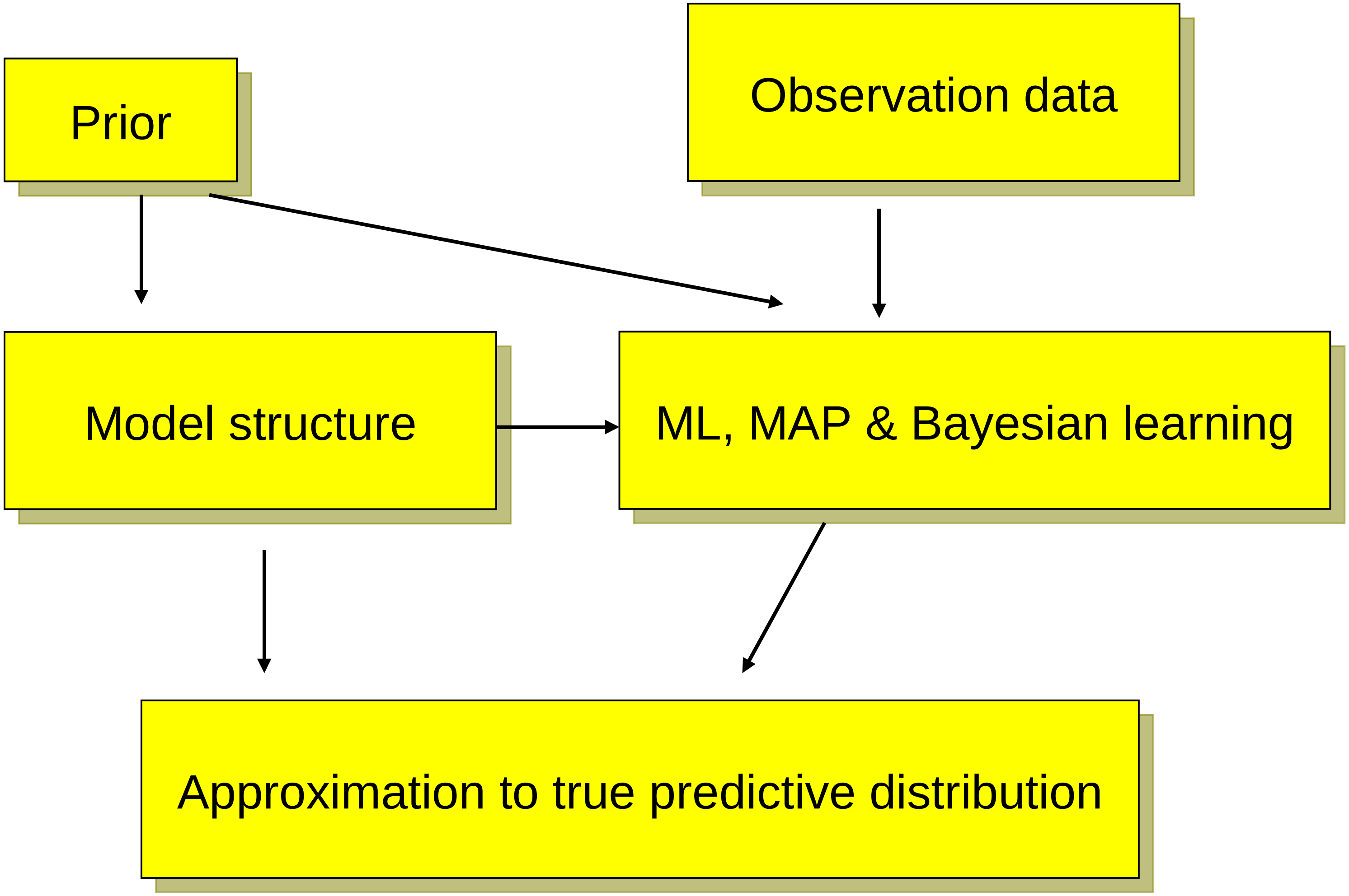
- Optimal decision surface is still a line

- Optimal decision surface changes abruptly



- Can we integrate prior knowledge about data, confidence, or willingness to take risk?

ML, MAP and Bayesian Prediction


Prior
Observation data
Model structure
ML, MAP & Bayesian learning
Approximation to true predictive distribution

ML vs. Bayesian inference

- Maximum Likelihood (ML)

$$\theta_{ML} = \arg\max_{\theta} P(D \mid \theta) \qquad P(x \mid D) \approx P(x \mid \theta_{ML})$$

- Maximum *a Posteriori* (MAP)

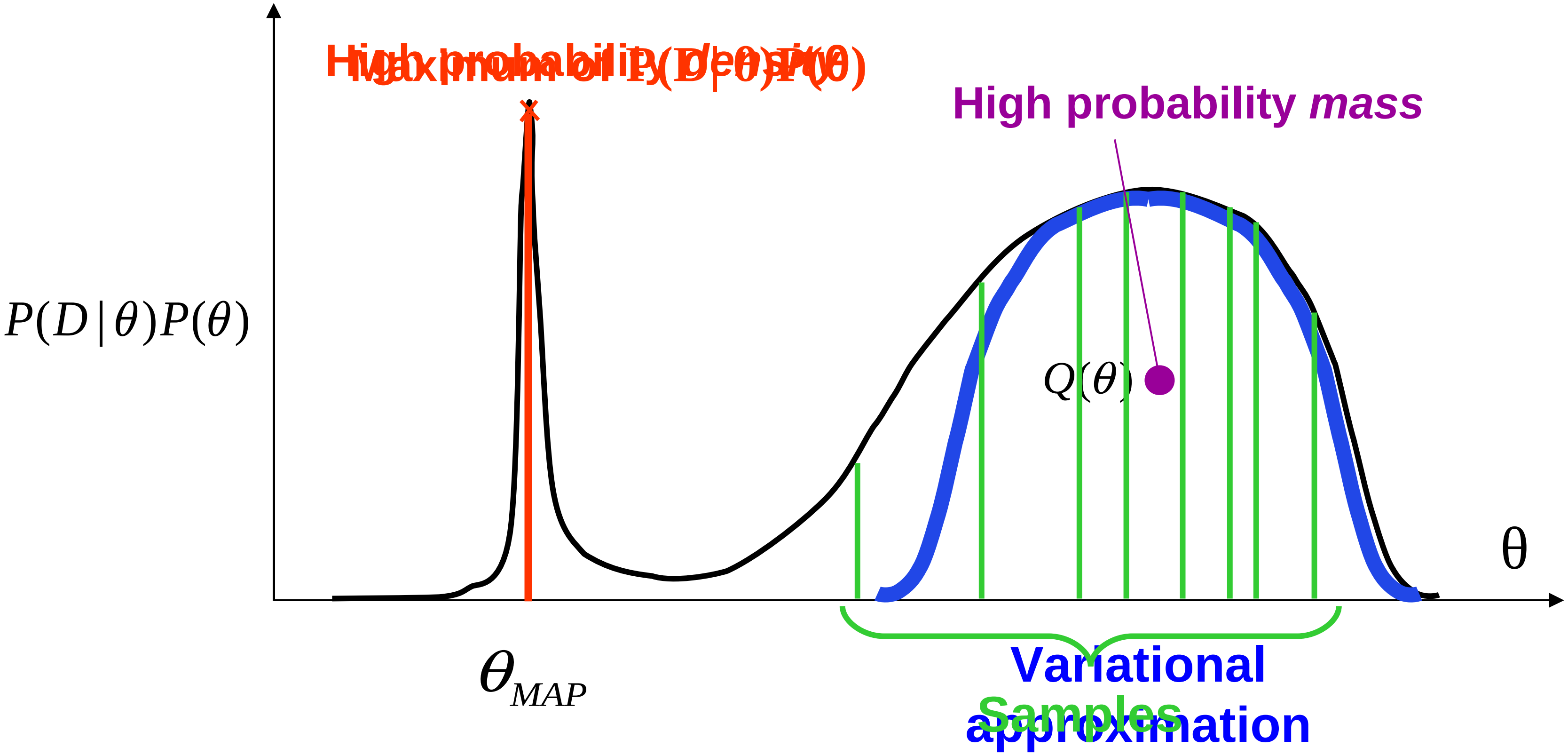
$$\theta_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} P(D \mid \theta) P(\theta) \qquad P(x \mid D) \approx P(x \mid \theta_{MAP})$$

- Bayesian Inference
  - avoid severe over-fitting problem in ML/MAP point estimates
  - allow model comparison

Predictive Distribution $P(x \mid D) = \int P(x \mid \theta) P(\theta \mid D) d\theta$

Bayesian inference

- Consider the learning of a parameter θ∈ **H.**



Model Complexity

- Model complexity is an important issue in statistical inference
  - too simple, poor prediction
  - too complex, poor prediction (and slow on test)
- Maximum likelihood always favors more complex models
  - *over-fitting*
- It is usual to resort to *cross validation*
  - extra data is required
  - computationally expensive
- *Bayesian inference* is performed for *model selection* from training data

Evidence Framework

- Inference using ML/MAP is conditional on the model being true
- We don't know if the model is true
  - affect reliability of posterior distribution, precision, etc.
- Model selection by *evidence framework*
  - *posterior probabilities*

$$p(M_i \mid D) \propto p(D \mid M_i) P(M_i)$$

  - *for equal priors, models are compared using the evidence*

$$p(D \mid M_i) = \int p(D \mid \theta, M_i) p(\theta \mid M_i) d\theta$$

  - *maximizing lower bound* $p(D \mid M_i)$ *for model inference*

Variational Inference

- Exact *marginalization* over uncertainty of parameters does not exist
- Goal: approximate the posterior $P(\theta \mid D)$ by a *variational distribution* $q(\theta)$ for which marginalization is tractable
- Posterior related to joint $P(\theta, D)$ in marginal likelihood

$$P(D) = \int P(D \mid \theta) P(\theta) d\theta$$

  - a good objective for model selection
- Three steps
  1. Choose a family of variational distributions $Q(H)$
  2. Calculate KL divergence between P and Q
  3. Find Q which *minimizes* $\mathrm{KL}(Q||P)$

Automatic Speech Recognition

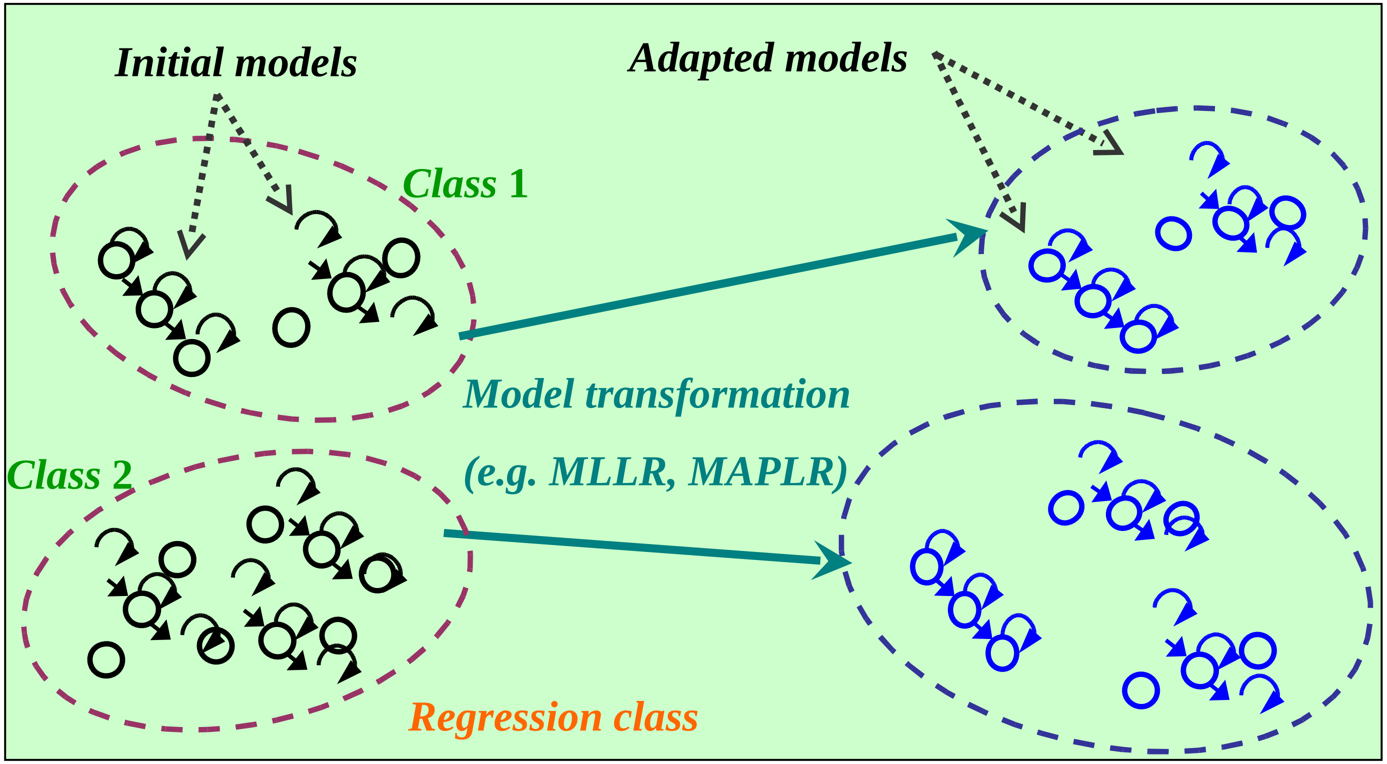
$$\hat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} p_{\Lambda}(X|W)\,p_{\Gamma}(W)$$

Research Topics

- Bayesian speaker adaptation
- Online adaptation
- Bayesian predictive classification
  - uncertainty decoding
- Model selection and clustering
  - evidence framework
- Bayesian large margin HMMs
- Bayesian language model
  - latent Dirichlet language model
  - latent Dirichlet segmentation

Bayesian Adaptation & Predictive Classification

Linear Regression Adaptation



Maximum Likelihood Linear Regression

- Linear regression transformation

$$\hat{\lambda} = G_\eta(\lambda) = \{\omega_{ik}, A_c\mu_{ik} + b_c, r_{ik}\} = \{\omega_{ik}, W_c\xi_{ik}, r_{ik}\}$$

- Maximum likelihood estimation

$$W_{ML} = \arg\max_{W} \; p(\mathbf{X}|W,\lambda)$$

where $p(\mathbf{x}_t|W_c,\mu_{ik},\Sigma_{ik}) \propto |r_{ik}|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{x}_t - W_c\xi_{ik})^T r_{ik}(\mathbf{x}_t - W_c\xi_{ik})\right]$

and $W = \{W_c\}$ $\quad \xi_{ik} = [1,\mu_{ik}^T]^T$

Quasi-Bayes Linear Regression

- ML estimate often leads to biased estimate in case of sparse data.
- *MAPLR* is to estimate the regression matrix by

$$W_{MAP} = \arg\max_{W} p(W \big| \mathbf{X}, \lambda) = \arg\max_{W} p(\mathbf{X} \big| W, \lambda)\, p(W \big| \varphi)$$

- In *online adaptation* using *QBLR* , we estimate the χ^n regression matrix from sequentially observed data . At the *n*th learning epoch, we perform

$$W_{QB}^{(n)} = \arg\max_{W} p(W \mid \chi^{n}, \lambda) = \arg\max_{W} p(\mathbf{X}_n \mid W, \lambda)\, p(W \mid \chi^{n-1}, \lambda)$$

$$\cong \arg\max_{W} p(\mathbf{X}_n \big| W, \lambda)\, p(W \mid \varphi^{(n-1)})$$

Reproducible Prior/Posterior Pair

- Prior density of regression matrix $W_c^{(n)} = \{W_c^{(n)}(i)\}$ can be modeled by a *matrix variate normal distribution*

$$p(W_c^{(n)} | \varphi_c^{(n-1)}) \propto \left|\Delta_c^{(n-1)}\right|^{-1/2} q\left(\sum_{i=1}^{d} (W_c^{(n)}(i) - M_c^{(n-1)}(i)) \Sigma_{ci}^{(n-1)^{-1}} (W_c^{(n)}(i) - M_c^{(n-1)}(i))^T \right)$$

  hyperparameters $M_c^{(n)} = \{M_c^{(n)}(i)\}$, $\Delta_c^{(n-1)} = diag(\Sigma_{c1}^{(n-1)}, \cdots, \Sigma_{cd}^{(n-1)})$

- Expectation function of the posterior distribution in E-step is yielded by a new *matrix variate normal distribution* with new hyperparameters.

Bayesian Predictive Classification

- *Plug-in Bayesian classifier* - regression parameter $\hat{\eta}$ acts as true value to fulfill Bayes decision rule

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

- We consider the *uncertainty* of regression parameters and construct a new decision rule.

- *Linear Regression Bayesian predictive classifier (LRBPC)* - replace the likelihood in plug-in Bayesian classifier using a *predictive distribution*

$$p(\mathbf{X}|W,\hat{\eta},\lambda) \Longrightarrow \tilde{p}_{\eta}(\mathbf{X}|W,\lambda) = \int p(\mathbf{X}|W,\eta,\lambda)\,p(\eta|\varphi)\,d\eta$$

LRBPC

- In case of *single variable linear regression*, the transformation $\hat{\mu}_{ik} = W_c \xi_{ik} = \mathbf{A}_c \mu_{ik} + \mathbf{b}_c$ with $\mathbf{A}_c = \mathrm{diag}\{a_{cl}\}$ becomes independent adaptation for each HMM mean component.

$$\hat{\mu}_{ikl} = a_{cl}\mu_{ikl} + b_{cl}$$

- *Multivariate* frame-based predictive pdf $f_{ik}(\mathbf{x}_t)$ is fulfilled by individually computing *univariate* predictive pdf

$$f_{ik}(x_{tl}) = \int p(x_{tl} \mid \theta_{cl}, \mu_{ikl}, \sigma_{ikl}^2)\, p(\theta_{cl} | \varphi_{cl})\, d\theta_{cl}$$

$$= \int \left(\int p(x_{tl} \mid a_{cl}, b_{cl}, \mu_{ikl}, \sigma_{ikl}^2)\, p(a_{cl} | b_{cl}, \varphi_{cl})\, da_{cl} \right) p(b_{cl} | \varphi_{cl})\, db_{cl}$$

Frame-Based Predictive PDF

- Prior density of $\theta_{cl} = [a_{cl}, b_{cl}]^T$ is defined by a *joint Gaussian pdf*

$$g(\theta_{cl} | \varphi_{cl}) = g(a_{cl}, b_{cl} | \varphi_{cl} = (\mathbf{m}_{\theta_{cl}}, \Sigma_{\theta_{cl}}))$$

$$= \frac{1}{2\pi} \left| \begin{bmatrix} \sigma^2_{a_{cl}} & \sigma^2_{a_{cl} b_{cl}} \\ \sigma^2_{a_{cl} b_{cl}} & \sigma^2_{b_{cl}} \end{bmatrix} \right|^{-1/2} \exp\left\{ -\frac{1}{2} \begin{bmatrix} a_{cl} - m_{a_{cl}} & b_{cl} - m_{b_{cl}} \end{bmatrix} \begin{bmatrix} \sigma^2_{a_{cl}} & \sigma^2_{a_{cl} b_{cl}} \\ \sigma^2_{a_{cl} b_{cl}} & \sigma^2_{b_{cl}} \end{bmatrix}^{-1} \begin{bmatrix} a_{cl} - m_{a_{cl}} \\ b_{cl} - m_{b_{cl}} \end{bmatrix} \right\}$$

- Predictive pdf $f_{ik}(x_{tl})$ is derived as a *Gaussian distribution* of x_{tl} with new mean and new variance given by

$$\hat{\mu}_{x_l} = m_{a_{cl}} \mu_{ikl} + m_{b_{cl}} \quad \longleftarrow$$

Affine function

$$\hat{\sigma}^2_{x_l} = \sigma^2_{b_{cl}} \left(1 + \frac{\sigma^2_{a_{cl} b_{cl}}}{\sigma^2_{b_{cl}}} \mu_{ikl} \right)^2 + \mu^2_{ikl} \left(\sigma^2_{a_{cl}} - \frac{\sigma^4_{a_{cl} b_{cl}}}{\sigma^2_{b_{cl}}} \right) + \sigma^2_{ikl}$$

BAYESIAN
MODEL COMPARISON

An Evidence Framework For Bayesian Learning of Continuous-Density Hidden Markov Models, ICASSP 2009

Motivation

- The *ill-posed* conditions severely hamper the trained HMMs to recognize test data robustly.
- In an *evidence framework*, we build the *regularized* HMMs with given finite data, hence more robust recognition performance.
- In this study, we
  - apply evidence framework to *exponential family distribution* estimation.
  - extend it to estimating CDHMMs with naturally built-in model *uncertainty*.

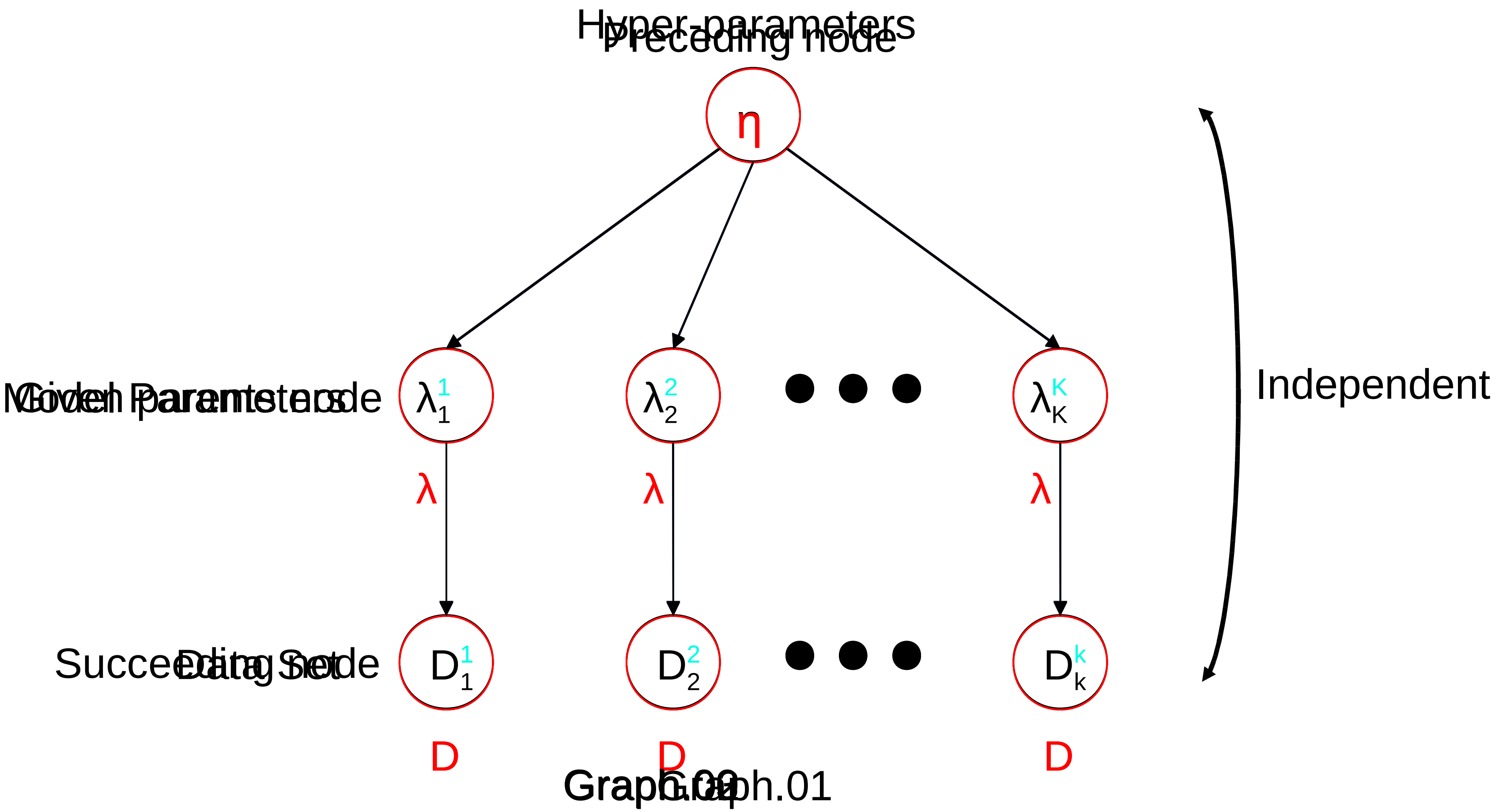
Evidence Framework

- Notations
  - η : hyperparameter of the model
  - $\{\lambda_i\}$: distribution parameters
  - $\{D_i\}$: set of training data

- *Model evidence* is used as the *objective function*

$$
\begin{aligned}
\hat{\eta} &= \arg\max_{\eta} p(D_1, \ldots, D_K \mid \eta) \\
&= \arg\max_{\eta} \prod_{i=1}^{K} \int p(D_i \mid \lambda_i) p(\lambda_i \mid \eta) d\lambda_i
\end{aligned}
$$

Graphical Representation



EM Solution

- Key idea: treat λ_i as *hidden* variable.
- E-Step:

$$Q(\eta,\eta^{old}) = \sum_{i=1}^{K} \int p(\lambda_i \mid D_i \eta^{old}) \ln p(D_{i,} \lambda_i \mid \eta) d\lambda_i$$

- M-Step: find the solutions to all hyperparameters in the *exponential family*.

Exponential Family & Conjugate Prior

- Exponential family

$$p(x_i \mid \lambda_i) = h(x_i) g(\lambda_i) \exp[\lambda_i^T u(x_i)]$$

- Sufficient statistics

$$\sum_{x \in D} u(x)$$

- Conjugate prior

$$p(\lambda_i \mid \chi_0, v_0) = f(\chi_0, v_0) g(\lambda_i)^{v_0} \exp(v_0 \lambda_i^T \chi_0)$$

Bayesian Learning

- Using two properties
  - with *conjugate prior*, the posterior can have the same functional form as its prior.
  - D_i is *conditionally independent* of η_i given λ_i $(D_i \perp \eta_i \mid \lambda_i)$

we get ⇒

$$Q(\eta, \eta^{old}) = \sum_{i=1}^{K} \int p(\lambda_i \mid \tilde{\eta}_i^{old}) \ln p(\lambda_i \mid \eta) d\lambda_i + C$$

EM Steps for Bayesian Learning

- E-step

$$\tilde{v}_i = v_0 + \gamma_i$$

$$\tilde{\chi}_i = \frac{\sum_{n=1}^{\gamma_i} u(x_{i,n}) + v_0 \chi_0}{\tilde{v}_i}$$

- M-step

$$\langle \lambda, \ln[g(\lambda)] \rangle_{\eta^{new}} = \frac{1}{K} \sum_{i=1}^{K} \langle \lambda, \ln[g(\lambda)] \rangle_{\tilde{\eta}_i^{old}}$$

Concavity Analysis

- The auxiliary function $Q(\eta, \eta^{old})$ is *concave* ⇒ we can obtain its global optimum in the M-step.
- In general, the objective function F (the evidence) is not concave.

$$F(\eta) = p(D_1, \ldots, D_K \mid \eta)$$

- Good news: $\nabla^2 F$ is proportional to $\sum_i \{\mathrm{cov}_{\tilde{\eta}_i} - \mathrm{cov}_{\eta}\}$ (Note: posterior is usually sharper than its prior)

Variational Inference

- We could hardly evaluate the joint *posterior distribution* of hidden variables.
  - For example, when training Bayesian HMMs empirically, we need to evaluate $p(\lambda, s \mid D)$ in the E-Step. where λ is the HMM parameters and *s* is the state sequence.
- Computationally feasible approach is to select a proper $q(\lambda, s)$ to approximate $p(\lambda, s \mid D)$.

Variational Bayesian

- Factorization assumption: $q(\lambda, s) = q(\lambda)q(s)$
- We can get a new *lower bound* of the log marginal likelihood

$$F_m(q(\lambda), q(s)) = \int \sum_s q(\lambda)q(s) \ln \frac{p(\lambda, s, D \mid m)}{q(\lambda)q(s)} d\lambda$$

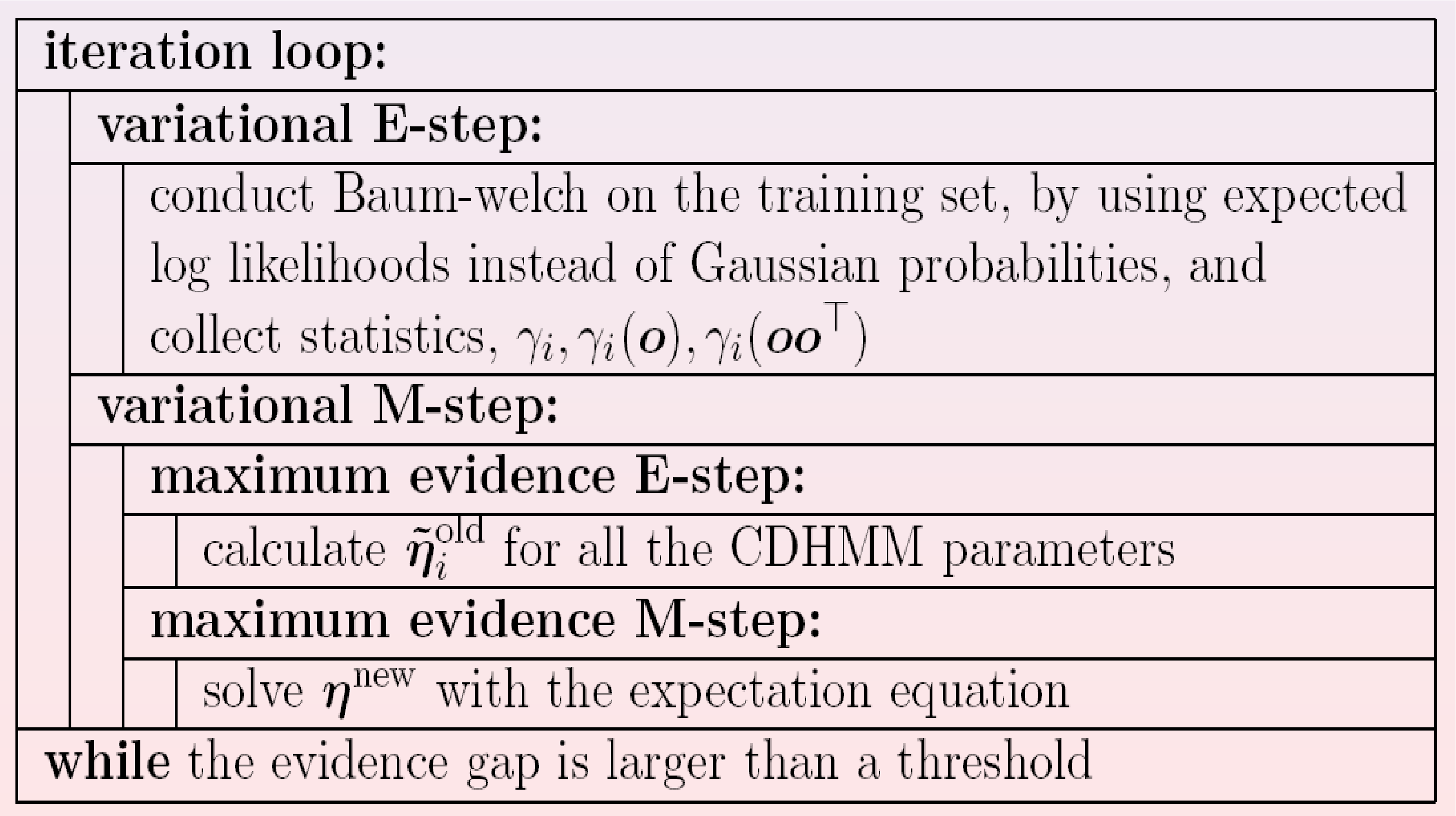
- It can be iteratively optimized

$$q^{new}(\lambda) \propto \exp < \ln p(D, s \mid \lambda) >_{q^{old}(s)}$$

$$q^{new}(s) \propto \exp < \ln p(D, s \mid \lambda) >_{q^{old}(\lambda)}$$

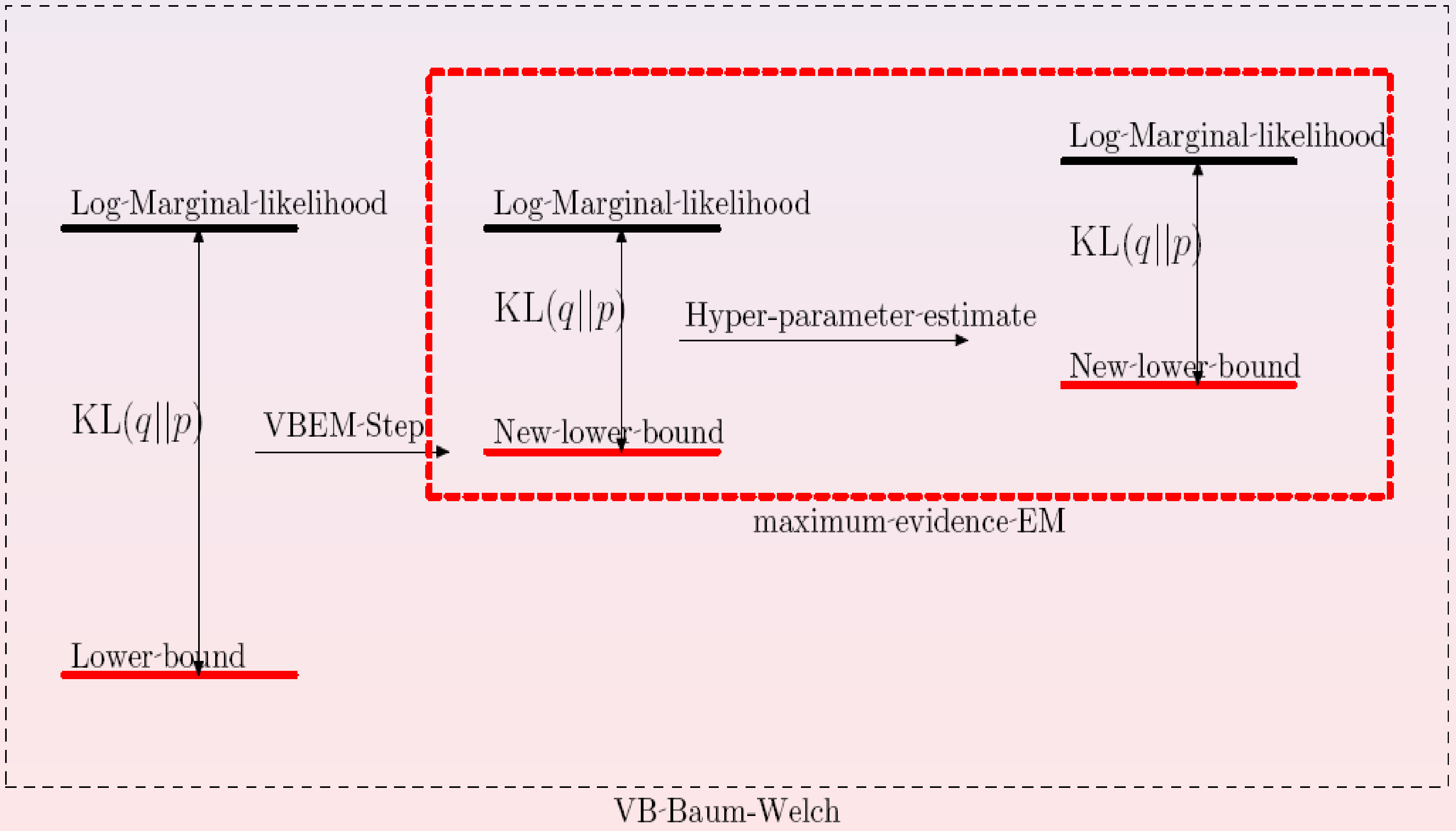
- We have the closed-form solutions to $q^{new}(\lambda)$, $q^{new}(s)$ in CDHMM case.

Evidence Framework for CDHMM Training



```
iteration loop:
    variational E-step:
        conduct Baum-welch on the training set, by using expected
        log likelihoods instead of Gaussian probabilities, and
        collect statistics, $\gamma_i, \gamma_i(\boldsymbol{o}), \gamma_i(\boldsymbol{o}\boldsymbol{o}^\top)$
    variational M-step:
        maximum evidence E-step:
            calculate $\tilde{\boldsymbol{\eta}}_i^{\mathrm{old}}$ for all the CDHMM parameters
        maximum evidence M-step:
            solve $\boldsymbol{\eta}^{\mathrm{new}}$ with the expectation equation
while the evidence gap is larger than a threshold
```

Optimization Procedure



Experimental Results on AURORA2

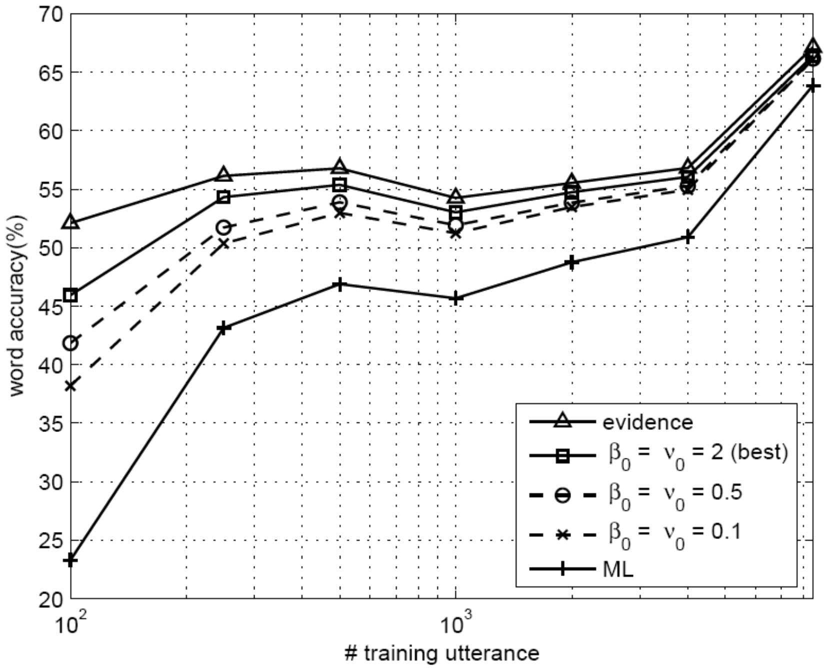


Figure 1: Recognition accuracy of model trained with different sized clean training data

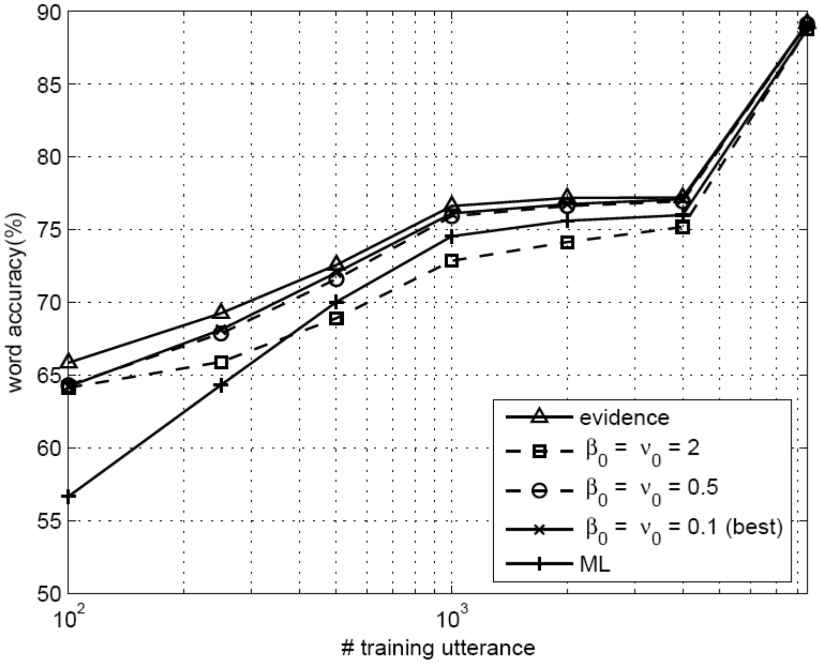
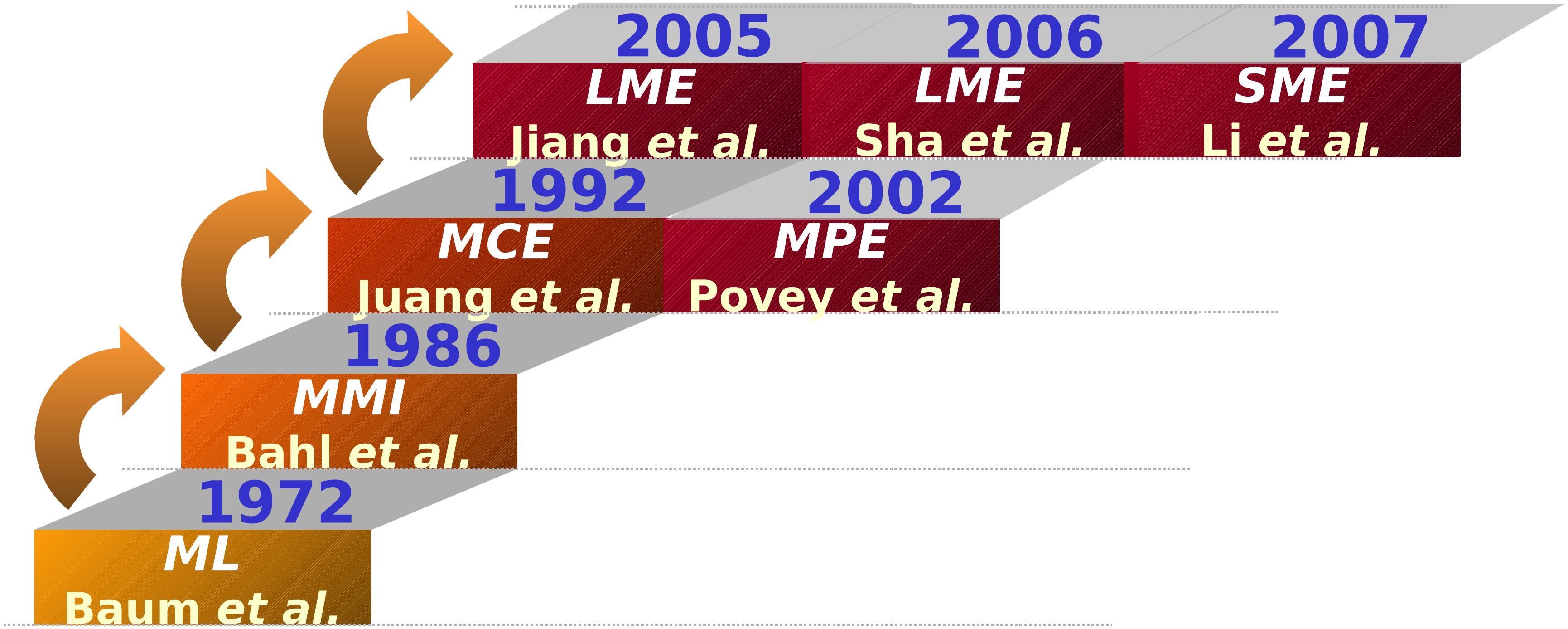


Figure 2: Recognition accuracy of model trained with different sized multi-conditional training data

BAYESIAN LARGE MARGIN HMMS

Bayesian Large Margin Hidden Markov Models for Speech Recognition, ICASSP 2009

History of HMM Training



1972 ML Baum *et al.*

1986 MMI Bahl *et al.*

1992 MCE Juang *et al.*

2002 MPE Povey *et al.*

2005 LME Jiang *et al.*

2006 LME Sha *et al.*

2007 SME Li *et al.*

Vapnik's Risk Bound

$$R(\Lambda) \le R_{emp}(\Lambda) + \sqrt{\frac{١}{N}\left(\mathrm{VC}_{\mathrm{dim}} \cdot \left(\log\left(\frac{٢N}{\mathrm{VC}_{\mathrm{dim}}}\right) + ١\right) - \log\left(\frac{\delta}{۴}\right)\right)}$$

- We should minimize the empirical risk as well as the *generalization* error.
- Increasing number of parameters suffers from *over-fitting* problem. Model generalization is degraded.
- *VC dimension* is closely related to the number of parameters and can be reduced by increasing the *margin*.

Motivation

- Generalization problem in SVM was tackled due to the *sparse learning* and *VC dimension*.
- The *static* LM-HMM parameters are not well fitted to the unknown variations in test environments.
- *Bayesian large margin* (BLM) classifier is presented to build the BLM-HMMs.
- We improve *model generalization* via Bayesian learning and cope with the *uncertainty* in large margin classifier.
- Speech recognition system has the capabilities of model *selection* and model *adaptation*.

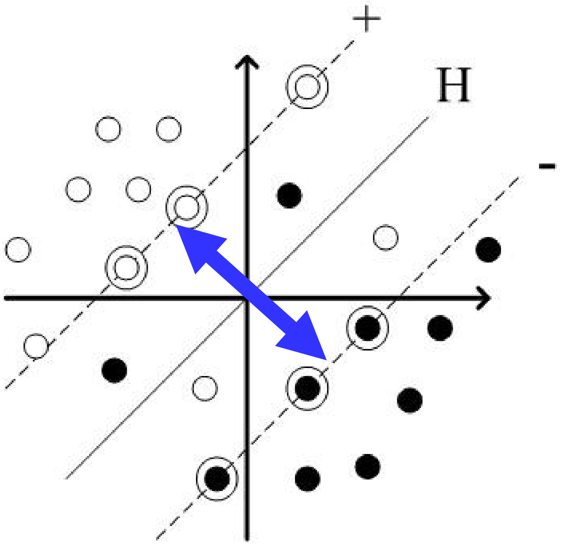
Large Margin Classifier

- Support Vector Machines (SVMs)

$$\min_{\mathbf{w}} Q(\mathbf{w}) \equiv \frac{1}{2}\|\mathbf{w}\|^2 + \boxed{C}\sum_{i=1}^{N}\zeta_i$$

C *is a trade-off*

$$\text{subject to} : y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \zeta_i, i = 1, \ldots, N$$



Margin Maximization

Hard Margin

Soft Margin

Large Margin Estimation

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

- Discriminant function & *separation margin* for an utterance

$$d_{\mathrm{LM}}(X_i, \lambda) = \log p(X_i \mid \lambda_{W_i}) - \max_{W_j \in \Omega_W, j \neq i} \log p(X_i \mid \lambda_{W_j})$$

- Support token set

$$\Psi_{\mathrm{LM}} = \{ X_i \mid X_i \in D \text{ and } 0 \leq d_{\mathrm{LM}}(X_i, \lambda) \leq \varepsilon \}$$

Correctly Classified Utterances

- Objective: maximize the minimum margin of support tokens

$$\lambda_{\mathrm{LM}} = \arg\max_{\lambda} \min_{X_i \in \Psi_{\mathrm{LM}}} d_{\mathrm{LM}}(X_i, \lambda)$$

Soft Margin Estimation

- Separation measure for an utterance

$$d_{\mathrm{SM}}(X_i) = \frac{1}{n_i}\sum_k \log\left[\frac{P(\mathbf{x}_{ik} \mid \lambda_{W_i})}{P(\mathbf{x}_{ik} \mid \lambda_{W_j})}\right] I(\mathbf{x}_{ik} \in F_i)$$

- *Hinge error loss function*

$$(\rho - d_{\mathrm{SM}}(X_i))_+ = \begin{cases} \rho - d_{\mathrm{SM}}(X_i), & \text{if } \rho - d_{\mathrm{SM}}(X_i) > 0 \\ 0, & \text{otherwise} \end{cases}$$

- Objective function

$$L^{\mathrm{SM}}(\Lambda) = \frac{\lambda}{\rho} + \frac{1}{N}\sum_{i=1}^{N}(\rho - d_{\mathrm{SM}}(X_i))_+$$

$$= \frac{\lambda}{\rho} + \frac{1}{N}\sum_{i=1}^{N}(\rho - d_{\mathrm{SM}}(X_i))\, l(X_i \in U)$$

Bayesian Large Margin Estimation

- From Bayesian viewpoint, the *model uncertainty* is considered in expressing the separation margin.
- The uncertainty is characterized by a *prior density*.
- *Posterior separation margin* is yielded by

$$\sum_{X_i \in \Psi_{\mathrm{BLM}}, W_j \in \Omega_W, j \neq i} \exp[\log p(\lambda_{W_j} \mid X_i) - \log p(\lambda_{W_i} \mid X_i)]$$

- *Variational Bayesian* is applied to approximate the true distribution $p(\lambda_W \mid X)$ by using a variational distribution $q(\lambda_W \mid X)$. *VB-EM* algorithm is performed.

Variational Inference

- Variational distribution is estimated through maximization of a *lower bound* of logarithm of *marginal likelihood*

$$
\begin{aligned}
\log p(X) &= \log \int \sum_{S,L} p(X,S,L \mid \lambda_W)\, p(\lambda_W)\, d\lambda_W \\
&\ge \int \sum_{S,L} q(S,L,\lambda_W \mid X) \log \frac{p(X,S,L \mid \lambda_W)\, p(\lambda_W)}{q(S,L,\lambda_W \mid X)}\, d\lambda_W \\
&= \int q(\lambda_W \mid X) \left[\sum_{S,L} q(S,L \mid X) \log \frac{p(X,S,L \mid \lambda_W)\, p(\lambda_W)}{q(\lambda_W \mid X)} \right] d\lambda_W \\
&\quad - \sum_{S,L} q(S,L,X) \log q(S,L \mid X).
\end{aligned}
$$

variational distributions

LM-HMM Parameters and Their Priors

- LM-HMM model *parameters* $\{\pi_i, a_{im}, \omega_{ik}, (\mu_{ik}, r_{ik})\}$
- We specify the prior of probability parameter to be *Dirichlet* density and the prior of Gaussian mean and precision to be a *normal-Wishart density*

$$p(\mu_{ik}, r_{ik} \mid m_{ik}, \tau_{ik}, \alpha_{ik}, u_{ik}) = \left| r_{ik} \right|^{(\alpha_{ik}-d)/2}$$

$$\times \exp\left[-\frac{\tau_{ik}}{2} (\mu_{ik} - m_{ik})^T r_{ik} (\mu_{ik} - m_{ik}) \right] \exp\left[-\frac{1}{2} \mathrm{tr}(u_{ik} r_{ik}) \right]$$

where $\tau_{ik} > 0$, $\alpha_{ik} > d-1$, μ_{ik} is $d \times 1$ vector,

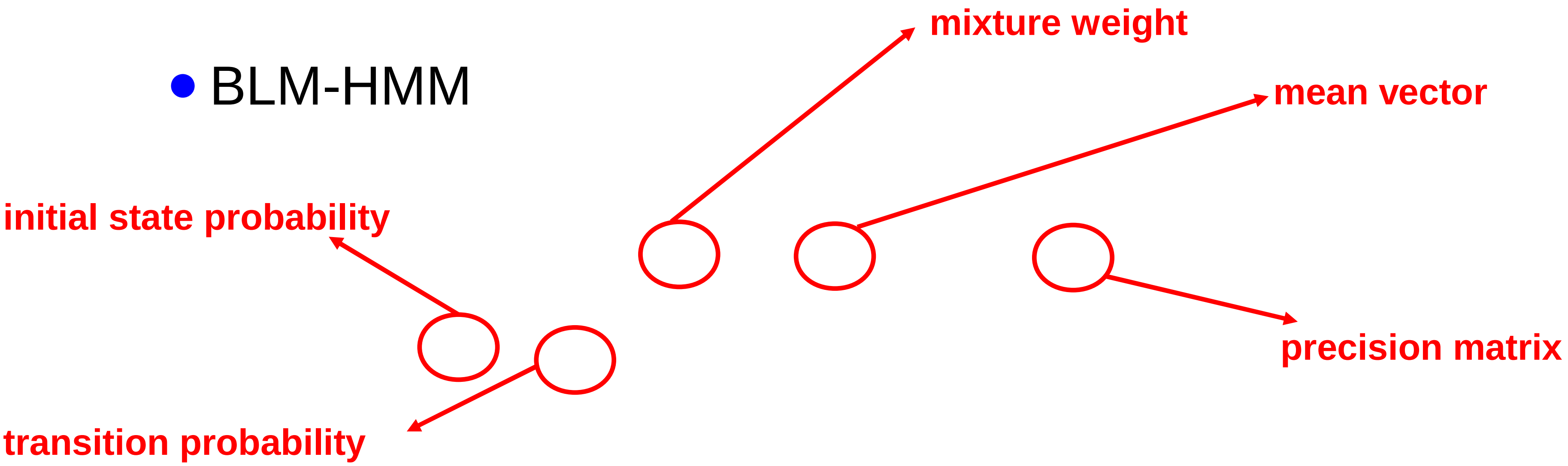
u_{ik} is a $d \times d$ positive definite matrix.

- *Hyperparameters* in LM-HMMs

$$\{\varpi_i, \phi_{im}, \varphi_{ik}, m_{ik}, \tau_{ik}, \alpha_{ik}, u_{ik}\}$$

Graphical Representation

- BLM-HMM



- Variational BLM-HMM

Variational Distribution

- *VB posterior distributions* $\tilde{q}(\lambda \mid X)$ and $\tilde{q}(S,L \mid X)$ are alternatively estimated

$$\tilde{q}(\lambda \mid X) \propto p(\lambda \mid \{\varpi_i, \phi_{im}, \varphi_{ik}, m_{ik}, \tau_{ik}, \alpha_{ik}, u_{ik}\})$$

$$\times \exp\left[\sum_{S,L} \tilde{q}(S,L \mid X) \log p(X,S,L \mid \lambda)\right]$$

$$= \prod_{i,m,k} \tilde{q}(\{\pi_i\} \mid X)\, \tilde{q}(\{a_{im}\} \mid X\}\, \tilde{q}(\{\omega_{ik}\} \mid X\}\, \tilde{q}(\{\mu_{ik}, r_{ik}\} \mid X)$$

$$= \prod_{i,m,k} p(\{\pi_i\} \mid \{\tilde{\varpi}_i\})\, p(\{a_{im}\} \mid \{\tilde{\phi}_{im}\})\, p(\{\omega_{ik}\} \mid \{\tilde{\varphi}_{ik}\})$$

$$\times\, p(\{\mu_{ik}, r_{ik}\} \mid \{\tilde{m}_{ik}, \tilde{\tau}_{ik}, \tilde{\alpha}_{ik}, \tilde{u}_{ik}\})$$

where $\tilde{q}(\{\mu_{ik}, r_{ik}\} \mid X) \propto p(\{\mu_{ik}, r_{ik}\} \mid \{m_{ik}, \tau_{ik}, \alpha_{ik}, u_{ik}\})$

$$\times \exp\left[\sum_{i,k,t \in \Psi_{\mathrm{BLM}}} \tilde{\xi}_{tik} \log p(\mathbf{x}_t \mid \mu_{ik}, r_{ik})\right]$$

Relation to SVM Objective Function

- We make the approximation

$$\tilde{q}(s_t = i, l_t = k \mid \mathbf{x}_{it}) \cong \exp(-[-d^{ij}_{\mathrm{BLM}}(\mathbf{x}_{it})]_+) = \exp(-\tilde{\xi}_t)$$

where $[b]_+ = b \text{ if } b > 0$ and $[b]_+ = 0 \text{ if } b < 0$.

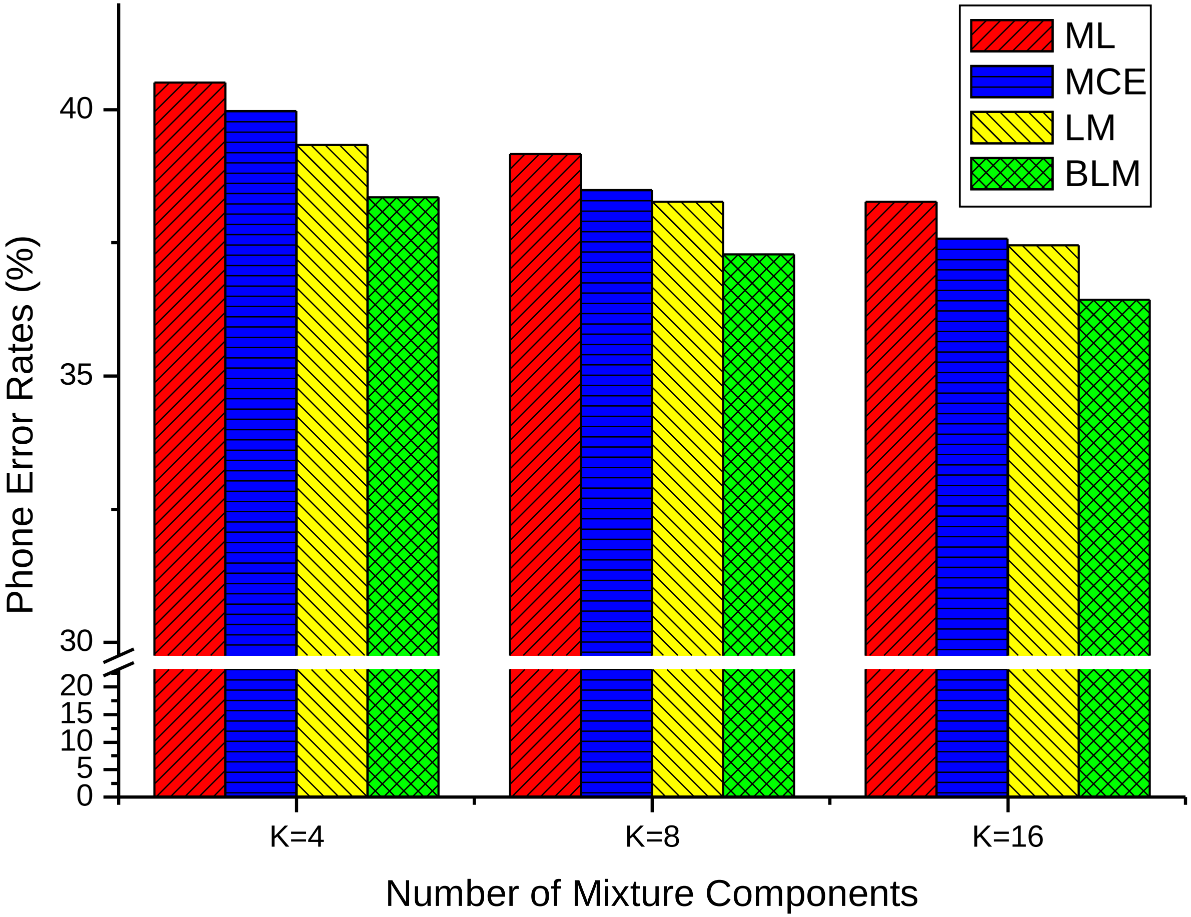
- Substitute this approximate probability into $-\log \tilde{q}(S, L, \mu_{ik}, r_{ik} \mid X_i)$, we obtain

$$-\log \tilde{q}(S, L, \mu_{ik}, r_{ik} \mid X_i) = \underbrace{\frac{\tilde{\tau}_{ik}}{2}(\mu_{ik} - \tilde{m}_{ik})^T r_{ik} (\mu_{ik} - \tilde{m}_{ik})}_{\text{Negative Class Margin}} + \underbrace{\sum_t \tilde{\xi}_t}_{\text{Sum of Errors}} + \text{constant}$$

Comparison

	MCE	LME	SME	BLME
Generalization		○	○	○ ○
Separation Measure	Utterance LLR	Utterance LLR	LLR with frame selection	Log Posterior Ratio with frame selection
Parameters	All Parameters	Mean	Mean	Mean & Precision
Parameter Solution	GPD	GPD	GPD	Closed form
Model Comparison & Adaptation	□	□	□	○

Experimental Results on TIMIT



Bayesian Topic Language Model

Latent Dirichlet Language Model for Speech Recognition, IEEE SLT Workshop 2008

N-Grams

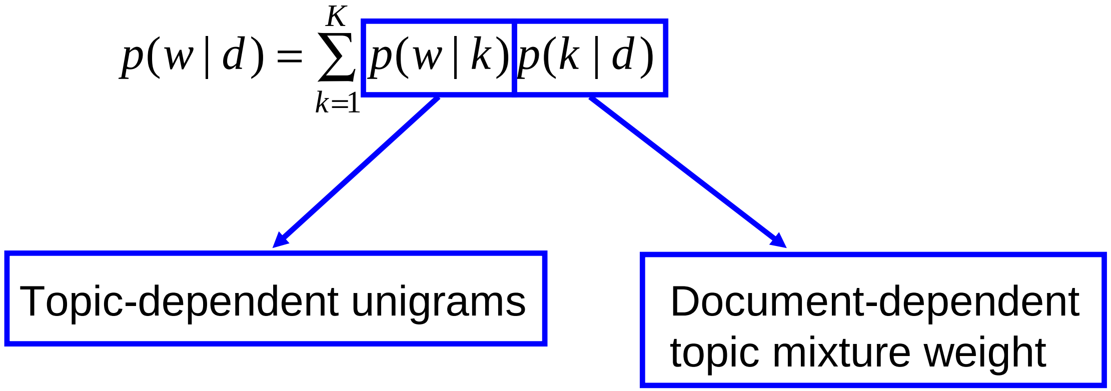
$$\Pr(W)=\Pr(w_1,\ldots,w_T)=\prod_{i=1}^{T}\Pr(w_i\,|\,w_1,w_2,\ldots,w_{i-1})\cong\prod_{i=1}^{T}\Pr(w_i\,|\,w_{i-n+1}^{i-1})$$

Two important issues:

- *Data sparseness* problem
  - Model smoothing
    - Backoff method
    - Continuous space LM
- *Insufficient long-distance* regularity
  - Topic information
    - Probabilistic latent semantic analysis (PLSA)
    - Latent Dirichlet allocation (LDA)

Probabilistic LSA LM [Gildea & Hofmann,1999]

- Document probability



$$p(w\,|\,d)=\sum_{k=1}^{K}\boxed{p(w\,|\,k)}\boxed{p(k\,|\,d)}$$

Topic-dependent unigrams

Document-dependent topic mixture weight

- *Online EM* algorithm was used.

$$p(k\,|\,w_1^{i-1})=\frac{1}{i+1}\frac{p(w_{i-1}\,|\,k)\,p(k\,|\,w_1^{i-2})}{\sum_{j=1}^{K}p(w_{i-1}\,|\,j)\,p(j\,|\,w_1^{i-2})}+\frac{i}{i+1}\,p(k\,|\,w_1^{i-2})$$

$$p(k\,|\,w_1)=p(k)=\frac{\sum_{w,d}N_{wd}\,p(k\,|\,d)}{\sum_{w,d}N_{wd}}$$

Latent Dirichlet Allocation [Blei et al., 2003]

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

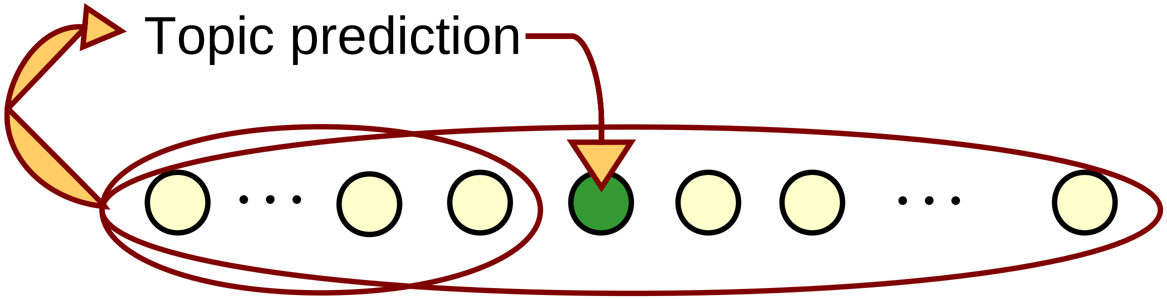
- Document probability

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

- *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.

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

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

Direct Topic Model for ASR

- Document-level topic model (PLSA, LDA)
  - bag-of-words scheme
  - *document clustering*
  - *indirect* model for speech recognition
- *N*-gram-level topic model (LDLM)
  - word orders are considered.
  - *history clustering*
  - *direct* model for speech recognition

Model Construction

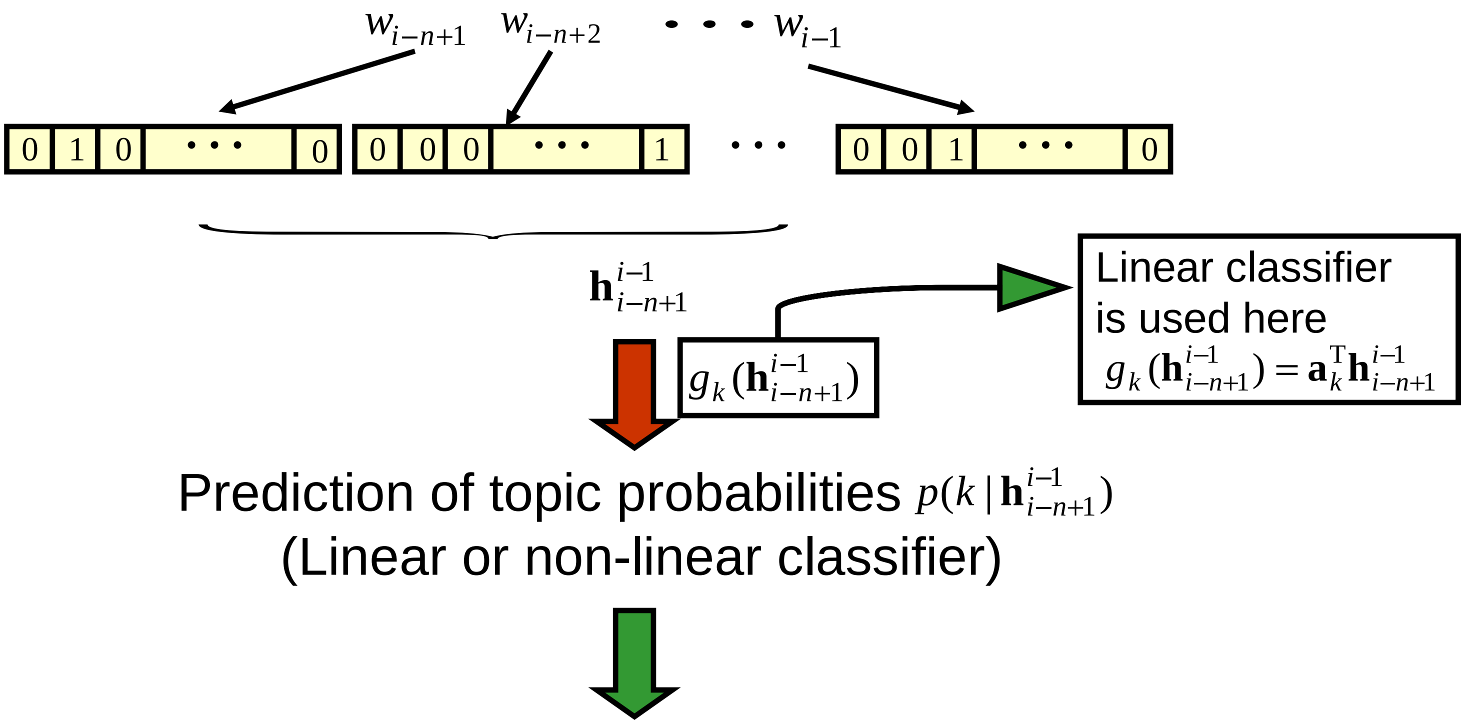
- Topic model is directly built from *n*-gram events.
- LDLM acts as a new *Bayesian topic language model* in which the prior density of the topic variable is involved.

H: number of histories in the training data

N_h: number 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.



Linear classifier is used here

$g_k(\mathbf{h}_{i-n+1}^{i-1})=\mathbf{a}_k^{\mathrm{T}}\mathbf{h}_{i-n+1}^{i-1}$

Prediction of topic probabilities $p(k\,|\,\mathbf{h}_{i-n+1}^{i-1})$
(Linear or non-linear classifier)

Prior density of topic mixture

$$\boldsymbol{\theta}=[\theta_1,\cdots,\theta_K]^{\mathrm{T}}\sim\mathrm{Dir}(\mathbf{g}(\mathbf{h}_{i-n+1}^{i-1}))$$

Latent Dirichlet Language Model

- Probability of an *n*-gram event

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

$$= \sum_{k=1}^{K} \beta_{ik} \frac{\mathbf{a}_k^{\mathrm{T}} \mathbf{h}_{i-n+1}^{i-1}}{\sum_{j=1}^{K} \mathbf{a}_j^{\mathrm{T}} \mathbf{h}_{i-n+1}^{i-1}}.$$

- LDLM performed the *unsupervised learning* and found the classes or latent topics through the VB-EM procedure.

Variational Inference

- Likelihood function of a data set *D*

$$\log p(D\,|\,\mathbf{A},\boldsymbol{\beta}) = \sum_{(w_i,\mathbf{h}_{i-n+1}^{i-1})\in D} \log p(w_i\,|\,\mathbf{h}_{i-n+1}^{i-1},\mathbf{A},\boldsymbol{\beta})$$

$$= \sum_{\mathbf{h}_{i-n+1}^{i-1}} \log\left\{\int p(\boldsymbol{\theta}\,|\,\mathbf{h}_{i-n+1}^{i-1},\mathbf{A})\left[\prod_{i=1}^{N_h}\sum_{k_i=1}^{K} p(w_i\,|\,k_i,\boldsymbol{\beta})\,p(k_i\,|\,\boldsymbol{\theta})\right]d\boldsymbol{\theta}\right\}$$

- True posterior probability

$$p(\boldsymbol{\theta},\mathbf{k}_h\,|\,\mathbf{w}_h,\mathbf{h}_{i-n+1}^{i-1},\mathbf{A},\boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}\,|\,\mathbf{h}_{i-n+1}^{i-1},\mathbf{A})\prod_{i=1}^{N_h} p(w_i\,|\,k_i,\boldsymbol{\beta})\,p(k_i\,|\,\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}\,|\,\mathbf{h}_{i-n+1}^{i-1},\mathbf{A})\prod_{i=1}^{N_h}\sum_{k_i=1}^{K} p(w_i\,|\,k_i,\boldsymbol{\beta})\,p(k_i\,|\,\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

- Variational distribution

$$q(\boldsymbol{\theta},\mathbf{k}_h\,|\,\boldsymbol{\gamma}_h,\boldsymbol{\varphi}_h) = \underbrace{\boxed{q(\boldsymbol{\theta}\,|\,\boldsymbol{\gamma}_h)}}_{\text{Dirichlet}}\prod_{i=1}^{N_h}\underbrace{\boxed{q(k_i\,|\,\boldsymbol{\varphi}_{h,i})}}_{\text{Multinomial}}$$

VB-E Step

- Lower bound of log marginal likelihood

$$L(\mathbf{A},\boldsymbol{\beta};\boldsymbol{\gamma},\boldsymbol{\varphi}) = \sum_{\mathbf{h}_{i-n+1}^{i-1}} \{E_q[\log p(\boldsymbol{\theta} \mid \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})] + E_q[\log p(\mathbf{k}_h \mid \boldsymbol{\theta})]$$

$$+ E_q[\log p(\mathbf{w}_h \mid \mathbf{h}_{i-n+1}^{i-1}, \mathbf{k}_h, \boldsymbol{\beta})] - E_q[\log q(\boldsymbol{\theta} \mid \boldsymbol{\gamma}_h)] - E_q[\log q(\mathbf{k}_h \mid \boldsymbol{\varphi}_h)]\}$$

- VB-E step (updating of variational parameters)

$$\hat{\gamma}_{h,k} = \mathbf{a}_k^{\mathrm{T}} \mathbf{h}_{i-n+1}^{i-1} + \sum_{i=1}^{N_h} \phi_{h,ik}$$

$$\hat{\phi}_{h,ik} = \frac{\beta_{ik} \exp[\Psi(\gamma_{h,k}) - \Psi(\sum_{j=1}^{K} \gamma_{h,j})]}{\sum_{l=1}^{K} \beta_{il} \exp[\Psi(\gamma_{h,l}) - \Psi(\sum_{j=1}^{K} \gamma_{h,j})]}$$

VB-M Step

- Updating of model parameters
  - word probabilities in different topics

$$\hat{\beta}_{vk} = \frac{\sum_{\mathbf{h}_{i-n+1}^{i-1}} \sum_{i=1}^{N_h} \hat{\phi}_{h,ik} \delta(w_v, w_i)}{\sum_{m=1}^{V} \sum_{\mathbf{h}_{i-n+1}^{i-1}} \sum_{i=1}^{N_h} \hat{\phi}_{h,ik} \delta(w_m, w_i)}$$

  - gradient function for updating transformation matrix

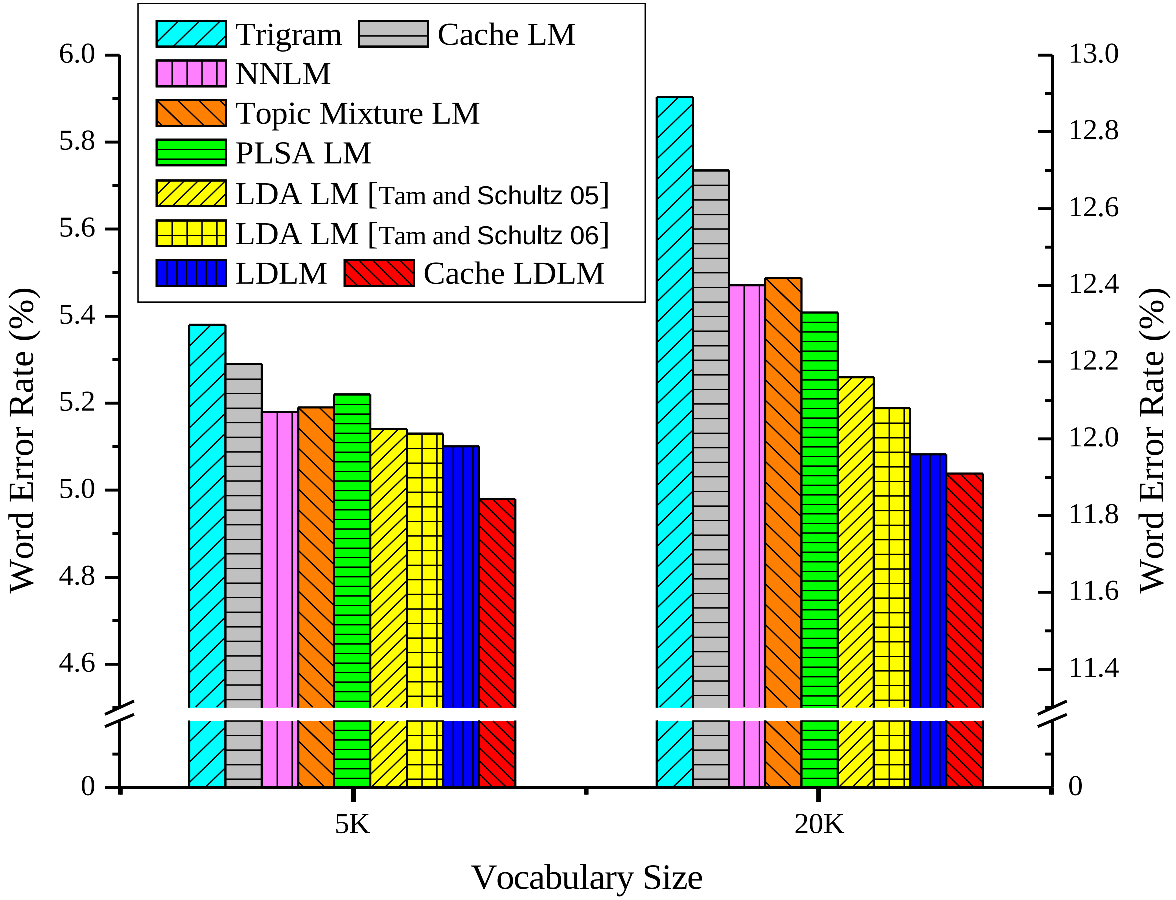
$$\nabla_{\mathbf{a}_k} L(\mathbf{A}, \hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}})$$

$$= \sum_{\mathbf{h}_{i-n+1}^{i-1}} [\Psi(\sum_{j=1}^{K} \mathbf{a}_j^{\mathrm{T}} \mathbf{h}_{i-n+1}^{i-1}) - \Psi(\mathbf{a}_k^{\mathrm{T}} \mathbf{h}_{i-n+1}^{i-1}) + \Psi(\hat{\gamma}_{h,k}) - \Psi(\sum_{j=1}^{K} \hat{\gamma}_{h,j})] \cdot \mathbf{h}_{i-n+1}^{i-1}$$

WER with Different Sizes of Training Data

	Size of Training Data			
	6M	12M	18M	38M
Baseline LM	**39.19 (-)**	**21.25 (-)**	**15.79 (-)**	**12.89 (-)**
Cache LM	**38.13** ***(2.1)***	**20.92** ***(1.6)***	**15.56** ***(1.5)***	**12.74** ***(1.4)***
PLSA LM	**35.96 (*8.2*)**	**19.77 (*7.0*)**	**14.96 (*5.2*)**	**12.33 (*4.5*)**
LDA LM	**38.86 (*8.5*)**	**19.67 (*7.4*)**	**14.73 (*6.7*)**	**12.16 (*5.7*)**
LDLM	**35.91 (*8.4*)**	**19.59 (*7.8*)**	**14.61 (*7.5*)**	**11.96 (*7.2*)**
Cache LDLM	**34.15 (*12.9*)**	**19.32 (*9.1*)**	**14.47 (*8.4*)**	**11.91 (*7.6*)**

WER Using Different Vocabularies

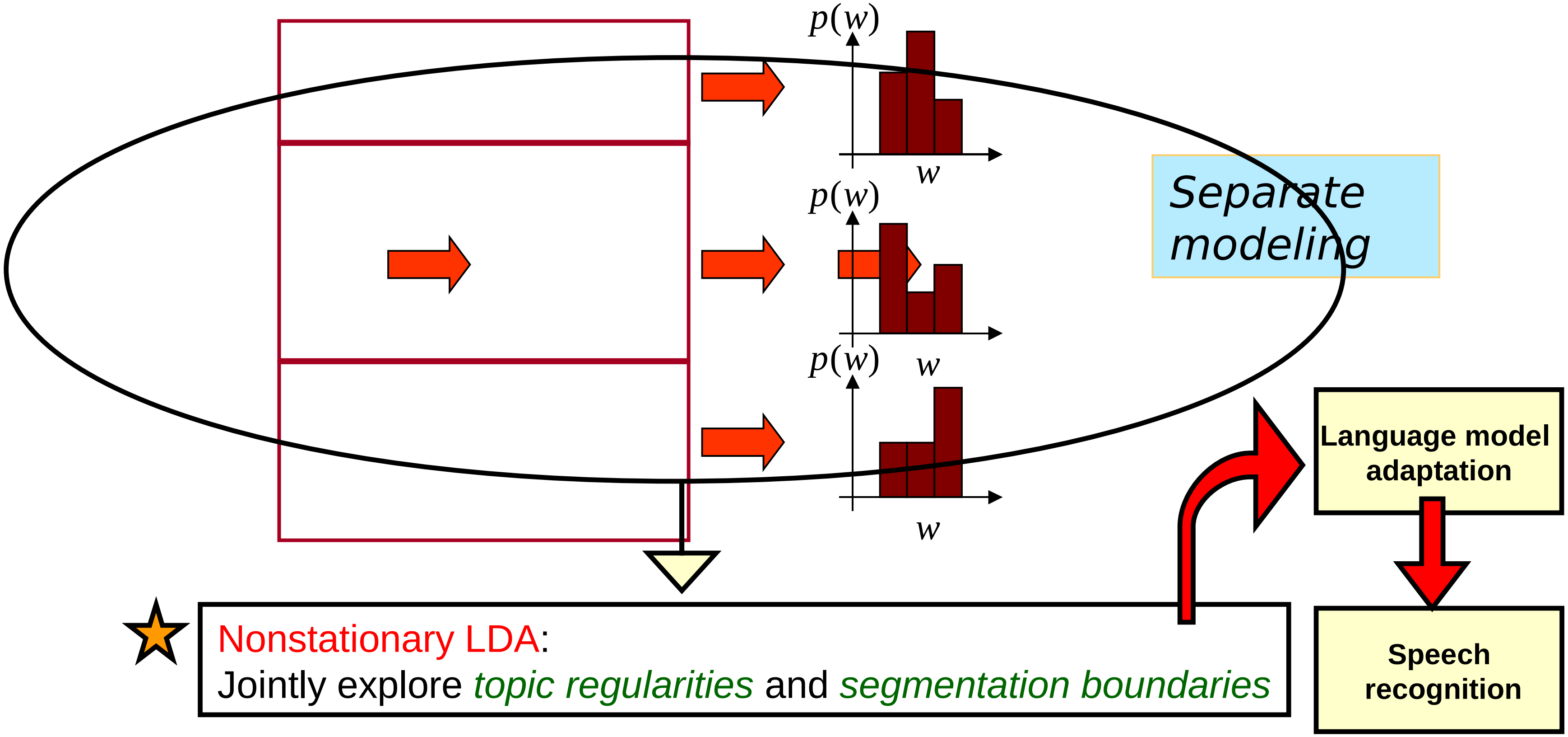


Bayesian Topic Language Model

Nonstationary Latent Dirichlet Allocation for Speech Recognition, INTERSPEECH 2009

Motivation

- *Words* in a document should be *non-stationary*.
  - The style of the same words is varied in different segments.



Separate modeling

Nonstationary LDA:
Jointly explore *topic regularities* and *segmentation boundaries*

Language model adaptation

Speech recognition

New Speech Recognition

$$\hat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} p_{\Lambda}(X|W)\boxed{p_{\text{composite}}(W)}$$

$$p_{n-\text{gram}}(W)$$

$$p_{\text{topic}}(W)$$

Model Construction

- Generation process of a document

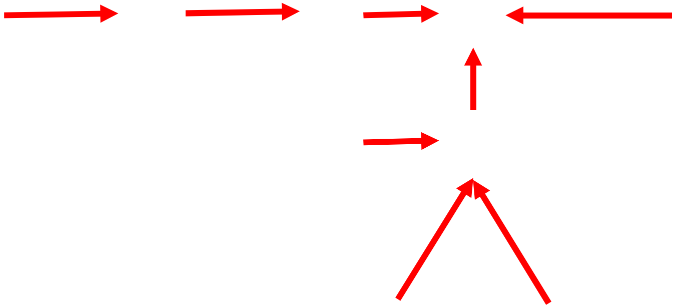
  1. Choose a topic mixture vector $\boldsymbol{\theta} \sim Dir(\boldsymbol{\alpha})$

  2. For each of the *N* words w_n :

     - Choose a topic $z_n \sim multinomial(\boldsymbol{\theta})$
     - If $n = 1$, choose a state $s_1 \sim multinomial(\boldsymbol{\pi})$
     - else choose a state $s_n \sim multinomial(\mathbf{a}_{s_{n-1}})$
     - Choose a word w_n from $p(w_n | z_n, s_n, \mathbf{B})$

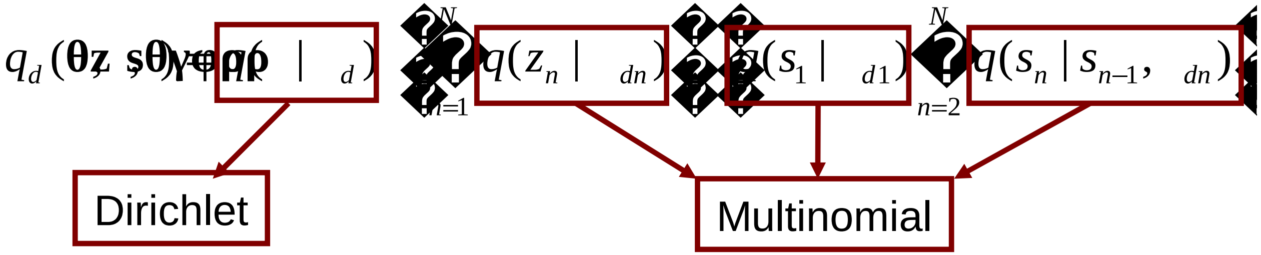
$$p(\mathbf{w} | \boldsymbol{\alpha}, \mathbf{B}, \boldsymbol{\pi}, \mathbf{A})$$

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



Model Inference

- Marginal likelihood is intractable.
  - variational inference
- True posterior $p(\boldsymbol{\theta},\mathbf{z},\mathbf{s}\,|\,\mathbf{w}_d,\boldsymbol{\alpha},\mathbf{B},\boldsymbol{\pi},\mathbf{A})$ is approximated by the variational distribution



$$q_d(\boldsymbol{\theta}\mathbf{z},\mathbf{s}\boldsymbol{\theta}\boldsymbol{\gamma}\boldsymbol{\varphi}\boldsymbol{\rho}\ \boxed{q(\ \ |\ \ _d)}\ \text{�}\prod_{n=1}^{N}\text{�}\ \boxed{q(z_n\,|\ \ _{dn})}\ \text{��}\ \boxed{q(s_1\,|\ \ _{d1})}\ \text{�}\prod_{n=2}^{N}\ \boxed{q(s_n\,|\,s_{n-1},\ \ _{dn})}\ \text{�}$$

- Lower bound of log marginal likelihood is calculated by

$$\begin{aligned} L(\boldsymbol{\alpha},\mathbf{B},\boldsymbol{\pi},\mathbf{A};\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho}) = \overset{D}{\underset{d=1}{\mathrm{P}}}\{ &< \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}\mathbf{z}\boldsymbol{\theta}) >_{q_d} + < \log p(\ \ |\ \ ) >_{q_d} \\ &< \log p(\mathbf{w}_d\,|\,\mathbf{z},\mathbf{s},\mathbf{B}) >_{q_d} + < \log p(\mathbf{s}\boldsymbol{\pi}|\mathbf{A},\ \ ) >_{q_d} \\ &- < \log q(\boldsymbol{\theta}|\boldsymbol{\gamma}_d\mathbf{z}\boldsymbol{\varphi}\mathbf{s}\boldsymbol{\rho}) >_{q_d} - < \log q(\ \ |\ \ _d) >_{q_d} - < \log q(\ \ |\ \ _d) >_{q_d} \} \end{aligned}$$

Variational Viterbi Decoding

- The best state sequence $\hat{\mathbf{s}}_d$ of a document $\mathbf{w}_d$ is obtained by

$$\hat{\mathbf{s}}_d = \arg\max_{\mathbf{s}} p(\mathbf{w}_d, \mathbf{s} | \alpha, \pi, \mathbf{A}, \mathbf{B})$$

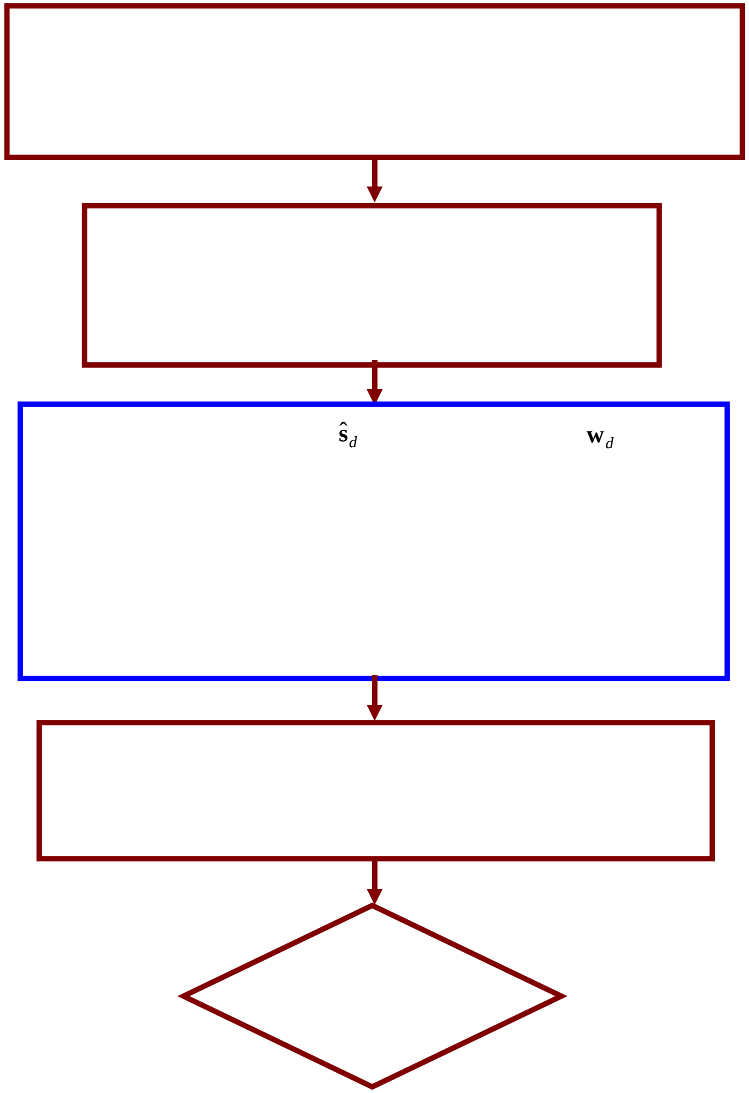
$$= \arg\max_{\mathbf{s}} < \log\{ p(\mathbf{w}_d | \mathbf{z}, \mathbf{s}, \mathbf{B}) p(\mathbf{s} | \pi, \mathbf{A}) \} >_{q(\mathbf{z})}$$

$$= \arg\max_{\mathbf{s}} \{ \log p(\mathbf{s} | \pi, \mathbf{A}) + < \log p(\mathbf{w}_d | \mathbf{z}, \mathbf{s}, \mathbf{B}) >_{q(\mathbf{z})} \}$$

$$= \arg\max_{\mathbf{s}} \left\{ \log \pi_{s_1} + \sum_{n=2}^{N} \log a_{s_{n-1} s_n} + \sum_{n=1}^{N} \sum_{k=1}^{K} \boxed{\phi_{dkn} \log b_{s_n k w_n}} \right\}$$

new output probability

Viterbi VB-EM Procedure



NLDA for Speech Recognition

- Using the best state sequence $\hat{\mathbf{s}}$ and the estimated variational parameter $\hat{\boldsymbol{\gamma}}$ for test document, we calculate NLDA unigram and use it for language model adaptation.

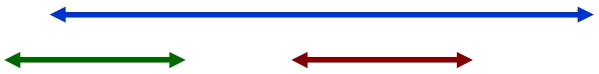
$$p_{\hat{s}}(w) = \int \sum_{k=1}^{K} p(w\,|\,k,\hat{s},\mathbf{B})\,p(k\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})d\boldsymbol{\theta}$$

$$\approx \int \sum_{k=1}^{K} b_{\hat{s}kw}\theta_k q(\theta_k\,|\,\hat{\gamma}_k)d\theta_k = \sum_{k=1}^{K} b_{\hat{s}kw} E_q[\theta_k\,|\,\hat{\gamma}_k] = \frac{\sum_{k=1}^{K} b_{\hat{s}kw}\hat{\gamma}_k}{\sum_{j=1}^{K}\hat{\gamma}_j}$$

$$\Rightarrow \quad \hat{p}(w\,|\,h) = \lambda p_{n-\mathrm{gram}}(w\,|\,h) + (1-\lambda)p_{\hat{s}}(w)$$

Experimental Results on WSJ

- NLDA for calculating sentence probability



- Relax the limitation of starting and ending states when searching the best state sequence.
- Comparison of perplexities and WERs

`	Baseline	LDA	NLDA
Perplexity	46.6	45.1	43.3
WER (%)	5.38	5.17	5.14

Conclusions

- *Online adaptation* was performed to continuously learn the unknown variations in speech recognition.
- Adopting *conjugate prior* was feasible to obtain the *closed-form solution* and perform the *hyperparameter evolution*.
- Robustness of a decision rule was strengthened by applying *BPC decision* rule. Ill-posed problem is tackled.
- We applied the *evidence framework* to HMM training, which automatically learnt the *priors* and their posteriors from data.
- *Bayesian sparse learning* was performed to establish the regularized large margin HMMs.

Conclusions

- A latent Dirichlet language model was developed for Bayesian topic modeling in *n-gram* level rather than in document level.
- A Markov chain was embedded in NLDA to characterize the *temporal word variations* in a document. Document segmentation was performed.
- A new NLDA document model was built for language model adaptation.
- Bayesian learning approaches are not only feasible to speech recognition but also to *other pattern recognition* applications.

Future Works

- We are extending the evidence framework for construction of different *probabilistic models* with/without latent variables.
- We are developing *kernel* method for Bayesian large margin HMMs. The *evidence framework* will be further developed for higher level inference.
- A Bayesian topic *cache* language model will be constructed.
- Conduct extensive experiments on a large-scale corpora consisting of spoken documents.
- Apply NLDA for spoken document *retrieval* and *summarization*.

Thank You!