Collapsed Variational Inference for HDP

Yee W. Teh, David Newman and Max Welling

Published on NIPS 2007

Discussion led by
Iulian Pruteanu

Outline

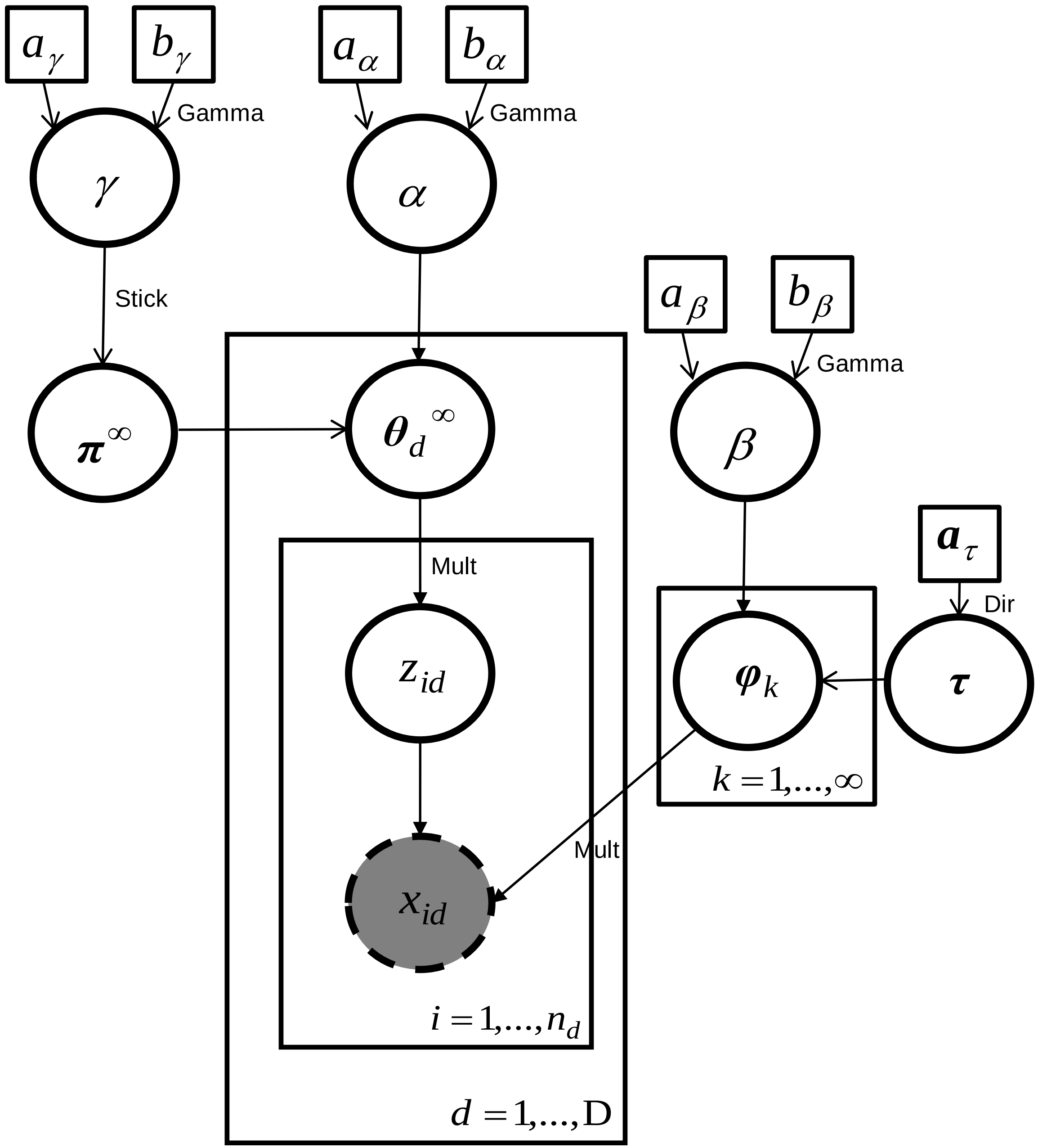
- Introduction
- Hierarchical Bayesian model for LDA
- Collapsed VB inference for HDP = CV-HDP
- Experiments
- Discussion

Introduction

- For Dirichlet-multinomial models (such as LDA or HDP) the inference method of choice is 'typically' collapsed Gibbs sampling but it seems to be necessary to consider alternatives to sampling.
- Teh *et. al.* (NIPS 2006) proposed an improved VB approximation for LDA (CV-LDA) based on the idea of collapsing (integrating out model parameters while assuming other latent variables independent).
- Previous work on collapsed variational Latent Dirichlet Allocation (LDA) did not consider model selection and inference for hyperparameters.
- Advantages of CV-HDP over CV-LDA:
  - the optimal number of variational components is not finite (the number of topics is unlimited);
  - the posterior distribution over hyperparameters of Dirichlet variables is treated exactly.

- The algorithm is fully Bayesian; the only assumptions made are independencies among latent topic variables and hyperparameters.
- CVB algorithm, making use of some approximations, is easy to implement and more accurate than standard VB (by collapsing model variables, the uncertainty upon the model is reduced).

Hierarchical Bayesian model for LDA(1/3)



Hierarchical Bayesian model for LDA(2/3)

$\boldsymbol{x} = \{x_{id}\}$ - observed words

$\boldsymbol{z} = \{z_{id}\}$ - latent variables (topic indices)

$\boldsymbol{\theta}_d = \{\theta_{dk}\}$ - mixing proportions

$\boldsymbol{\varphi}_k = \{\phi_{kw}\}$ - topic parameters

D - number of documents

K - number of topics

$$\boldsymbol{\theta}_j \mid \boldsymbol{\pi}, \alpha \sim Dir(\alpha \boldsymbol{\pi})$$

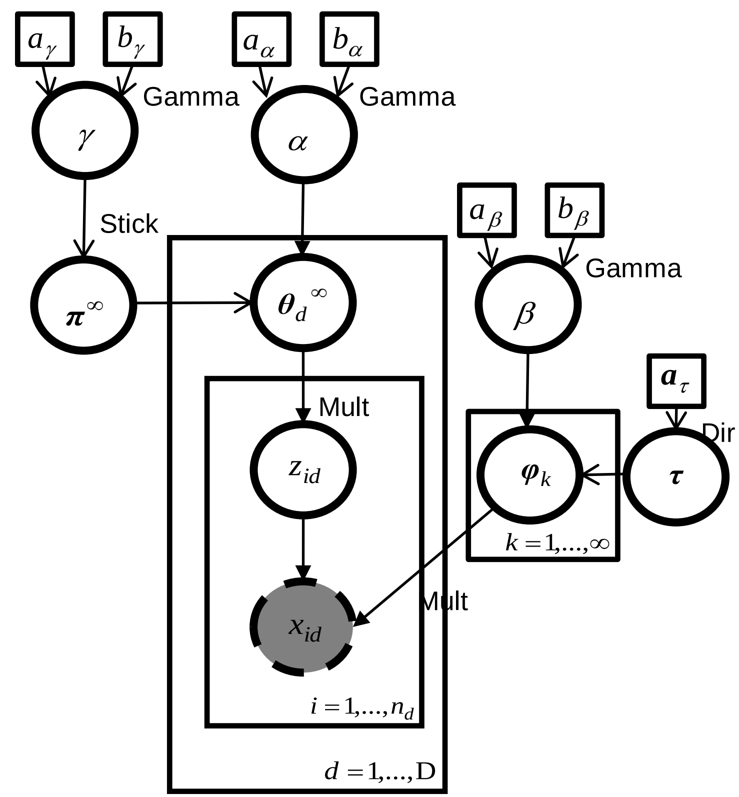
$$\boldsymbol{\varphi}_k \mid \boldsymbol{\tau}, \beta \sim Dir(\beta \boldsymbol{\tau}) \equiv \mathrm{H}(base\ distribution)$$

$$z_{id} \mid \boldsymbol{\theta}_d \sim Mult(\boldsymbol{\theta}_d)$$

$$x_{id} \mid z_{id}, \phi_{z_{id}} \sim Mult(\phi_{z_{id}})$$

$$\boldsymbol{\pi} = Stick(\gamma)$$

$$\boldsymbol{\tau} = Dir(\boldsymbol{a}_{\boldsymbol{\tau}})$$



Hierarchical Bayesian model for LDA(3/3)

In the normal Dirichlet process notation, we would equivalently have:

$$G_0 \sim DP(\gamma, \boldsymbol{H}) \qquad G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

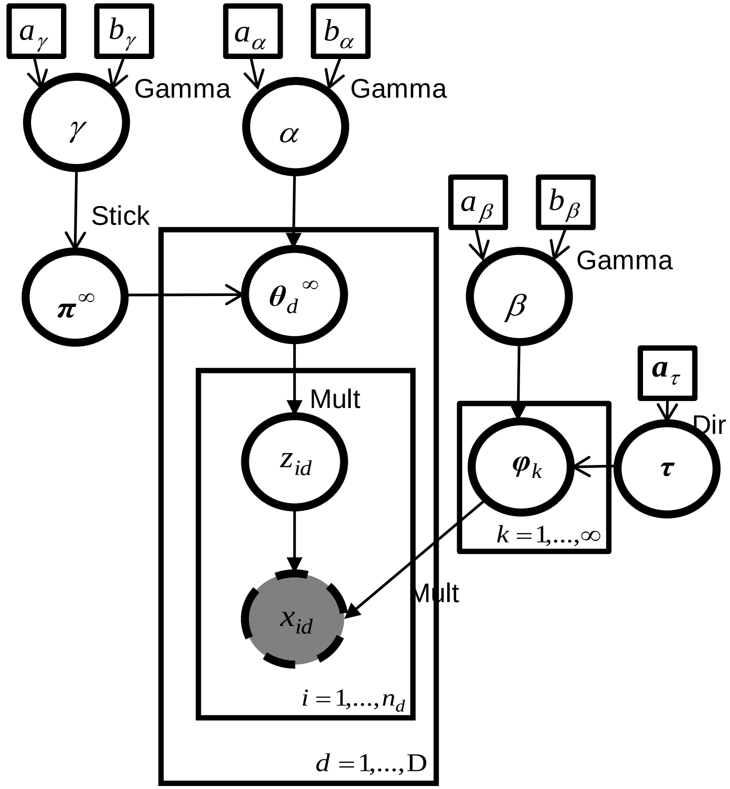
$$G_d \sim DP(\alpha, G_0) \qquad G_d = \sum_{k=1}^{\infty} \theta_{dk} \delta_{\phi_k}$$

$$\boldsymbol{H} \equiv Dir(\beta \boldsymbol{\tau}) \qquad \boldsymbol{\pi} = Stick(\gamma)$$

$$\boldsymbol{\theta}_d \sim Dir(\alpha \boldsymbol{\pi})$$

$$\boldsymbol{\varphi}_k \sim Dir(\beta \boldsymbol{\tau})$$

$$\boldsymbol{\tau} \sim Dir(\boldsymbol{a}_\tau)$$



CV inference for HDP (1/3)

In variational Bayesian approximation, we assume a factorized form for the posterior approximating distribution. However it is not a good assumption since changes in model parameters ($\boldsymbol{\theta}, \boldsymbol{\varphi}$) will have a considerable impact on latent variables ($\boldsymbol{z}$).

CVB is equivalent to marginalizing out the model parameters $\boldsymbol{\theta}, \boldsymbol{\varphi}$ before approximating the posterior over the latent variable $\boldsymbol{z}$.

The exact implementation of CVB has a closed form and it seems to be computationally practical.

The authors use the **Gaussian approximation** (which worked very accurately in the CV-LDA paper, as well).

CV inference for HDP (2/3)

$$p(\boldsymbol{z}, \boldsymbol{x} \mid \alpha, \beta, \gamma, \boldsymbol{\pi}, \boldsymbol{\tau}) = \prod_{d=1}^{D} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{d\cdot\cdot})} \prod_{k=1}^{K} \frac{\Gamma(\alpha\pi_k + n_{nd\cdot})}{\Gamma(\alpha\pi_k)} \cdot \prod_{k=1}^{K} \frac{\Gamma(\beta)}{\Gamma(\beta + n_{\cdot k\cdot})} \prod_{w=1}^{W} \frac{\Gamma(\beta\tau_w + n_{\cdot kw})}{\Gamma(\beta\tau_w)}$$

$$n_{dkw} = \#\{i : x_{id} = w, z_{id} = k\}$$

K = index such that $z_{id} \leq \mathrm{K}$

$$p(\mathbf{z}, \mathbf{x}, \boldsymbol{\eta}, \boldsymbol{\xi}, \mathbf{s}, \mathbf{t} | \alpha, \beta, \gamma, \pi, \tau)$$

$$= \prod_{d=1}^{D} \frac{\eta_d^{\alpha-1}(1-\eta_d)^{n_{d\cdot\cdot}-1} \prod_{k=1}^{K} \left[{n_{dk\cdot} \atop s_{dk}} \right] (\alpha\pi_k)^{s_{dk}}}{\Gamma(n_{d\cdot\cdot})} \prod_{k=1}^{K} \frac{\xi_k^{\beta-1}(1-\xi_k)^{n_{\cdot k\cdot}-1} \prod_{w=1}^{W} \left[{n_{\cdot kw} \atop t_{kw}} \right] (\beta\tau_w)^{t_{kw}}}{\Gamma(n_{\cdot k\cdot})}$$

The form of the variational posterior:

$$q(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \mathbf{s}, \mathbf{t}, \alpha, \beta, \gamma, \tau, \pi) = q(\alpha)q(\beta)q(\gamma)q(\tau)q(\pi)q(\boldsymbol{\eta}, \boldsymbol{\xi}, \mathbf{s}, \mathbf{t} | \mathbf{z}) \prod_{d=1}^{D} \prod_{i=1}^{n_{d\cdot\cdot}} q(z_{id})$$

CV inference for HDP (3/3)

Variational updates of the hyperparameters:

$$q(\alpha) \propto \alpha^{a_\alpha + \mathbb{E}[s_{\cdot\cdot}] - 1} e^{-\alpha(b_\alpha - \sum_d \mathbb{E}[\log \eta_d])}$$

$$q(\beta) \propto \beta^{a_\beta + \mathbb{E}[t_{\cdot\cdot}] - 1} e^{-\beta(b_\beta - \sum_k \mathbb{E}[\log \xi_k])}$$

$$q(\gamma) \propto \gamma^{a_\gamma + K - 1} e^{-\gamma(b_\gamma - \sum_{k=1}^{K} \mathbb{E}[\log(1-\tilde{\pi}_k)]}$$

$$q(\tilde{\pi}_k) \propto \tilde{\pi}_k^{\mathbb{E}[s_{\cdot k}]}(1-\tilde{\pi}_k)^{\mathbb{E}[\gamma] + \mathbb{E}[s_{\cdot > k}] - 1}$$

$$q(\tau) \propto \prod_{w=1}^{W} \tau_w^{a_\tau + \mathbb{E}[t_{\cdot w}] - 1}$$

Variational updates of auxiliary variables:

$$q(\eta_d|\mathbf{z}) \propto \eta_d^{\mathbb{E}[\alpha]-1}(1-\eta_d)^{n_{d\cdot\cdot}-1}$$

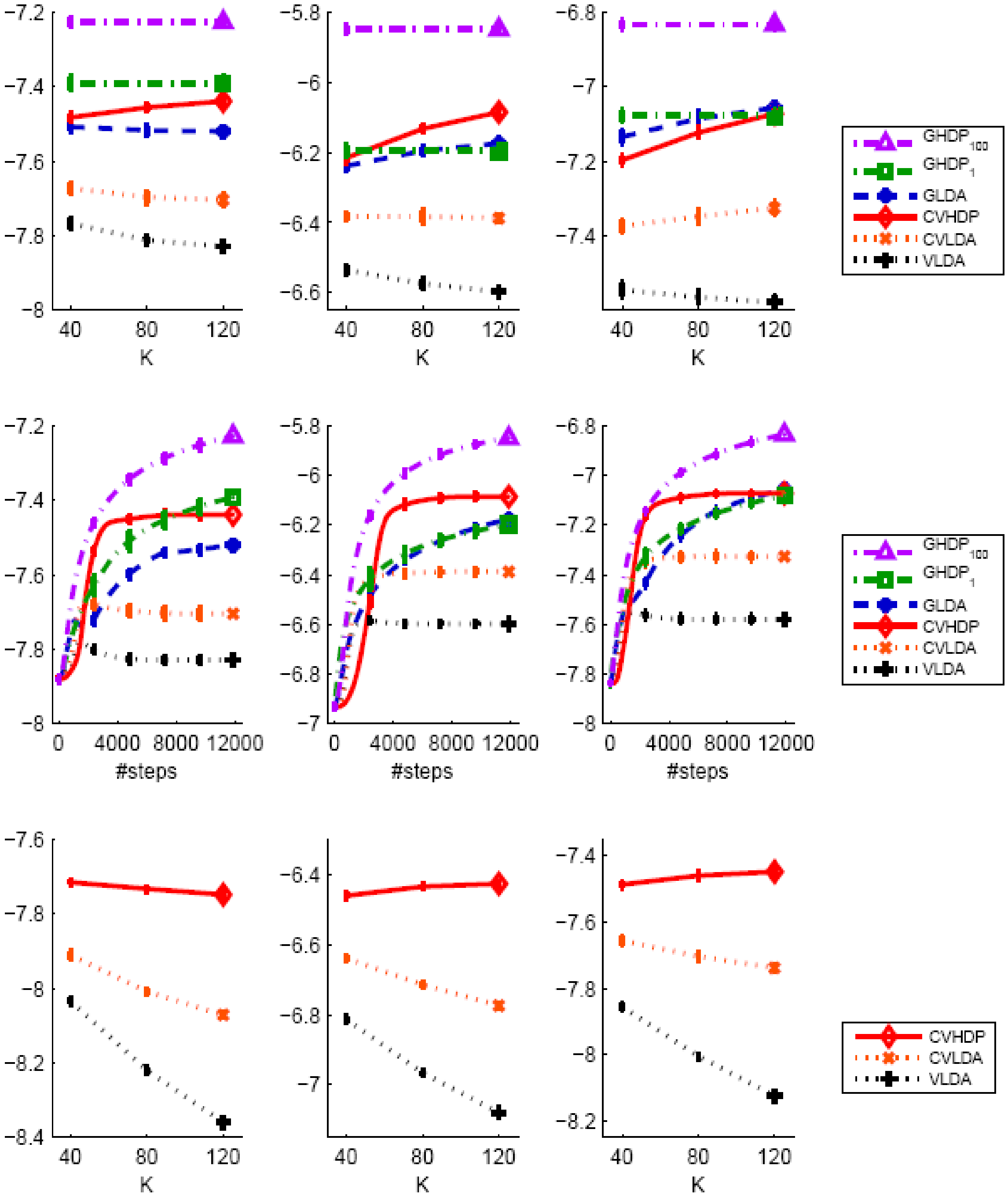
$$q(\xi_k|\mathbf{z}) \propto \xi_k^{\mathbb{E}[\beta]-1}(1-\xi_k)^{n_{\cdot k \cdot}-1}$$

$$q(s_{dk} = m|\mathbf{z}) \propto \left[{n_{dk\cdot} \atop m}\right] (\mathbb{G}[\alpha\pi_k])^m$$

$$q(t_{kw} = m|\mathbf{z}) \propto \left[{n_{\cdot kw} \atop m}\right] (\mathbb{G}[\beta\tau_w])^m$$

$$\begin{aligned} q(z_{id} = k) \propto & \mathbb{G}\left[\mathbb{G}[\alpha\pi_k] + n_{dk\cdot}^{\neg id}\right] \mathbb{G}\left[\mathbb{G}[\beta\tau_{x_{id}}] + n_{\cdot k x_{id}}^{\neg id}\right] \mathbb{G}\left[\mathbb{E}[\beta] + n_{\cdot k \cdot}^{\neg id}\right]^{-1} \\ \approx\propto & \left(\mathbb{G}[\alpha\pi_k] + \mathbb{E}[n_{dk\cdot}^{\neg id}]\right)\left(\mathbb{G}[\beta\tau_{x_{id}}] + \mathbb{E}[n_{\cdot k x_{id}}^{\neg id}]\right)\left(\mathbb{E}[\beta] + \mathbb{E}[n_{\cdot k \cdot}^{\neg id}]\right)^{-1} \\ & \exp\left(-\frac{\mathbb{V}[n_{dk\cdot}^{\neg id}]}{2(\mathbb{G}[\alpha\pi_k] + \mathbb{E}[n_{dk\cdot}^{\neg id}])^2} - \frac{\mathbb{V}[n_{\cdot k x_{id}}^{\neg id}]}{2(\mathbb{G}[\beta\tau_{x_{id}}] + \mathbb{E}[n_{\cdot k x_{id}}^{\neg id}])^2} + \frac{\mathbb{V}[n_{\cdot k \cdot}^{\neg id}]}{2(\mathbb{E}[\beta] + \mathbb{E}[n_{\cdot k \cdot}^{\neg id}])^2}\right) \end{aligned}$$

Experiments



A: results for KOS.

D=3,430 documents;

W=6,906; N=467,714 words

B: results for Reuters dataset.

D=8,433 documents;

W=4,593; N=566,298 words

10% for testing; 50 random runs

C: results for 20 Newsgroups.

D=4,726 documents;

W=8,424; N=437,850 words

VLDA=VB-LDA

GLDA=C-Gibbs-LDA

GHDP=Gibbs-HDP

K=truncation level

#steps=#iterations*K

Conclusions

- CV-HDP presents an improvement over CV-LDA by taking the infinite topic limit in the generative model and truncating the variational posterior and by inferring posterior distributions over the higher level variables.