

# Variational inference in a truncated Dirichlet process

David M. Blei and Michael I. Jordan  
U.C. Berkeley

December 4, 2003

## 1 The truncated Dirichlet process

The  $N$ -component truncated Dirichlet process ( $\text{DP}_N$ ) is defined in Ishwaran and James [2001] and converges almost surely to a true Dirichlet process ( $\text{DP}_\infty$ ). Like a full Dirichlet process, this distribution can be used as a nonparametric Bayesian prior in a mixture model. Ishwaran and James show that this approximation allows a blocking strategy in the corresponding Gibbs sampler which can be faster than the classical Gibbs samplers developed for the  $\text{DP}_\infty$  prior [Escobar and West, 1995]. In this paper, we develop a variational inference algorithm to approximate the posterior in a Bayesian mixture model with a  $\text{DP}_N$  prior.

An exponential family mixture model with  $\text{DP}_N$  prior on the natural parameter of the mixture component is illustrated in Figure 2. The random variables are distributed as follows:

$$\begin{aligned} p(V_n | \alpha) &= \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)} (1 - V_i)^{\alpha-1} \text{ for } n \in [1, N-1] \\ p(V_N = 1) &= 1 \\ p(\eta_n | \lambda) &= h(\eta_n) \exp\{\lambda_1 \eta_n + \lambda_2 (-a(\eta_n)) - a(\lambda)\} \\ p(K_d^1 = 1 | \mathbf{V}) &= V_1 \\ p(K_d^n = 1 | \mathbf{V}) &= (1 - V_1)(1 - V_2) \cdots (1 - V_{n-1})V_n \text{ for } n \in [2, N] \\ p(X_d | K_d, \boldsymbol{\eta}) &= \prod_{n=1}^N (h(X_d) \exp\{\boldsymbol{\eta}_n^T X_d - a(\boldsymbol{\eta}_n)\})^{K_d^n} \end{aligned}$$

Note that in the standard conjugate exponential set-up,  $\lambda$  has dimension  $\dim(\boldsymbol{\eta}) + 1$  and  $-a(\boldsymbol{\eta})$  is the last component of the sufficient statistic of  $\boldsymbol{\eta}$ .

## 2 Variational inference

Consider the log likelihood of a dataset  $\mathbf{X} = \{X_d\}_{d=1}^D$ :

$$\log p(\mathbf{X}) = \log \int \int p(\mathbf{V})p(\boldsymbol{\eta}) \left( \prod_{d=1}^D \sum_K p(K | \mathbf{V})p(X_d | K) \right) d\mathbf{V}d\boldsymbol{\eta}$$

This quantity can be bounded with Jensen's inequality as follows:

$$\begin{aligned} \log p(\mathbf{X}) &\geq \text{E} [\log p(\mathbf{V} | \alpha)] + \text{E} [\log p(\boldsymbol{\eta} | \lambda)] \\ &\quad + \sum_{d=1}^D \text{E} [\log p(K_d | \mathbf{V})] + \text{E} [\log p(X_d | K)] + H(q). \end{aligned} \tag{1}$$

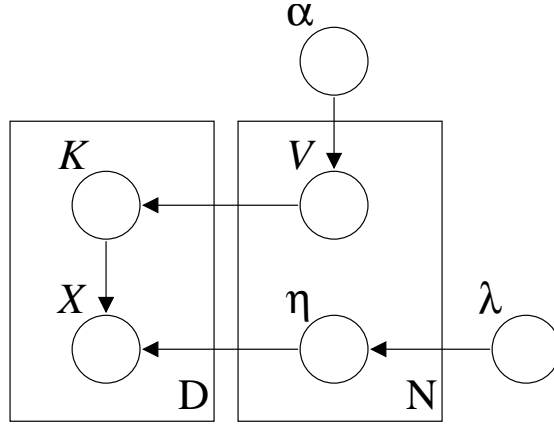


Figure 1: Graphical model representation of a mixture model with  $DP_N$  prior.

All expectations are taken with respect to a variational distribution on the latent variables  $\mathbf{V}$ ,  $\boldsymbol{\mu}$ , and  $\mathbf{K}$ :

$$q(\mathbf{V}, \boldsymbol{\mu}, \mathbf{K}) = \prod_{n=1}^N q(V_n | \gamma_n) \prod_{n=1}^N q(\eta_n | \tau_n) \prod_{d=1}^D q(K_d | \phi_d), \quad (2)$$

where  $\gamma_n$  are the Beta parameters for the distribution on  $V_n$ ,  $\tau_n$  are natural parameters for the distribution on  $\eta_n$ , and  $\phi_d$  are multinomial parameters for the distribution on  $K_d$ .

In variational inference, we optimize the bound in Eq. (1) with respect to the variational parameters. As shown in Jordan et al. [1999], this finds the setting of the parameters that minimizes the KL divergence between  $q$  and the true posterior on the latent variables.

The first, second, fourth, and fifth terms in Eq. (1) correspond to standard quantities in a directed graphical model with exponential family distributions. We rewrite the third term with indicator random variables:

$$\mathbb{E} [\log p(K_d | \mathbf{V})] = \mathbb{E} \left[ \log \left( \prod_{n=1}^N (1 - V_n)^{\mathbf{1}[K_d > n]} (V_n)^{\mathbf{1}[K_d = n]} \right) \right].$$

This expectation simplifies to:

$$\mathbb{E} [\log p(K_d | \mathbf{V})] = \sum_{n=1}^N q(K_d > n) \mathbb{E} [\log(1 - V_n)] + q(K_d = n) \mathbb{E} [\log V_n],$$

where:

$$\begin{aligned} q(K_d = n) &= \phi_{d,n} \\ q(K_d > n) &= \sum_{m=n+1}^N \phi_{d,m} \\ \mathbb{E} [\log V_n] &= \Psi(\gamma_{n,1}) - \Psi(\gamma_{n,1} + \gamma_{n,2}) \\ \mathbb{E} [\log(1 - V_n)] &= \Psi(\gamma_{n,2}) - \Psi(\gamma_{n,1} + \gamma_{n,2}) \end{aligned}$$

In optimization, the updates for  $\tau_n$  and  $\gamma_n$  follow the standard recipe for coordinate ascent variational inference with exponential family distributions in a conjugate setting [Ghahramani and Beal, 2001, Attias,

2000].

$$\begin{aligned}\gamma_{n,1} &= 1 + \sum_d \phi_{d,n} \\ \gamma_{n,2} &= \alpha + \sum_d \sum_{m=n+1}^N \phi_{d,m} \\ \tau_{n,1} &= \lambda_1 + \sum_d \phi_{d,n} X_d \\ \tau_{n,2} &= \lambda_2 + \sum_d \phi_{d,n},\end{aligned}$$

The update for  $\phi_{d,n}$  is:

$$\phi_{d,n} \propto \exp \left\{ \mathbb{E} [\log V_n | \gamma_n] + \mathbb{E} [\eta_n | \tau_n]^T X_d - \mathbb{E} [a(\eta_n) | \tau_n] - \sum_{m=n}^N \mathbb{E} [\log(1 - V_n) | \gamma_n] \right\},$$

### 3 Gibbs sampling

Gibbs sampling in a full Dirichlet process (i.e.,  $N = \infty$ ) under a conjugate prior is straightforward. One can integrate out all the parameters except  $K_d$ :

$$p(K_d = k | \mathbf{X}, \mathbf{K}_{-d}) \propto p(X_d | \mathbf{X}_{-d}, \mathbf{K}_{-d}, K_d = k) p(K_d = k | \mathbf{K}_{-d})$$

Let  $\hat{\lambda}_1 = \lambda_1 + \sum_j 1(K_j = k)X_j$  and  $\hat{\lambda}_2 = \lambda_2 + N - 1$ . The first term is:

$$\begin{aligned}p(X_d | \mathbf{X}_{-d}, \mathbf{K}_{-d}, K_d = k) &= \prod_{k'} \int p(X_d | \eta_{k'}) p(\eta_{k'} | \mathbf{X}_{-d}, \mathbf{K}_{-i}) d\eta_{k'} \\ &\propto \int \exp \left\{ \eta_k X_d - a(\eta_k) + \hat{\lambda}_1 \eta_k + \hat{\lambda}_2 (-a(\eta_k)) - a(\hat{\lambda}_1, \hat{\lambda}_2) \right\} d\eta_k \\ &= \exp \left\{ a(\hat{\lambda}_1 + X_d, \hat{\lambda}_2 + 1) - a(\hat{\lambda}_1, \hat{\lambda}_2) \right\},\end{aligned}$$

which is simply a ratio of normalizing constants. The second term comes from the partition structure of the Dirichlet process. When  $k$  is a previously seen component:

$$p(K_d = k | \mathbf{K}_{-d}) = \frac{n_k}{\alpha + N - 1}.$$

When  $k$  is an unvisited component:

$$p(K_d = k | \mathbf{K}_{-d}) = \frac{\alpha}{\alpha + N - 1}.$$

### 4 Experiments

In an experiment, we sample a training set of 100 data points from a Dirichlet process mixture model. Then, we sample a test set of another 100 “101st” data points (i.e., a next point in the process which is only dependent on the previous 100 points).

We apply three inference techniques: the collapsed Gibbs sampler described in Section 3, the blocked Gibbs sampler for a truncated Dirichlet process described in Ishwaran and James [2001], and the variational approximation to the truncated Dirichlet process described in Section 1.

In each case, we run the variational approximation to convergence and compute an approximate predictive distribution for the next data point. We then run the Gibbs samplers until the approximate likelihood

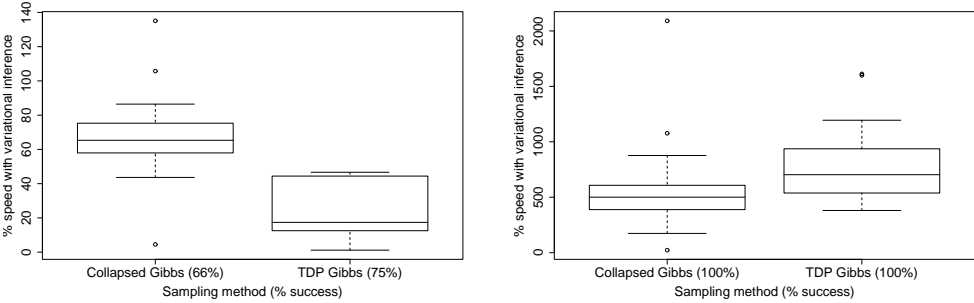


Figure 2: (Left) Results on Dirichlet-Multinomial set-up with  $N=100$ . (Right) Results on one-dimensional Normal-Gamma-Normal set-up.

assigned to the test set is the same as that for the variational approach. If a Gibbs sampler cannot reach this likelihood after 1000 times the convergence time for the variational algorithm, then it is considered a failure.

Some results are illustrated in Figure 4 for a Normal-Gamma-Normal mixture model and a Dirichlet-Multinomial mixture model. These box plots show 50 experiments in each setting and we have plotted how much faster/slower (percentage) the variational algorithm is compared with the samplers. A preliminary observation is that in low dimensional data (e.g., one dimensional Gaussian), the samplers are faster. However, in high dimensional data (e.g., multinomial with  $N=100$ ), the variational algorithm is faster. Note that we only plot successful trials as defined above — the success rate for the samplers is indicated in parenthesis.

## References

H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, 2000.

M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

Z. Ghahramani and M. Beal. Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems 13*, pages 507–513, 2001.

J. Ishwaran and L. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–174, 2001.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.