Supplement to Truncation-free Online Variational Inference for Bayesian Nonparametric Models

Chong Wang[*]
Machine Learning Department
Carnegie Mellon University
chongw@cs.cmu.edu

David M. Blei
Computer Science Department
Princeton Univeristy
blei@cs.princeton.edu

S.1 Explanation using Expectation Propagation (EP)

Our goal is to approximate the posterior distribution

$$p(\beta, z_{1:n} \,|\, x_{1:n}, \eta) \propto p(\beta, z_{1:n}, x_{1:n} \,|\, \eta)$$

using a fully factorized distribution

$$q(\beta, z_{1:n}) = q(\beta) \textstyle\prod_{i=1}^{n} q(z_i).$$

Different from the mean-field approach, Expectation Propagation (EP) tries to minimize the following KL-divergence [4, 5],

$$\mathrm{KL}_{\mathrm{EP}}(p||q) = \int \sum_{z_{1:n}} p(\beta, z_{1:n} \,|\, x_{1:n}, \eta) \log \frac{p(\beta, z_{1:n} \,|\, x_{1:n}, \eta)}{q(\beta, z_{1:n})} \mathrm{d}\beta.$$

First, taking the derivative of $\mathrm{KL}_{\mathrm{EP}}(p||q)$ w.r.t. $q(z_i)$ and setting it to zero gives

$$q(z_i) \propto \int \sum_{z_{-i}} p(\beta, z_{1:n} \,|\, x_{1:n}, \eta) \mathrm{d}\beta = \int p(z_i, x_i \,|\, \beta) p(\beta \,|\, x_{-i}, \eta) \mathrm{d}\beta,$$

where z_{-i} indicates $\{z_j, j = 1, \ldots, n, \text{but } j \neq i\}$. ($x_{-i}$ is similarly defined.) This is intractable. If we use $q(\beta)$ as an approximation to the true marginal posterior $p(\beta \,|\, x_{-i}, \eta)$, and this gives,

$$q(z_i) \propto \int p(z_i, x_i \,|\, \beta) p(\beta \,|\, x_{-i}, \eta) \mathrm{d}\beta \approx \int p(z_i, x_i \,|\, \beta) q(\beta) \mathrm{d}\beta = \mathbb{E}_{q(\beta)} \left[p(x_i, z_i \,|\, \beta) \right],$$

which is precisely the definition of $q(z_i)$ as in Eq. 6 in the main paper.

Next taking the derivative of $\mathrm{KL}_{\mathrm{EP}}(p||q)$ w.r.t. $q(\beta)$ and setting it to zero gives

$$q(\beta) = \sum_{z_{1:n}} p(\beta, z_{1:n} \,|\, x_{1:n}, \eta) = \sum_{z_{1:n}} p(\beta \,|\, z_{1:n}, x_{1:n}, \eta) p(z_{1:n} \,|\, x_{1:n}, \eta),$$

This is intractable. We thus use $q(z_{1:n}) = \prod_{i=1}^{n} q(z_i)$ as an approximation to the true marginal posterior $p(z_{1:n} \,|\, x_{1:n}, \eta)$, and this gives,

$$\begin{aligned} q(\beta) &\approx \sum_{z_{1:n}} p(\beta \,|\, z_{1:n}, x_{1:n}, \eta) q(z_{1:n}) = \mathbb{E}_{q(z_{1:n})} \left[p(\beta \,|\, z_{1:n}, x_{1:n}, \eta) \right] \\ &= \exp\left\{ \log \mathbb{E}_{q(z_{1:n})} \left[p(\beta \,|\, z_{1:n}, x_{1:n}, \eta) \right] \right\} \\ &\leq \exp\left\{ \mathbb{E}_{q(z_{1:n})} \left[\log p(\beta \,|\, z_{1:n}, x_{1:n}, \eta) \right] \right\} \\ &\propto \exp\left\{ \mathbb{E}_{q(z_{1:n})} \left[\log p(z_{1:n}, x_{1:n}, \beta, \eta) \right] \right\}, \end{aligned}$$

[*]Work was done when the author was with Princeton University.

where the inequality comes from the concavity of the log function. Note that if we assumed $q(z_i)$ is a peaky distribution; the inequality is almost an equality.[1] We finally have

$$q(\beta) \propto \exp\left\{\mathbb{E}_{q(z_{1:n})}\left[\log p(z_{1:n}, x_{1:n}, \beta, \eta)\right]\right\},$$

which is the update for $q(\beta)$ as in Eq. 7 in the main paper.

S.2 Truncation-free online variational inference for the HDP

Hierarchical Dirichlet process (HDP) topic models [6] can be summarized using the stick-breaking construction as follows,

1. Draw top-level topics θ_k and sticks π_k for $k = 1, 2, \cdots$,

$$\theta_k \sim \mathrm{Dirichlet}(\eta),$$
$$\pi_k = \bar{\pi}_k \textstyle\prod_{l=1}^{k-1}(1-\bar{\pi}_l), \;\; \bar{\pi}_k \sim \mathrm{theta}(1, a)$$

2. For each document t, draw document-level topic proportions ϕ_t,[2]

$$\phi_t \sim \mathrm{Dirichlet}(b\pi).$$

   For each word x_{tn} in document t,
   (a) Draw the topic index $z_{tn} \sim \mathrm{Mult}(\phi_t)$.
   (b) Draw the word $x_{tn} \sim \mathrm{Mult}(\theta_{z_{tn}})$.

Unfortunately, topic proportions ϕ_t is not conjugate to sticks π. We adopt an auxiliary variable approach proposed in [6]. The conditional distribution of of $z_t \triangleq z_{t,1:n_t}$,

$$p(z_t|\pi) = \int p(z_t|\phi_t)p(\phi_t|\pi)\mathrm{d}\phi_t = \frac{\Gamma(b)}{\Gamma(b+n_t)}\prod_k \frac{\Gamma(b\pi_k + n_{tk})}{\Gamma(b\pi_k)}, \tag{1}$$

where n_{tk} is the number of the words assigned to topic k in document t and n_t is the number of the words in document t. By introducing a random variable s_{tk}, the random number of occupied tables in a Chinese restaurant process with n_{tk} customers and concentration parameter $b\pi_k$, we have

$$p(z_t, s_t|\pi) = \frac{\Gamma(b)}{\Gamma(b+n_t)}\prod_k S(n_{tk}, s_{tk})(b\pi_k)^{s_{tk}}, \tag{2}$$

where $S(n,m)$ are unsigned Stirling numbers of the first kind [1]. Integrating out variable s_t in Eq. 2 gives the marginal distribution of z_t given sticks π in Eq. 1. Furthermore, variable s_t is conjugate to sticks π. Given the formulation in Eq. 2, we can sample s_{tk} given n_{tk} using,

$$p(s_{tk}|n_{tk}, b\pi_k) = \frac{\Gamma(b\pi_k)}{\Gamma(b\pi_k + n_{tk})} S(n_{tk}, s_{tk})(b\pi_k)^{s_{tk}}. \tag{3}$$

S.2.1 Online variational updates

The variational distribution for the global hidden variables θ_k, $\bar{\pi}_k$ is $k = 1, 2, \cdots$.

$$q(\theta, \bar{\pi}) = \textstyle\prod_k q(\theta_k|\lambda_k)q(\bar{\pi}_k|u_k, v_k),$$

where λ_k is the Dirichlet parameter and (u_k, v_k) is the theta parameter. Suppose we have obtained one sample[3] of the hidden variables s_t and z_t for document t. Then we have

$$\begin{aligned}
\lambda_{kw} &\leftarrow \lambda_{kw} + \rho_t(-\lambda_{kw} + \eta + Dn_{tkw}) \\
u_k &\leftarrow u_k + \rho_t(-u_k + 1 + Ds_{tk}) \\
v_k &\leftarrow v_k + \rho_t(-v_k + a + D\textstyle\sum_{j=k+1}^{\infty} s_{tj}).
\end{aligned}$$

[1] This is usually satisfied in practice—in mixture modeling, most data points belong to one mixture; in topic modeling, words in a document only belong to a very small set of topics [9].

[2] Here the Dirichlet distribution is a generalized version of its finite counterpart [7].

[3] The case with more than one samples can be similarly derived.

where n_{tkw} is number of times of word w assigned to topic k in document t and D is the total number of documents. Note that s_{tk} and n_{tkw} will be always 0 after k is larger than the current number of topics. This results in a property that $q(\theta_k)$ will remain as the prior distribution $p(\theta_k)$, if there is no word assignment. We therefore do not need to store those topics until they are instantiated. This also applies to the sticks parameter (u_k, v_k).

S.2.2 Gibbs sampling for the local variables

We use a collapsed Gibbs sampler similar to that in [6] to obtain samples for s_t and z_t. The idea is to use the variational distribution $q(\bar{\pi}|u, v)$ and $q(\theta|\lambda)$ as "priors". Note that θ can be marginalized out while $\bar{\pi}$ can not. We thus sample $\bar{\pi}$ jointly with s_t and z_t. We denote the vocabulary size as W.

Sampling z_{tn}. The conditional distribution for z_{tn} (word w) as follows,

$$p(z_{tn} = k|z_{-tn}, \lambda, \pi) \propto (n_{tk,-tn} + b\pi_k)\frac{n_{kw,-tn} + \lambda_{kw}}{n_{k,-tn} + \sum_w \lambda_{kw}}$$

When $k > T$, where T is the current number of topics, this becomes

$$p(z_{tn} = k|z_{-tn}, \lambda, \pi) \propto b\pi_k/W.$$

This implies

$$p(z_{tn} > T|z_{-tn}, \lambda, \pi) \propto b(1 - \textstyle\sum_{k=1}^{T} \pi_k)/W.$$

This indicates that we only need to sample z_{tn} up to $T + 1$. When a new topic is generated, we set $k = T + 1$, and sample $\bar{\pi}_{T+1} \sim \text{Beta}(1, a)$ and set $\pi_{T+1} = \bar{\pi}_{T+1} \prod_{k=1}^{T}(1 - \bar{\pi}_k)$.

Sampling s_{tk}. Sampling s_{tk} can be done using Eq. 3.

Sampling π. We sample π given the following conditional distribution,

$$p(\bar{\pi}_k) \propto \bar{\pi}_k^{u_k-1+\sum_{t\in\mathbb{S}} s_{tk}}(1 - \bar{\pi}_k)^{v_k-a+\sum_{t\in\mathbb{S}}\sum_{j=k+1}^{\infty} s_{tj}}.$$

We do not need to sample sticks $\bar{\pi}_k$ when $k > T$; they just come from the prior distribution.

S.3 Computing the held-out likelihood

The likelihood we want to compute is defined as

$$\text{likelihood}_{\text{pw}} \triangleq \log p(\mathcal{D}_{\text{test}} \,|\, \mathcal{D}_{\text{train}})/\textstyle\sum_{x_i\in\mathcal{D}_{\text{test}}} |x_i|.$$

Since this is intractable for both DP and HDP, we use following approximations. For both DP and HDP, we use the mean of the global variational distribution $q(\theta, \bar{\pi})$ to represent the inferred model. We ignore the unused components in our algorithm. This results in $\hat{\theta} = \mathbb{E}_{q(\theta)}[\theta]$ and $\hat{\pi} = \mathbb{E}_{q(\bar{\pi})}[\pi]$, both with finite dimensions. In other words, DP mixtures reduce to a finite mixture model and HDP mixtures reduce to LDA [2]. Then $\log p(\mathcal{D}_{\text{test}} \,|\, \mathcal{D}_{\text{train}})$ is approximated by

$$\log p(\mathcal{D}_{\text{test}} \,|\, \mathcal{D}_{\text{train}}) \approx \sum_{x_i\in\mathcal{D}_{\text{test}}} \log p(x_i \,|\, \hat{\theta}, \hat{\pi}).$$

For DP mixtures, the term $p(x_i \,|\, \hat{\theta}, \hat{\pi})$ is analytically tractable,

$$p(x_i \,|\, \hat{\theta}, \hat{\pi})_{\text{DP}} = \sum_k \hat{\pi}_k \prod_w \hat{\theta}_{kw}^{\sum_j 1[x_{ij}=w]}.$$

However, for HDP mixtures, the term $\log p(x_i \mid \hat{\theta}, \hat{\pi})$ is still intractable. We propose to use importance sampling. [8] shows that importance sampling can underestimate the probability, but usually gives the correct ranking of different models. To be concrete,

$$p(x_i \mid \hat{\theta}, \hat{\pi})_{\mathrm{HDP}} = \int p(\phi_i \mid \hat{\pi}) \prod_j \sum_{z_{ij}} p(z_{ij} \mid \phi_i) p(x_{ij} \mid \hat{\theta}, z_{ij}) \mathrm{d}\phi_i.$$

To approximate this integral, we first use a collapsed Gibbs sampler to sample topic assignments z_{ij}, then construct a proposal distribution over ϕ_i using these samples [3],

$$q(\phi_i) = \mathrm{Dirichlet}(\phi_i \mid \ldots, b\hat{\pi}_k + \sum_j 1[z_{ij} = k], \ldots).$$

Samples from $q(\phi_i)$ are used for importance sampling to approximate $p(x_i \mid \hat{\theta}, \hat{\pi})_{\mathrm{HDP}}$.

References

[1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[3] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences (PNAS)*, 2004.

[4] T. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence (UAI)*, pages 362–369, 2001.

[5] T. Minka. Divergence measures and message passing. Technical Report TR-2005-173, Microsoft Research, 2005.

[6] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2007.

[7] Y. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[8] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *International Conference on Machine Learning (ICML)*, 2009.

[9] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.