Latent IBP compound Dirichlet Allocation

Cedric Archambeau
Xerox Research Centre Europe
cedric.archambeau@xrce.xerox.com

Balaji Lakshminarayanan
Gatsby Computational Neuroscience Unit, UCL
balaji@gatsby.ucl.ac.uk

Guillaume Bouchard
Xerox Research Centre Europe
guillaume.bouchard@xrce.xerox.com

1 Introduction

Probabilistic topic models such as latent Dirichlet allocation (LDA) are widespread tools to analyse and explore large document corpora. Consider a corpus of D documents. LDA models these documents as a mixture of K discrete distributions over vocabulary words, which are called topics. Let $w_{id} \in \{1, \ldots, V\}$ denote the i^{th} word observed in document d and $z_{id} \in \{1, \ldots, K\}$ indicate the topic associated with this word. The generative model of LDA ignores the sequential structure of text and is defined as follows:

$$\begin{aligned} z_{id}|\boldsymbol{\theta}_d &\sim \text{Discrete}(\boldsymbol{\theta}_d), & \boldsymbol{\theta}_d &\sim \text{Dirichlet}(\alpha \mathbf{1}_K), \\ w_{id}|z_{id}, \{\boldsymbol{\phi}_k\} &\sim \text{Discrete}(\boldsymbol{\phi}_{z_{id}}), & \boldsymbol{\phi}_k &\sim \text{Dirichlet}(\beta \mathbf{1}_V), \end{aligned} \tag{1}$$

where $d = \{1, \ldots, D\}$, $k = \{1, \ldots, K\}$ and $i = \{1, \ldots, N_d\}$. To circumvent the model selection problem, its nonparametric Bayesian extension, which is known as the *hierarchical Dirichlet process* (HDP), [2], can be considered.

Recently sparsity-enforcing priors have been proposed to enable topics to be defined by a small subset of the vocabulary. Sparsity enforcing priors lead to compression as well an easier interpretation of the topics. A suitable candidate in the Bayesian nonparametric domain is the *IBP-compound-Dirichlet* distribution (ICD)[4], which has another interest beyond the simple sparsity-promoting advantage: it enables to decouple the topic inter-document frequency and intra-document frequency. Hence, unlike the HDP, the ICD can lead to very specific topics that might be very rare in a document corpus overall, but relate to a lot of words in the few documents that address this topic. The ICD assumes that a random infinite binary matrix generated by an *Indian Buffet Process*[1] prior "selects" a subset of the components before applying a symmetric Dirichlet prior on the subset of activated components. The ICD has been applied as a prior for the document-topic distribution in a model called the *Focused Topic Model* (FTM)[4] to enable a small number of topics allocated per document; it has also been applied as prior for the topic-word matrix in the *Sparse Topic Model* (STM)[3] to obtain topics with fewer words describing them.

In this work, we propose a novel unified inference algorithm for the two-parameter ICD model, which unlike previous methods is based on collapsed Gibbs sampling. Based on the degenerate Dirichlet we are able to alternatively sample activation variables and topic assignment variables. Currently, we are evaluating the advantages of ICD when infering sparse representation of documents in terms of topics (i.e. FTM), words (i.e. STM) or both on several benchmark data.

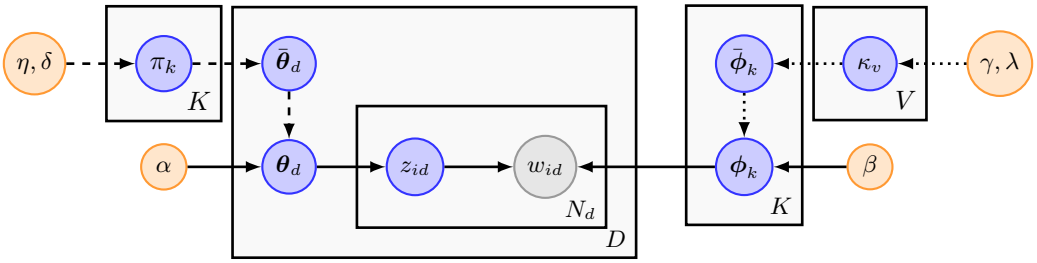

Figure 1: Graphical models for the different configurations (fixed K): LDA: solid arrows only, STM: solid + dotted arrows, FTM: solid + dashed arrows, LIDA: all arrows.

2 Two-parameter IBP compound Dirichlet prior

Let $\bar{\boldsymbol{\Theta}}$ be a binary matrix. We assume $\bar{\boldsymbol{\Theta}}$ serves as a prior for $\boldsymbol{\Theta}$ such that they share the same sparsity profile. The prior for $\boldsymbol{\Theta}$ can be formalised as follows:

$$\boldsymbol{\Theta}|\bar{\boldsymbol{\Theta}} \sim \prod_d \text{Dirichlet}_{\bar{\boldsymbol{\theta}}_d}(\alpha \mathbf{1}_K), \qquad \bar{\theta}_{kd}|\pi_k \sim \text{Bernoulli}(\pi_k), \qquad \pi_k \sim \text{Beta}(\tfrac{\eta\delta}{K}, \delta), \tag{2}$$

where the Dirichlet distribution is degenerate; it is defined over the simplex of dimension $\sum_k \bar{\theta}_{kd} - 1$:

$$\boldsymbol{\theta}_d|\bar{\boldsymbol{\theta}}_d \sim \text{Dirichlet}_{\bar{\boldsymbol{\theta}}_d}(\alpha \mathbf{1}_K) = \frac{\Gamma(\bar{\theta}_{\cdot d}\alpha)}{\prod_k \Gamma(\bar{\theta}_{kd}\alpha)^{\bar{\theta}_{kd}}} \prod_k \theta_{kd}^{(\alpha-1)\bar{\theta}_{kd}},$$

where $\Gamma(\cdot)$ is the gamma function. By convention we assume $\theta_{kd} = 0$ if it does not belong to the support (i.e. if $\bar{\theta}_{kd} = 0$).

The prior for π_k in (2) is a truncated (finite-dimensional) two-parameter IBP. The two-parameter IBP is a generalisation of the one-parameter IBP [1]. We can interpret $\delta > 0$ as a repulsion parameter; when it increases, the number of different features will increase for a given number of expected active features. When $\delta = 1$, we recover the one parameter IBP. In contrast to the one-parameter IBP, the two-parameter IBP decouples the expected number of active elements (topics) per row and the overall number of active elements (words). We claim this is a more realistic bag-of-words generative model for documents.

The IBP is obtained by integrating out $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and letting $K \to \infty$[1]. The two parameter IBP compound Dirichlet prior is given by

$$p(\boldsymbol{\Theta}|\alpha, \eta, \delta) = \textstyle\sum_{\bar{\boldsymbol{\Theta}}} p(\boldsymbol{\Theta}|\bar{\boldsymbol{\Theta}}, \alpha) P(\bar{\boldsymbol{\Theta}}|\eta, \delta). \tag{3}$$

From this expression we see that the prior is a mixture of degenerate Dirichlet distributions over simplices of different dimensions.

3 Latent IBP compound Dirichlet Allocation (LIDA)

We obtain the *latent IBP compound Dirichlet allocation* (LIDA) model by replacing the Dirichlet prior in LDA by a truncated IBP compound Dirichlet prior. The generative model is given by

$$\begin{aligned} &\boldsymbol{\theta}_d|\bar{\boldsymbol{\theta}}_d \sim \text{Dirichlet}_{\bar{\boldsymbol{\theta}}_d}(\alpha \mathbf{1}_K), && \bar{\theta}_{kd}|\pi_k \sim \text{Bernoulli}(\pi_k), && \pi_k \sim \text{Beta}(\tfrac{\eta\delta}{K}, \delta), \\ &\boldsymbol{\phi}_k|\bar{\boldsymbol{\phi}}_k \sim \text{Dirichlet}_{\bar{\boldsymbol{\phi}}_k}(\beta \mathbf{1}_V), && \bar{\phi}_{vk}|\kappa_v \sim \text{Bernoulli}(\kappa_v), && \kappa_v \sim \text{Beta}(\tfrac{\gamma\lambda}{V}, \lambda), \end{aligned} \tag{4}$$

For appropriate values of $\eta, \delta, \gamma, \lambda$, the ICD prior reduces to the non-degenerate Dirichlet distribution; hence, we can recover FTM, STM and LDA as special cases of LIDA (see Fig.1). Note that unlike [4], we use a two-parameter IBP in our FTM.

3.1 Inference

Integrating out $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_V)$, $\{\boldsymbol{\theta}_d\}$ and $\{\boldsymbol{\phi}_k\}$ leads to the following marginals:

$$P(\bar{\boldsymbol{\Theta}}) \propto \prod_k \frac{\mathcal{B}(\bar{\theta}_{k\cdot} + \frac{\eta\delta}{K}, D - \bar{\theta}_{k\cdot} + \delta)}{\mathcal{B}(\frac{\eta\delta}{K}, \delta)},$$

$$P(\bar{\boldsymbol{\Phi}}) \propto \prod_v \frac{\mathcal{B}(\bar{\phi}_{v\cdot} + \frac{\gamma\lambda}{V}, K - \bar{\phi}_{v\cdot} + \lambda)}{\mathcal{B}(\frac{\gamma\lambda}{V}, \lambda)},$$

$$P(\mathbf{z}|\bar{\boldsymbol{\Theta}}) \propto \prod_d \frac{\Gamma(\bar{\theta}_{\cdot d}\alpha)}{\Gamma(\bar{\theta}_{\cdot d}\alpha + n_{\cdot\cdot d})} \prod_k \left(\frac{\Gamma(\bar{\theta}_{kd}\alpha + n_{\cdot kd})}{\Gamma(\bar{\theta}_{kd}\alpha)} \right)^{\bar{\theta}_{kd}},$$

$$P(\mathbf{w}|\mathbf{z}, \bar{\boldsymbol{\Phi}}) \propto \prod_k \frac{\Gamma(\bar{\phi}_{\cdot k}\beta)}{\Gamma(\bar{\phi}_{\cdot k}\beta + n_{\cdot k\cdot})} \prod_v \left(\frac{\Gamma(\bar{\phi}_{vk}\beta + n_{vk\cdot})}{\Gamma(\bar{\phi}_{vk}\beta)} \right)^{\bar{\phi}_{vk}},$$

where $\mathcal{B}(\cdot,\cdot)$ is the beta function and n_{vkd} is the number of times token v was assigned to topic k in document d. The notation $\cdot$ means we sum over the corresponding index. We use the convention $0^0 = 1$. Note also that $n_{\cdot kd} = 0$ if $\bar{\theta}_{kd} = 0$ (as $\theta_{kd} = 0$) and $n_{vk\cdot} = 0$ if $\bar{\phi}_{vk} = 0$ (as $\phi_{vk} = 0$).

The collapsed Gibbs sampler can be derived using Bayes' rule and exchangeability:

$$P(\bar{\theta}_{kd} = 1|\mathbf{z}, \bar{\boldsymbol{\Theta}}^{\setminus kd}) = \frac{P(\mathbf{z}, \bar{\boldsymbol{\Theta}})}{P(\mathbf{z}, \bar{\boldsymbol{\Theta}}^{\setminus kd})} \propto \begin{cases} \frac{\mathcal{B}(\bar{\theta}_{\cdot d}^{\setminus kd}\alpha + n_{\cdot\cdot d}, \alpha)(\bar{\theta}_{k\cdot}^{\setminus kd} + \frac{\eta\delta}{K})}{\mathcal{B}(\bar{\theta}_{\cdot d}^{\setminus kd}\alpha, \alpha)(D - 1 - \bar{\theta}_{k\cdot}^{\setminus kd} + \delta)} & \text{if } n_{\cdot kd} = 0, \\ 1 & \text{if } n_{\cdot kd} > 0, \end{cases}$$

$$P(\bar{\phi}_{vk} = 1|\mathbf{w}, \bar{\boldsymbol{\Phi}}^{\setminus vk}, \mathbf{z}) = \frac{P(\mathbf{w}, \bar{\boldsymbol{\Phi}}|\mathbf{z})}{P(\mathbf{w}, \bar{\boldsymbol{\Phi}}^{\setminus vk}|\mathbf{z})} \propto \begin{cases} \frac{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta + n_{\cdot k\cdot}, \beta)(\bar{\phi}_{v\cdot}^{\setminus vk} + \frac{\gamma\lambda}{V})}{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta, \beta)(K - 1 - \bar{\phi}_{v\cdot}^{\setminus vk} + \lambda)} & \text{if } n_{vk\cdot} = 0, \\ 1 & \text{if } n_{vk\cdot} > 0, \end{cases}$$

$$P(z_{id} = k|\mathbf{w}, \mathbf{z}^{\setminus id}, \bar{\boldsymbol{\Phi}}, \bar{\boldsymbol{\Theta}}) = \frac{P(\mathbf{w}, \mathbf{z}|\bar{\boldsymbol{\Phi}}, \bar{\boldsymbol{\Theta}})}{P(\mathbf{w}, \mathbf{z}^{\setminus id}|\bar{\boldsymbol{\Phi}}, \bar{\boldsymbol{\Theta}})} \propto \frac{(\alpha + n_{\cdot kd}^{\setminus id})(\beta + n_{vk\cdot}^{\setminus id})}{\bar{\phi}_{\cdot k}\beta + n_{\cdot k\cdot}^{\setminus id}} \mathbb{I}\{\bar{\theta}_{kd} = 1 \wedge \bar{\phi}_{vk} = 1\},$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. The variables $\bar{\theta}_{kd}$ and $\bar{\phi}_{vk}$ need only to be resampled when $n_{\cdot kd} = 0$ and $n_{vk\cdot} = 0$, respectively. We obtain the updates for the nonparametric version by letting $K \to \infty$ in $P(\bar{\theta}_{kd} = 1|\mathbf{z}, \bar{\boldsymbol{\Theta}}^{\setminus kd})$; in the posterior $P(\bar{\phi}_{vk} = 1|\mathbf{w}, \bar{\boldsymbol{\Phi}}^{\setminus vk}, \mathbf{z})$ we only need to replace K by K_{obs} as the actual number of observed topics is finite. The prior on the number of new topics is given by $\text{Poisson}(\eta\delta/(D-1+\delta))$. The sampler for the one-parameter IBP is recovered by setting $\delta = 1$ and $\lambda = 1$.

3.2 Special cases

In the focussed topic model (FTM), there is no sampling of $\{\bar{\phi}_{vk}\}$ and the topic assignments are sampled as follows:

$$P(z_{id} = k|\mathbf{z}^{\setminus id}, \mathbf{w}, \bar{\boldsymbol{\Theta}}) = \frac{P(\mathbf{w}, \mathbf{z}|\bar{\boldsymbol{\Theta}})}{P(\mathbf{w}, \mathbf{z}^{\setminus id}|\bar{\boldsymbol{\Theta}})} \propto \frac{(\bar{\theta}_{kd}\alpha + n_{\cdot kd}^{\setminus id})(\beta + n_{vk\cdot}^{\setminus id})}{V\beta + n_{\cdot k\cdot}^{\setminus id}} \mathbb{I}\{\bar{\theta}_{kd} = 1\}.$$

Similarly, in the sparse-smooth topic model (SSTM) there is no sampling of $\{\bar{\theta}_{kd}\}$ and the topics assignments are sampled as follows:

$$P(z_{id} = k|\mathbf{z}^{\setminus id}, \mathbf{w}, \bar{\boldsymbol{\Phi}}) = \frac{P(\mathbf{w}, \mathbf{z}|\bar{\boldsymbol{\Phi}})}{P(\mathbf{w}, \mathbf{z}^{\setminus id}|\bar{\boldsymbol{\Phi}})} \propto \frac{(\alpha + n_{\cdot kd}^{\setminus id})(\bar{\phi}_{vk}\beta + n_{vk\cdot}^{\setminus id})}{\bar{\phi}_{\cdot k}\beta + n_{\cdot k\cdot}^{\setminus id}} \mathbb{I}\{\bar{\phi}_{vk} = 1\}.$$

Finally, in standard LDA there is no sampling of $\{\bar{\phi}_{vk}\}$ or $\{\bar{\theta}_{kd}\}$, which leads to the well known collapsed Gibbs sampler:

$$P(z_{id} = k|\mathbf{z}^{\setminus id}, \mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{P(\mathbf{w}, \mathbf{z}^{\setminus id})} \propto \frac{(\alpha + n_{\cdot kd}^{\setminus id})(\beta + n_{vk\cdot}^{\setminus id})}{V\beta + n_{\cdot k\cdot}^{\setminus id}}.$$

3.3 Evaluation

Let $\mathbf{w}^*$ denote the test corpus. We use perplexity as a performance measure:

$$Perplexity(\mathbf{w}^*) = \exp(-\frac{\ln P(\mathbf{w}^*|\mathbf{w})}{\sum_d N_d^*}), \tag{5}$$

where the test log likelihood $\ln P(\mathbf{w}^*|\mathbf{w})$ is approximated as

$$\ln P(\mathbf{w}^*|\mathbf{w}) \approx \sum_d \sum_v n_{v\cdot d}^* \ln \frac{1}{P} \sum_p \sum_k \mathbb{E}\left[\phi_{vk}^{(p)}|\mathbf{w}, \mathbf{z}^{(p)}\right] \mathbb{E}\left[\theta_{kd}^{(p)}|\mathbf{z}^{(p)}\right]. \tag{6}$$

The posterior expectations are approximated as follows

$$\mathbb{E}\left[\theta_{kd}|\mathbf{z}\right] \approx \frac{\mathbb{E}\left[\bar{\theta}_{kd}|\mathbf{z}, \bar{\mathbf{\Theta}}^{\setminus kd}\right] \alpha + n_{\cdot kd}}{\sum_k \mathbb{E}\left[\bar{\theta}_{kd}|\mathbf{z}, \bar{\mathbf{\Theta}}^{\setminus kd}\right] \alpha + n_{\cdot\cdot d}}, \tag{7}$$

$$\mathbb{E}\left[\phi_{vk}|\mathbf{w}, \mathbf{z}\right] \approx \frac{\mathbb{E}\left[\bar{\phi}_{vk}|\mathbf{w}, \mathbf{z}, \bar{\mathbf{\Phi}}^{\setminus vk}\right] \beta + n_{vk\cdot}}{\sum_v \mathbb{E}\left[\bar{\phi}_{vk}|\mathbf{w}, \mathbf{z}, \bar{\mathbf{\Phi}}^{\setminus vk}\right] \beta + n_{\cdot k\cdot}}. \tag{8}$$

For the test documents, the topics $\mathbf{z}$ are sampled until convergence and finally, the test perplexity is computed.

4 Discussion

We proposed LIDA, a new model that subsumes FTM, STM and LDA, and naturally extends to its nonparametric counterpart. We believe that our sampler is simpler than previously proposed samplers for sparse topic models. We are currently empirically evaluating the performance (perplexity, sparsity, number of topics for the nonparametric version) of the different benchmark data sets including the 20 Newsgroups and Reuters-21578 datasets.

References

[1] T. Griffiths and Z. Ghahramani. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

[2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[3] C. Wang and D. M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *Advances in Neural Information Processing Systems 23 (NIPS)*. MIT Press, 2010.

[4] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP-compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010.