

Vectorized Bayesian Inference for Latent Dirichlet-Tree Allocation

Zheng Wang

ZHENG.WANG.20211@MAIL.CONCORDIA.CA

Concordia Institute for Information Systems Engineering

Concordia University

Montreal, QC H3G 1M8, Canada

Nizar Bouguila

NIZAR.BOUGUILA@CONCORDIA.CA

Concordia Institute for Information Systems Engineering

Concordia University

Montreal, QC H3G 1M8, Canada

Abstract

Latent Dirichlet Allocation (LDA) is a foundational model for discovering latent thematic structure in discrete data, but its Dirichlet prior cannot represent the rich correlations and hierarchical relationships often present among topics. We introduce the framework of Latent Dirichlet-Tree Allocation (LDTA), a generalization of LDA that replaces the Dirichlet prior with an arbitrary Dirichlet-Tree (DT) distribution. LDTA preserves LDA’s generative structure but enables expressive, tree-structured priors over topic proportions. To perform inference, we develop universal mean-field variational inference and Expectation Propagation, providing tractable updates for all DT. We reveal the vectorized nature of the two inference methods through theoretical development, and perform fully vectorized, GPU-accelerated implementations. The resulting framework substantially expands the modeling capacity of LDA while maintaining scalability and computational efficiency. The source code is available at <https://github.com/intaiyiw/ldta>.

Keywords: Bayesian Computing, Latent Dirichlet Allocation, Dirichlet-Tree, Mean-Field Variational Inference, Expectation Propagation

1 Introduction

Probabilistic topic modeling has become one of the central frameworks for understanding latent structure in large, unorganized collections of discrete data. As digital text, image annotations, scientific literature, biomedical data, and social media content continue to grow at unprecedented scale, tools for discovering hidden themes and interpretable patterns have become indispensable in machine learning and data mining (Blei, 2012). Among these tools, Latent Dirichlet Allocation (LDA) has played a foundational role. Since its introduction by Blei et al. (2003), LDA has remained one of the most widely used generative models for unsupervised learning of thematic structure. Its influence extends far beyond text analysis: LDA has been applied in genetics (Liu et al., 2016), marketing (Hofmann, 1999), computer vision (Fei-Fei and Perona, 2005), recommendation systems (Wang and Blei, 2011), and many other domains where observed data can be interpreted as arising from mixtures of latent components.

The original LDA model assumes that each document is represented by a probability vector over latent topics, and each topic is a probability distribution over a vocabulary of

terms (Blei et al., 2003). Crucially, LDA places a Dirichlet prior over the document-level topic proportions. This choice of prior has several attractive properties: it is conjugate to the multinomial distribution, it facilitates tractable posterior updates, and it induces a simple geometric interpretation on the simplex (Bishop, 2006; Gelman et al., 2013). However, the Dirichlet prior also imposes structural limitations. Although its parameters can control concentration and sparsity, the Dirichlet distribution cannot express fine-grained correlations among topic components beyond the inherent negative dependence due to the simplex constraint (Wallach et al., 2009). As a result, while LDA is simple and effective, it is often too rigid to capture the complex hierarchical or clustered topic structures that arise in real data.

This limitation has led to numerous extensions and generalizations of LDA. Correlated Topic Models (Blei and Lafferty, 2007) introduced a logistic-normal prior to encode richer covariance structures, while hierarchical topic models such as the Hierarchical Pachinko Allocation Model (Li and McCallum, 2006) and the nested Chinese restaurant process (Blei et al., 2010) allow for topic hierarchies and tree structures. Although these models provide greater flexibility, they typically sacrifice conjugacy and introduce substantial computational overhead. Many such models require non-conjugate variational inference or sampling procedures, which can be computationally expensive and difficult to scale (Hoffman et al., 2013; Teh et al., 2006).

A powerful yet underutilized alternative to the Dirichlet distribution is the Dirichlet-Tree (DT) distribution, conceptualized for the first time by Connor and Mosimann (1969) and later revived in works by Dennis III (1991) and Minka (1999). The Dirichlet-Tree distribution provides a natural way to encode structured correlations through a tree of nested Dirichlet random variables. Each internal node of the tree defines a sub-Dirichlet distribution governing how probability mass is split among its children, and the full distribution over the leaves (which lie on the simplex) arises from the multiplicative combination of these local branching decisions. This formulation has two particularly attractive properties: (1) it can express arbitrarily rich correlation patterns and hierarchical structures by choosing appropriate tree shapes, and (2) it retains many of the computational conveniences of the ordinary Dirichlet distribution, including local conjugacy and analytical tractability.

Given the advantages of the Dirichlet-Tree distribution, it is natural to ask whether the traditional LDA model can be generalized by replacing the Dirichlet prior with a Dirichlet-Tree prior. Doing so would allow topic models to encode complex prior structures—such as that certain topics should co-occur, that some topics form meaningful groups, or that certain topics follow hierarchical relationships—without abandoning the general framework of LDA. Indeed, the special cases of Dirichlet-Tree, including Beta-Liouville (Gupta and Richards, 1987; Bakhtiari and Bouguila, 2016) and Generalized Dirichlet (Fang et al., 1990; Wong, 1998; Bakhtiari and Bouguila, 2014), have been explored and utilized in recent topic modeling literature. This idea motivates the framework developed in this paper, which we call Latent Dirichlet-Tree Allocation (LDTA). LDTA generalizes LDA by allowing an arbitrary Dirichlet-Tree to govern the topic proportions. It preserves the generative semantics of LDA, retains interpretability, and supports a broad class of structured priors suitable for modeling realistic topic correlations. However, generalizing LDA to LDTA also introduces significant challenges in posterior inference. While LDA benefits from the conjugacy of the Dirichlet and multinomial distributions, LDTA introduces a hierarchical and potentially

deep tree structure whose internal dependencies complicate updates. Efficient and scalable inference algorithms are therefore essential. Two influential approaches to approximate Bayesian inference are particularly relevant: Mean-Field Variational Inference (MFVI) and Expectation Propagation (EP).

Mean-Field Variational Inference (MFVI) (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017), introduced in the late 1990s and now widely used in modern machine learning, replaces the exact posterior with a factorized approximation chosen to optimize a lower bound on the log marginal likelihood (the evidence lower bound, or ELBO). Variational inference is deterministic, often fast, and well-suited for models where conjugacy provides closed-form updates. It has become a standard tool for scalable Bayesian modeling and is used in diverse settings ranging from Bayesian neural networks to probabilistic graphical models. For LDTA, variational inference enables the derivation of structured coordinate ascent algorithms that update each factor of the variational distribution in closed or semi-closed form.

On the other hand, Expectation Propagation (EP), introduced by Minka (2001); Minka and Lafferty (2002), takes a different approach. Rather than optimizing a lower bound, EP iteratively approximates each non-conjugate factor in the model by projecting it onto a chosen exponential family distribution through moment matching. EP often provides more accurate approximations than MFVI, especially in models where the posterior is not well approximated by fully factorized distributions (mean-field) (Minka, 2005). EP has been used in Gaussian process classification (Hernández-Lobato and Hernández-Lobato, 2016), Bayesian logistic regression (Vehtari et al., 2020), and many other settings where non-conjugacy arises. For LDTA, EP offers a principled way to approximate the interactions between different nodes of the Dirichlet-tree, yielding improved posterior estimates at the cost of increased algorithmic complexity.

As modern machine learning applications frequently involve very large corpora, it is critical that inference algorithms for topic models leverage vectorized computation and GPU acceleration. Frameworks such as PyTorch (Paszke et al., 2019) have made it easy to express complex probabilistic updates in terms of batched tensor operations that can run efficiently on CUDA-capable GPUs. While traditional implementations of LDA rely on nested loops and sequential updates, a vectorized implementation can exploit parallelism across documents, topics, and tree nodes. Such acceleration is essential for large datasets where non-vectorized methods would be prohibitively slow.

This paper develops both mean-field variational inference and Expectation Propagation algorithms for LDTA and implements them in a fully vectorized fashion that exploits GPU parallelism. The resulting algorithms can perform efficient inference even for large corpora and complex Dirichlet-tree structures. Compared with scalar or loop-based implementations, the vectorized approach drastically reduces runtime and allows the algorithm to scale to modern datasets. The remainder of the paper proceeds as follows. We develop an improved and comprehensive formulation of Dirichlet-Tree in Section 2, revealing the properties of Dirichlet-Tree and preparing it to integrate into LDA, mean-field variational inference and expectation propagation. In Section 3, we describe the LDTA model, which is a generalization of LDA by extending the Dirichlet prior to an arbitrary Dirichlet-Tree distribution. In Section 4, a vectorized universal mean-field variational inference method is derived for any LDTA model. In Section 5, we interpret the idea and demonstrate the

implementation steps of the vectorized Expectation Propagation for any LDFA model. In Section 6, we choose three typical Dirichlet-Tree priors: the conventional Dirichlet, the Beta-Liouville and the Generalized Dirichlet to implement our models. We validate and compare our methods by a range of experiments including document modeling, document and image classification, and RNA-sequencing in bioinformatics. Finally, in Section 7, we conclude our paper and discuss the potential future directions.

2 Dirichlet-Tree

This section serves as a formal and comprehensive introduction to the Dirichlet-Tree distribution. Dirichlet-Tree was developed as a generalization of Dirichlet distribution by Dennis III (1991) and further explored by Minka (1999). We further prove the exponential form of Dirichlet-Tree distribution, and the triple forms of Dirichlet-Tree are concluded: the node form, the general form and the exponential form. As a result, we show that Dirichlet-Tree remains a conjugate prior to the multinomial likelihood with a transformed parameter updating. Finally, several important concepts including Bayesian operator are proposed to further assist the derivation of vectorized inferences for Latent Dirichlet-Tree Allocation.

2.1 Tree

The tree is a well-studied data structure in computer science. It is useful to introduce several tree-related notations before deriving the Dirichlet-Tree distribution. Figure 1 shows a general tree with height values of 3. A tree consists of multiple nodes and branches (or edges). We refer to any node λ that has multiple child nodes, including the root node, as an internal node; and the set for all internal nodes is denoted by $\mathbf{\Lambda}$. In particular, $\mathbf{\Lambda}^{\setminus r}$ denotes the set of all internal nodes except the root node. Any node ω that does not have a child node is called a terminal node (or leaf) and the set of all terminal nodes is $\mathbf{\Omega}$.

Between a parent node and one of its child nodes is a branch (or edge). We use $t | s$ to denote the branch starting from node s and pointing to node t . An indicator function $\delta_{t|s}(\omega)$ whose domain is $\mathbf{\Omega}$ is defined and assigned to each branch:

$$\delta_{t|s}(\omega) = \begin{cases} 1, & \text{if the branch } t|s \text{ leads to leaf } \omega, \\ 0, & \text{otherwise} \end{cases} \quad \omega \in \mathbf{\Omega}, s \in \mathbf{\Lambda}$$

Sometimes, the starting node of a branch does not matter, and we denote this branch that directly points to node s as $s|*$. We define the function $l(s)$, $s \in \mathbf{\Lambda} \cup \mathbf{\Omega}$ as the number of terminal nodes that a node s travels to. When s is the root node, we have:

$$l(r) = \sum_{t|r} \sum_{\omega \in \mathbf{\Omega}} \delta_{t|r}(\omega) = K$$

where $K = |\mathbf{\Omega}|$ is the number of all terminal nodes. For $s \in \mathbf{\Omega}$, we have:

$$l(s) = \sum_{\omega \in \mathbf{\Omega}} \delta_{s|*}(\omega) = 1, \quad s \in \mathbf{\Omega}$$

For any $s \in \mathbf{\Lambda}^{\setminus r}$:

$$l(s) = \sum_{\omega \in \Omega} \delta_{s|*}(\omega) = \sum_{t|s} \sum_{\omega \in \Omega} \delta_{t|s}(\omega), \quad s \in \mathbf{\Lambda}^{\setminus r}$$

We further define the function $c(s)$, $s \in \mathbf{\Lambda}$ as the number of direct child nodes (or branches) under an internal node. For example, in Figure 1, $c(\lambda_2) = 3$. The following theorem is obvious and useful to mention:

Theorem 1 *In a tree structure, for any internal node $s \in \mathbf{\Lambda}$,*

$$l(s) = c(s) + \sum_{t|s} (l(t) - 1), \quad s \in \mathbf{\Lambda}. \quad (1)$$

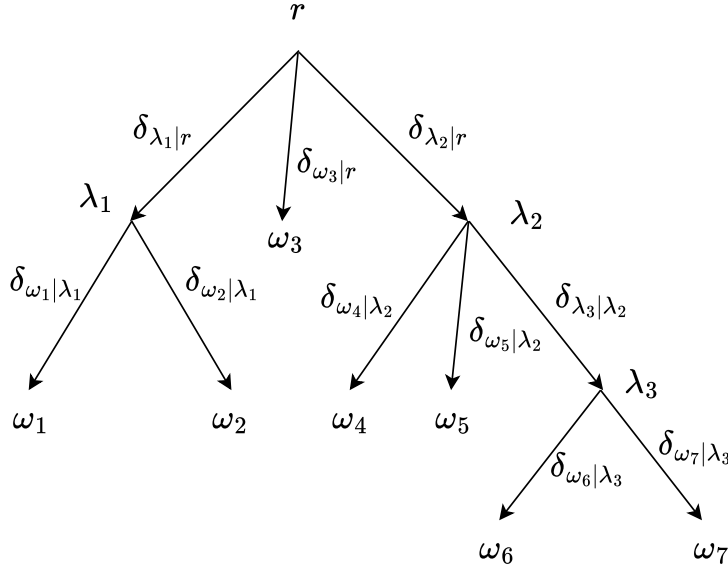


Figure 1: Notation for a general tree structure

2.2 Hierarchical Multinomial and Dirichlet-Tree

The Dirichlet distribution has been widely used as a conjugate prior to multinomial under conventional parameterization (Minka, 1999; Blei et al., 2003; Bishop, 2006). Minka (1999) described the sample of a conventional multinomial as the outcome of a K -sided die. In this analogy, a K -sided die is rolled once and one out of K possible outcomes is produced, where each outcome is assigned to a possibility matter p_k , $k = 1, 2, \dots, K$. As in Figure 2 (a), we can formally represent the conventional multinomial by assigning each p_k to a branch in a tree with height 1. The probability mass function (PMF) is given by:

$$p(\omega | \mathbf{p}) = \prod_{\omega|r} p_{\omega|r}^{\delta_{\omega|r}(\omega)}, \quad \omega \in \Omega, \mathbf{p} \in \Delta^{K-1}, r \text{ is root}$$

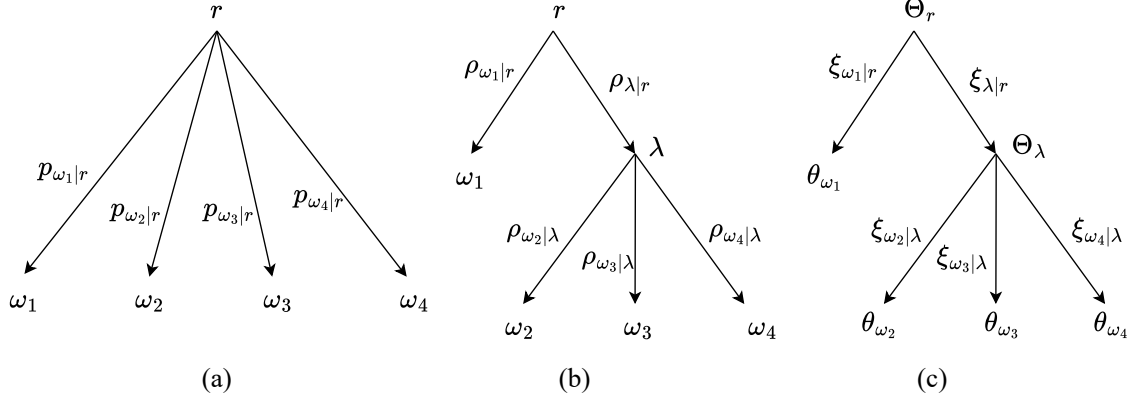


Figure 2: (a) Conventional Multinomial; (b) Hierarchical Multinomial; (c) Dirichlet-Tree

Correspondingly, its conjugate Dirichlet can be represented by the same tree with each branch assigned a parameter $\alpha_{\omega|r}$, and each leaf assigned the component of the $(K - 1)$ -dimensional simplex. The probability density function (PDF) follows:

$$\text{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_r) = \frac{1}{B(\boldsymbol{\alpha}_r)} \prod_{\omega \in \Omega} \theta_{\omega}^{\alpha_{\omega|r}-1}, \quad \boldsymbol{\theta} \in \Delta^{|\Omega|-1}, \quad \boldsymbol{\alpha}_r \in \mathbb{R}_{>0}^K, \quad r \text{ is root}$$

$$B(\boldsymbol{\alpha}_r) = \frac{\prod_{\omega|r} \Gamma(\alpha_{\omega|r})}{\Gamma\left(\sum_{\omega|r} \alpha_{\omega|r}\right)}$$

where $\Delta^{|\Omega|-1}$ is the $(|\Omega| - 1)$ -dimensional simplex defined by:

$$\Delta^{|\Omega|-1} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{|\Omega|} \mid \theta_{\omega} \geq 0, \sum_{\omega \in \Omega} \theta_{\omega} = 1 \right\}$$

In the case of Dirichlet, $|\Omega| = K$. Despite its convenience as a conjugate prior for multinomial, Dirichlet suffers from the following limitations: (i) all components share a common variance parameter although each component has its own mean (Minka, 1999), and (ii) the covariance between any two components is strictly negative (Mosimann, 1962). To solve these limitations, Dennis III (1991) developed the Dirichlet-Tree distribution as a generalization of Dirichlet.

Instead of rolling a K -dimensional die once, the sample of a multinomial can also be thought of as rolling several different dice for finite times, as shown in Figure 2 (b). This finite stochastic process results in a hierarchical multinomial under a tree-like parameterization with the following PMF:

$$p(\omega \mid \boldsymbol{\rho}) = \prod_{s \in \mathbf{\Lambda}} \prod_{t|s} \rho_{t|s}^{\delta_{t|s}(\omega)}, \quad \omega \in \Omega, \quad \boldsymbol{\rho}_s \in \Delta_s^{c(s)-1}, \quad s \in \mathbf{\Lambda} \quad (2)$$

Consequently, the corresponding prior for the hierarchical multinomial is based on the collection of $|\mathbf{\Lambda}|$ independent Dirichlets endowed with a tree structure (Figure 2 (c)). And

the PDF is given by:

$$p(\boldsymbol{\rho} \mid \boldsymbol{\xi}) = \prod_{s \in \boldsymbol{\Lambda}} \frac{1}{B(\boldsymbol{\xi}_s)} \prod_{t|s} \rho_{t|s}^{\xi_{t|s}-1}, \quad \boldsymbol{\rho}_s \in \Delta_s^{c(s)-1}, \quad \boldsymbol{\xi}_s \in \mathbb{R}_{>0}^{c(s)}, \quad s \in \boldsymbol{\Lambda} \quad (3)$$

This is not the final form of Dirichlet-Tree distribution yet. What we want is a distribution over $(|\boldsymbol{\Omega}| - 1)$ -dimensional simplex assigned to the set of terminal nodes. A transformation from $\boldsymbol{\rho} \in \prod_{s \in \boldsymbol{\Lambda}} \Delta_s^{c(s)-1}$ to $\boldsymbol{\theta} \in \Delta^{|\boldsymbol{\Omega}|-1}$ is needed to achieve this distribution. As (2) implies, the component θ_ω assigned to a terminal node is equal to the product of all probability mass assigned to the branches that lead to the terminal node. We denote this mapping as follows:

$$T : \prod_{s \in \boldsymbol{\Lambda}} \Delta_s^{c(s)-1} \rightarrow \Delta^{|\boldsymbol{\Omega}|-1}, \quad \boldsymbol{\rho} \mapsto T(\boldsymbol{\rho})$$

$$\theta_\omega = \prod_{s \in \boldsymbol{\Lambda}} \prod_{t|s} \rho_{t|s}^{\delta_{t|s}(\omega)}, \quad \omega \in \boldsymbol{\Omega} \quad (4)$$

$$T(\boldsymbol{\rho}) = \bigcirc_{s \in \boldsymbol{\Lambda}} T_s(\boldsymbol{\rho}_s)$$

where T is the composition of all sub-mappings T_s under each internal node. Conversely, to show the inverse transform T^{-1} , we first define the node mass Θ_s :

$$\Theta_s(\boldsymbol{\theta}) = \begin{cases} 1, & s \text{ is the root,} \\ \sum_{\omega \in \boldsymbol{\Omega}} \delta_{s|*}(\omega) \theta_\omega, & s \in \boldsymbol{\Lambda}^{\setminus r} \cup \boldsymbol{\Omega} \end{cases}$$

Obviously, $\Theta_\omega = \theta_\omega$ when $\omega \in \boldsymbol{\Omega}$. Thus, we have the inverse transform as follows:

$$T^{-1} : \Delta^{|\boldsymbol{\Omega}|-1} \rightarrow \prod_{s \in \boldsymbol{\Lambda}} \Delta_s^{c(s)-1}, \quad \boldsymbol{\theta} \mapsto T^{-1}(\boldsymbol{\theta})$$

$$\rho_{t|s} = \frac{\Theta_t}{\Theta_s}, \quad s \in \boldsymbol{\Lambda} \quad (5)$$

Now, we have a pair of inverse transforms and assume the transform to be smooth. We go back to (4) and derive the Jacobian of T . For $\forall s \in \boldsymbol{\Lambda}$, we denote the sub-transform by free-coordinates representation:

$$T_s(\boldsymbol{\rho}_s) : (\rho_{1|s}, \dots, \rho_{c(s)-1|s}, \Theta_s) \mapsto (\Theta_1, \dots, \Theta_{c(s)})$$

Note that we enumerate the set of immediate child nodes of node s by $1, \dots, c(s)$. For $i = 1, \dots, c(s) - 1$ and $j = 1, \dots, c(s) - 1$:

$$\frac{\partial \Theta_i}{\partial \rho_{j|s}} = \frac{\partial (\rho_{i|s} \Theta_s)}{\partial \rho_{j|s}} = \Theta_s \mathbf{1}_{\{i=j\}}, \quad \frac{\partial \Theta_{c(s)}}{\partial \rho_{j|s}} = \frac{\partial \left(\left(1 - \sum_{l=1}^{c(s)-1} \rho_{l|s} \right) \Theta_s \right)}{\partial \rho_{j|s}} = -\Theta_s$$

For the column corresponding to Θ_s :

$$\frac{\partial \Theta_i}{\partial \Theta_s} = \rho_{i|s}, \quad i = 1, \dots, c(s)$$

Thus, we have the Jacobian with the form:

$$J_{T_s} = \begin{pmatrix} \Theta_s I_{c(s)-1} & b \\ c^\top & d \end{pmatrix}$$

where $I_{c(s)-1}$ is $(c(s)-1)$ -dimensional identity, $b = (\rho_{1|s}, \dots, \rho_{c(s)-1|s})^\top$, $c^\top = (-\Theta_s, \dots, -\Theta_s)$, $d = \rho_{c(s)|s}$. Thus, the determinant derives as follows:

$$\det(J_{T_s}) = \det(\Theta_s I_{c(s)-1}) \left(d - c^\top (\Theta_s I_{c(s)-1})^{-1} b \right) = \Theta_s^{c(s)-1} \quad (6)$$

According to the change-of-variable theorem, what we call the node form of Dirichlet-Tree is derived as follows by combining (3), (5) and (6):

$$\begin{aligned} \text{DT}(\boldsymbol{\theta} \mid \boldsymbol{\xi}) &= p(T^{-1}(\boldsymbol{\theta}) \mid \boldsymbol{\xi}) |\det(J_T)|^{-1} \\ &= \prod_{s \in \boldsymbol{\Lambda}} \frac{1}{B(\boldsymbol{\xi}_s)} \left(\frac{1}{\Theta_s} \right)^{c(s)-1} \prod_{t|s} \left(\frac{\Theta_t}{\Theta_s} \right)^{\xi_{t|s}-1}, \quad \boldsymbol{\theta} \in \Delta^{|\boldsymbol{\Omega}|-1}, \boldsymbol{\xi} \in \mathbb{R}_{>0}^{\sum_{s \in \boldsymbol{\Lambda}} c(s)} \end{aligned} \quad (7)$$

It is easy to derive a more general form of Dirichlet-Tree from (7). Based on this general form, we give a formal definition of Dirichlet-Tree distribution.

Definition 2 (General form of Dirichlet-Tree) *With respect to a tree structure where r is the root node, $\boldsymbol{\Omega}$ is the set of terminal nodes and $\boldsymbol{\Lambda}^{\setminus r}$ is the set of the remaining nodes, a Dirichlet-Tree distribution is a probability measure on $\Delta^{|\boldsymbol{\Omega}|-1}$ with PDF*

$$\text{DT}(\boldsymbol{\theta} \mid \boldsymbol{\xi}) = \frac{1}{B(\boldsymbol{\xi}_r)} \prod_{\omega \in \boldsymbol{\Omega}} \theta_\omega^{\xi_{\omega|*}-1} \left(\prod_{s \in \boldsymbol{\Lambda}^{\setminus r}} \frac{1}{B(\boldsymbol{\xi}_s)} \Theta_s^{\xi_{s|*}-\xi_{*|s}} \right) \quad (8)$$

where $\xi_{t|s} \in \mathbb{R}_{>0}$ is assigned to each branch for $s \in \boldsymbol{\Lambda}$, and

$$\xi_{*|s} = \sum_{t|s} \xi_{t|s}$$

and $B(\cdot)$ is the multivariate beta function.

2.3 Exponential Form and Conjugacy

We further prove that a Dirichlet-Tree distribution belongs to the exponential family. The exponential form of a Dirichlet-Tree is given as follows:

Theorem 3 (Exponential form of Dirichlet-Tree)

$$\begin{aligned} \text{DT}(\boldsymbol{\theta} \mid \boldsymbol{\xi}) &= \prod_{s \in \boldsymbol{\Lambda}} \frac{1}{B(\boldsymbol{\xi}_s)} \prod_{t|s} \left(\frac{\Theta_t}{\Theta_s} \right)^{\xi_{t|s}-l(t)} \\ &= \exp \left(\sum_{s \in \boldsymbol{\Lambda}} \sum_{t|s} (\xi_{t|s} - l(t)) \log \left(\frac{\Theta_t}{\Theta_s} \right) - \log \prod_{s \in \boldsymbol{\Lambda}} B(\boldsymbol{\xi}_s) \right) \end{aligned} \quad (9)$$

where $l(t)$ is the number of leaf nodes that the node t leads to.

Proof Dividing (7) by (9) yields

$$\begin{aligned}
 & \prod_{s \in \Lambda} \left(\frac{1}{\Theta_s} \right)^{c(s)-1} \prod_{t|s} \left(\frac{\Theta_t}{\Theta_s} \right)^{l(t)-1} \\
 &= \left(\prod_{s \in \Lambda} \left(\frac{1}{\Theta_s} \right)^{c(s) + \sum_{t|s} (l(t)-1) - 1} \right) \left(\prod_{t \in \Lambda \setminus r \cup \Omega} \Theta_t^{l(t)-1} \right) \\
 &= \left(\prod_{s \in \Lambda} \left(\frac{1}{\Theta_s} \right)^{l(s)-1} \right) \left(\prod_{t \in \Lambda \setminus r} \Theta_t^{l(t)-1} \right) \\
 &= 1
 \end{aligned}$$

The derivation uses Theorem 1 and the fact that

$$\Theta_{\text{root}} = 1$$

■

The natural parameter, sufficient statistics, normalizer, and base measure are respectively denoted as:

$$\begin{aligned}
 \eta_{t|s}(\boldsymbol{\xi}) &= \xi_{t|s} - l(t), \quad s \in \Lambda \\
 u_{t|s}(\boldsymbol{\theta}) &= \log \left(\frac{\Theta_t}{\Theta_s} \right), \quad s \in \Lambda \\
 g(\boldsymbol{\eta}(\boldsymbol{\xi})) &= \prod_{s \in \Lambda} B(\boldsymbol{\xi}_s) \\
 h(\boldsymbol{\theta}) &= 1
 \end{aligned}$$

Given the exponential form, all properties of the exponential family listed in Appendix A will apply.

Corollary 4 (Expectation of sufficient statistics)

$$\begin{aligned}
 \mathbb{E}_{\text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi})} \left[\log \left(\frac{\Theta_t}{\Theta_s} \right) \right] &= \frac{\partial (\log \prod_{s \in \Omega} B(\boldsymbol{\xi}_s))}{\partial \xi_{t|s}} = \psi(\xi_{t|s}) - \psi(\xi_{*|s}), \quad s \in \Lambda \\
 \xi_{*|s} &= \sum_{t|s} \xi_{t|s}
 \end{aligned}$$

We proceed to show the conjugacy of Dirichlet-Tree to any conventional $|\Omega|$ -dimensional multinomial. Combining the pair of the transforms (4) and (5), we can represent each component θ_ω assigned to a terminal node by node mass and indicator functions assigned to non-terminal nodes and branches as follows:

$$\theta_\omega = \prod_{s \in \Lambda} \prod_{t|s} \left(\frac{\Theta_t}{\Theta_s} \right)^{\delta_{t|s}(\omega)}, \quad \omega \in \Omega \tag{10}$$

Let a Dirichlet-Tree distribution with $\boldsymbol{\Omega}$ be the prior to a $|\boldsymbol{\Omega}|$ -dimensional multinomial. Given a collection of multinomial samples $\mathbf{n} = (n_1, \dots, n_K)$ where $K = |\boldsymbol{\Omega}|$, the posterior is given as:

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{n}) &\propto p(\mathbf{n} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&\propto \left(\prod_{\omega \in \boldsymbol{\Omega}} \theta_{\omega}^{n_{\omega}} \right) \left(\prod_{s \in \boldsymbol{\Lambda}} \prod_{t|s} \left(\frac{\Theta_t}{\Theta_s} \right)^{\xi_{t|s} - l(t)} \right) \\
&= \exp \left\{ \sum_{\omega \in \boldsymbol{\Omega}} n_{\omega} \log \theta_{\omega} + \sum_{s \in \boldsymbol{\Lambda}} \sum_{t|s} (\xi_{t|s} - l(t)) \log \left(\frac{\Theta_t}{\Theta_s} \right) \right\} \\
&= \exp \left\{ \sum_{\omega \in \boldsymbol{\Omega}} n_{\omega} \sum_{s \in \boldsymbol{\Lambda}} \sum_{t|s} \delta_{t|s}(\omega) \log \left(\frac{\Theta_t}{\Theta_s} \right) + \sum_{s \in \boldsymbol{\Lambda}} \sum_{t|s} (\xi_{t|s} - l(t)) \log \left(\frac{\Theta_t}{\Theta_s} \right) \right\} \\
&= \exp \left\{ \sum_{s \in \boldsymbol{\Lambda}} \sum_{t|s} \left(\xi_{t|s} + \sum_{\omega \in \boldsymbol{\Omega}} n_{\omega} \delta_{t|s}(\omega) - l(t) \right) \log \left(\frac{\Theta_t}{\Theta_s} \right) \right\}
\end{aligned}$$

Now it is obvious that the posterior remains the same kind of Dirichlet-Tree with a transformed parameter updating. For convenience, we denote the posterior Dirichlet-Tree as

$$\text{DT}(\boldsymbol{\theta} \mid \boldsymbol{\xi}') = \text{DT}(\boldsymbol{\theta} \mid \boldsymbol{\xi} \oplus_p \mathbf{n}) = \text{DT}(\boldsymbol{\theta} \mid \boldsymbol{\xi} + \mathbf{D}_p \mathbf{n})$$

where we call \oplus_p the Bayesian addition; and \mathbf{D}_p the Dirichlet selection operator as \mathbf{D}_p selects the components from the partitioned samples to add to the Dirichlet-Tree parameter corresponding to each branch in the tree structure; and

$$\xi'_{t|s} = \xi_{t|s} + \sum_{\omega \in \boldsymbol{\Omega}} n_{\omega} \delta_{t|s}(\omega), \quad s \in \boldsymbol{\Lambda}$$

We have the following corollary by combining Corollary 4 and (10):

Corollary 5 (Log-space expectation)

$$\mathbb{E}_{\text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi})} [\log \theta_{\omega}] = \mathbb{E}_{\text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi})} \left[\sum_{s \in \boldsymbol{\Lambda}} \sum_{t|s} \delta_{t|s}(\omega) \log \left(\frac{\Theta_t}{\Theta_s} \right) \right] = \sum_{s \in \boldsymbol{\Lambda}} \sum_{t|s} \delta_{t|s}(\omega) (\psi(\xi_{t|s}) - \psi(\xi_{*|s}))$$

Intuitively, the expectation of $\log \theta_{\omega}$ is equal to the sum of $\mathbb{E}_{\text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi})} [u_{t|s}]$ corresponding to all the branches that leads to ω .

2.4 Bayesian operator

In this section, we propose several concepts based on the exponential form and conjugacy of Dirichlet-Tree to further assist the derivations of two approximate Bayesian inference algorithms introduced in the following sections. In the Bayesian inference of generative

graphical models like Latent Dirichlet Allocation (Blei et al., 2003), it is common to have a substantial number of parameter updating processes. Bayesian operator is proposed to represent this kind of processes in a compact way. We first give the formal definition of Bayesian operator:

Definition 6 (Bayesian operator) *Let \mathcal{T} be the statistical manifold of Dirichlet-Tree. In conjugacy to multinomial, the Bayesian operator for $\forall p \in \mathcal{T}$ transforms the p as a prior to a posterior $p' \in \mathcal{T}$ given the multinomial observation \mathbf{n} :*

$$p' = \mathbf{B}(\mathbf{n}) p$$

It is easy to derive:

$$\mathbf{B}(\mathbf{n}) = \left(\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p \mathbf{n}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \right)^{-1} \exp \left(\mathbf{n}^\top \log \boldsymbol{\theta} \right)$$

The Bayesian operator exhibits a closure nature and provides a new expression of conjugacy. The mathematical expression in our following derivation of Expectation Propagation is largely simplified with the Bayesian operator. Given the Bayesian operator, the relationship between the Dirichlet selection operator and the Bayesian operator is built, which is our central theorem:

Theorem 7 (The central theorem)

$$p(\boldsymbol{\theta} \mid \boldsymbol{\xi} + \mathbf{D}_p \mathbf{n}) = \mathbf{B}(\mathbf{n}) p(\boldsymbol{\theta} \mid \boldsymbol{\xi})$$

Even though the multinomial observation \mathbf{n} is non-negative integers, \mathbf{n} can be naturally extended to any real numbers in Bayesian operator that makes the posterior meaningful. The Bayesian operator satisfied the following properties:

- **Additivity and Commutativity**

$$\mathbf{B}(\mathbf{n}_1 + \mathbf{n}_2) = \mathbf{B}(\mathbf{n}_1) \mathbf{B}(\mathbf{n}_2) = \mathbf{B}(\mathbf{n}_2) \mathbf{B}(\mathbf{n}_1)$$

Proof

$$\text{RHS} = \left(\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p(\mathbf{n}_1 + \mathbf{n}_2)))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \right)^{-1} \exp \left((\mathbf{n}_2 + \mathbf{n}_1)^\top \log \boldsymbol{\theta} \right) = \text{LHS}$$

- **Inverse operator**

$$\mathbf{B}^{-1}(\mathbf{n}) = \mathbf{B}(-\mathbf{n})$$

- **Base posteriors**

For $\forall p(\boldsymbol{\theta} \mid \boldsymbol{\xi}) \in \mathcal{T}$, there exists a group of base posteriors $\{p(\boldsymbol{\theta} \mid \boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)})\}_K$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)}) = \mathbf{B}(\mathbf{1}_{(k)}) p(\boldsymbol{\theta} \mid \boldsymbol{\xi}), \quad k = 1, \dots, K$$

where $K = |\boldsymbol{\Omega}|$ and $\mathbf{1}_{(k)}$ is the K -dimensional binary vector where k^{th} element equals 1 and 0 otherwise. We also use $p(\boldsymbol{\theta} \mid \boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)})$ to represent the group of derived distributions $\{p(\boldsymbol{\theta} \mid \boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)})\}_K$.

Corollary 8 (Expectation operator)

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi})} \left[\prod_{k=1}^K \theta_k^{n_k} \right] &= \int_{\Delta^{K-1}} \exp \left(\mathbf{n}^\top \log \boldsymbol{\theta} \right) p(\boldsymbol{\theta}|\boldsymbol{\xi}) d\boldsymbol{\theta} \\
&= \int_{\Delta^{K-1}} \left[\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p \mathbf{n}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \mathbf{B}(\mathbf{n}) \right] p(\boldsymbol{\theta}|\boldsymbol{\xi}) d\boldsymbol{\theta} \\
&= \frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p \mathbf{n}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \int_{\Delta^{K-1}} p(\boldsymbol{\theta}|\boldsymbol{\xi} + \mathbf{D}_p \mathbf{n}) d\boldsymbol{\theta} \\
&= \frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p \mathbf{n}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \\
&= \prod_{s \in \Lambda} \frac{B(\boldsymbol{\xi}_s + \mathbf{D}_p \mathbf{n})}{B(\boldsymbol{\xi}_s)} \\
&= \prod_{s \in \Lambda} \left(\frac{\prod_{t|s} \Gamma(\xi_{t|s}')}{\Gamma(\sum_{t|s} \xi_{t|s}')} \right) \left(\frac{\Gamma(\sum_{t|s} \xi_{t|s})}{\prod_{t|s} \Gamma(\xi_{t|s})} \right) \\
&= \prod_{s \in \Lambda} \frac{\Gamma(\sum_{t|s} \xi_{t|s})}{\Gamma(\sum_{t|s} \xi_{t|s} + \sum_{t|s} \sum_{\omega \in \Omega} n_\omega \delta_{t|s}(\omega))} \prod_{t|s} \frac{\Gamma(\xi_{t|s} + \sum_{\omega \in \Omega} n_\omega \delta_{t|s}(\omega))}{\Gamma(\xi_{t|s})}
\end{aligned}$$

In particular,

$$\begin{aligned}
\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi})} [\theta_k] &= \frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \\
&= \prod_{s \in \Lambda} \frac{\Gamma(\sum_{t|s} \xi_{t|s})}{\Gamma(\sum_{t|s} \xi_{t|s} + \sum_{t|s} \sum_{\omega \in \Omega} \mathbf{1}_{(k)} \delta_{t|s}(\omega))} \prod_{t|s} \frac{\Gamma(\xi_{t|s} + \sum_{\omega \in \Omega} \mathbf{1}_{(k)} \delta_{t|s}(\omega))}{\Gamma(\xi_{t|s})} \\
&= \prod_{s \in \Lambda} \frac{\Gamma(\sum_{t|s} \xi_{t|s})}{\Gamma(\sum_{t|s} \xi_{t|s} + \sum_{t|s} \delta_{t|s}(k))} \prod_{t|s} \left(\frac{\Gamma(\xi_{t|s} + \delta_{t|s}(k))}{\Gamma(\xi_{t|s})} \right) \\
&= \prod_{s \in \Lambda} \left(\frac{1}{\sum_{t|s} \xi_{t|s}} \right)^{\sum_{t|s} \delta_{t|s}(k)} \prod_{t|s} \xi_{t|s}^{\delta_{t|s}(k)}, \quad k = 1, \dots, K; \quad s \in \Lambda
\end{aligned}$$

The derivation of the corollary uses the Gamma function's property:

$$\Gamma(x+1) = x\Gamma(x)$$

Intuitively, the expectation of each component θ_k of Dirichlet-Tree is equal to the product of parameter proportions corresponding to all branches that lead to the component.

2.5 Derived Distributions of Dirichlet-Tree

It is useful to define the group of derived distributions to assist the Bayesian inference in the following sections.

Definition 9 (Derived distributions of Dirichlet-Tree) *Let $p(\boldsymbol{\theta}|\boldsymbol{\xi}) \in \mathcal{T}$ be a Dirichlet-Tree. The corresponding group of derived distributions $\{p_{t|s}(\boldsymbol{\theta})\}_{|\boldsymbol{\Lambda}|}$ is defined as:*

$$p_{t|s}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi})} [u_{t|s}(\boldsymbol{\theta})]^{-1} u_{t|s}(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\xi}), \quad s \in \boldsymbol{\Lambda}$$

For notational simplicity, and when no ambiguity arises, we replace the label $t | s$ for each branch with d , and $|\boldsymbol{\Lambda}| = D$. We also denote $|\boldsymbol{\Omega}| = K$.

Corollary 10 (Expectation matrix of derived distributions) *By stacking the expectation of each derived distribution, the $(K \times D)$ -dimensional expectation matrix is derived, and each column gives:*

$$\begin{aligned} \mathbb{E}_{p_d(\boldsymbol{\theta})} [\boldsymbol{\theta}] &= \left\{ \int_{\Delta^{K-1}} \left(\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \mathbf{B}(\mathbf{1}_{(k)}) \right) \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi})} [u_d(\boldsymbol{\theta})]^{-1} u_d(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\xi}) d\boldsymbol{\theta} \right\}_{k=1}^K \\ &= \left\{ \frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi})} [u_d(\boldsymbol{\theta})]^{-1} \int_{\Delta^{K-1}} u_d(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)}) d\boldsymbol{\theta} \right\}_{k=1}^K \\ &= \frac{\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi})} [\boldsymbol{\theta}] \odot_k \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi} + \mathbf{D}_p \mathbf{1}_{(k)})} [u_d(\boldsymbol{\theta})]}{\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{\xi})} [u_d(\boldsymbol{\theta})]}, \quad d = 1, \dots, D \end{aligned} \tag{11}$$

where $D = |\boldsymbol{\Lambda}|$ and \odot_k is the Hadamard (element-wise) product with respect to the subscript k .

3 Latent Dirichlet-Tree Allocation

In this section, we introduce Latent Dirichlet-Tree Allocation (LDТА), which generalizes the Latent Dirichlet Allocation model of Blei et al. (2003) by replacing the Dirichlet prior with an arbitrary Dirichlet-Tree. Similar to LDA, LDТА is a generative probabilistic graphical model. For clarity, we adopt column-major order in the mathematical formulation of the model, whereas the implementation uses row-major order to adhere to standard programming conventions.

3.1 The Generative Process

Two sets with known cardinalities V and K are postulated before constructing our model: (i) the collection of V words (the vocabulary, or visual words in the case of images):

$$\mathbb{W} = \{w_1, \dots, w_v, \dots, w_V\}$$

and (ii) the collection of K topics:

$$\mathbb{Z} = \{z_1, \dots, z_k, \dots, z_K\}$$

The observed data set is a sequence of M documents (the corpus, or images):

$$\mathcal{D} = (\mathbf{w}_1, \dots, \mathbf{w}_m, \dots, \mathbf{w}_M)$$

where each document is a sequence of N_m words:

$$\mathbf{w}_m = (w_{1m}, \dots, w_{nm}, \dots, w_{N_m m}), \quad w_{nm} \in \mathbb{W}$$

In LDТА, a latent topic is assumed with respect to each observed word:

$$\mathbf{z}_m = (z_{1m}, \dots, z_{nm}, \dots, z_{N_m m}), \quad z_{nm} \in \mathbb{Z}$$

Sometimes, we ignore the subscript m to simply keep our notation uncluttered when we refer to any document in \mathcal{D} . According to the ex-changeability assumption (Blei et al., 2003), we only care about the occurring numbers of the words that appeared in a document, instead of the order of the word sequence. Therefore, we indicate each word as a binary vector $\mathbf{1}_{nm}^{(v)} \in \{0, 1\}^V$ defined over \mathbb{W} , where the v^{th} element equals 1 if the v^{th} word is chosen, and 0 otherwise. Consequently, we obtain the word counting vector of a document by summing up all the word indicator vector:

$$\mathbf{n}_m = (n_{1m}, \dots, n_{vm}, \dots, n_{Vm})^\top = \sum_{i=1}^{N_m} \mathbf{1}_{im}^{(v)}$$

Taking the latent topics into account, the binary indicator vector for each word is lifted to a binary matrix $\mathbf{1}_{nm}^{(vk)} \in \{0, 1\}^{V \times K}$, defined over $\mathbb{W} \times \mathbb{Z}$, to represent word–topic co-occurrence. Similarly, we obtain the word–topic counting matrix for a document by summing up all the word–topic indicator matrices:

$$\mathbf{n}_m = \{n_{vk}\}_{(V \times K)} = \begin{pmatrix} n_{11} & \cdots & n_{1k} & \cdots & n_{1K} \\ \vdots & & \vdots & & \vdots \\ n_{v1} & \cdots & n_{vk} & \cdots & n_{vK} \\ \vdots & & \vdots & & \vdots \\ n_{V1} & \cdots & n_{Vk} & \cdots & n_{VK} \end{pmatrix} = \sum_{i=1}^{N_m} \mathbf{1}_{im}^{(vk)}$$

Moreover, the document–word–topic counting tensor is constructed by stacking counting matrices of all documents in the data set (Figure 3). Similar to LDA, LDТА is probabilistic graphic model. The generative process for each document is described as follows:

1. Draw $N \sim \text{Poisson}(\mu)$.
2. Draw $\boldsymbol{\theta} \sim \text{DT}(\boldsymbol{\xi})$.
3. For each word w_n for $n = 1, \dots, N$:
 - (a) Draw $z_n \sim \text{Mult}(\boldsymbol{\theta})$.
 - (b) Draw $w_n \sim \text{Mult}(\boldsymbol{\varphi}_{z_n})$.

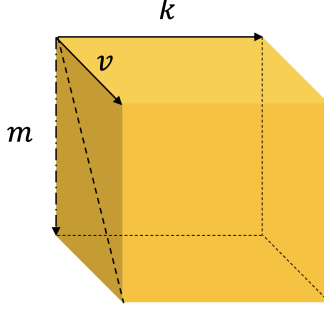


Figure 3: The document–topic–word counting tensor.

And the whole generation process produced our data set by repeating above procedures for M times (Figure 4). Among the parameters, ξ is the hyper-parameter for the Dirichlet-Tree prior; $\theta \in \Delta^{K-1}$ is the topic proportion with respect to \mathbb{Z} , and we denote the square diagonal form of each document’s proportion as $M(\theta_m)$; φ is a word–topic conditional probability matrix defined as the conditional probability of generating v^{th} word given the k^{th} topic:

$$\varphi = \{\varphi_{vk}\}_{V \times K} = \begin{pmatrix} \varphi_{11} & \cdots & \varphi_{1k} & \cdots & \varphi_{1K} \\ \vdots & & \vdots & & \vdots \\ \varphi_{v1} & \cdots & \varphi_{vk} & \cdots & \varphi_{vK} \\ \vdots & & \vdots & & \vdots \\ \varphi_{V1} & \cdots & \varphi_{Vk} & \cdots & \varphi_{VK} \end{pmatrix}$$

and we assume φ to be non-generated. By multiplying the word–topic conditional probability matrix and the square diagonal form of each document’s topic proportion, we obtain the joint probability matrix of the co-occurrence of each word–topic couple in each document:

$$\Phi_m = \begin{pmatrix} \varphi_{11}\theta_1 & \cdots & \varphi_{1k}\theta_k & \cdots & \varphi_{1K}\theta_K \\ \vdots & & \vdots & & \vdots \\ \varphi_{v1}\theta_1 & \cdots & \varphi_{vk}\theta_k & \cdots & \varphi_{vK}\theta_K \\ \vdots & & \vdots & & \vdots \\ \varphi_{V1}\theta_1 & \cdots & \varphi_{Vk}\theta_k & \cdots & \varphi_{VK}\theta_K \end{pmatrix} = \varphi M(\theta)$$

The relationship of the variables in LDFA expressed in tensors is shown in Figure 5.

3.2 The Joint and the Marginalized Distributions

Given the hyper-parameters ξ and φ , as well as the conditional independence assumption of the variables, the joint distribution of the latent parameters and the observed counting vector for a single document is given by:

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}, \theta | \varphi, \xi) &= p(\mathbf{w} | \mathbf{z}, \varphi) p(\mathbf{z} | \theta) p(\theta | \xi) \\ &= p(\theta | \xi) \prod_{n=1}^{N_m} p(z_n | \theta) p(w_n | z_n, \varphi) \end{aligned}$$

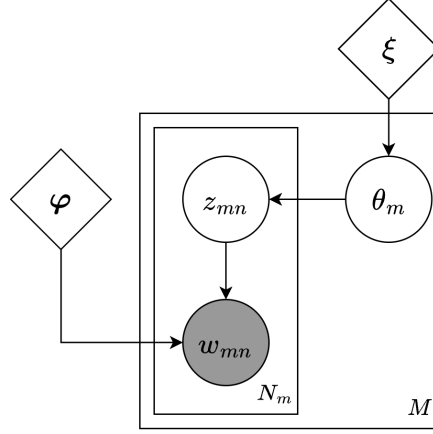


Figure 4: Graphical representation (plate notation) of LDFA model. The shaded circle represents the observed variable; the blank circles represent the latent variables and diamonds indicate hyper-parameters. The arrows imply dependent generative relationship and the rectangles with number at its right-lower corner represent repetitions.

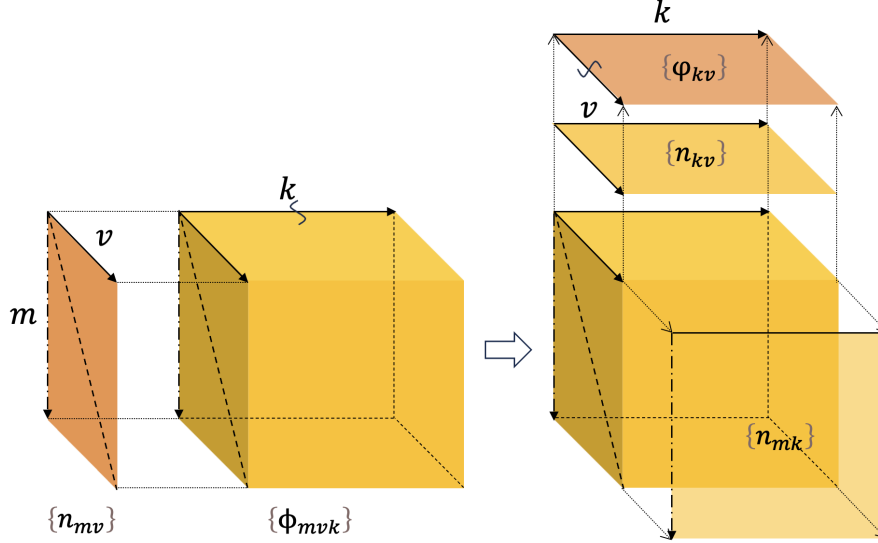


Figure 5: Tensor expression of relationship of the variables

where \mathbf{w} and \mathbf{z} are sequences of words and topics respectively. Based on above discussion, the order of the word sequence is trivial and can simply be replaced by counting matrix

$\{n_{vk}\}_{(V \times K)}$. Hence, each term is respectively represented as:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{\xi}) &= \text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi}) \\ p(\mathbf{z}|\boldsymbol{\theta}) &= \prod_{k=1}^K \theta_k^{n_k} = \prod_{k=1}^K \prod_{v=1}^V \theta_k^{n_{vk}} \\ p(\mathbf{w}|\mathbf{z}, \boldsymbol{\varphi}) &= \prod_{k=1}^K \prod_{v=1}^V \varphi_{vk}^{n_{vk}} \end{aligned}$$

Therefore, we arrive at the product form of prior multiplying likelihood:

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\varphi}, \boldsymbol{\xi}) = \text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi}) \prod_{k=1}^K \prod_{v=1}^V (\varphi_{vk} \theta_k)^{n_{vk}}$$

By collapsing the topic sequence, we derive the joint distribution of observed \mathbf{w} and latent $\boldsymbol{\theta}$:

$$\begin{aligned} p(\mathbf{w}, \boldsymbol{\theta}|\boldsymbol{\varphi}, \boldsymbol{\xi}) &= \sum_{\mathbf{z}} p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\varphi}, \boldsymbol{\xi}) \\ &= \text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi}) \prod_{v=1}^V \sum_{\{n_{vk}\}_K} \binom{n_v}{n_{v1}, \dots, n_{vK}} \prod_{k=1}^K (\varphi_{vk} \theta_k)^{n_{vk}} \\ &= \text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi}) \prod_{v=1}^V \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right)^{n_v} \end{aligned}$$

It is noteworthy to mention that a document's probability vector over word collection \mathbb{W} denoted as $\{t_v(\boldsymbol{\theta}) = \sum_{k=1}^K \varphi_{vk} \theta_k\}_V$ is determined given the document-topic proportion vector $\boldsymbol{\theta}$ and topic-word conditional probability matrix $\boldsymbol{\varphi}$:

$$\begin{aligned} \{t_v(\boldsymbol{\theta})\}_{v=1}^V &= \begin{pmatrix} \varphi_{11} & \cdots & \varphi_{1k} & \cdots & \varphi_{1K} \\ \vdots & & \vdots & & \vdots \\ \varphi_{v1} & \cdots & \varphi_{vk} & \cdots & \varphi_{vK} \\ \vdots & & \vdots & & \vdots \\ \varphi_{V1} & \cdots & \varphi_{Vk} & \cdots & \varphi_{VK} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \\ \vdots \\ \theta_K \end{pmatrix} \\ &= \left(\sum_{k=1}^K \varphi_{1k} \theta_k, \dots, \sum_{k=1}^K \varphi_{vk} \theta_k, \dots, \sum_{k=1}^K \varphi_{Vk} \theta_k \right)^\top \end{aligned}$$

By integrating out the latent variable $\boldsymbol{\theta}$, the evidence, or the marginal distribution of observed word counting vector is given:

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\varphi}, \boldsymbol{\xi}) &= \int_{\Delta^{K-1}} p(\mathbf{w}, \boldsymbol{\theta}|\boldsymbol{\varphi}, \boldsymbol{\xi}) d\boldsymbol{\theta} \\ &= \int_{\Delta^{K-1}} \text{DT}(\boldsymbol{\theta}|\boldsymbol{\xi}) \prod_{v=1}^V (t_v(\boldsymbol{\theta}))^{n_v} d\boldsymbol{\theta} \end{aligned}$$

Finally, we compute the total probability of corpus \mathcal{D} containing M documents:

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\varphi}, \boldsymbol{\xi}) &= \prod_{m=1}^M p_m(\mathbf{w}_m|\boldsymbol{\varphi}, \boldsymbol{\xi}) \\ &= \prod_{m=1}^M \int_{\boldsymbol{\theta}_m} \text{DT}(\boldsymbol{\theta}_m|\boldsymbol{\xi}) \prod_{v=1}^V (t_{vm}(\boldsymbol{\theta}_m))^{n_{vm}} d\boldsymbol{\theta}_m \end{aligned}$$

4 Mean-Field Variational Inference

The mean-field variational inference employs a global mean-field distribution to perform approximation for each latent node in the probabilistic graph. The approximated posterior is achieved by minimizing the KL-divergence:

$$\begin{aligned} q(\mathbf{z}|\boldsymbol{\eta}^*) &= \arg \min_{q \in \mathcal{Q}} \text{KL}(q(\mathbf{z}|\boldsymbol{\eta}) || p(\mathbf{z}|\mathbf{w})) \\ &= \arg \min_{q \in \mathcal{Q}} \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\eta}) \log \frac{q(\mathbf{z}|\boldsymbol{\eta})}{p(\mathbf{z}|\mathbf{w})} d\mathbf{z} \end{aligned}$$

which is equivalent to maximizing Evidence Lower Bound (ELBO):

$$\text{ELBO} = \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\eta}) \log \frac{p(\mathbf{w}, \mathbf{z})}{q(\mathbf{z}|\boldsymbol{\eta})} d\mathbf{z}$$

Since we will deal with both continuous variables (topic proportions) and discrete variables (topics) in our model, we introduce a new notation:

$$\left\langle f(x), g(x) \right\rangle_x = \begin{cases} \sum_i f(x_i)g(x_i), & \text{if } x \text{ is discrete,} \\ \int_x f(x)g(x) dx, & \text{if } x \text{ is continuous} \end{cases}$$

The feasibility of variational mean-field as an approximating distribution lies on the independency assumption between its random variables, which leads to an important property:

$$\left\langle y(x_n), q(x_1)q(x_2) \cdots q(x_n) \cdots q(x_N) \right\rangle_{\mathbf{x}} = \left\langle y(x_n), q(x_n) \right\rangle_{x_n}$$

This enables us to expand the joint distribution over every individual latent or observed variable node in the generative probabilistic graph, and deal with each ELBO term with respect to the corresponding latent nodes separately from an external viewpoint. By taking the probabilistic graph in Figure 4 as demonstrating example, we illustrate the derivation of mean-field variational inference and provide a general algorithm regardless of the specific Dirichlet-Tree prior. More inference algorithms for the generalized probabilistic graphs can be easily derived in a similar manner.

4.1 The Evidence Lower Bound

Suppose $p(\boldsymbol{\theta}|\boldsymbol{\xi})$ is a Dirichlet-Tree distribution with selection matrix \mathbf{D}_p and $p(\boldsymbol{\theta}|\boldsymbol{\xi})$ is employed as the prior distribution in the generative plate notation in Figure 4. We already

prove that $p(\boldsymbol{\theta}|\boldsymbol{\xi})$ is both a member of exponential family and a conjugate prior to the multinomial:

$$p(\boldsymbol{\theta}|\boldsymbol{\xi}) = \exp\{\boldsymbol{\eta}(\boldsymbol{\xi})^T \mathbf{u}(\boldsymbol{\theta}) - \log g(\boldsymbol{\eta}(\boldsymbol{\xi}))\}$$

Without the need to specify the concrete expression, the universal form of approximation algorithm and key steps in the optimization is demonstrated by only utilizing Dirichlet-Tree's properties of exponential family and conjugacy to Multinomial. Our goal is to approximate the joint posterior distribution $p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{w}, \boldsymbol{\varphi}, \boldsymbol{\xi})$ of all latent variables of interest given the observed sequences \mathbf{w} , and the global mean-field approximating distribution is defined as:

$$q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\zeta}, \boldsymbol{\phi}) = \prod_{m=1}^M q(\boldsymbol{\theta}|\boldsymbol{\zeta}_m) \prod_{n=1}^{N_m} q(z_{nm}|\boldsymbol{\phi}_{nm})$$

It is obvious that the global approximation wipes out the original dependency between variables such as the dependency between $\boldsymbol{\theta}$ and \mathbf{z} . Given the approximating distribution, the optimized parameters are obtained by minimizing $\text{KL}(q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\zeta}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\varphi}, \boldsymbol{\xi}))$, which is equivalent to maximizing the Evidence Lower Bound (ELBO):

$$\begin{aligned} L(\boldsymbol{\zeta}, \boldsymbol{\phi}; \boldsymbol{\xi}, \boldsymbol{\varphi}) &= \left\langle \log \frac{p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\xi}, \boldsymbol{\varphi})}{q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\zeta}, \boldsymbol{\phi})}, q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\zeta}, \boldsymbol{\phi}) \right\rangle_{(\mathbf{z}, \boldsymbol{\theta})} \\ &= \sum_{m=1}^M \left\{ \left\langle \log p(\boldsymbol{\theta}|\boldsymbol{\xi}) \right\rangle_{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_m)} + \left\langle \log p(\mathbf{z}|\boldsymbol{\theta}) \right\rangle_{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_m)q(\mathbf{z}|\boldsymbol{\phi}_m)} \right. \\ &\quad + \left\langle \log p(\mathbf{w}_m|\mathbf{z}, \boldsymbol{\varphi}) \right\rangle_{q(\mathbf{z}|\boldsymbol{\phi}_m)} - \left\langle \log q(\boldsymbol{\theta}|\boldsymbol{\zeta}_m) \right\rangle_{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_m)} \\ &\quad \left. - \left\langle \log q(\mathbf{z}|\boldsymbol{\phi}_m) \right\rangle_{q(\mathbf{z}|\boldsymbol{\phi}_m)} \right\} \\ &= \sum_{m=1}^M \left\{ \left\langle \log p(\boldsymbol{\theta}_m|\boldsymbol{\xi}) \right\rangle_{q(\boldsymbol{\theta}_m|\boldsymbol{\zeta}_m)} + \sum_{n=1}^{N_m} \left\langle \log p(z_{nm}|\boldsymbol{\theta}_m) \right\rangle_{q(\boldsymbol{\theta}_m|\boldsymbol{\zeta}_m)q(z_{nm}|\boldsymbol{\phi}_{nm})} \right. \\ &\quad + \sum_{n=1}^{N_m} \left\langle \log p(w_{nm}|z_{nm}, \boldsymbol{\varphi}) \right\rangle_{q(z_{nm}|\boldsymbol{\phi}_{nm})} - \left\langle \log q(\boldsymbol{\theta}_m|\boldsymbol{\zeta}_m) \right\rangle_{q(\boldsymbol{\theta}_m|\boldsymbol{\zeta}_m)} \\ &\quad \left. - \sum_{n=1}^{N_m} \left\langle \log q(z_{nm}|\boldsymbol{\phi}_{nm}) \right\rangle_{q(z_{nm}|\boldsymbol{\phi}_{nm})} \right\} \end{aligned}$$

Without loss of generality, we consider ELBO for any one of the documents and drop the subscript m . We have:

$$\begin{aligned}
L_m(\zeta, \phi; \xi, \varphi) &= \boldsymbol{\eta}(\xi)^T \mathbb{E}_{q(\boldsymbol{\theta}|\zeta)} [\mathbf{u}(\boldsymbol{\theta})] - \log g(\boldsymbol{\eta}(\xi)) \\
&\quad + \sum_{n=1}^N \boldsymbol{\phi}_n^T \mathbb{E}_{q(\boldsymbol{\theta}|\zeta)} [\log \boldsymbol{\theta}] \\
&\quad + \sum_{n=1}^N \boldsymbol{\phi}_n^T \log \varphi_{v(n)} \\
&\quad - \boldsymbol{\eta}(\zeta)^T \mathbb{E}_{q(\boldsymbol{\theta}|\zeta)} [\mathbf{u}(\boldsymbol{\theta})] + \log g(\boldsymbol{\eta}(\zeta)) \\
&\quad - \sum_{n=1}^N \boldsymbol{\phi}_n^T \log \boldsymbol{\phi}_n \\
&= \left[\boldsymbol{\eta}(\xi) - \boldsymbol{\eta}(\zeta) + \mathbf{D}_q \left(\sum_{n=1}^N \boldsymbol{\phi}_n \right) \right]^T \mathbb{E}_{q(\boldsymbol{\theta}|\zeta)} [\mathbf{u}(\boldsymbol{\theta})] \\
&\quad + \log g(\boldsymbol{\eta}(\zeta)) - \log g(\boldsymbol{\eta}(\xi)) \\
&\quad + \sum_{n=1}^N \boldsymbol{\phi}_n^T \log \varphi_{v(n)} - \sum_{n=1}^N \boldsymbol{\phi}_n^T \log \boldsymbol{\phi}_n
\end{aligned}$$

where the subscript v indicates the v^{th} choice from the dictionary at the n^{th} observed word, and \mathbf{D}_q represents Dirichlet selection matrix which depends only on the structure of approximator $q(\boldsymbol{\theta}|\zeta)$ regardless of the value of ζ . The tensor expression of ELBO is shown in Figure 6. It is noteworthy to mention that the form of this specific expression of ELBO depends only on the specific generative probabilistic graph with the corresponding nodes replaced by the choice of the specific Dirichlet-Tree distribution.

4.2 E-Step

• Optimization in terms of ζ

Parameter ζ belongs to $q(\boldsymbol{\theta}|\zeta)$ over a document's topic proportion $\boldsymbol{\theta}$, one term of the global approximating mean-field. We differentiate the ELBO with respect to ζ :

$$\begin{aligned}
\nabla_{\zeta} L(\zeta) &= \nabla_{\zeta} \left\{ \left[\boldsymbol{\eta}(\xi) - \boldsymbol{\eta}(\zeta) + \mathbf{D}_q \left(\sum_{n=1}^N \boldsymbol{\phi}_n \right) \right]^T \mathbb{E}_{q(\boldsymbol{\theta}|\zeta)} [\mathbf{u}(\boldsymbol{\theta})] + \log g(\boldsymbol{\eta}(\zeta)) + C \right\} \\
&= \left[\xi - \zeta + \mathbf{D}_q \left(\sum_{n=1}^N \boldsymbol{\phi}_n \right) \right]^T \nabla_{\zeta} \mathbb{E}_{q(\boldsymbol{\theta}|\zeta)} [\mathbf{u}(\boldsymbol{\theta})] \\
&= \left\{ \left[\xi - \zeta + \mathbf{D}_q \left(\sum_{n=1}^N \boldsymbol{\phi}_n \right) \right]^T \left(\frac{\partial}{\partial \zeta_d} \mathbb{E}_{q(\boldsymbol{\theta}|\zeta)} [\mathbf{u}(\boldsymbol{\theta})] \right) \right\}_{d=1}^D
\end{aligned}$$

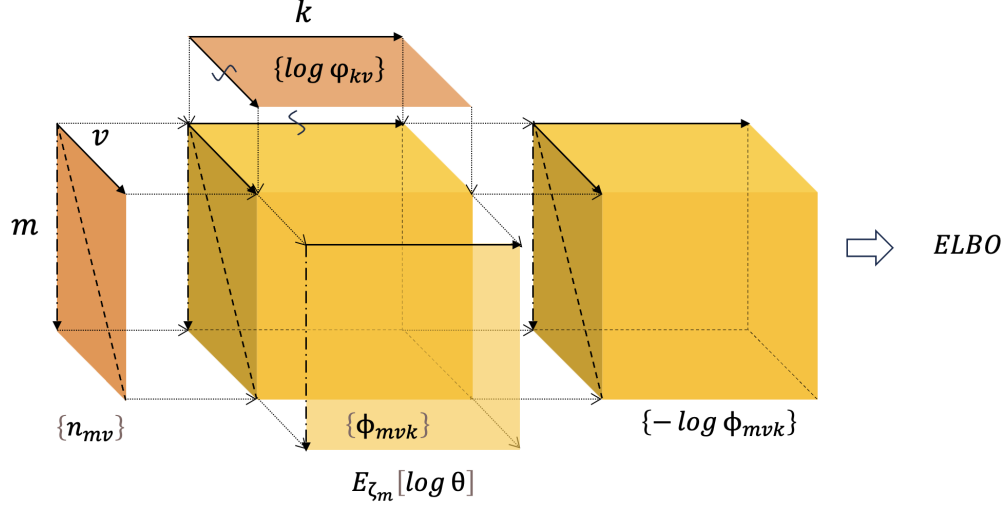


Figure 6: Tensor expression of ELBO

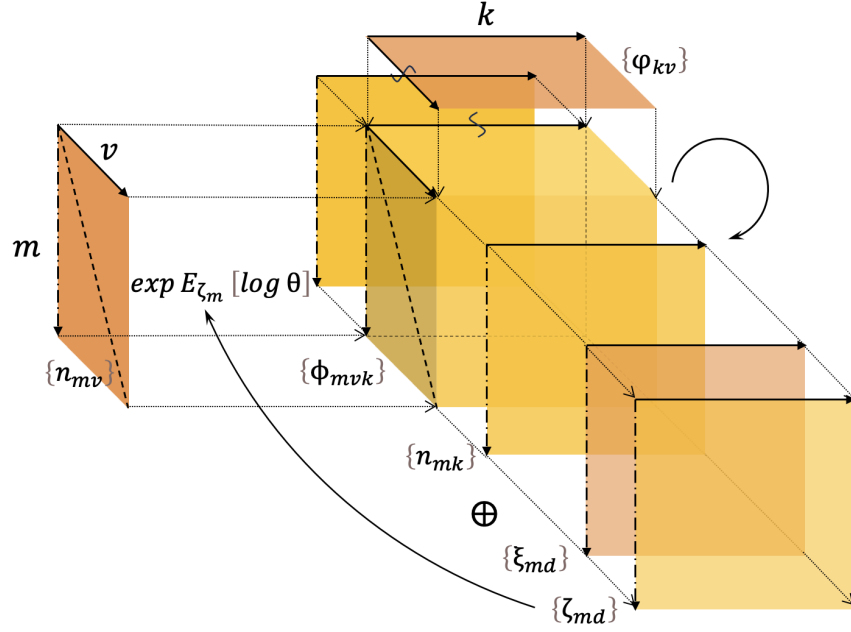


Figure 7: Expectation-Maximization iteration process of mean-field variational inference

By setting derivatives to $\mathbf{0}$, we have:

$$\zeta^* = \xi + \mathbf{D}_q \left(\sum_{n=1}^N \phi_n \right) = \xi \oplus_q \left(\sum_{n=1}^N \phi_n \right)$$

Notice that ϕ_n only depends on the observed words w_v in this single document, the optimization step can be expressed in terms of matrix multiplication:

$$\zeta^* = \xi \oplus_q \{\phi_{kv} n_v\}_K$$

• Optimization in terms of ϕ

Parameter ϕ is matrix $\{\phi_{kn}\}_{(K \times N)}$, the multinomial parameters of $\prod_{n=1}^N q(z_n | \phi_n)$ given the observed word w_n at each location n . The optimized ϕ^* is obtained by maximizing the ELBO, which is a constrained optimization problem with respect to ϕ where there are N constraints:

$$\sum_{l=1}^K \phi_{ln} = 1, \quad n = 1, 2, \dots, N$$

Without loss of generality, we consider the Lagrangian function with respect to ϕ_{kn} of ϕ_n :

$$\mathcal{L}(\phi_n) = \phi_n^T \mathbb{E}_{q(\theta|\zeta)} [\log \theta] + \phi_n^T \log \varphi_{v(n)} - \phi_n^T \log \phi_n + \sum_{n=1}^N \lambda_n \left(\sum_{l=1}^K \phi_{ln} - 1 \right) + C$$

$$\frac{\partial \mathcal{L}}{\partial \phi_{kn}} = \mathbb{E}_{q(\theta|\zeta)} [\log \theta_k] + \log \varphi_{vk} - \log \phi_{kn} - 1 + \lambda_n$$

By setting partial derivatives with respect to ϕ_{nk} zero, we have:

$$\phi_{kn}^* = \varphi_{vk} \exp \{ \mathbb{E}_{q(\theta|\zeta)} [\log \theta_k] + \lambda_n - 1 \}$$

where $\exp\{\lambda_n - 1\}$ is the normalizer.

Since $\{\phi_n\}_N$ depends only on which word $\{w_{vn}\}_N$ out of vocabulary $\{w_v\}_V$ is chosen, the matrix $\{\phi_{kn}\}_{(K \times N)}$ is transformed into $\{\phi_{kv}\}_{(K \times V)}$ for $\prod_{v=1}^V q(z | \phi_v)^{n_v}$, where n_v is the count of the v^{th} word in the document. Therefore, the optimization step can be expressed in terms of matrices:

$$\begin{aligned} \{\phi_{kv}\}_{(K \times V)}^* &= \{z_v^{-1} \varphi_{vk} \exp \{ \mathbb{E}_{q(\theta|\zeta)} [\log \theta_k] \} \}_{(K \times V)} \\ &= \left\{ z_v^{-1} \varphi_{vk} \exp \{ (\mathbf{D}_q^T) \mathbb{E}_{q(\theta|\zeta)} [\mathbf{u}(\theta)] \}_k \right\}_{(K \times V)} \end{aligned}$$

where z_v is the normalizer. The E-step of mean-field variational inference is concluded in Algorithm 1.

4.3 M-Step

Given optimized $\{q(\theta | \zeta_m^*)\}_M$ and $\{\phi_{kvm}^*\}_{(K \times V \times M)}$ after an inference step, we proceed to optimize the model's hyper-parameters ξ and $\{\varphi_{vk}\}_{V \times K}$ by maximizing ELBO on all documents \mathcal{D} :

$$L(\xi, \varphi) = \sum_{m=1}^M \left\{ \eta(\xi)^T \mathbb{E}_{q(\theta|\zeta_m^*)} [\mathbf{u}(\theta)] - \log g(\eta(\xi)) + \sum_{n=1}^{N_m} \phi_{nm}^{*T} \log \varphi_{v(nm)} \right\} + C$$

Algorithm 1: Variational mean-field inference for a single document \mathbf{w}_m

Input: Prior ξ , $\{\varphi_{vk}\}_{V \times K}$; Observation count $\{n_v\}_V$
Output: Optimized ζ , $\{\phi_{kv}\}_{K \times V}$
 Initialize each element in $\{\phi_{kv}\}_{K \times V}$ as $1/K$;
 Initialize $\zeta = \xi \oplus \{\phi_{kv} n_v\}_K$;
while not convergence **do**
 for w_v in \mathbb{W} in parallel **do**
 Compute $\{\phi_{kv}\}_{(K \times V)}^{new} = \{\varphi_{vk} \exp\{\mathbb{E}_{q(\theta|\zeta)}[\log \theta_k]\}\}_{(K \times V)}$;
 Normalize $\{\phi_{kv}\}_{(K \times V)}^{new}$ such that $\sum_{k=1}^K \phi_{kv} = 1$;
 Update $\zeta^{new} = \xi \oplus \{\phi_{kv}^{new} n_v\}_K$;

• Optimization of φ

ELBO with respect to φ is given as:

$$L(\varphi) = \sum_{m=1}^M \sum_{n=1}^{N_m} \phi_{nm}^* \log \varphi_{v(nm)} + C$$

Notice that φ has K constraints:

$$\sum_{l=1}^V \varphi_{lk} = 1, \quad k = 1, 2, \dots, K$$

and the Lagrangian is given as:

$$\begin{aligned} \mathcal{L}(\varphi) &= \sum_{m=1}^M \sum_{n=1}^N \phi_{nm}^* \log \varphi_{v(nm)} + \sum_{k=1}^K \lambda_k \left(\sum_{v=1}^V \varphi_{vk} - 1 \right) \\ &= \sum_{v=1}^V \sum_{k=1}^K \left(\sum_{m=1}^M \phi_{kvm}^* n_{vm} \right) \log \varphi_{vk} + \sum_{k=1}^K \lambda_k \left(\sum_{v=1}^V \varphi_{vk} - 1 \right) \end{aligned}$$

Without loss of generality, the partial derivatives of Lagrangian with respect to φ_{vk} is written as:

$$\frac{\partial \mathcal{L}}{\partial \varphi_{vk}} = \left(\sum_{m=1}^M \phi_{kvm}^* n_{vm} \right) \frac{1}{\varphi_{vk}} + \lambda_k$$

By setting the partial derivatives to zero, we have:

$$\varphi_{vk}^* = -\frac{1}{\lambda_k} \left(\sum_{m=1}^M \phi_{kvm}^* n_{vm} \right)$$

Therefore, the optimized $\{\varphi_{vk}\}_{(V \times K)}^*$ is derived as:

$$\{\varphi_{vk}\}_{(V \times K)}^* = \left\{ z_k^{-1} \sum_{m=1}^M \phi_{kvm}^* n_{vm} \right\}_{(V \times K)}$$

where $\{z_k\}_K$ is the normalizer.

- **Optimization of ξ**

ELBO with respect to ξ is given as:

$$L(\xi) = \boldsymbol{\eta}(\xi)^T \left(\sum_{m=1}^M \mathbb{E}_{q(\boldsymbol{\theta}|\zeta_m^*)} [\mathbf{u}(\boldsymbol{\theta})] \right) - M \log g(\boldsymbol{\eta}(\xi)) + C$$

The first-order derivative is written as:

$$\nabla_{\xi} L(\xi) = \left\{ \sum_{m=1}^M \mathbb{E}_{q(\boldsymbol{\theta}|\zeta_m)} [u_d(\boldsymbol{\theta})] - M \mathbb{E}_{p(\boldsymbol{\theta}|\xi)} [u_d(\boldsymbol{\theta})] \right\}_{d=1}^D$$

By setting the derivative to zero, we can see that the optimized ξ is obtained by matching the moment of hyper-Dirichlet with parameter ξ to the average of those of the inferred posteriors, which is a similar process to the message passing in Expectation Propagation. Thus, we proceed with Newton's method, and the Hessian is derived as:

$$\mathbf{H}_{\xi} L(\xi) = -M \left\{ \frac{\partial}{\partial \xi_i} \mathbb{E}_{p(\boldsymbol{\theta}|\xi)} [u_j(\boldsymbol{\theta})] \right\}_{(D \times D)}$$

The optimized ξ^* can be approximated iteratively by Newton's method:

$$\xi^{(t+1)} := \xi^{(t)} - \mathbf{H}(\xi^{(t)})^{-1} \nabla(\xi^{(t)})$$

The tensor expression of the whole Expectation-Maximization is shown in Figure 7.

5 Expectation Propagation

In Bayesian inference, we compute the posterior distribution on latent variables given the observation. In mean-field variational inference, the algorithm employs a global mean-field distribution to perform approximation for each latent node in the probabilistic graph. In contrast, Expectation Propagation (EP) separates target distributions into terms, approximate each one and combine together to obtain a global approximation.

5.1 Formal Alignment of Approximate Distribution to Target Posterior

In EP for LDTA, we focus on the marginal posterior on $\boldsymbol{\theta}$ given the observed sequence of words incorporating the model's hyper-parameters. For a single document we have the target marginal posterior:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{w}, \boldsymbol{\varphi}, \xi) &= \frac{p(\mathbf{w}, \boldsymbol{\theta}|\boldsymbol{\varphi}, \xi)}{p(\mathbf{w}|\boldsymbol{\varphi}, \xi)} \\ &\propto p(\boldsymbol{\theta}|\xi) \prod_{v=1}^V \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right)^{n_v} \\ &= p(\boldsymbol{\theta}|\xi) \exp \left\langle n_v, \log t_v(\boldsymbol{\theta}; \boldsymbol{\varphi}) \right\rangle_v \end{aligned}$$

where

$$t_v(\boldsymbol{\theta}; \boldsymbol{\varphi}) = \sum_{k=1}^K \varphi_{vk} \theta_k$$

We can see that the hierarchical probabilistic graph leads to the transform of likelihood breaking the conjugacy, where the compound of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ makes it impossible to revise the parameters directly from prior, and eventually causes the whole posterior to be intractable. Therefore, we need to employ approximating methods such as Variational Inference or Expectation Propagation with tractable distributions to estimate the intractable target posterior. The mean-field variational inference is introduced in the last section, and we demonstrate the Expectation Propagation routine in the following.

• Dual conjugate expressions

In EP, we try to approximate the target distribution by an approximating distribution with the similar form:

$$q(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \left\langle n_v, \log \tilde{t}_v(\boldsymbol{\theta}) \right\rangle_v$$

In contrast to variational mean-field where we simply optimize the approximating distribution's parameter by maximizing the ELBO, EP performs a more in-depth fine-grained calibration based on the structure of target distribution by constraining the functional space of the approximating distributions. In addition to the requirements of tractability such as pre-normalization and independence that variational mean-field complies, EP further expects the production form of the approximating distribution, in accordance with the prior-likelihood-points product form of the target distribution. Moreover, due to the nature of likelihood, the observation $\mathbf{n}_{(v)}$ is incorporated into the approximator, which in effect imposes a constraint on the functional space of the approximating distribution. Given the bijective correspondence of the terms between the approximator and the target, distributed term-to-term approximation is achieved by moment matching in a serial or parallel manner, while the simplicity of global normalization is maintained. Based on the further requirements on approximator in EP, Dirichlet-Tree with the same selection matrix \mathbf{D} as the prior's is chosen as the approximator of target posterior, thanks to its dual conjugate expressions:

$$\begin{aligned} q(\boldsymbol{\theta}|\boldsymbol{\zeta}_{\mathbf{n}}) &= q(\boldsymbol{\theta}|\boldsymbol{\xi} + \mathbf{D}_q \mathbf{n}_{(\mathbf{k})}) \\ &= \mathbf{B}_q(\mathbf{n}_{(\mathbf{k})}) p(\boldsymbol{\theta}|\boldsymbol{\xi}) \\ &= \left\{ \left[\frac{g(\boldsymbol{\eta}(\boldsymbol{\zeta}_{\mathbf{n}}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \right]^{-1} \exp \left\langle n_k, \log \theta_k \right\rangle_k \right\} p(\boldsymbol{\theta}|\boldsymbol{\xi}) \\ &\propto p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \left\langle n_k, \log \theta_k \right\rangle_k \end{aligned}$$

where \mathbf{D}_q is Dirichlet selection matrix and \mathbf{B}_q the Bayesian operator. We call the first expression whose parameters are directly revised the explicit form of approximator, and the second expression conforming the target the implicit form of approximator. It is noteworthy to mention that the global normalizer is determined if latent $\{n_k\}_K$ on the set of topic is given.

- **Explicitization of pseudo-observation**

Even though it closely resembles the target distribution, the inner product expression for the likelihood terms is with respect to the topic dimension k which relies on the latent variable \mathbf{z} , instead of the word dimension v in those of the target distribution which is drawn directly from the observed \mathbf{w} . To continue constructing our approximate distribution, we introduce the transition matrix $\{\phi_{kv}\}_{(K \times V)}$, which leads to the counting matrix of pseudo-observation $\{\phi_{kv}n_v\}_{(K \times V)}$, so that the vector $\{\phi_{kv}n_v\}_K$ approximates to $\{n_k\}_K$. Observation transition matrix ϕ can roughly be viewed as the pseudo-proportions of k^{th} topic in the authentic count of v^{th} word in a document, therefore the normalization is required. And we call the reverse transform procedure $\{\phi_{kv}n_v\}_{(K \times V)}$, which formulates an explicit counting matrix representing both the observed and the latent variables, the explicitization of pseudo-observation on latent variable—in our case the latent topics as direct results of the stochastic procedure subject to the multinomial whose parameters are dominated by the Dirichlet-Tree prior. We say “pseudo” because the reverse transform does not strictly reflect the genuine integer counts as described in our model, but gives the decimal estimations instead. Therefore, we have:

$$\begin{aligned}
q(\boldsymbol{\theta}|\boldsymbol{\zeta}_n) &\propto p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \left\langle \sum_v \phi_{kv} n_v, \log \theta_k \right\rangle_k \\
&= p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \sum_v n_v \left\langle \phi_{kv}, \log \theta_k \right\rangle_k \\
&= p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \left\langle n_v, \left\langle \phi_{kv}, \log \theta_k \right\rangle_k \right\rangle_v \\
&= p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \left\langle n_v, \log \prod_k \theta_k^{\phi_{kv}} \right\rangle_v
\end{aligned}$$

It is easy to tell that the explicitization of pseudo-observation is the process of reversely transforming from the authentic observed variable to some approximated latent variables, which are usually the unseen results of random process controlled by the parameters of interest, so that the information from authentic observation can be passed to an object which helps build a tractable explicit approximate posterior on those parameters of interest. As desired by the inference, the resulted approximate posterior remains a similar implicit product form in compliance with the target prior-likelihood-points posterior. We call the object the pseudo-observation.

- **Term-wise synchronization**

Now, we have a quite similar form to the target distribution with $t(\boldsymbol{\theta}; \boldsymbol{\varphi})$ replaced by $\prod_k \theta_k^{\phi_{kv}}$. According to Jensen’s inequality:

$$\prod_{k=1}^K \theta_k^{\phi_{kv}} = \exp \left\{ \sum_{k=1}^K \phi_{kv} \log \theta_k \right\} \leq \sum_{k=1}^K \phi_{kv} \theta_k$$

where $\sum_{k=1}^K \phi_{kv} = 1$. However, we have:

$$0 \leq \sum_{k=1}^K \varphi_{vk} \leq K$$

It is likely that the corresponding approximate term will be smaller than the target term—the bigger the topic cardinality K is, the more likely $\prod_k \theta_k^{\phi_{kv}}$ will be smaller than $t(\boldsymbol{\theta}; \boldsymbol{\varphi})$. To solve this inherent limitation, we add a coefficient for each approximate term to synchronize with the target term, which will be removed by global normalizer later. As a result, we obtain our final form of approximating likelihood terms:

$$\tilde{t}_v(\boldsymbol{\theta}; \boldsymbol{\phi}) = s_v \prod_{k=1}^K \theta_k^{\phi_{kv}}$$

where s_v is the synchronization coefficient such that each $\tilde{t}_v(\boldsymbol{\theta})$ approximates $t_v(\boldsymbol{\theta})$. With each approximating term obtained, we derive our final approximating distribution expressions:

$$\begin{aligned} q(\boldsymbol{\theta}|\boldsymbol{\zeta}_n) &= q(\boldsymbol{\theta}|\boldsymbol{\xi} + \mathbf{D}_q \boldsymbol{\phi} \mathbf{n}_{(v)}) \\ &= \mathbf{B}_q(\boldsymbol{\phi} \mathbf{n}_{(v)}) p(\boldsymbol{\theta}|\boldsymbol{\xi}) \\ &= \left[\frac{g(\boldsymbol{\eta}(\boldsymbol{\zeta}_n))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \exp \left\langle n_v, \log s_v \right\rangle_v \right]^{-1} p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \left\langle n_v, \log s_v \prod_{k=1}^K \theta_k^{\phi_{kv}} \right\rangle_v \\ &\propto p(\boldsymbol{\theta}|\boldsymbol{\xi}) \exp \left\langle n_v, \log \tilde{t}_v(\boldsymbol{\theta}; \boldsymbol{\phi}) \right\rangle_v \end{aligned}$$

Now, it is clear that the transition matrix $\{\phi_{kv}\}_{K \times V}$ works as a transformer which converts the observation over vocabulary to pseudo-observation over latent topics to achieve a conjugate posterior on the latent variable of interest, and to build a series of cell likelihoods in compliance with the likelihood of observations conforming the target posterior as a whole without the loss of global normalization. Given the transition matrix $\{\phi_{kv}\}_{K \times V}$, observation $\mathbf{n}_{(v)}$, and hyper-parameter $\boldsymbol{\xi}$, the parameter $\boldsymbol{\zeta}_n$ for the explicit approximating distribution is given as follows:

$$\boldsymbol{\zeta}_n = \boldsymbol{\xi} \oplus \boldsymbol{\phi} \mathbf{n}_{(v)}$$

It is obvious to tell that the implicit form of approximate distribution is now in alignment with the target posterior which holds a prior-likelihood-points product form. The process of achieving this bijectively corresponding alignment from approximate distribution to the target posterior is called the alignment of approximate distribution. Figure 8 illustrates the whole picture over the relationships among the target posterior, the explicit approximate distribution, and implicit approximate distribution.

5.2 Cavity and Tilted Distributions

Given the explicit and implicit expression of the approximating distribution, the approximator's statistical manifold is fixed and the task now is to determine an optimized group of the

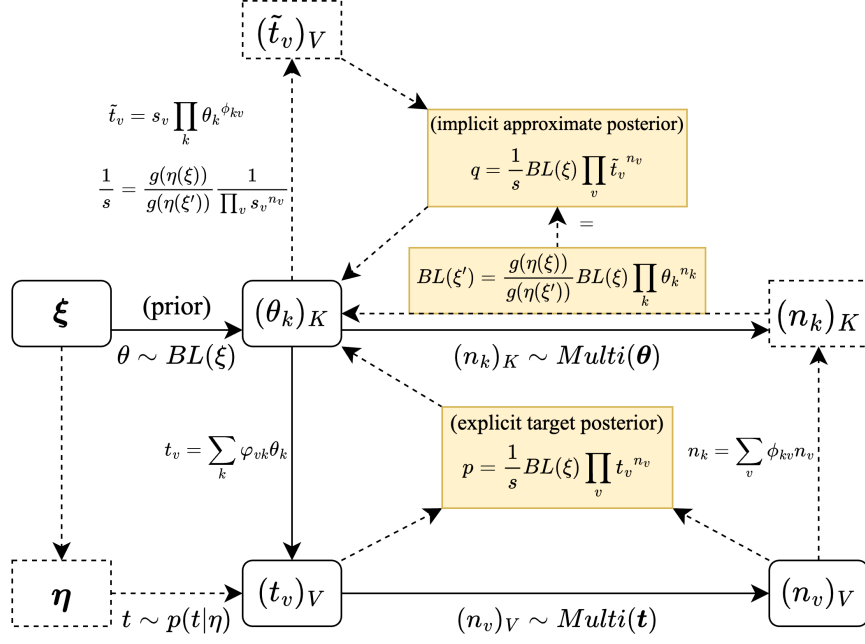


Figure 8: Construction of approximate posterior distribution and its relationship with other parameters. The $(n_v)_V$ and $(n_k)_K$ represents observed count vector of words, and latent count vector of topics, respectively. In explicit posterior distribution, the normalizer, or the evidence $s(\xi, \varphi)$ is intractable, which is the reason why we have to use approximate method. Notice $\prod_k \theta_k^{\sum_v \phi_{kv} n_v} = \prod_v (\prod_k \theta_k^{\phi_{kv}})^{n_v}$. ξ' is the revision of prior parameter ξ which incorporates the latent count of topics $(n_k)_K$ based on the property of conjugacy.

parameter $(\phi^*, s_{(v)}^*)$ as the final approximation to target posterior. With the term-to-term bijective correspondence between the approximating distribution and the target posterior, the problem of global approximation is transformed into the approximation of each term $\tilde{t}_v(\theta; \phi)$ to $t_v(\theta; \varphi)$ individually. A series of cavity and tilted distributions, $\{q(\theta|\zeta^{\setminus v})\}_V$ and $\{p_v^*(\theta)\}_V$ with respect to each term $t_v(\theta; \varphi)$ or $\tilde{t}_v(\theta; \phi)$ are built to perform the proceeding approximations in a distributed manner.

Theoretically, for an ideal approximator $q(\theta|\zeta^*)$ where $\{\tilde{t}_v(\theta; \phi^*) = t_v(\theta; \varphi)\}_V$, it would cause no effect if any $\tilde{t}_v(\theta; \phi^*)$ is divided out and replaced by $t_v(\theta; \varphi)$. However, for any initialized value $\phi^{(0)}$, the statement will not hold in general, meaning that the resulted new distributions are neither normalized nor proportional to the original approximator. The key of EP is that, due to the nature of the approximating distribution we choose, a group of normalized distributions $\{p_v^*(\theta)\}_V$ with respect to every individual term, which we call the titled distributions, can be formulated by unloading the approximate term and loading the target term, so that one local individual term is separated and focused on each time and further optimizations can be performed with respect to each term iteratively. Suppose we have the approximator $q(\theta|\zeta^{(t)})$ in an iteration. Normalized as they are, the corresponding $\{p_v^{*(t)}(\theta)\}_V$ are not proportional to $q(\theta|\zeta^{(t)})$ because there is still a distance between $q(\theta|\zeta^{(t)})$ and target $p(\theta|n_{(v)}, \varphi, \xi)$. Besides, the tilted distributions containing informa-

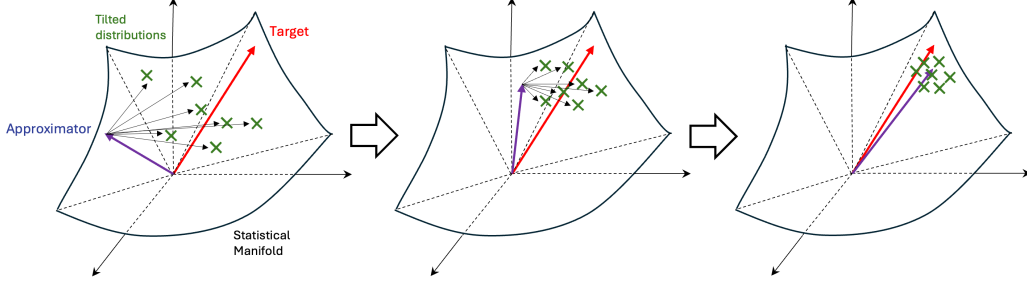


Figure 9: EP inference in each iteration. The statistical manifold is conceptually illustrated with three dimensions. The target, the approximator and the tilted distributions are all located on the statistical Riemann manifold thanks to the tractability of the approximating distribution we choose. The approximator is drawn to the target by the group of tilted distributions in each iteration.

tion from both $t_v(\boldsymbol{\theta}; \boldsymbol{\varphi})$ and $q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{(t)})$ locate between the target and current approximator. Starting from the “imperfect” initial approximator $q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{(0)})$, the approximator is drawn closer to its titled distributions in each iteration, and therefore is drawn closer to the target posterior, until the iteration converges when $q(\boldsymbol{\theta}|\boldsymbol{\zeta}^*)$, $\{p_v^*(\boldsymbol{\theta})\}_V$, and $p(\boldsymbol{\theta}|\mathbf{n}_{(v)}, \boldsymbol{\varphi}, \boldsymbol{\xi})$ almost overlap (Figure 9).

The update of the approximator is further elaborated as follows. In one iteration, suppose we have the transition matrix $\boldsymbol{\phi}^{(t)}$, and the cavity distributions $\{q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v})\}_V$ playing the role as a group of background distributions which are formulated as:

$$q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v}) = \mathbf{B}^{-1}(\boldsymbol{\phi}_v)\mathbf{B}(\boldsymbol{\phi}\mathbf{n}_{(v)})p(\boldsymbol{\theta}|\boldsymbol{\xi})$$

Notice that the cavity distributions are naturally normalized due to tractability of approximating distribution. Subsequently, the normalized tilted distributions are derived as:

$$p_v^*(\boldsymbol{\theta}) = z_v^{-1} \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v})$$

where the normalizer $\{z_v\}_V$ is trivial to compute:

$$z_v = \int_{\boldsymbol{\theta}} \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v}) d\boldsymbol{\theta} = \boldsymbol{\varphi}_v^T \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v})} [\boldsymbol{\theta}]$$

The messages from the target term contained in the titled distributions are subsequently passed to a group of new separate approximators, which are of same selection matrix \mathbf{D}_q as the current approximator, by minimizing the KL-divergence:

$$\boldsymbol{\zeta}_v^{sep} := \arg \min_{\boldsymbol{\zeta}_v^{sep}} \text{KL}(p_v^*(\boldsymbol{\theta}) || q(\boldsymbol{\theta}|\boldsymbol{\zeta}_v^{sep})) \quad (12)$$

This process is equivalent to match the moment of two distributions and will be introduced in the next section. After obtaining the optimized $\{\boldsymbol{\zeta}_v^{sep}\}_V$, the parameter $\boldsymbol{\phi}$ is updated

such that:

$$q(\boldsymbol{\theta}|\boldsymbol{\zeta}_v^{sep}) = \mathbf{B}(\phi_v^{(t+1)})q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v}) = q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v} + \mathbf{D}_q\phi_v^{(t+1)})$$

and

$$\phi_v^{(t+1)} := \mathbf{D}_q^{-1}(\boldsymbol{\zeta}_v^{sep} - \boldsymbol{\zeta}^{\setminus v})$$

where \mathbf{D}_q^{-1} is the inverse selection operator. In practice, we only focus on the leaf nodes of \mathbf{D}_q to obtain the updated $\phi^{(t+1)}$.

The optimization of synchronization coefficient $\{s_v\}_V$ is the next step. Suppose we have $\phi^{(t)}$ in an iteration, and so we have the approximate distribution (regardless of explicit or implicit form): $q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{(t)}) = \mathbf{B}(\phi^{(t)})\mathbf{n}_{(v)}p(\boldsymbol{\theta}|\boldsymbol{\xi})$. Ideally, if $q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{(t)})$ approximates target posterior enough, the following equation of the corresponding tilted distributions should hold:

$$\begin{aligned} q_v^{*(t)}(\boldsymbol{\theta}) &\equiv q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{(t)}) \left(s_v \prod_{k=1}^K \theta_k^{\phi_{kv}^{(t)}} \right)^{-1} \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) \\ &= \mathbf{B}^{-1}(\phi_v) \mathbf{B}(\phi \mathbf{n}_{(v)}) p(\boldsymbol{\theta}|\boldsymbol{\xi}) \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) \frac{g(\boldsymbol{\eta}(\boldsymbol{\zeta}^{\setminus v}))}{g(\boldsymbol{\eta}(\boldsymbol{\zeta}))} s_v^{-1} \\ &\approx \mathbf{B}^{-1}(\phi_v) \mathbf{B}(\phi \mathbf{n}_{(v)}) p(\boldsymbol{\theta}|\boldsymbol{\xi}) \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) z_v^{-1} \end{aligned}$$

where $v = 1, 2, \dots, V$. Consequently, we obtain the update of $\{s_v\}_V$:

$$s_v^{(t)} := \frac{g(\boldsymbol{\eta}(\boldsymbol{\zeta}^{\setminus v(t)}))}{g(\boldsymbol{\eta}(\boldsymbol{\zeta}^{(t)}))} z_v^{(t)}$$

According to the above discussion, the parameter update as preparation for message passing in each iteration is concluded as Algorithm 2. The whole inference procedure for a single document is concluded as Algorithm 3. The relationship between approximator, target, cavity and tilted distributions is illustrated in Figure 10.

5.3 Parallel Message Passing

This section elaborates in detail the process of message passing, which is implemented by moment matching as indicated in (12). The purpose of message passing is to pass the information of target terms $\{t_v(\boldsymbol{\theta}; \boldsymbol{\varphi})\}_V$ from the tilted distributions $\{p_v^*(\boldsymbol{\theta})\}_V$ to a group of separate approximators $\{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_v^{sep})\}_V$ respectively, so that the updated $\phi^{(t+1)}$ can subsequently be obtained by $\{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_v^{sep})\}_V$ peeling off the background $\{q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\setminus v})\}_V$ to support the newer approximation $\{t_v(\boldsymbol{\theta}; \phi^{(t+1)})\}_V$ to $\{t_v(\boldsymbol{\theta}; \boldsymbol{\varphi})\}_V$. The moment matching is achieved in parallel with respect to index v .

According to Theorem 13 in Appendix A, the optimized separate approximators $\{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_v^{sep})\}_V$ defined in (12) are the ones whose expectations on sufficient statistics equal to those expectations of the titled distributions. In other words, the optimization process of $\{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_v^{sep})\}_V$ is equivalent to matching the moment of $\{q(\boldsymbol{\theta}|\boldsymbol{\zeta}_v^{sep})\}_V$ to $\{p_v^*(\boldsymbol{\theta})\}_V$. This parallel optimization process can be expressed as follows. On the one hand, thanks to the property of

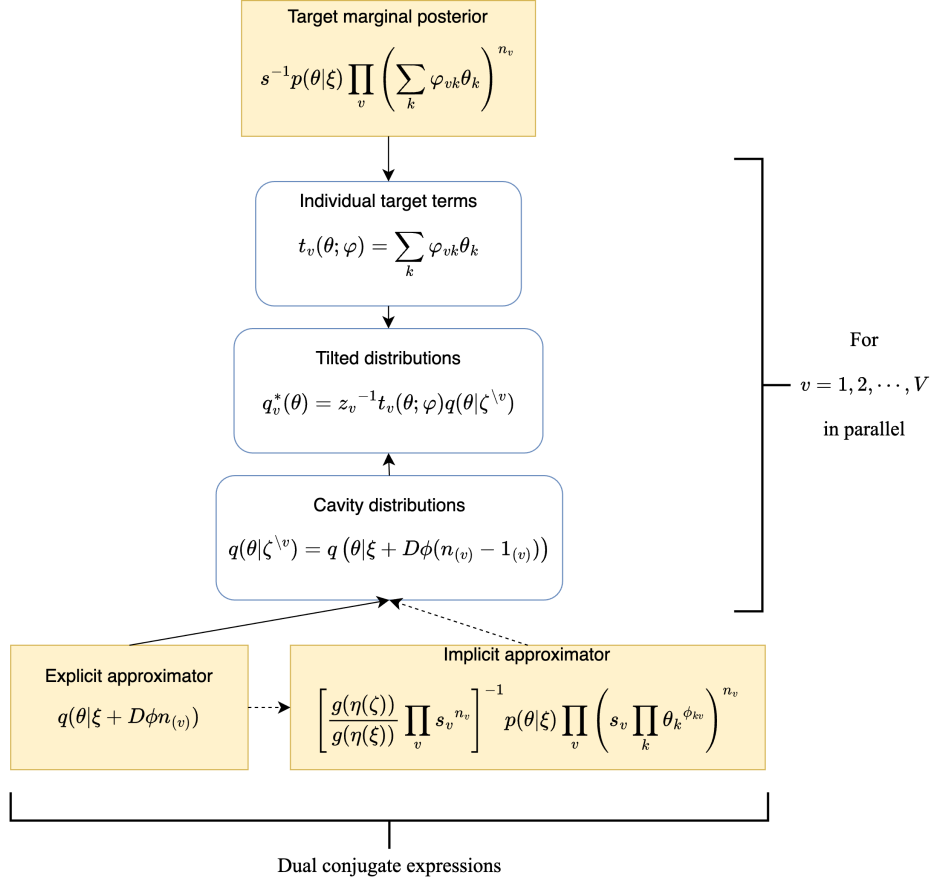


Figure 10: Relationships of Cavity and Tilted distributions

exponential family, the expectations on sufficient statistics of separate approximators are easily derived, which span a $(D \times V)$ matrix $\mathbf{Q}(\zeta^{sep})$ with unknowns ζ^{sep} :

$$\mathbf{Q} = \begin{pmatrix} \mathbb{E}_{q_1^{sep}}[u_1(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{q_v^{sep}}[u_1(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{q_V^{sep}}[u_1(\boldsymbol{\theta})] \\ \vdots & & \vdots & & \vdots \\ \mathbb{E}_{q_1^{sep}}[u_d(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{q_v^{sep}}[u_d(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{q_V^{sep}}[u_d(\boldsymbol{\theta})] \\ \vdots & & \vdots & & \vdots \\ \mathbb{E}_{q_1^{sep}}[u_D(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{q_v^{sep}}[u_D(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{q_V^{sep}}[u_D(\boldsymbol{\theta})] \end{pmatrix}$$

On the other hand, the $(D \times V)$ -dimensional expectation matrix $\mathbf{P}(\phi^{(t)})$ on the same group of sufficient statistics with respect to the tilted distributions $\{p_v^*(\boldsymbol{\theta})\}_V$ is formulated as:

$$\mathbf{P} = \begin{pmatrix} \mathbb{E}_{p_1^*}[u_1(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{p_v^*}[u_1(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{p_V^*}[u_1(\boldsymbol{\theta})] \\ \vdots & & \vdots & & \vdots \\ \mathbb{E}_{p_1^*}[u_d(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{p_v^*}[u_d(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{p_V^*}[u_d(\boldsymbol{\theta})] \\ \vdots & & \vdots & & \vdots \\ \mathbb{E}_{p_1^*}[u_D(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{p_v^*}[u_D(\boldsymbol{\theta})] & \cdots & \mathbb{E}_{p_V^*}[u_D(\boldsymbol{\theta})] \end{pmatrix}$$

Therefore, the moment matching is more specifically expressed by solving a system of $(D \times V)$ non-linear equations:

$$\mathbf{Q} \left(\zeta^{sep} \left(\phi^{(t+1)} \right) \right) := \mathbf{P} \left(\phi^{(t)} \right) \quad (13)$$

This system of equations can be more explicitly reduced to several sub-groups of equations with respect to the internal nodes of the tree structure in the following form:

$$\psi(\zeta_{t|s}) - \psi \left(\sum_{t|s} \zeta_{t|s} \right) = C_{t|s}, \quad s \in \Lambda$$

where $\{\zeta_{t|s}\}$ is the sub-group of child nodes of node s and $C_{t|s}$ the constant from \mathbf{P} . Each individual sub-group of equations can be solved numerically by fix-point iteration or Newton's method. Notice that \mathbf{P} is determined given $\phi^{(t)}$ of current iteration:

$$\begin{aligned} \mathbb{E}_{p_v^*}[u_d(\boldsymbol{\theta})] &= \int_{\boldsymbol{\theta}} \frac{u_d(\boldsymbol{\theta})}{z_v} \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v}) d\boldsymbol{\theta} \\ &= \frac{\mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[u_d(\boldsymbol{\theta})]}{z_v} \int_{\boldsymbol{\theta}} \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) \frac{1}{\mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[u_d(\boldsymbol{\theta})]} u_d(\boldsymbol{\theta}) q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v}) d\boldsymbol{\theta} \\ &= \mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[u_d(\boldsymbol{\theta})] \frac{\boldsymbol{\varphi}_v^T \mathbb{E}_{q_d(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[\boldsymbol{\theta}]}{\boldsymbol{\varphi}_v^T \mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[\boldsymbol{\theta}]} \\ &= \frac{\boldsymbol{\varphi}_v^T \left(\mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[\boldsymbol{\theta}] \odot_k \mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v} + \mathbf{D}_q \mathbf{1}_{(k)})}[u_d(\boldsymbol{\theta})] \right)}{\boldsymbol{\varphi}_v^T \mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[\boldsymbol{\theta}]} \\ &= \frac{\left(\boldsymbol{\varphi}_v \odot_k \mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[\boldsymbol{\theta}] \right)^T \mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v} + \mathbf{D}_q \mathbf{1}_{(k)})}[u_d(\boldsymbol{\theta})]}{\left(\boldsymbol{\varphi}_v \odot_k \mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[\boldsymbol{\theta}] \right)^T \mathbf{1}} \end{aligned} \quad (14)$$

The (14) uses the result from (11), where $\{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v} + \mathbf{D}_q \mathbf{1}_{(k)})\}_{(K \times V)}$ is the group of base posteriors of $\{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})\}_V$. We can see that the right-hand side of the non-linear equations is reduced to quite a concise and elegant closed-form solution, meaning that the moment of a tilted distribution equals to the weighted average of the counterparts of its corresponding cavity distribution's base posteriors, where the weights are given by the Hadamard product of $\boldsymbol{\varphi}$ and $\{\mathbb{E}_{q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\setminus v})}[\boldsymbol{\theta}]\}_{(V \times K)}$.

5.4 Parameter Estimation

We embedded the EP as the E-step into Expectation-Maximization framework. The optimized model parameters $\boldsymbol{\xi}$ and $\{\varphi_{vk}\}_{V \times K}$ measured by Kullback-Leibler divergence are esti-

Algorithm 2: Essential parameters update

Input: Prior ξ , $\{\varphi_{vk}\}_{V \times K}$; Observation $\{n_v\}_V$; $\{\phi_{kv}\}_{K \times V}$ of current iteration
Output: ζ ; $\{\zeta^{\setminus v}\}_V$, $\{z_v\}_V$ and $\{s_v\}_V$; evidence $p(\mathbf{w}_m)$ of current iteration
 Compute $\zeta = \xi \oplus \phi_m \mathbf{n}_m$;
for w_v *in* \mathbb{W} *in parallel* **do**
 Compute $\zeta^{\setminus v} = \zeta \ominus \phi_v$;
 Compute $z_v = \varphi_v^T \mathbb{E}_{q(\theta|\zeta^{\setminus v})}[\theta]$;
 Compute $s_v = z_v g(\eta(\zeta^{\setminus v}))/g(\eta(\zeta))$;
 Compute $p(\mathbf{w}_m) = \exp\langle n_v, \log s_v \rangle_v g(\eta(\zeta))/g(\eta(\xi))$;

Algorithm 3: EP inference for a single document \mathbf{w}_m

Input: Prior ξ , $\{\varphi_{vk}\}_{V \times K}$; Observation count $\{n_v\}_V$
Output: Optimized $\{\phi_{kv}^*\}_{(K \times V)}$, $\{s_v^*\}_V$; Estimated evidence $p(\mathbf{w}_m)$
 Initialize each element in $\{\phi_{kv}\}_{K \times V}$ as $1/K$;
 Initialize essential parameters as Algorithm 2;
while *not convergence* **do**
 for w_v *in* \mathbb{W} *in parallel* **do**
 Compute ζ_v^{sep} by moment matching as Algorithm 4;
 Compute $\phi_v^{(t+1)} = \zeta_v^{sep} \ominus \zeta^{\setminus v}$;
 Normalize $\phi_v^{(t+1)}$ such that $\sum_k \phi_{kv} = 1$;
 Update essential parameters as Algorithm 2;

Algorithm 4: Moment matching

Input: Essential parameters obtained in Algorithm 2
Output: Optimized separate approximators $\{\zeta_v^{sep}\}_V$
for w_v *in* \mathbb{W} *in parallel* **do**
 Compute $\mathbb{E}_{q(\theta|\zeta^{\setminus v})}[\theta]$;
 for $d = 1$ *to* D *in parallel* **do**
 for $k = 1$ *to* K *in parallel* **do**
 Compute $\mathbb{E}_{q(\theta|\zeta^{\setminus v} + \mathbf{D}_q \mathbf{1}_{(k)})}[u_d(\theta)]$;
 Compute $\mathbb{E}_{p_v^*}[u_d(\theta)]$ according to (14);
 Compute $\{\zeta_v^{sep}\}_V$ by Newton's method according to (13);

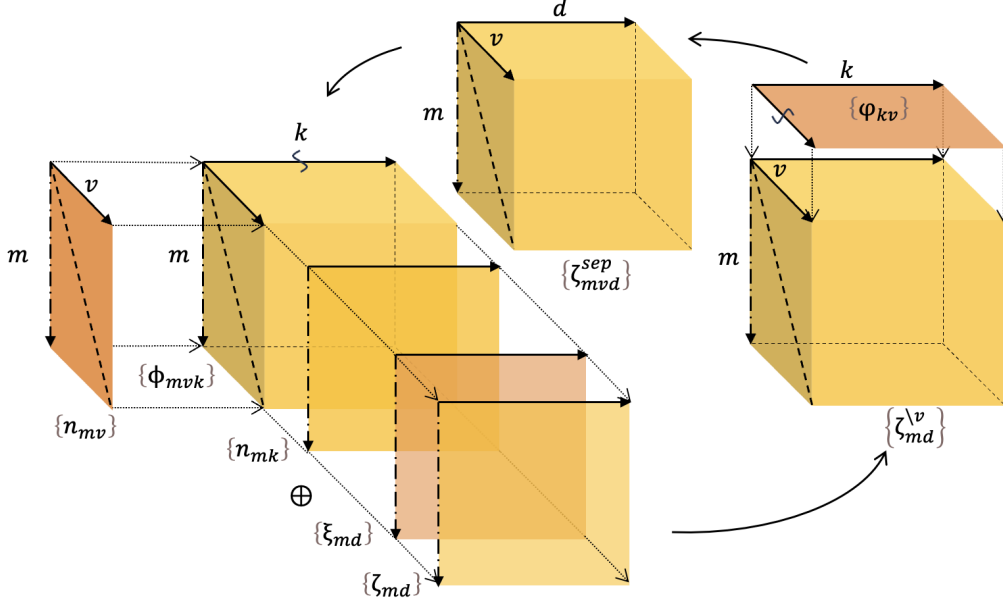


Figure 11: Tensor expression of Expectation Propagation

mated by maximizing the Evidence Lower Bound (ELBO) with respect to ξ and $\{\varphi_{vk}\}_{V \times K}$:

$$\begin{aligned}
 L(\xi, \varphi) &= \sum_{m=1}^M \mathbb{E}_{q(\theta|\zeta_m)} [\log p(\theta|\xi)] + \sum_{m=1}^M \mathbb{E}_{q(\theta|\zeta_m)} [\log p(w_m|\varphi, \theta)] + C \\
 &= \sum_{m=1}^M \left\{ \eta(\xi)^T \mathbb{E}_{q(\theta|\zeta_m)} [\mathbf{u}(\theta)] - \log g(\eta(\xi)) + \mathbb{E}_{q(\theta|\zeta_m)} \left[\sum_{v=1}^V n_{vm} \log \left(\sum_{k=1}^K \varphi_{vk} \theta_k \right) \right] \right\} + C
 \end{aligned} \tag{15}$$

The optimization is performed over the collection of all documents \mathcal{D} . The whole learning algorithm is depicted in Algorithm 5, and the tensor expression is shown in Figure 11.

6 Experiments

In this section, we test and evaluate LDTA with three different kinds of Dirichlet-Tree priors: the Dirichlet, the Beta-Liouville and the generalized Dirichlet. The details of these three distributions are listed in Appendix C. Both two inference methods are examined in three problem domains: document modeling, document and image classification, and a bioinformatics task. Our experiments are performed on several publicly available and well-known datasets: NIPS, Reuters-21578, 20 Newsgroups, 15 Scene Categories and Peripheral Blood Mononuclear Cells (PBMC).

Algorithm 5: EP-embedded EM algorithm

Input: Document collection \mathcal{D}

Output: Approximated distribution parameters $\{\phi_{kvm}\}_{K \times V \times M}$, $\{s_{vm}\}_{V \times M}$,
estimated evidence $\{\hat{p}(\mathbf{w}_m)\}_M$; Model parameters φ, ξ

Initialize φ, ξ ;

while *not convergence* **do**

 // EP-Inference

for \mathbf{w}_m *in* \mathcal{D} **do**

 Given model parameters, approximate the posterior by EP as Algorithm 3;

 // Maximization

 Given the approximated posteriors, optimize model parameters according to
 (15)

6.1 Document modeling

Dataset In document modeling, we focus on NIPS dataset, which contains a corpus of 1740 research articles accepted by Neural Information Processing Systems conference from 1987-2016.

Preprocessing As a necessary preprocessing procedure for raw documents, we first tokenize the documents and remove numbers, punctuations, and stop words such as those with less than two letters. Next, all words are reduced to their root form by a standard lemmatizer from NLTK library. To capture more specific and identifiable semantic objects, bigrams are additionally recognized and added to the token list. Finally, we remove the words that are too rare or too common and generate the vocabulary.

Models and experiment description After preprocessing and tokenizing the raw documents, we train a series of topic models with different model settings: (i) three typical Dirichlet-Tree priors: original Dirichlet, Beta-Liouville and Generalized Dirichlet; (ii) two vectorized inference methods: Mean-Field Variational Inference and Expectation Propagation; and (iii) a range of topic quantities from 10, 20, to 80.

Therefore, in total, we train and examine $3 \times 2 \times 8 = 48$ different topic models. We evaluate the obtained models by metrics including (i) convergence of log-likelihood, (ii) predictive perplexity, (iii) topic coherence and (iv) topic diversity.

6.1.1 CONVERGENCE

The ELBO values are computed and recorded in each training iteration for all models. Equally for all model settings, the training process is judged as converged when the ELBO's change rate is less than $1e-4$. Fig. 12 shows the ELBO values with respect to each iteration in different models. We can see that the training process of all models converges successfully after several iterations. The models with variational inference generally achieves a little advantage in final ELBO values, while the models with expectation propagation converges in less iterations.

6.1.2 PREDICTIVE PERPLEXITY

Perplexity measures the quality of a topic model by evaluating its predictive capability on hold-out test set. We split the dataset into 90% training set and 10% test set to check the models' predictive perplexity. We first train a model on the training set, and then perform inference on the test set. In practice, the perplexity in our experiment is computed as follows:

$$\text{Perplexity}(D_{test}) = \exp \left\{ - \frac{\sum_{m=1}^M \sum_{v=1}^V n_{mv} \log \sum_{k=1}^K \varphi_{vk} \mathbb{E}_{q_m(\zeta)} [\theta_k]}{\sum_{m=1}^M \sum_{v=1}^V n_{mv}} \right\}$$

Fig. 13 plots the perplexity of different model settings.

6.1.3 TOPIC COHERENCE AND DIVERSITY

Topic coherence and topic diversity are two complementary metrics to measure the quality of learned topics.

Topic coherence targets at each individual topic and evaluate the semantic similarity within the top words assigned to a topic. Popular topic coherence measures include UCI Coherence, UMass Coherence, and CV Coherence. In our experiments, we use a mixing topic coherence measure provided by Gensim library. Top ten words are chosen for each topic, and one model's coherence value is computed by averaging the coherence values of each topic.

On the contrary, topic diversity measures how "orthogonal", or dissimilar the generated topics are against each other. A good model should have diverse topics to convey abundant information dimensions. In practice, our topic diversity is computed as follows:

$$\text{Diversity}(\varphi) = \frac{U(N)}{KN}$$

where K is the number of topics, N is number of top words chosen for each topic, and U is number of unique words within these top words. We choose $N = 10$ for our experiments. Fig. 14 and Fig. 15 shows the results of topic coherence and topic diversity, respectively. We observe that Variational Inference generally has better topic coherence, while expectation propagation has better topic diversity.

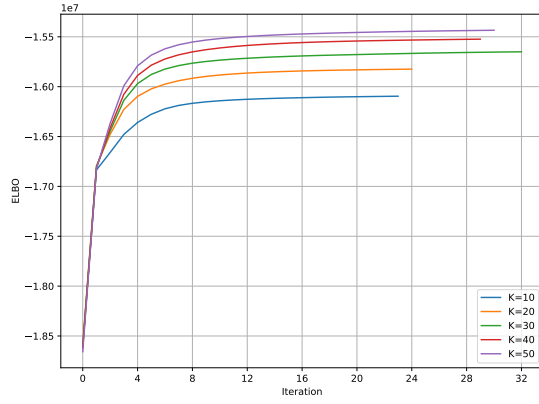
6.2 Document Classification

In the tasks of document classification, a collection of documents are categorized into binary or multiple classes, where each document is represented by a fixed set of features (Blei 2003).

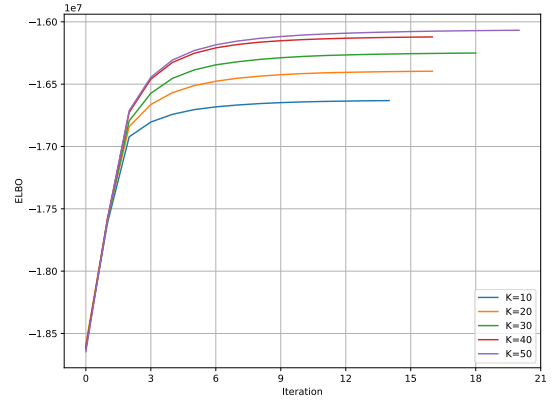
A natural thinking is to use the counts of all the words that appear in a document as the representative features, yet leading to an extremely large feature space. Topic models play the role of dimension reduction in document classification, where the large vocabulary is reduced to K -dimensional simplex.

In our experiment, we use Reuters-21578 corpus set which is a collection of labeled newswire articles. After necessary preprocessing, we choose the six categories that contain the most documents. Finally, we obtain a vocabulary of 4562 unique words, and 7460 documents in total. The numbers of documents in each category are shown in Table 1.

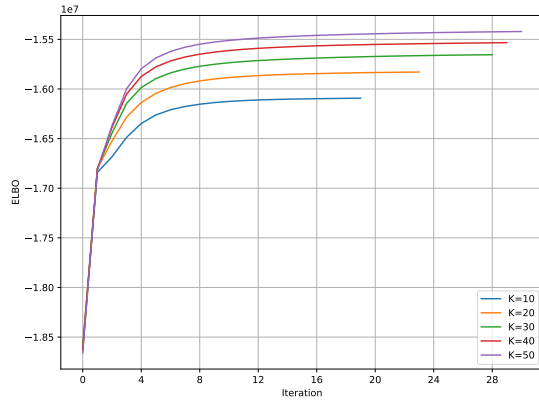
VECTORIZED BAYES FOR LDFA



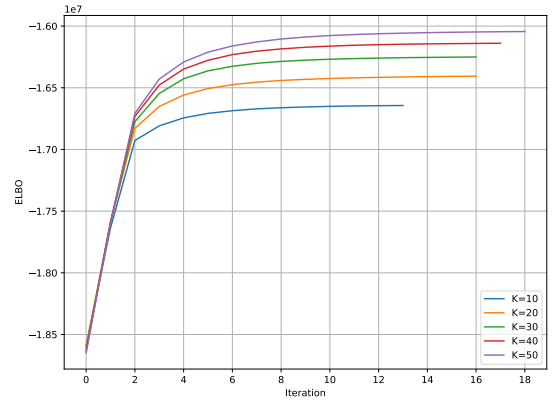
(a) Dir-VI



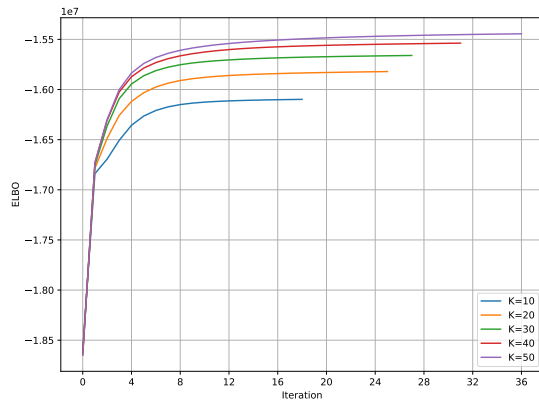
(b) Dir-EP



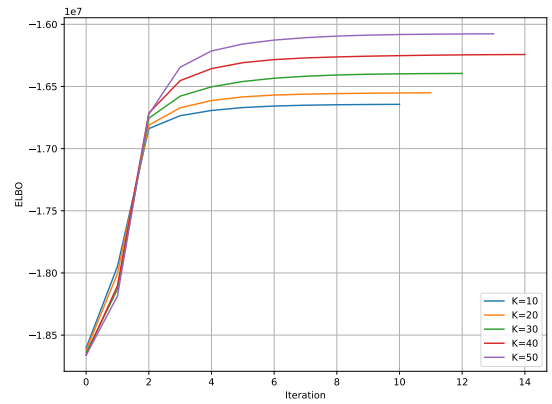
(c) BL-VI



(d) BL-EP



(e) GDir-VI



(f) GDir-EP

Figure 12: ELBO convergence with respect to iterations; iterations stop when ELBO change less than 0.0001

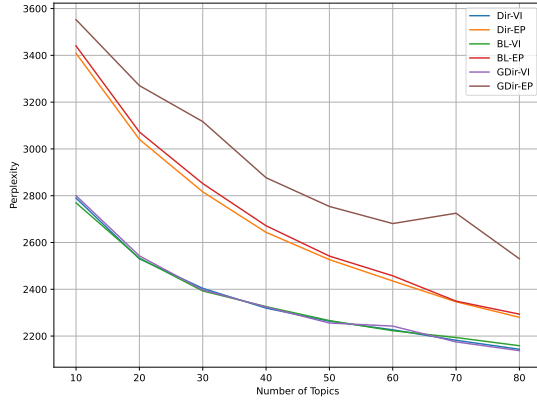


Figure 13: Perplexity

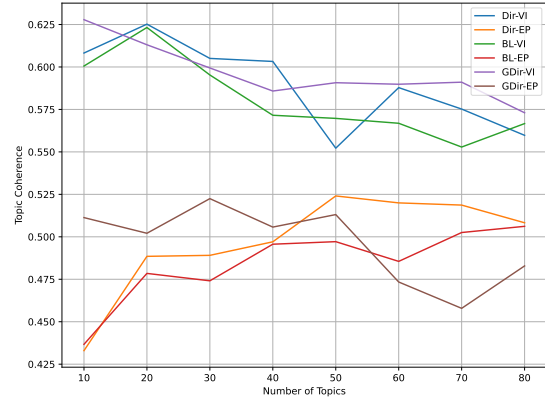


Figure 14: Topic Coherence

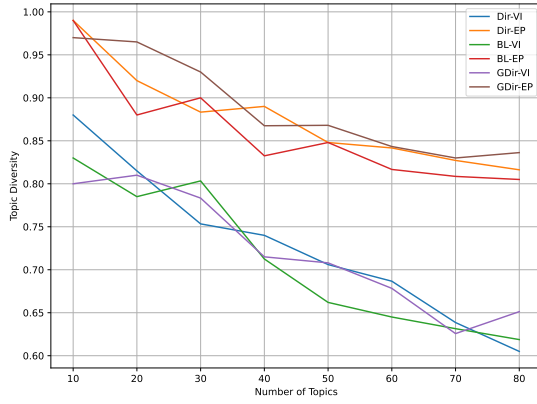


Figure 15: Topic Diversity

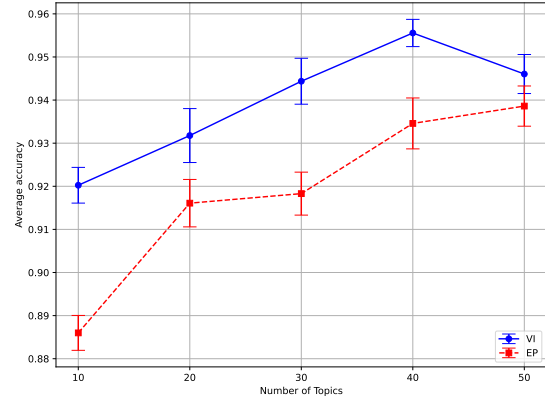


Figure 16: Average classification accuracy of Dirichlet prior

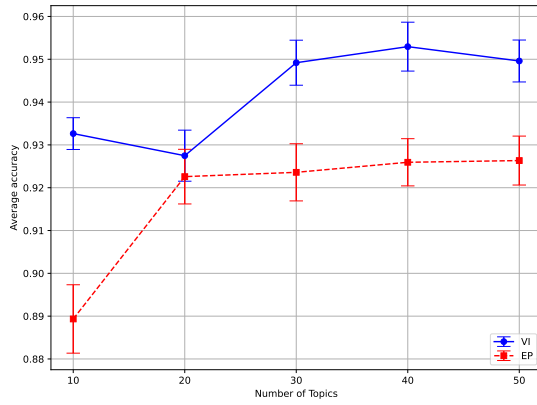


Figure 17: Average classification accuracy of Beta-Liouville prior

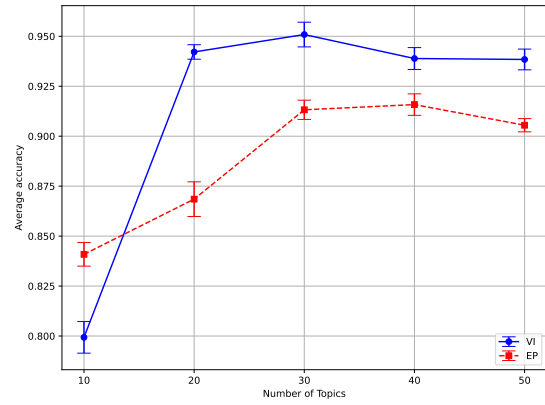


Figure 18: Average classification accuracy of Generalized Dirichlet prior

Table 1: Number of documents of chosen categories

Category	# Documents
earn	3900
acq	2289
crude	370
trade	324
money-fx	306
interest	271

We train multiple models with different settings for the whole corpus, and compute the expectation of each document’s posterior distribution as its representative feature. Subsequently, we train a logistic regression classifier based on the learned topic proportions and the original labels. For each topic model setting, we randomly split the corpus into 80% training set and 20% test set for 10 times. The average classification accuracy, as well as its standard deviation, is computed and employed as its final accuracy result.

Figure 16, Figure 17 and Figure 18 are, respectively, the results of classification accuracy of Dirichlet prior, Beta Liouville prior, and Generalized Dirichlet prior. Table 2 shows the specific accuracy value of each model setting. For each model setting, we choose the topic number that generates the highest average accuracy and demonstrate its confusion matrix in Figure 19.

Models	Accuracy				
	K=10	K=20	K=30	K=40	K=50
Dir-VI	92.0%	93.2%	94.4%	95.6%	94.6%
Dir-EP	88.6%	91.6%	91.8%	93.5%	93.9%
BL-VI	93.3%	92.7%	94.9%	95.3%	95.0%
BL-EP	88.9%	92.3%	92.4%	92.5%	92.6%
GDir-VI	79.9%	94.2%	95.1%	93.9%	93.8%
GDir-EP	84.1%	86.8%	91.3%	91.6%	90.5%

Table 2: Average accuracy of classification

6.3 Image Classification

We further examine our models with an image classification task. We choose 15 Scene Categories dataset, and pick five categories out of them: forest, highway, kitchen, office, and tallbuilding.

After loading the five categories of images, their feature descriptors are extracted with SIFT(Scale-Invariant Feature Transform), which are roughly distributed in a continuous space. Thus, we use K-MEANS on all obtained descriptors to form bag-of-visualword (BoVW) vocabulary. Then, a topic model with a certain prior-inference combination is trained on the BoVM, and the topic proportions of each image are obtained. Subsequently,

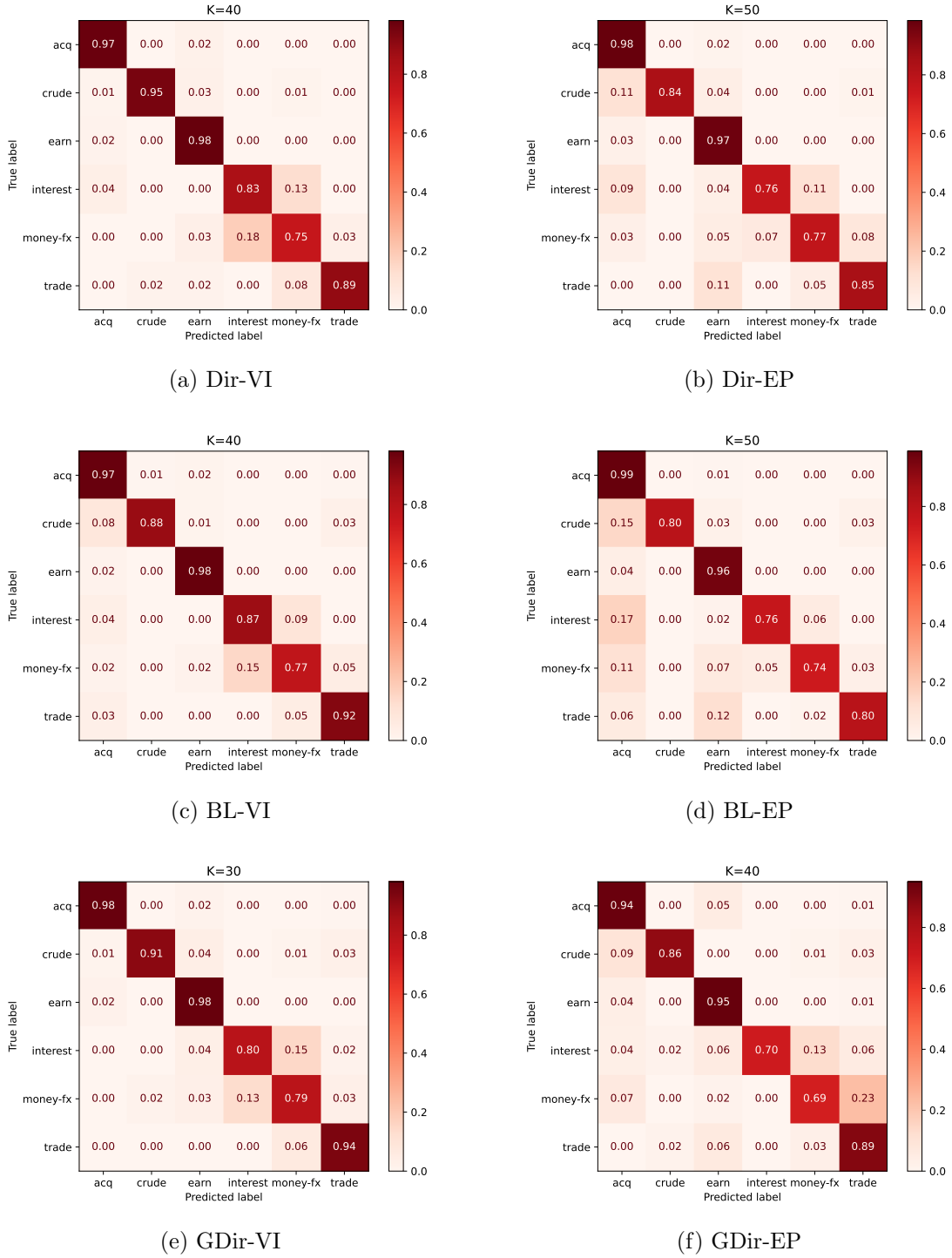
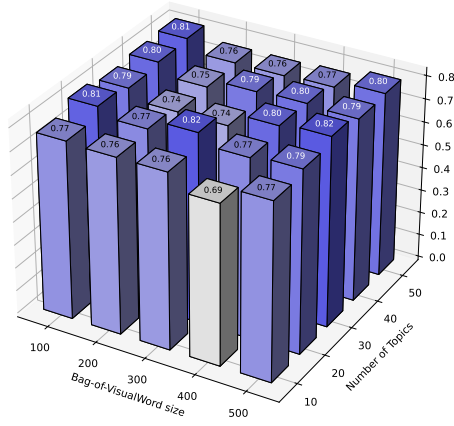
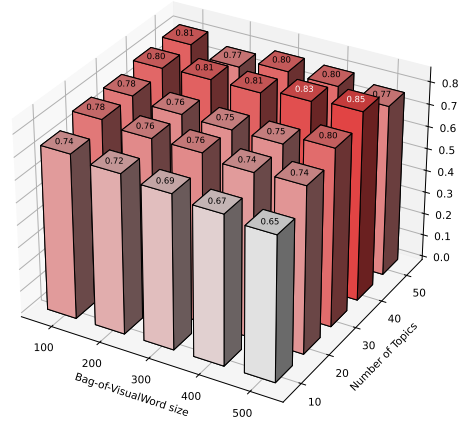


Figure 19: Best confusion matrix for each model settings; the titles indicating the topic number

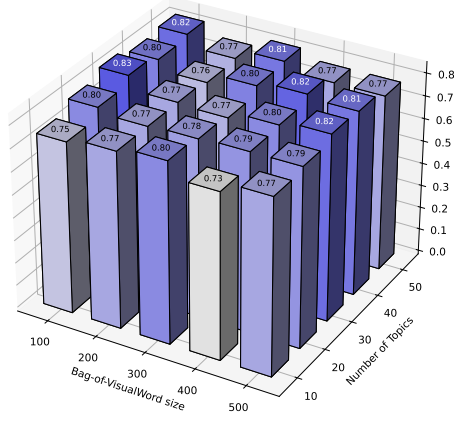
VECTORIZED BAYES FOR LDFA



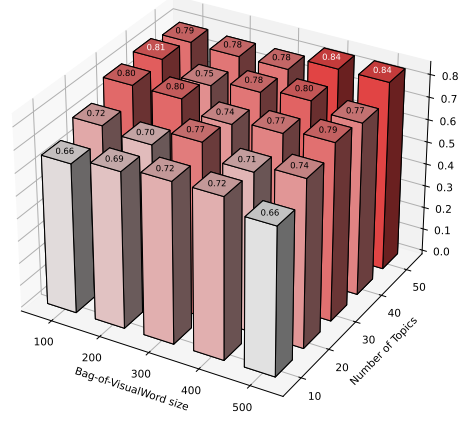
(a) Dir-VI



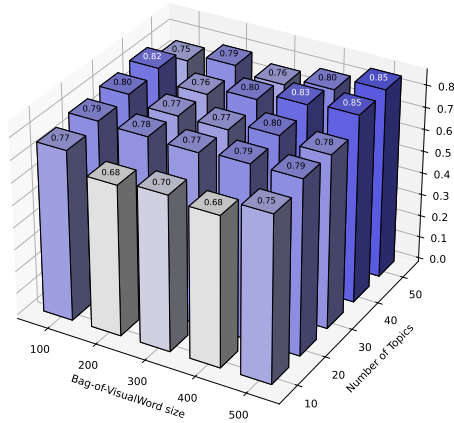
(b) Dir-EP



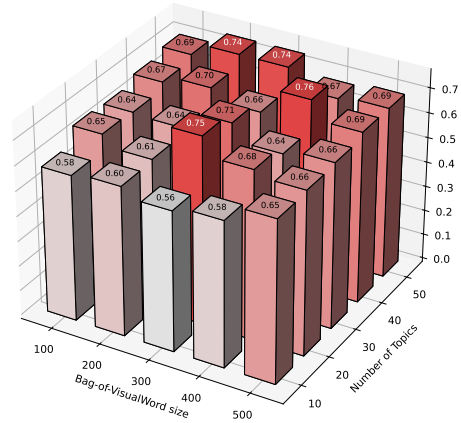
(c) BL-VI



(d) BL-EP



(e) GDir-VI



(f) GDir-EP

Figure 20: Image classification accuracies

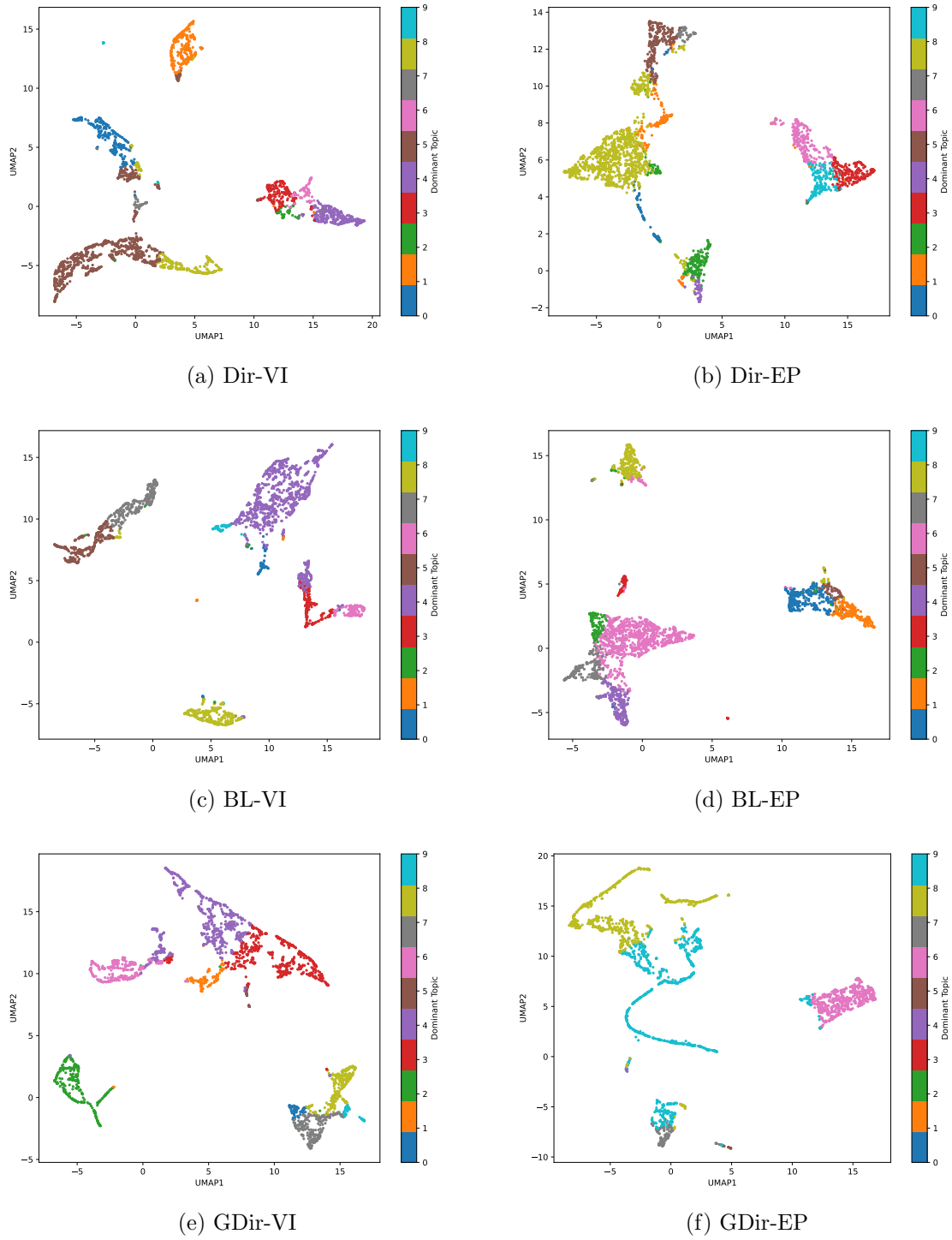


Figure 21: UMAP of topic proportion of single-cells

an SVM classifier is employed to perform classifications based on the topic proportions of all images.

In our experiment, we set the grid search based on the size of the BoVM and the topic number. We choose from 100 to 500 for the size of the BoVW, and 10 to 50 for the topic number. Finally, we obtain a total of 150 models of different settings, and compute their classification accuracies. Figure 20 is the 3D-plot of the image classification accuracy results. We achieve a maximum accuracy of 85% from our models.

6.4 An Application in Bioinformatics

In this section, we test our models using a well-known task in bioinformatics, namely single-cell RNA sequencing (scRNA-seq). We use the famous dataset of Peripheral Blood Mononuclear Cells (PBMC) from 10X Genomics, which contains 2700 single cells that were sequenced on the Illumina NextSeq 500. Each cell is equivalent to a document in our textual experiments, and the collection of genes of a single cell is equivalent to the collection of words in a document. We firstly preprocess the PBMC dataset, and train multiple models with number of topics set to 10, as it is reported as the effective setting. Figure 21 is the UMAPs of topic proportion of all single-cells obtained from our models.

7 Discussion

This paper focuses on developing a general framework, called Latent Dirichlet-Tree Allocation (LDТА), to extend the conventional Latent Dirichlet Allocation (LDA) model both theoretically and practically. Despite being overwhelmingly successful in modeling all kinds of discrete data, LDA employs the classical Dirichlet distribution as the prior for the topic proportions, which has intrinsic limitations such as strictly negative covariance matrix. This shortcoming consequently imposes a structural restriction on LDA’s capability of capturing the correlations and hierarchical relationships among topics. To alleviate this restriction, we generalize the Dirichlet distribution to a broader family of Dirichlet-Tree distributions. While maintaining the properties of conjugacy and exponential form, the Dirichlet-Tree’s expressiveness is significantly enhanced thanks to its flexible hierarchical structure. LDТА therefore extends the modeling capacity of LDA while preserving its generative semantic and interpretability.

In Section 2, we formally derive the Dirichlet-Tree distribution and demonstrate the key properties that underpin the subsequent derivation of Bayesian inference algorithm. In particular, we prove the exponential form of Dirichlet-Tree for the first time, and we summarize the three equivalent representations: the node form, the general form, and the exponential form. Under the Dirichlet-Tree framework, the classical Dirichlet, the Generalized Dirichlet, and the Beta-Liouville proposed in earlier literature are all unified as special cases of Dirichlet-Tree corresponding to different tree-structure. To further prepare the development of the inference algorithm, we introduce several new concepts including Dirichlet selection operator, Bayesian operator and the collection of derived distributions. These concepts provides new insight into the conjugacy and Bayesian updating while greatly simplifying the description of approximate Bayesian inference procedures.

In Section 3, 4 and 5, we introduce the generative process of LDТА, and derive and compare two universal inference algorithms, and Mean-Field Variational Inference (MFVI)

and Expectation Propagation (EP). We use the word “universal” because the algorithms are derived and formulated for any arbitrary Dirichlet-Tree prior. The two algorithms aim to approximate the intractable posteriors from different perspectives. MFVI employs a global mean-field approximating distribution and optimizes the approximating parameters by maximizing the lower bound of evidence, while EP, in contrast, factorizes the target posterior into multiple terms and approximates them by a corresponding group of approximators step by step. Throughout the formulation, we particularly highlight the vectorized nature of the two algorithms by expressing their computational steps using tensor operations. And finally in Section 6, we implement LDTA with three different Dirichlet-Tree priors and both inference algorithms across three different application scenarios. The applications and experiments are accelerated using Pytorch on GPU, and the experimental results demonstrate feasibility and scalability of our model.

Despite these advantages, LDTA also brings further challenges and opportunities for improvement. One limitation of the current work is that the tree structure must be specified in advance. While this gives the model users the flexibility to choose the desired tree structure, it remains an important question of how in general the tree should “grow” itself given data where no obvious tree structure appears. Automatic structure learning, such as inferring the tree topology from data or using a flexible nonparametric prior, represents a promising direction for future research. Another future direction calls the improvement and generalization of KL-divergence to measure the similarity between the approximating distributions and the target posteriors. Both MFVI and EP are based on KL-divergence which is known to be second-order approximation of Fisher-Rao metric (Amari, 2016). Moreover, the global mean-field approximating distribution in MFVI can have more generalized and adaptive inner structure based on statistical optimal transport (Chewi et al., 2025; Wu and Blei, 2024). In this viewpoint, the mean-field is a special case of a more general family of approximating distributions. In summary, this thesis demonstrates a meaningful step in this research field, and more potential improvements are waiting to be discovered and developed.

Acknowledgments and Disclosure of Funding

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A. Exponential Family and KL-Divergence

The exponential family of distributions has the following form (Bishop, 2006):

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x}) - \log g(\boldsymbol{\eta})\}$$

where $\mathbf{u}(\mathbf{x})$ is the sufficient statistics; $\boldsymbol{\eta}$ is the natural parameter; $h(\mathbf{x})$ is the underlying measure; and $g(\boldsymbol{\eta})$ is the normalizer:

$$g(\boldsymbol{\eta}) = \int_{\mathbf{x}} \exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\} h(\mathbf{x}) d\mathbf{x}$$

Theorem 11 (Expectation of sufficient statistics) *Let probability distribution $p(\mathbf{x}|\boldsymbol{\eta})$ belong to exponential family. The expectation of its sufficient statistics $\mathbf{u}(\mathbf{x})$ is equal to the gradient of its log-normalizer $\log g(\boldsymbol{\eta})$ w.r.t. its natural parameter $\boldsymbol{\eta}$.*

$$\nabla_{\boldsymbol{\eta}} \log g(\boldsymbol{\eta}) = \frac{1}{g(\boldsymbol{\eta})} \nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta}) = \int_{\mathbf{x}} \frac{1}{g(\boldsymbol{\eta})} \exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\} \mathbf{u}(\mathbf{x}) h(\mathbf{x}) d\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Theorem 12 (ELBO Maximization) *Suppose $p(\mathbf{z}|\mathbf{w})$ is a target posterior given observation \mathbf{w} ; $q(\mathbf{z}|\boldsymbol{\eta})$ is the approximating distribution. The optimized approximating distribution $q(\mathbf{z}|\boldsymbol{\eta}^*)$ to the target posterior $p(\mathbf{z}|\mathbf{w})$ measured by minimizing KL-Divergence $\text{KL}(q(\mathbf{z}|\boldsymbol{\eta})||p(\mathbf{z}|\mathbf{w}))$ is the one that maximizes the Evidence Lower Bound (ELBO) given as:*

$$\text{ELBO} = \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\eta}) \log \frac{p(\mathbf{w}, \mathbf{z})}{q(\mathbf{z}|\boldsymbol{\eta})} d\mathbf{z}$$

Proof Given that $p(\mathbf{z}|\mathbf{w})$ is the target posterior and $q(\mathbf{z}|\boldsymbol{\eta})$ belongs to exponential family, the Kullback-Leibler divergence $\text{KL}(q(\mathbf{z}|\boldsymbol{\eta})||p(\mathbf{z}|\mathbf{w}))$ w.r.t. $q(\mathbf{z}|\boldsymbol{\eta})$ has the following form:

$$\begin{aligned} \text{KL}(q(\mathbf{z}|\boldsymbol{\eta})||p(\mathbf{z}|\mathbf{w})) &= \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\eta}) \log \frac{q(\mathbf{z}|\boldsymbol{\eta})}{p(\mathbf{z}|\mathbf{w})} d\mathbf{z} \\ &= \log p(\mathbf{w}) - \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\eta}) \log \frac{p(\mathbf{w}, \mathbf{z})}{q(\mathbf{z}|\boldsymbol{\eta})} d\mathbf{z} \\ &= \log p(\mathbf{z}) - \text{ELBO} \end{aligned}$$

where $\log p(\mathbf{w})$ is independent of $\boldsymbol{\eta}$, and $\text{KL}(q(\mathbf{z}|\boldsymbol{\eta})||p(\mathbf{z}|\mathbf{w})) \geq 0$. The theorem is thus proved. ■

Theorem 13 (Moment Matching) *Let $q(\mathbf{x}|\boldsymbol{\eta})$ belong to the exponential family. The optimized approximating distribution $q(\mathbf{x}|\boldsymbol{\eta}^*)$ to a target distribution $p(\mathbf{x})$ measured by minimizing KL-Divergence $\text{KL}(p(\mathbf{x})||q(\mathbf{x}|\boldsymbol{\eta}))$ is the one under which the expectation of its own sufficient statistics is equal to the expectation of that sufficient statistics under the target distribution.*

Proof Given that $p(\mathbf{x})$ is a target distribution and $q(\mathbf{x}|\boldsymbol{\eta})$ is the approximating distribution and $q(\mathbf{x}|\boldsymbol{\eta})$ belongs to the exponential family. The Kullback-Leibler divergence $\text{KL}(p(\mathbf{x})||q(\mathbf{x}|\boldsymbol{\eta}))$ w.r.t. $q(\mathbf{x}|\boldsymbol{\eta})$ has the following simple form:

$$\begin{aligned}\text{KL}(p(\mathbf{x})||q(\mathbf{x}|\boldsymbol{\eta})) &= \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x}|\boldsymbol{\eta})} d\mathbf{x} \\ &= \log g(\boldsymbol{\eta}) - \int_{\mathbf{x}} p(\mathbf{x}) \left\{ \boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x}) \right\} d\mathbf{x} + \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{h(\mathbf{x})} d\mathbf{x} \\ &= \log g(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top \mathbb{E}_{p(\mathbf{x})} [\mathbf{u}(\mathbf{x})] + \text{const.}\end{aligned}$$

where the constant is independent of $\boldsymbol{\eta}$. The gradient is given as:

$$\nabla_{\boldsymbol{\eta}} \text{KL}(\boldsymbol{\eta}) = \mathbb{E}_{q(\mathbf{x})} [\mathbf{u}(\mathbf{x})] - \mathbb{E}_{p(\mathbf{x})} [\mathbf{u}(\mathbf{x})]$$

Consider minimizing the KL divergence w.r.t. $\boldsymbol{\eta}$ by setting the gradient to zero. This complete the proof of the theorem. \blacksquare

Appendix B. Details of Moment Matching

In previous sections, we demonstrate that both maximization of ELBO with respect to $\boldsymbol{\xi}$ in mean-field variational inference and message passing in expectation propagation inference are reduced to moment matching. More specifically, we want to “recover” a Dirichlet distribution’s parameters given its expectation of sufficient statistics. Minka (2000) introduces the more general case of maximum likelihood estimate of a Dirichlet, and fix-point iteration and Newton-Raphson algorithm for solving the problem. Sklar (2014) tackles the problem with a fast Newton method. In this appendix, we formulate the problem and introduce the fast Newton method that we used in the implementation of this paper, based on Sklar (2014) with slight differences in formulation and implementation. In the end, we show that the same method can be generalized to any kind of Dirichlet-Tree.

Suppose that we have a Dirichlet distribution $\text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ with unknown parameter $\boldsymbol{\alpha}$, and we want to compute the parameter from the known expectation of sufficient statistics \mathbf{u} . For convenience, we define the function $\xi : \mathbb{R}_{>0}^K \rightarrow \mathbb{R}_{>0}^K$, such that $\boldsymbol{\alpha} = \xi(\mathbf{u})$. This is equivalent to solving a system of non-linear equations:

$$\left\{ \psi(\alpha_k) - \psi \left(\sum_{l=1}^K \alpha_l \right) - u_k = 0 \right\}_{k=1}^K$$

A natural idea is to use fixed-point iteration, as indicated in (Minka, 2000). Starting from an initial guess $\boldsymbol{\alpha}_0$, the optimized $\boldsymbol{\alpha}$ can be computed iteratively:

$$\left\{ \alpha_k^{(t+1)} = \psi^{-1} \left(\psi \left(\sum_{l=1}^K \alpha_l^{(t)} \right) + u_k \right) \right\}_{k=1}^K$$

This approach is computationally inefficient and may require many iterations in some cases. Therefore, we employ a fast Newton-Raphson method in our implementation. We denote $\alpha_0 = \sum_{l=1}^K \alpha_l$ and:

$$\begin{aligned}
 d(\boldsymbol{\alpha}) &= \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_K \end{pmatrix} = \begin{pmatrix} \psi(\alpha_1) - \psi(\alpha_0) - u_1 \\ \psi(\alpha_2) - \psi(\alpha_0) - u_2 \\ \vdots \\ \psi(\alpha_K) - \psi(\alpha_0) - u_K \end{pmatrix} \\
 J(\boldsymbol{\alpha}) &= \begin{pmatrix} \frac{\partial}{\partial \alpha_1} d_1 & \frac{\partial}{\partial \alpha_1} d_2 & \cdots & \frac{\partial}{\partial \alpha_1} d_K \\ \frac{\partial}{\partial \alpha_2} d_1 & \frac{\partial}{\partial \alpha_2} d_2 & \cdots & \frac{\partial}{\partial \alpha_2} d_K \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \alpha_K} d_1 & \frac{\partial}{\partial \alpha_K} d_2 & \cdots & \frac{\partial}{\partial \alpha_K} d_K \end{pmatrix} \\
 &= \begin{pmatrix} \psi'(\alpha_1) - \psi'(\alpha_0) & -\psi'(\alpha_0) & \cdots & -\psi'(\alpha_0) \\ -\psi'(\alpha_0) & \psi'(\alpha_2) - \psi'(\alpha_0) & \cdots & -\psi'(\alpha_0) \\ \vdots & \vdots & \ddots & \vdots \\ -\psi'(\alpha_0) & -\psi'(\alpha_0) & \cdots & \psi'(\alpha_K) - \psi'(\alpha_0) \end{pmatrix} \\
 &= \text{diag}(\psi'(\boldsymbol{\alpha})) - \psi'(\alpha_0) \mathbf{1}_K \mathbf{1}_K^\top
 \end{aligned}$$

where $\psi'(\cdot)$ is the trigamma function. The Newton's method is given as:

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - J(\boldsymbol{\alpha}^{(t)})^{-1} d(\boldsymbol{\alpha}^{(t)})$$

Here, the key is to compute the inverse of Jacobian (or Hessian in the problem of MLE of a Dirichlet). It should be noted that $J(\boldsymbol{\alpha})$ equals a diagonal matrix plus a constant matrix. Sherman-Morrison formula, which is a special case of the Woodbury matrix inversion lemma, provides a perfect analytic solution for the inversion of this special kind of matrix. According to the Sherman-Morrison formula, if $A \in \mathbb{R}^{K \times K}$ is an invertible matrix, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^K$ are two vectors, we have:

$$(A + \mathbf{u} \mathbf{v}^\top)^{-1} = A^{-1} - \frac{A^{-1} \mathbf{u} \mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1} \mathbf{u}}$$

We replace A with $\text{diag}(\psi'(\boldsymbol{\alpha}))$, and $\mathbf{u} \mathbf{v}^\top$ with $-\psi'(\alpha_0) \mathbf{1}_K \mathbf{1}_K^\top$. After a few steps, we arrive at the following expression:

$$J(\boldsymbol{\alpha})^{-1} = \text{diag}(\boldsymbol{\pi}(\boldsymbol{\alpha})) - \frac{\boldsymbol{\pi}(\boldsymbol{\alpha}) \boldsymbol{\pi}(\boldsymbol{\alpha})^\top}{\sigma(\boldsymbol{\alpha})}$$

where

$$\begin{aligned}
 \boldsymbol{\pi}(\boldsymbol{\alpha}) &= \left(\frac{1}{\psi'(\alpha_1)}, \frac{1}{\psi'(\alpha_2)}, \cdots, \frac{1}{\psi'(\alpha_K)} \right)^\top \\
 \sigma(\boldsymbol{\alpha}) &= \left(\sum_{l=1}^K \frac{1}{\psi'(\alpha_l)} \right) - \frac{1}{\psi'(\alpha_0)}
 \end{aligned}$$

We proceed to derive the update in each iteration:

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \left(d(\boldsymbol{\alpha}^{(t)}) - \frac{\boldsymbol{\pi}(\boldsymbol{\alpha}^{(t)})^\top d(\boldsymbol{\alpha}^{(t)})}{\sigma(\boldsymbol{\alpha}^{(t)})} \mathbf{1}_K \right) \odot \boldsymbol{\pi}(\boldsymbol{\alpha}^{(t)})$$

In Dirichlet-Tree distribution, the parameter $\{\xi_{t|s}\}_{c(s)}$, $s \in \mathbf{\Lambda}$ under each internal node is treated as an independent sub-group of Dirichlet parameters that results in a sub-group of equations, and solved locally.

Appendix C. Examples of Dirichlet-Tree

In this section, we introduce three typical Dirichlet-Tree distributions: the classical Dirichlet, the Beta-Liouville, and the Generalized Dirichlet.

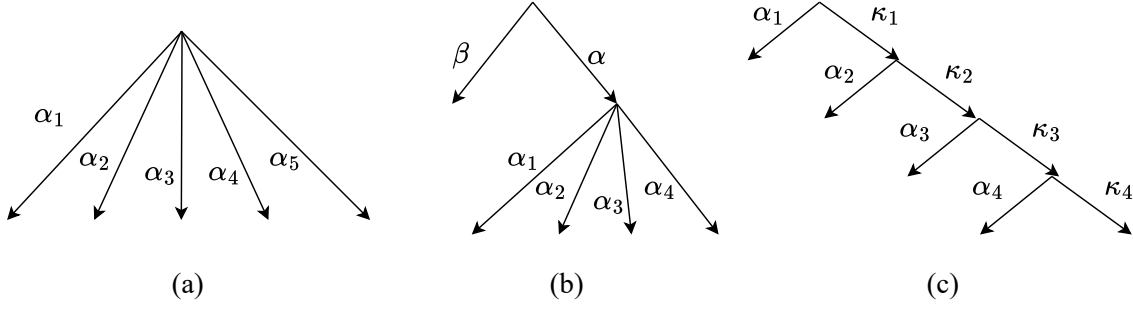


Figure 22: (a) Classical Dirichlet; (b) Beta-Liouville; (c) Generalized Dirichlet

C.1 Dirichlet

The Dirichlet distribution is the most basic Dirichlet-Tree (Figure 22 (a)). Given the parameter $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$, the PDF of Dirichlet is given as:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{l=1}^K \alpha_l\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k-1}$$

Dirichlet's mean, variance, and covariance are given as:

$$\begin{aligned} \mathbb{E}[\theta_k] &= \frac{\alpha_k}{\sum_{l=1}^K \alpha_l} \\ \text{Var}(\theta_k) &= \frac{\alpha_k \left(\sum_{l=1}^K \alpha_l - \alpha_k \right)}{\left(\sum_{l=1}^K \alpha_l \right)^2 \left(\sum_{l=1}^K \alpha_l + 1 \right)} \\ \text{Cov}(\theta_i, \theta_j) &= - \frac{\alpha_i \alpha_j}{\left(\sum_{l=1}^K \alpha_l \right)^2 \left(\sum_{l=1}^K \alpha_l + 1 \right)} \end{aligned}$$

C.1.1 EXPONENTIAL FORM

The natural parameters follows:

$$\eta_k = \alpha_k - 1, \quad k = 1, \dots, K$$

The sufficient statistics:

$$u_k = \log \theta_k, \quad k = 1, \dots, K$$

The log-normalizer:

$$\log g(\boldsymbol{\eta}) = \sum_{k=1}^K \log \Gamma(\alpha_k) - \log \Gamma\left(\sum_{l=1}^K \alpha_l\right)$$

The expectation of sufficient statistics:

$$\mathbb{E}[u_k] = \psi(\alpha_k) - \psi\left(\sum_{l=1}^K \alpha_l\right), \quad k = 1, \dots, K$$

C.1.2 CONJUGACY

The Dirichlet distribution is conjugate to multinomial. Given multinomial observation $\mathbf{n} = (n_1, \dots, n_K)$, the parameters of the Dirichlet posterior is updated as:

$$\alpha_k' = \alpha_k + n_k, \quad k = 1, \dots, K$$

C.2 Beta-Liouville

Given the parameter $\boldsymbol{\xi} = (\alpha, \beta, \alpha_1, \dots, \alpha_{K-1})$, the PDF of Beta-Liouville (Figure 22 (b)) is given as:

$$p(\boldsymbol{\theta}|\boldsymbol{\xi}) = \frac{\Gamma(\sum_{l=1}^{K-1} \alpha_l) \Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \prod_{k=1}^{K-1} \frac{\theta_k^{\alpha_k-1}}{\Gamma(\alpha_k)} \left(\sum_{l=1}^{K-1} \theta_l\right)^{\alpha - \sum_{l=1}^{K-1} \alpha_l} \left(1 - \sum_{l=1}^{K-1} \theta_l\right)^{\beta-1}$$

Beta-Liouville also belongs to Liouville family of distributions whose general form is:

$$p(\boldsymbol{\theta}|\boldsymbol{\xi}) = \frac{f(u|\zeta)}{u^{\sum_{l=1}^{K-1} \alpha_l - 1}} \frac{\Gamma(\sum_{l=1}^{K-1} \alpha_l)}{\prod_{k=1}^{K-1} \Gamma(\alpha_k)} \prod_{k=1}^{K-1} \theta_k^{\alpha_k-1}$$

where $u = \sum_{l=1}^{K-1} \theta_l$ and $f(u|\zeta)$ is the generating density with parameter ζ . Particularly, Beta-Liouville employs Beta distribution as $f(u|\zeta)$. When $\alpha = \sum_{k=1}^{K-1} \alpha_k$ and $\beta = \alpha_K$, Beta-Liouville descends to the conventional Dirichlet distribution. Beta-Liouville's mean, variance, and covariance matrix are given as:

$$\begin{aligned} \mathbb{E}[\theta_k] &= \frac{\alpha}{\alpha + \beta} \frac{\alpha_k}{\sum_{l=1}^{K-1} \alpha_l} \\ \text{Var}(\theta_k) &= \left(\frac{\alpha}{\alpha + \beta}\right)^2 \frac{\alpha_k(\alpha_k + 1)}{(\sum_{l=1}^{K-1} \alpha_l)(\sum_{l=1}^{K-1} \alpha_l + 1)} - \left(\frac{\alpha_k \mathbb{E}(\theta_k)}{\sum_{l=1}^{K-1} \alpha_l}\right)^2 \\ \text{Cov}(\theta_p, \theta_q) &= \frac{\alpha_p \alpha_q}{\sum_{l=1}^K \alpha_l} \left(\frac{\frac{\alpha+1}{\alpha+\beta+1} \frac{\alpha}{\alpha+\beta}}{\sum_{l=1}^{K-1} \alpha_l + 1} - \frac{\frac{\alpha}{\alpha+\beta}}{\sum_{l=1}^{K-1} \alpha_l} \right) \end{aligned}$$

C.2.1 EXPONENTIAL FORM

The natural parameters are as follows:

$$\begin{aligned}\eta_k &= \alpha_k - 1, \quad k = 1, \dots, K-1 \\ \eta_K &= \alpha - K + 1 \\ \eta_{K+1} &= \beta - 1\end{aligned}$$

The sufficient statistics:

$$\begin{aligned}u_k &= \log \theta_k - \log \left(\sum_{l=1}^{K-1} \theta_l \right), \quad k = 1, \dots, K-1 \\ u_K &= \log \left(\sum_{l=1}^{K-1} \theta_l \right) \\ u_{K+1} &= \log \left(1 - \sum_{l=1}^{K-1} \theta_l \right)\end{aligned}$$

The log-normalizer:

$$\log g(\boldsymbol{\eta}) = \sum_{k=1}^{K-1} \log \Gamma(\alpha_k) + \log \Gamma(\alpha) + \log \Gamma(\beta) - \log \Gamma \left(\sum_{l=1}^{K-1} \alpha_l \right) - \log \Gamma(\alpha + \beta)$$

The expectation of sufficient statistics:

$$\begin{aligned}\mathbb{E}[u_k] &= \psi(\alpha_k) - \psi \left(\sum_{l=1}^{K-1} \alpha_l \right), \quad k = 1, \dots, K-1 \\ \mathbb{E}[u_K] &= \psi(\alpha) - \psi(\alpha + \beta) \\ \mathbb{E}[u_{K+1}] &= \psi(\beta) - \psi(\alpha + \beta)\end{aligned}$$

And:

$$\begin{aligned}\mathbb{E}[\log \theta_k] &= \psi(\alpha_k) - \psi \left(\sum_{l=1}^{K-1} \alpha_l \right) + \psi(\alpha) - \psi(\alpha + \beta), \quad k = 1, \dots, K-1 \\ \mathbb{E}[\log \theta_K] &= \psi(\beta) - \psi(\alpha + \beta)\end{aligned}$$

C.2.2 CONJUGACY

The Beta-Liouville distribution is conjugate to multinomial. Given multinomial observation $\mathbf{n} = (n_1, \dots, n_K)$, the parameters of the Beta-Liouville posterior is updated as:

$$\begin{aligned}\alpha'_k &= \alpha_k + n_k, \quad k = 1, \dots, K-1 \\ \alpha' &= \alpha + \sum_{l=1}^{K-1} n_l \\ \beta' &= \beta + n_K\end{aligned}$$

C.3 Generalized-Dirichlet

Given the parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{K-1})$ and $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_{K-1})$, the PDF of the Generalized-Dirichlet (Figure 22 (c)) is given as:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\kappa}) = \prod_{k=1}^{K-1} \frac{\Gamma(\alpha_k + \kappa_k)}{\Gamma(\alpha_k)\Gamma(\kappa_k)} \theta_k^{\alpha_k-1} \left(1 - \sum_{l=1}^k \theta_l\right)^{\gamma_k}$$

where $\gamma_k = \kappa_k - \alpha_{k+1} - \kappa_{k+1}$ for $k = 1, \dots, K-2$ and $\gamma_{K-1} = \kappa_{K-1} - 1$. Generalized-Dirichlet's mean, variance, and covariance are given as:

$$\begin{aligned}\mathbb{E}[\theta_k] &= \frac{\alpha_k}{\alpha_k + \kappa_k} \prod_{l=1}^{k-1} \frac{\kappa_l}{\alpha_l + \kappa_l} \\ \text{Var}(\theta_k) &= \mathbb{E}[\theta_k] \left(\frac{\alpha_k + 1}{\alpha_k + \kappa_k + 1} \prod_{l=1}^{k-1} \frac{\kappa_l + 1}{\alpha_l + \kappa_l + 1} - \mathbb{E}[\theta_k] \right) \\ \text{Cov}(\theta_m, \theta_n) &= \mathbb{E}[\theta_n] \left(\frac{\alpha_m}{\alpha_m + \kappa_m + 1} \prod_{l=1}^{m-1} \frac{\kappa_l + 1}{\alpha_l + \kappa_l + 1} - \mathbb{E}[\theta_m] \right)\end{aligned}$$

C.3.1 EXPONENTIAL FORM

The natural parameters are as follows:

$$\begin{aligned}\eta_{1k} &= \alpha_k - 1, \quad k = 1, \dots, K-1 \\ \eta_{2k} &= \kappa_k - K + k, \quad k = 1, \dots, K-1\end{aligned}$$

The sufficient statistics:

$$\begin{aligned}u_{1k} &= \log \theta_k - \log \sum_{l=k}^K \theta_l, \quad k = 1, \dots, K-1 \\ u_{2k} &= \log \sum_{l=k+1}^K \theta_l - \log \sum_{j=k}^K \theta_j, \quad k = 1, \dots, K-1\end{aligned}$$

The log-normalizer:

$$\log g(\boldsymbol{\eta}) = \sum_{k=1}^{K-1} \log \Gamma(\alpha_k) + \sum_{k=1}^{K-1} \log \Gamma(\kappa_k) - \sum_{k=1}^{K-1} \log \Gamma(\alpha_k + \kappa_k)$$

The expectation of sufficient statistics:

$$\mathbb{E}[u_{1k}] = \psi(\alpha_k) - \psi(\alpha_k + \kappa_k), \quad k = 1, \dots, K-1$$

$$\mathbb{E}[u_{2k}] = \psi(\kappa_k) - \psi(\alpha_k + \kappa_k), \quad k = 1, \dots, K-1$$

And:

$$\mathbb{E}[\log \theta_1] = \psi(\alpha_1) - \psi(\alpha_1 + \kappa_1)$$

$$\mathbb{E}[\log \theta_k] = \psi(\alpha_k) - \psi(\alpha_k + \kappa_k) + \sum_{l=1}^{k-1} \psi(\kappa_l) - \sum_{j=1}^{k-1} \psi(\alpha_j + \kappa_j), \quad k = 2, \dots, K-1$$

$$\mathbb{E}[\log \theta_K] = \sum_{l=1}^{K-1} \psi(\kappa_l) - \sum_{j=1}^{K-1} \psi(\alpha_j + \kappa_j)$$

C.3.2 CONJUGACY

The Generalized-Dirichlet distribution is conjugate to the multinomial. Given multinomial observation $\mathbf{n} = (n_1, \dots, n_K)$, the parameters of the Generalized-Dirichlet posterior are updated as:

$$\alpha'_k = \alpha_k + n_k, \quad k = 1, \dots, K-1$$

$$\kappa'_k = \kappa_k + \sum_{l=k+1}^K n_l, \quad k = 1, \dots, K-1$$

Appendix D. Matrix form of Bayesian Theorem

The matrix form of Bayesian theory provides an intuition for Expectation-Maximization algorithm. The basic form of Bayesian theorem is defined over binary sample space $\{A, A^C\}$ and $\{B, B^C\}$:

$$\begin{aligned} \begin{pmatrix} P(A|B) & P(A|B^C) \\ P(A^C|B) & P(A^C|B^C) \end{pmatrix} \begin{pmatrix} P(B) & 0 \\ 0 & P(B^C) \end{pmatrix} = \\ \begin{pmatrix} P(AB) & P(AB^C) \\ P(A^CB) & P(A^CB^C) \end{pmatrix} = \\ \begin{pmatrix} P(A) & 0 \\ 0 & P(A^C) \end{pmatrix} \begin{pmatrix} P(B|A) & P(B|A^C) \\ P(B^C|A) & P(B^C|A^C) \end{pmatrix}^T \end{aligned}$$

Therefore, we have:

$$\begin{pmatrix} P(A|B) & P(A|B^C) \\ P(A^C|B) & P(A^C|B^C) \end{pmatrix} = \begin{pmatrix} P(A) & 0 \\ 0 & P(A^C) \end{pmatrix} \begin{pmatrix} P(B|A) & P(B|A^C) \\ P(B^C|A) & P(B^C|A^C) \end{pmatrix}^T \begin{pmatrix} P(B) & 0 \\ 0 & P(B^C) \end{pmatrix}^{-1}$$

Now, we generalize Bayesian theorem to multiple sample space. Suppose we have two finite sample spaces, $\mathbb{Z} = \{z_1, \dots, z_k, \dots, z_K\}$, and $\mathbb{W} = \{w_1, \dots, w_v, \dots, w_V\}$; and we assign a probabilistic distribution to each of the sample spaces:

$$\boldsymbol{\theta} = (\theta_k)_K = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \\ \vdots \\ \theta_K \end{pmatrix} = \begin{pmatrix} P(z_1) \\ \vdots \\ P(z_k) \\ \vdots \\ P(z_K) \end{pmatrix} \quad \mathbf{t} = (t_v)_V = \begin{pmatrix} t_1 \\ \vdots \\ t_v \\ \vdots \\ t_V \end{pmatrix} = \begin{pmatrix} P(w_1) \\ \vdots \\ P(w_v) \\ \vdots \\ P(w_V) \end{pmatrix}$$

Let $M(\cdot)$ be the transform from a vector to a square diagonal matrix with each of its diagonal component being the corresponding component of the vector, and other components being 0.

$$M(\boldsymbol{\theta}) = \begin{pmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_K \end{pmatrix} \quad M(\mathbf{t}) = \begin{pmatrix} t_1 & 0 & \cdots & 0 \\ 0 & t_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_V \end{pmatrix}$$

Let $\boldsymbol{\varphi} = \{\varphi_{vk}\}_{V \times K}$ and $\boldsymbol{\phi} = \{\phi_{kv}\}_{K \times V}$ be two conditional probability matrices where $\varphi_{vk} = P(w_v|z_k)$ and $\phi_{kv} = P(z_k|w_v)$ for $k = 1, 2, \dots, K$ and $v = 1, 2, \dots, V$; and let $\Phi = \{P(w_v, z_k)\}_{V \times K}$ be the joint probabilistic matrix where $P(w_v, z_k) = P(w_v|z_k)P(z_k)$. Similarly, $\Phi^T = \{P(z_k, w_v)\}_{K \times V}$ where $P(z_k, w_v) = P(z_k|w_v)P(w_v)$. Obviously, we have:

$$\mathbf{t} = \boldsymbol{\varphi}\boldsymbol{\theta}$$

$$\boldsymbol{\theta} = \boldsymbol{\phi}\mathbf{t}$$

$$\Phi = \boldsymbol{\varphi}M(\boldsymbol{\theta})$$

$$\Phi^T = \boldsymbol{\phi}M(\mathbf{t})$$

Let $\mathbf{1}_n$ be an n -dimensional vector whose components all equal to 1. Consequently, we have:

$$\Phi \mathbf{1}_K = \mathbf{t}$$

$$\Phi^T \mathbf{1}_V = \boldsymbol{\theta}$$

Theorem 14 (Matrix form of Bayesian theorem) *Given the marginal probability distributions $\boldsymbol{\theta}$ and \mathbf{t} respectively corresponding to the conditional probability matrices $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$, the following transforms exit:*

$$\boldsymbol{\varphi} = M(\mathbf{t})\boldsymbol{\phi}^T M(\boldsymbol{\theta})^{-1}$$

$$\boldsymbol{\phi} = M(\boldsymbol{\theta})\boldsymbol{\varphi}^T M(\mathbf{t})^{-1}$$

Proof

$$\Phi = \boldsymbol{\varphi} M(\boldsymbol{\theta}) = (\boldsymbol{\phi} M(\mathbf{t}))^T$$

therefore

$$\boldsymbol{\varphi} = M(\mathbf{t})\boldsymbol{\phi}^T M(\boldsymbol{\theta})^{-1}$$

The proof of the other equation is analogous. ■

Appendix E. Details of Explicitization

For the first step, we define a new operation on matrices.

Definition 15 (Matrix Exponential) *Given two matrices $\mathbf{A} = \{a_{ml}\}_{M \times L}$ and $\mathbf{B} = \{b_{ln}\}_{L \times N}$, define \mathbf{A} to the power of \mathbf{B} as $\mathbf{C} = \mathbf{A}^{\mathbf{B}} = \{c_{mn}\}_{M \times N}$, where:*

$$c_{mn} = \prod_{l=1}^L a_{ml}^{b_{ln}}$$

The operation has the following properties:

1. $\mathbf{A}^{\mathbf{I}} = \mathbf{A}$
2. $\mathbf{A}^{\mathbf{B}+\mathbf{C}} = \mathbf{A}^{\mathbf{B}} \odot \mathbf{A}^{\mathbf{C}}$
3. $(\mathbf{A} \odot \mathbf{B})^{\mathbf{C}} = \mathbf{A}^{\mathbf{C}} \odot \mathbf{B}^{\mathbf{C}}$
4. $\Phi^{\mathbf{AB}} = (\Phi^{\mathbf{A}})^{\mathbf{B}}$

where \mathbf{I} is identity matrix, \odot is Hadamard product. The properties are obvious except the last one. Now, we prove the last one:

Proof Without the loss of generality, consider a special case where Φ is a row vector $\boldsymbol{\theta}^T = (\theta_1, \dots, \theta_k, \dots, \theta_K)$, \mathbf{B} is a column vector $\mathbf{n} = (n_v)_V$, and \mathbf{A} is matrix $\boldsymbol{\phi} = \{\phi_{kv}\}_{K \times V}$. We only need to prove that the property is correct for this special case and it can be naturally generalized.

According to the definition:

$$\begin{aligned} (\boldsymbol{\theta}^T)^{\boldsymbol{\phi}\mathbf{n}} &= \prod_{k=1}^K \theta_k^{\sum_{v=1}^V \phi_{kv} n_v} = \\ &= \prod_{k=1}^K \left(\prod_{v=1}^V \theta_k^{\phi_{kv} n_v} \right) = \prod_{v=1}^V \left(\prod_{k=1}^K \theta_k^{\phi_{kv} n_v} \right) = \\ &= \prod_{v=1}^V \left(\prod_{k=1}^K \theta_k^{\phi_{kv}} \right)^{n_v} = \left[(\boldsymbol{\theta}^T)^{\boldsymbol{\phi}} \right]^{\mathbf{n}} \end{aligned}$$

■

The definition here is different from the traditional definition of matrix exponentiation in Lie groups and Lie algebra, which requires the base to be square matrix. The proof explains the origin of implicit approximate posterior in Expectation Propagation.

Matrix exponential can be viewed as some form of “inner power”, which is similar in spirit to the definition of inner product. Now, we can show that the explicitization is more essentially interpreted as a decoration of Bayesian operator caused by the matrix exponential:

$$\begin{aligned}
 \mathbf{B}_D(\phi \mathbf{n}_{(v)}) &= \left[\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi} + \mathbf{D}\phi \mathbf{n}_{(v)}))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \right]^{-1} \prod_{k=1}^K \theta_k^{\sum_{v=1}^V \phi_{kv} n_v} \\
 &= \left[\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi}_n))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \right]^{-1} (\boldsymbol{\theta}^T)^{\phi \mathbf{n}} \\
 &= \left[\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi}_n))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \right]^{-1} ((\boldsymbol{\theta}^T)^{\phi})^{\mathbf{n}} \\
 &= \left[\frac{g(\boldsymbol{\eta}(\boldsymbol{\xi}_n))}{g(\boldsymbol{\eta}(\boldsymbol{\xi}))} \right]^{-1} \prod_{v=1}^V \left(\prod_{k=1}^K \theta_k^{\phi_{kv}} \right)^{n_v}
 \end{aligned}$$

References

- Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, 2016.
- Ali Shojaee Bakhtiari and Nizar Bouguila. A variational bayes model for count data learning and classification. *Engineering Applications of Artificial Intelligence*, 35:176–186, 2014.
- Ali Shojaee Bakhtiari and Nizar Bouguila. A latent beta-liouville allocation model. *Expert Systems with Applications*, 45:260–272, 2016.
- Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006.
- David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- David M. Blei and John D. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.
- David M. Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.
- David M. Blei, Thomas Griffiths, and Michael Jordan. The nested chinese restaurant process and bayesian nonparametric models of hierarchies. *Journal of the ACM*, 2010.
- David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

- Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport: École d'Été de Probabilités de Saint-Flour XLIX – 2019*, volume 2364 of *Lecture Notes in Mathematics*. Springer Cham, 2025. ISBN 978-3-031-85159-9. doi: 10.1007/978-3-031-85160-5.
- Robert Connor and James Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64:194–206, 1969.
- S. Y. Dennis III. On the hyper-dirichlet type 1 and hyper-liouville distributions. *Communications in Statistics—Theory and Methods*, 20(12):4069–4081, 1991.
- Kai-Wang Fang, Samuel Kotz, and Kai-Wang Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, 1990.
- Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005.
- Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2013.
- Rameshwar D. Gupta and Donald St. P. Richards. Multivariate liouville distributions. *Journal of Multivariate Analysis*, 23(2):233–256, 1987.
- Daniel Hernández-Lobato and José Miguel Hernández-Lobato. Scalable gaussian process classification via expectation propagation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *Proceedings of Machine Learning Research*, pages 168–176, 2016.
- Matthew Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.
- Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.
- Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.
- Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, 2006.
- Lianliang Liu, Liang Tang, Wei Dong, Shiyao Yao, and Wenbin Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608, 2016. doi: 10.1186/s40064-016-3252-8.
- Thomas P. Minka. The dirichlet-tree distribution. *Justsystem Pittsburgh Research Center*, 1999. URL <https://tminka.github.io/papers/dirichlet/minka-dirtree.pdf>.
- Thomas P. Minka. Estimating a dirichlet distribution. Technical report, Microsoft Research, 2000. URL <https://tminka.github.io/papers/dirichlet/>. Technical report.

- Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.
- Thomas P. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005. URL <https://tminka.github.io/papers/message-passing/minka-divergence.pdf>. Technical Report.
- Thomas P. Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 352–359. Morgan Kaufmann, 2002.
- J. E. Mosimann. On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. *Biometrika*, 49(1–2):65–82, 1962.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andy Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
- Max Sklar. Fast mle computation for the dirichlet multinomial. *arXiv preprint*, arXiv:1405.0099, 2014. URL <https://arxiv.org/abs/1405.0099>. Submitted May 2014; revised May 2023.
- Yee Whye Teh, Michael Jordan, Matthew Beal, and David M. Blei. Hierarchical dirichlet processes. In *Journal of the American Statistical Association*, 2006.
- Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P. Cunningham, David Schiminovich, and Christian P. Robert. Expectation propagation as a way of life: A framework for bayesian inference on partitioned data. *Journal of Machine Learning Research*, 21(17):1–53, 2020.
- Martin Wainwright and Michael Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, 2008.
- Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. *Advances in Neural Information Processing Systems*, 22, 2009.
- Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. *KDD*, pages 448–456, 2011.
- Tzu-Tung Wong. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, 97(2–3):165–181, 1998. doi: 10.1016/S0096-3003(97)10140-0.
- Bohan Wu and David Blei. Extending mean-field variational inference via entropic regularization: Theory and computation. *arXiv preprint*, arXiv:2404.09113, 2024.