Supervised Hierarchical Dirichlet Processes with Variational Inference

Cheng Zhang   Carl Henrik Ek   Xavi Gratal   Florian T. Pokorny   Hedvig Kjellström
Computer Vision and Active Perception Lab, Centre for Autonomous Systems
KTH Royal Institute of Technology
Stockholm, Sweden
{chengz, chek, javiergm, fpokorny, hedvig}@kth.se

Abstract

We present an extension to the Hierarchical Dirichlet Process (HDP), which allows for the inclusion of supervision. Our model marries the non-parametric benefits of HDP with those of Supervised Latent Dirichlet Allocation (SLDA) to enable learning the topic space directly from data while simultaneously including the labels within the model. The proposed model is learned using variational inference which allows for the efficient use of a large training dataset. We also present the online version of variational inference, which makes the method scalable to very large datasets. We show results comparing our model to a traditional supervised parametric topic model, SLDA, and show that it outperforms SLDA on a number of benchmark datasets.

1. Introduction

During the last decade, topic models have successfully been used for modelling data across several different domains such as information retrieval [1, 2, 13], computer vision [3, 5, 10, 16, 18], and robotics [19]. This success stems from the fact that I) the interpretation in terms of topics is a natural description for many types of data, and II) the theoretical foundation of the models provides a principled approach for learning and inference. The original work on topic models comes from the text-mining community [4], and with the introduction of Latent Dirichlet Allocation (LDA) [2], topic models were then later applied widely to other domains.

Many learning tasks, *e.g.*, in computer vision, make use of supervised models where data are associated with labels. There have been several extensions of the LDA model [1, 7, 11] to accommodate supervision in such a way that the model can be directly used for classification tasks. These can be roughly divided into two different approaches: upstream supervision, where the latent topics are dependent on the label, and downstream supervision, where the label is set as a response variable of the topics. The different LDA extensions use different learning procedures. Supervised LDA (SLDA) [1], which uses downstream supervision, optimizes the joint likelihood of the data and the label of the document. Given the recent success of SLDA [1, 14, 20], the model presented in this paper uses the same supervision approach.

The main drawback of LDA, and its predecessor Latent Semantic Analysis [4], is that the number of topics needs to be set manually. This is especially troublesome for applications where the topic space lacks a clear semantical interpretation, as often is the case in computer vision. Moreover, it was shown in [13] that the performance of the LDA model is very dependent on the number of topics. To rectify this problem, Teh *et al*. [13] proposed a non-parametric model capable of learning the number of topics, referred to as Hierarchical Dirichlet Processes (HDP).

In this paper, we describe a supervised version of HDP, SHDP. The contributions of this method are twofold:

- Firstly, its non-parametric property (that the topic space size is automatically learned from the data) gives it an advantage compared to current supervised topic models such as SLDA.
- Secondly, it allows us to use an HDP framework in a supervised setting, which increases the range of problems (not least in computer vision) to which HDP methods can be applied.

We provide experimental results which show that the SHDP model outperforms SLDA on a number of benchmark datasets.

2. Related Work

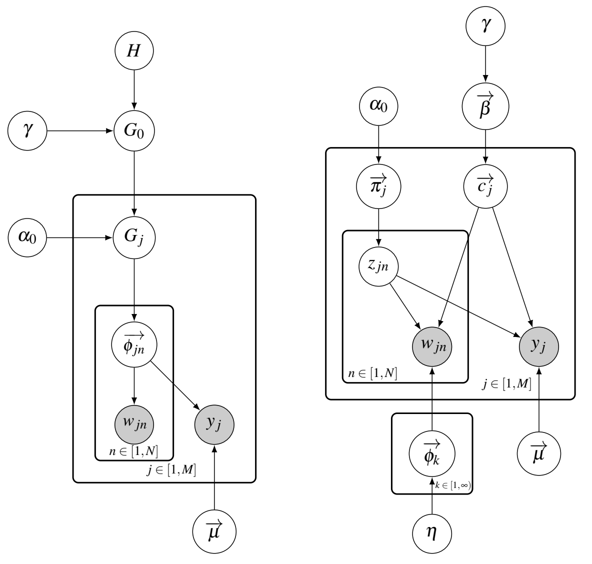
Latent Dirichlet Allocation [2] assumes a generative process where each document is modeled as a distribution over topics, and topics are modeled as distributions over the observed words. This hierarchical structure of the representation is natural for many types of data, *e.g.*, in computer vision, where an image corresponds to a document and the words are visual features extracted from the image. Blei and McAuliffe [1] firstly introduced supervision in a downstream manner as an additional response variable. Later, Wang *et al.* [14] accommodated a discrete response variable to achieve a model suitable for classification. Taking a different approach, Fei-Fei and Perona [5] used upstream supervision, where the topic space is directly affected by the labels, to perform natural scene classification. The success of these works has shown that LDA is a suitable framework for numerous computer vision tasks that require supervision. Supervised approaches to LDA have further been developed for computer vision. For example, Cao and Fei-Fei [3] proposed a spatially coherent LDA model which achieved promising results on both segmentation and classification, and Zhang *et al.* [18] introduced a model that learned a factorized topic space that separated signal (descriptive of class) and noise (not descriptive of class) into different topics, making the topics more interpretable and giving a better classification result.

When learning an LDA model, the number of topics needs to be set a-priori. Experimentally, and not surprisingly, it has been shown [13] that the performance of the model is influenced by the number of topics. For many types of data, e.g., image data, the appropriate number of topics is not obvious. To circumvent this problem, the notion of a Hierarchical Dirichlet Process (HDP) was proposed by Teh *et al.* [13]. An HDP is a non-parametric approach to topic modelling which automatically learns the number of topics from data. Applied to natural language processing, Xie and Rassoneau [17] proposed a semi-supervised HDP model, where the “label” is the distribution of topics of the words – effectively a word-level labeling. Thus, this model is not directly applicable to document classification tasks of the type common in computer vision. To our knowledge no fully supervised HDP models exists, which is why most topic models in computer vision are still based on LDA rather than HDP. Therefore, we expect our SHDP model to be of great use to the computer vision community.

Variational inference is in general an efficient method [2, 1, 14] in itself. However, to adapt variational inference to massive amounts of data, online variational inference methods have been developed. Hoffman *et al.* [6] accommodated online variational inference in the LDA framework, developing Online LDA. Building on this, Wang *et al.* [15] proposed Online HDP. In the experiments in this paper, we use the SHDP framework with variational inference. We also present an extension of SHDP to accommodate online variational learning which makes it useful for computer vision tasks with massive amounts of data.

3. Model

In coherence with most other research articles on topic modeling, we use the notion of a corpus, documents, topics and words to describe our model. In our experiments concerning action classification, a corpus is a set of video sequences, a document is a single video clip, and words are bag-of-STIP [8] features. Our model (Figure 1) is an extension of HDP [13]. In the following, we omit some details about HDP and focus on the extension required to introduce supervision.



(a) Baysian representation

(b) Stick-breaking representation for $H = Dir(\eta,\ldots,\eta)$

Figure 1. Graphic representation of Supervised HDP

3.1. The Stick Breaking Construction for HDP

The stick-breaking construction [12] is an intuitive way to construct a Dirichlet process $DP(H,\gamma)$ with base distribution H and concentration parameter γ. We use the same approach as Wang *et al.* [15], which we review now.

To sample $G \sim DP(H,\gamma)$, we draw, for $k \in \mathbb{N}$, $\phi_k \sim H$ and

$$\begin{aligned} \beta'_k &\sim Beta(1,\gamma), \\ \beta_k &= \beta'_k \prod_{l=1}^{k-1}(1-\beta'_l), \end{aligned}$$

where one can think of β'_k as proportions of a unit length stick which is recursively being broken into two, with one part being put aside. The resulting β_k sum to one and $G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$, where δ_{ϕ_k} denotes a delta function supported at ϕ_k and $G \sim DP(H,\gamma)$ yields a draw from $DP(H,\gamma)$. In a topic model, H is typically a symmetric Dirichlet distribution over the vocabulary simplex and with parameter $\eta > 0$. The ϕ_k then correspond to topics which are distributions over words and the β_k form topic weights. Here, γ is a parameter for the Beta process, and β_k can be interpreted as the length of the k^{th} resulting stick part.

There are different ways to derive an HDP using a stick breaking process. In [15], the HDP is constructed by applying two stick-breaking constructions successively, first on the corpus and then on the document level. On the corpus level, the above construction yields $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \sim DP(H,\gamma)$. Next a per-document stick breaking yields $G_j \sim DP(G_0,\alpha_0)$, which depends on the corpus level topics:

$$G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}, \tag{1}$$

where π_{jt} are the weights for topic ψ_{jt} of document j generated in a similar manner as the weights β_k on the corpus level by sampling $\pi'_{jt} \sim Beta(1,\alpha_0)$. The important difference is how the topics themselves are drawn. On the corpus level, topics are drawn from the prior distribution H, but on the document level, the topics ψ_{jt} are drawn from G_0 in the following way:

$$\begin{aligned} c_{jt} &\sim Mult(\beta), \\ \psi_{jt} &= \phi_{c_{jt}}, \end{aligned} \tag{2}$$

where c_{jt} are indicator variables which index the corpus-level topic corresponding to ψ_{jt}.

The document-level topics depend on the latent variable $\mathbf{c} = (c_{jt})$, which constitutes a link between the corpus and document topics, in the following way: let z_{jn} denote the document level topic indicator for document j and word n. Given a stick-breaking partitioning on both corpus and document level, the n^{th} word in the j^{th} document, w_{jn}, is generated by first drawing the indicator variable z_{jn} from a multinomial parameterized by the topic weights on the document level π_j as $z_{jn} \sim Mult(\pi_j)$. The word w_{jn} is then drawn from the corpus level topic space mapped by $\mathbf{c}$:

$$\begin{aligned} \theta_{jn} &= c_{jz_{jn}}, \\ w_{jn} &\sim Mult(\phi_{\theta_{jn}}) = Mult(\phi_{c_{jz_{jn}}}). \end{aligned} \tag{3}$$

Both z_{jn} and θ_{jn} have the role of a topic assignment indicator, z_{jn} by choosing a topic from the support of the document level topic distribution G_j and θ_{jn} by mapping from the document level topic assignment z_{jn} to the corpus level topics though $\mathbf{c}$. As Figure 1(b) shows, the generation of a word w_{jn} depends on the document level topic assignment z_{jn}, the indicator $\mathbf{c}_j$ which maps the document level topic assignment to the corpus level topic assignment, and the corpus level topic-words distribution π. Instead of using topic indicators in the stick-breaking representation of the model as Figure 1(b), the Bayesian representation uses the topic mixtures directly. In Figure 1(a), H is a symmetric Dirichlet distribution parametrized by η, defined over a V-dimensional simplex, where V is the vocabulary size; G_0 denotes the corpus level topic mixture; G_j denotes topic mixture for document j. Figure 1(a) yields a visualization which can be used to compare the model with HDP [13] while Figure 1(b) provides more intuition on the stick-breaking constructions of the model and for the variational inference of the model. We will now proceed to show how supervision can be incorporated with HDP using the above construction.

3.2. Supervised HDP

As discussed in the related work, there are generally two different ways to introduce supervision or labels into topic models, upstream or downstream. The model we present uses a downstream approach [1, 14] where the labels can be seen as an additional response variable to the topics. In this paper, we will focus on applications where the labels are discrete, *i.e.*, classification tasks. The goal is then to infer a discrete response y, given a set of words $\{w_n\}$ for a document.

The addition compared to standard HDP (Figure 1) is that y is generated as a response to the topics along with the words. The distribution over labels $\{y_1,\ldots,y_C\}$ within the model is implemented using a soft-max function [14]:

$$p(y_j|\bar{\theta}_j,\mu) = \frac{\exp(\mu_{y_j}^T \bar{\theta}_j)}{\sum_{l=1}^{C} \exp(\mu_l^T \bar{\theta}_j)}, \tag{4}$$

where μ is the parameter of the soft-max function that will be inferred from data as described in the next section. The parameter $\bar{\theta}_j$ is a vector representing the accumulative topic count:

$$\bar{\theta}_j = \frac{1}{N} \sum_{n=1}^{N} \Theta_{jn} = \frac{1}{N} \sum_{n=1}^{N} \mathfrak{c}_{jz_{jn}}, \tag{5}$$

where Θ_{jn} is the binary indicator vector representation of $\theta_{jn} \in \mathbb{N}$. Similarly, $\mathfrak{c}_{jt}$ denotes the binary indicator vector representation of c_{jt} which is 0 except at position $c_{jt} \in \mathbb{N}$ where it takes value 1. Note that, if $K \in \mathbb{N}$ denotes the largest index i of any topic ϕ_i giving rise to some word in the finite corpus, then all entries of $\bar{\theta}_j$ and hence also of $\mu_{y_j}^T \bar{\theta}_j$, for indices larger than K and all $j \in \{1,\ldots,M\}$ are zero, and we can then think of these as finite K-dimensional vectors. It is furthermore easy to exchange the soft-max function with any generalized linear model to model response variables with a different distribution.

We will now proceed to show how the parameters of this HDP model with additional supervision can be learned from data in an efficient manner.

4. Inference

In this section, we will describe how the parameters of the model are inferred from data. This will be done with a variational inference method. We will focus on the differences between the proposed model and the original HDP rather than deriving the full model. However, for the sake of clarity, we provide a more detailed description as supplementary material to the paper. Our derivation will first focus on a batch approach, based on [2, 15], which we will proceed to extend to an online version adopting the inference scheme used in [6, 15] and enabling us to use very large datasets.

4.1. Variational Inference

Given the model, we need to estimate all the latent variables: the corpus level stick breaking relevance proportion

β', per document stick breaking relevance proportion π', per document topic indices $\mathbf{c}$ (which map the document topics to corpus level topics), per word topic indices $\mathbf{z}$ (which are the document level topic indices for each word in the document), the topics ϕ (the word distribution for each topic), and the soft-max parameter μ. We will follow the standard mean field variational steps [15] and use a fully factorized variational distribution:

$$q(\beta',\pi',\mathbf{c},\mathbf{z},\phi) = q(\beta')q(\pi')q(\mathbf{c})q(\mathbf{z})q(\phi), \qquad (6)$$

where

1. $q(\beta') = \prod_{k=1}^{K} q(\beta'_k|v_{1k},v_{2k})$, where v_{1k}, v_{2k} are variational parameters for a beta distribution.
2. $q(\pi') = \prod_{j=1}^{M}\prod_{t=1}^{T} q(\pi_{jt}|a_{1jt},a_{2jt})$, where a_{1jt}, a_{2jt} are variational parameters for a beta distribution.
3. $q(\mathbf{c}) = \prod_j \prod_t q(c_{jt}|\rho_{jt})$, where ρ_{jt} is a variational parameter for a multinominal distribution.
4. $q(\mathbf{z}) = \prod_j \prod_n q(z_{jn}|\zeta_{jn})$, where ζ_{jn} is a variational parameter for a multinominal distribution.
5. $q(\phi) = \prod_k q(\phi_k|\lambda_k)$, where λ_k is a variational parameter for a Dirichlet distribution.

In SHDP, the number of topics, K, on the corpus level is unbounded. However, due to the nature of the stick breaking constructions, the weight β'_k tends towards 0 when k becomes large, simply because less and less of the stick remains to partition. We exploit this by truncating the weights; with a large K, we consider only β'_j for $j \le K$. A similar strategy can be deployed for the document level where we only consider the first T topics on the document level and discard the remainder, resulting in truncated π_j and c_j vectors. The parameters v_{1k}, v_{2k}, a_{1jt} and a_{2jt} control the beta distributions governing the stick breaking on the corpus and the document level respectively. ρ_{jt} is a K-dimensional multinomial parameter which controls the probability to select corpus level topics for the j^{th} document's t^{th} topic, while ζ_{jn} is a T-dimensional multinomial parameter which governs the probability of topic assignment from the document level topics.

In a standard variational inference, using Jensen's inequality, we get:

$$\begin{aligned}
&\log p(\mathbf{w},\mathbf{y}|\gamma,\alpha_0,\eta,\mu)\\
&= \log \int_{\beta',\pi',\phi} \sum_{\mathbf{c},\mathbf{z}} p(\beta',\pi',\mathbf{c},\mathbf{z},\phi,\mathbf{w},y|\gamma,\alpha_0,\eta,\mu)\\
&= \log \int_{\beta',\pi',\phi} \sum_{\mathbf{c},\mathbf{z}} p(\beta',\pi',\mathbf{c},\mathbf{z},\phi,\mathbf{w},y|\gamma,\alpha_0,\eta,\mu)\frac{q(\beta',\pi',\mathbf{c},\mathbf{z},\phi)}{q(\beta',\pi',\mathbf{c},\mathbf{z},\phi)}\\
&\ge \mathop{\mathbb{E}}_{q}[\log p(\beta',\pi',\mathbf{c},\mathbf{z},\phi,\mathbf{w},y|\gamma,\alpha_0,\eta,\mu)] - \mathop{\mathbb{E}}_{q}[\log q(\beta',\pi',\mathbf{c},\mathbf{z},\phi)].
\end{aligned} \qquad (7)$$

Hence, the right hand side above yields a lower bound, which is:

$$\begin{aligned}
\mathscr{L} &= \mathop{\mathbb{E}}_{q}[\log p(\beta',\pi',\mathbf{c},\mathbf{z},\phi,\mathbf{w},y|\gamma,\alpha_0,\eta)] - \mathbb{E}[\log q(\beta',\pi',\mathbf{c},\mathbf{z},\phi)]\\
&= \sum_{j=1}^{M}\Big(\mathop{\mathbb{E}}_{q}[\log p(w_j|\mathbf{c}_j,\mathbf{z}_j,\phi)] + \mathop{\mathbb{E}}_{q}[\log p(\mathbf{c}_j|\beta')] + \mathop{\mathbb{E}}_{q}[\log p(\mathbf{z}_j|\pi_j)]\\
&+ \mathop{\mathbb{E}}_{q}[\log p(\pi'_j|\alpha_0)] + \mathop{\mathbb{E}}_{q}[\log p(y_j|\mathbf{z}_j,\mathbf{c}_j,\mu)]\Big) + \mathop{\mathbb{E}}_{q}[\log p(\phi|\eta)]\\
&+ \mathop{\mathbb{E}}_{q}[\log p(\beta'|\gamma)] - \sum_{j=1}^{M}\Big(\mathop{\mathbb{E}}_{q}[\log q(\pi_j)] + \mathop{\mathbb{E}}_{q}[\log q(\mathbf{c}_j)] + \mathop{\mathbb{E}}_{q}[\log q(\mathbf{z}_j)]\Big)\\
&- \mathop{\mathbb{E}}_{q}[\log q(\beta')] - \mathop{\mathbb{E}}_{q}[\log q(\phi)].
\end{aligned} \qquad (8)$$

Computation details of Equation 8 can be found in the supplementary material to this paper.

Compared to standard variational inference on HDP, the extra terms we need to compute are of the form $\mathbb{E}_q[\log p(y_j|\mathbf{z},\mathbf{c},\mu)]$. These terms contain the per document class label y_j, as implemented with the softmax function in Equation (4). It depends on the average of the topic distribution for each document, and hence on the per word topic indices $\mathbf{z}_j$ and per document topic indices $\mathbf{c}_j$. The softmax parameter μ also needs to be updated. Using the same trick as [15], we get:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{q}[\log p(y_j|\mathbf{z},\mathbf{c},\mu)] \ge \mu_{y_j}^{T}\Big(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\rho_{jt}\zeta_{jnt}\Big)\\
&-\log\left(\sum_{l=1}^{C}\prod_{n=1}^{N}\left(\sum_{e=1}^{K}\sum_{i=1}^{T}\rho_{jie}\zeta_{jni}\exp\Big(\frac{1}{N}\mu_{le}\Big)\right)\right).
\end{aligned} \qquad (9)$$

Details on the computation of this inequality can be found in Appendix A.1. $\mathbb{E}_q[\log p(y_j|\mathbf{z},\mathbf{c},\mu)]$ is part of the lower bound. Compared to variational HDP, this is the only new term that we introduced to the bound. Hence, the update equation for ρ and ζ will be influenced. Other variational parameters will be updated in the same manner as HDP [15]. Using the right-hand side of the inequality, we get a new bound, which is used in the following.
To update ρ, we see that the part of the new lower bound with ρ terms is:

$$\begin{aligned}
\mathscr{L}_\rho &= \sum_{j=1}^{M}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\sum_{n=1}^{N}\zeta_{jnt}\rho_{jtk}\mathop{\mathbb{E}}_{q}[\log p(w_{jn}|\phi_k)]\right.\\
&\left.+\rho_{jtk}\mathop{\mathbb{E}}_{q}[\log\beta_k] - \rho_{jtk}\log\rho_{jtk}\right) + \sum_{j=1}^{M}\left(\mu_{y_j}^{T}\Big(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\rho_{jt}\zeta_{jnt}\Big)\right.\\
&\left.-\log\left(\sum_{l=1}^{C}\prod_{n=1}^{N}\left(\sum_{e=1}^{K}\sum_{i=1}^{T}\rho_{jie}\zeta_{jni}\exp\Big(\frac{1}{N}\mu_{le}\Big)\right)\right)\right).
\end{aligned} \qquad (10)$$

There is no closed form solution to compute ρ to maximize (10). Hence, optimizing ρ becomes a constrained non-linear optimization problem. There are many algorithms and libraries for solving this. For example, the conjugate gradient method, which only requires the computation of partial derivatives, can be used. The partial derivatives are given in Appendix A.2. In order to update ζ,

Algorithm 1: Batch Variational Inference for SHDP

```
Initialize all the variational parameters
while Not converged or within MAX iteration do
    E Step:
    For each document
    Update per document stick
        $a_{1jt} = 1 + \sum_{n}^{N} \zeta_{jnt}$        $a_{2jt} = \alpha_0 + \sum_{n}^{N} \sum_{s=t+1}^{T} \zeta_{jns}$
    Update per document topic indices $\rho$ numerically using
    Equations (10) and (18)
    Update per word topic indices $\zeta$ using Equation (23)
    M Step:
    Update corpus level stick
        $v_{1k} = 1 + \sum_{j}^{M} \sum_{t=1}^{T} \rho_{jtk}$        $v_{2k} = \gamma + \sum_{j}^{M} \sum_{t=1}^{T} \sum_{l=k+1}^{K} \rho_{jtl}$
    Update topic mixture
        $\lambda_{ki} = \sum_{j=1}^{M} \sum_{t=1}^{T} \rho_{jtk} (\sum_{n=1}^{N} \zeta_{jnt} [w_{jn} = i]) + \eta$
    Update the label parameter $\mu$ using Equation (12) and (24)
```

we observe that $\sum_{l=1}^{C} \prod_{n=1}^{N} \left(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N} \mu_{le}) \right)$ can be considered as a linear function of ζ_{jn}, for fixed j, n. We define h not involving ζ_{jn} (see A.3) so that $h^T \zeta_{jn} = \sum_{l=1}^{C} \prod_{n=1}^{N} \left(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N} \mu_{le}) \right)$. Following [14], we derive the fixed point update:

$$\begin{aligned} \zeta_{jnt} \propto \exp \Big(& \sum_{k=1}^{K} \rho_{jtk} \mathop{\mathbb{E}}_{q} [\log p(w_{jn} | \phi_k)] + \mathop{\mathbb{E}}_{q} [\log \pi_{jt}] \\ & + \frac{1}{N} \mu_{y_j}^T \rho_{jt} - (h^T \zeta_{jn}^{old})^{-1} h_t \Big). \end{aligned} \tag{11}$$

Details of this derivation is shown in Appendix A.3.

Finally, we need to estimate the soft-max parameter μ. The part of the lower bound which varies with μ is

$$\begin{aligned} \mathscr{L}_{\mu} = \sum_{j=1}^{M} \Bigg(& \mu_{y_j}^T (\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \rho_{jt} \zeta_{jnt}) \\ & - \log \Big(\sum_{l=1}^{C} \prod_{n=1}^{N} \Big(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N} \mu_{le}) \Big) \Big) \Bigg). \end{aligned} \tag{12}$$

Compared with Eq. 9, we can see that this is the labeling component that we introduced to the lower bound. The optimization of this term does not have a closed form solution either since it is highly non-linear in μ. Thus, we will estimate an optimal solution using the conjugate gradient method as well. The derivatives of this expression are given in Appendix A.4. We sum up the batch variational SHDP algorithm in Algorithm 1.

4.2. Online Variational Inference

We will now extend the batch variational inference to an online setting which scales well with large data sizes. Let M be the total number of documents. The basic idea is [15, 6]:

$$\mathscr{L} = \sum_{j}^{M} \mathscr{L}_j = \mathop{\mathbb{E}}_{j} [M \mathscr{L}_j]. \tag{13}$$

Given the corpus level parameters, the document level parameters ($a_{1j.}$, $a_{2j.}$, $\rho_{j.}$, $\zeta_{j.}$) are computed in the same way as in the batch variational inference. As in [15], we update the corpus level parameters using the gradient $D\mathscr{L}_j$ and follow the gradient with learning rate ω_{t_o}. ω_{t_o} is decreasing as t_0 increases, which denotes the number of the documents that the model has read. As in [6], $\lambda = (1 - \omega_{t_o}) \lambda + \omega_{t_o} \widetilde{\lambda}(j)$, $\mathbf{v}_1 = (1 - \omega_{t_o}) \mathbf{v}_1 + \omega_{t_o} \widetilde{\mathbf{v}}_1(j)$, $\mathbf{v}_2 = (1 - \omega_{t_o}) \mathbf{v}_2 + \omega_{t_o} \widetilde{\mathbf{v}}_2(j)$, and $\mu = (1 - \omega_{t_o}) \mu + \omega_{t_o} \widetilde{\mu}(j)$, where $\widetilde{\lambda}(j)$, $\widetilde{v}_1(j)$,$\widetilde{\mathbf{v}}_2(j)$,$\widetilde{\mu}(j)$ are the estimates for $D\mathscr{L}_j$. The online variational inference and its implementation is available for download on the first author's homepage[1], however, only the batch variational inference was evaluated in this paper.

4.3. Classification

As shown in Equation (4), the class labels $y_j \in \{1, \ldots, C\}$ are estimated with a softmax function. Classification is done using a variation approximation of $\bar{\theta}_j$ (the corpus level topic distribution of document j): $\bar{\theta}_j = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \rho_{jt} \zeta_{jnt}$. Since the denominator of Equation (4) is constant with respect to class, only the numerator has to be regarded when computing the MAP estimate of y:

$$\hat{y}_j = argmax_{y_j \in \{1, \ldots, C\}} \mathbb{E}[\mu_{y_j}^T \bar{\theta}_j] \tag{14}$$

corresponding to the MAP class estimate in SLDA [14].

5. Experiments

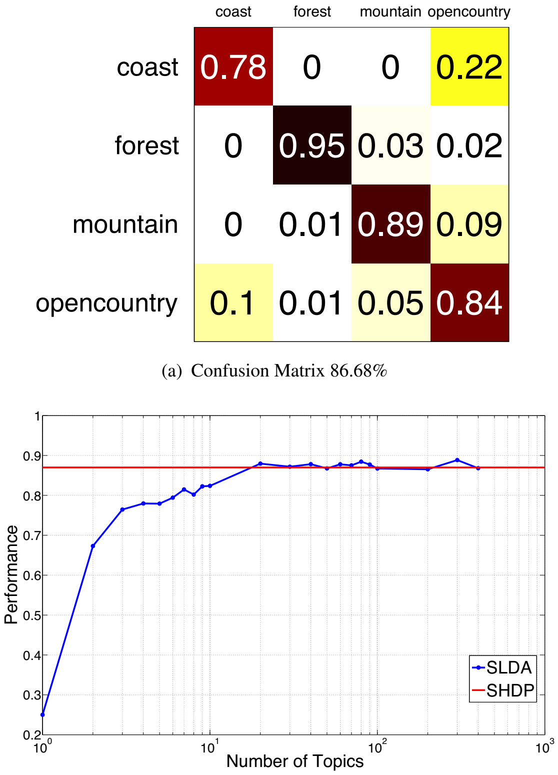
We tested our algorithm with two classification tasks: natural scene classification and action classification, which are typical computer vision tasks. As we will see, our experimental evaluation shows that our proposed SHDP is able to achieve the same level of performance as SLDA with an optimal number of topics. The open source code for SLDA from Wang[2] was used for our SLDA experiments. Batch variational inference is used in all the experiments below, since the considered datasets are sufficiently small. The source code for these experiments are available on the first author's homepage. Next, we describe the details of our experiments.

5.1. Natural Scene Classification

Our experiment on natural scene classification is performed with 4 common classes of natural scenes as in [5, 18]. Here, documents in the topic model are images. Words are features. We use a bag-of-visual-words presentation of SIFT [9] to adapt the discrete presentation of the

[1] http://www.csc.kth.se/ chengz/TopicModelCode.html

[2] http://www.cs.cmu.edu/ chongw/slda



	coast	forest	mountain	opencountry
coast	0.78	0	0	0.22
forest	0	0.95	0.03	0.02
mountain	0	0.01	0.89	0.09
opencountry	0.1	0.01	0.05	0.84

(a) Confusion Matrix 86.68%

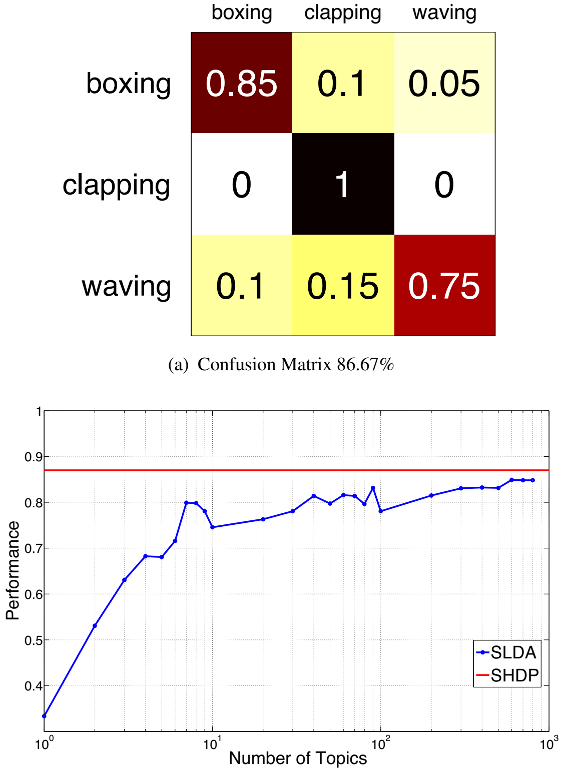
(b) Scene Classification performance

Figure 2. The top figure shows the confusion matrix for scene classification. The bottom figure displays a comparison of average performance over classes using SLDA and SHDP. The horizontal axis indicates the number of topics used for SLDA and is plotted on a log_{10} scale.

words. The class label for each document is the natural scene class on each image. There are more than 300 images for each class. For each class, a random selection of 80% the total images is used for training and 20% are used for testing. $\alpha_0 = 1$, $\gamma = 1$, $\eta = 0.5$ as in [15], $K = 150$, $T = 20$ are used in our experiment. A classification rate of 86.68% can be obtained by the proposed SHDP. Figure 2 (a) shows the confusion matrix. To compare, we ran SLDA with different numbers of topics with $\alpha = 0.1$. Figure 2 (b) shows the performance of SLDA with different numbers of topics and the SHDP performance. This result is consistent with the comparison of LDA and HDP shown in [13]. When the number of topics is too small, the result suffers from under-fitting. However, blindly increasing the number of topics could on the other hand make the computation cost extremely high without further improving the performance. SHDP is able to achieve the same level of performance as SLDA with an optimal number of topics in our experiments.

5.2. Action Classification

For action recognition, we used 3 actions from the KTH action dataset [8] as in [18]. The documents in this action classification task are video clips. Words are video features.



	boxing	clapping	waving
boxing	0.85	0.1	0.05
clapping	0	1	0
waving	0.1	0.15	0.75

(a) Confusion Matrix 86.67%

(b) Action Classification performance

Figure 3. The top figure shows the confusion matrix for action classification. The bottom figure displays the comparison of average performance over classes using SLDA and SHDP. The horizontal axis indicates the number of topics used for SLDA and is plotted on a log_{10} scale.

We use a bag-of-visual-words presentation of STIP features [8]. The class label for each document is the action label on each video clip. There are around 100 video clips for each class. A random selection of 80% of the video clips is used for training and the rest is used for testing. $\alpha_0 = 1$, Here, we set $\gamma = 1$, $\eta = 0.5$, $K = 80$, $T = 20$. An average of 86.67% test video clips can be correctly classified. Figure 3 (a) shows the confusion matrix and Figure 3 (b) displays the comparison of SLDA with $\alpha = 0.1$ and SHDP. The result is consistent with the unsupervised version of the models [13] and the previous natural scene classification result. SHDP is able to achieve better performance than the SLDA setting could achieve.

6. Discussion

In this paper, we presented Supervised HDP with variational inference and further extended it to an online setting. In our experiments, we show that SHDP can achieve the same level of performance as SLDA with an optimal setting for the number of topics. We have released our topic modeling code for public usage which includes variational LDA, variational SLDA, variational HDP, online HDP, variational SHDP and online SHDP.[3] Since online SHDP is a model which is intended to be used on large scale datasets, we intend to develop an incremental learning image classification system by fetching different classes of images, for example from Google image search. We furthermore intend to extend the model by *e.g.*, factorizing noise and useful information in the SHDP framework [18], and we would like to apply the model on multi-modal data for contextual modeling [19].

References

[1] D. M. Blei and J. D. McAuliffe. Supervised topic models, *arxiv:1003.0783*, 2010. 1, 2, 3

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 1, 2, 3

[3] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007. 1, 2

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, and G. W. Furnas. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, Sept. 1990. 1

[5] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 1, 2, 5

[6] M. D. Hoffman, D. M. Blei, and F. Bach. Online learning for Latent Dirichlet Allocation. In *NIPS*, 2010. 2, 3, 5

[7] S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008. 1

[8] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *ECCV*, 2004. 2, 6

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 5

[10] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *CVPR*, 2012. 1

[11] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Conference on Empirical Methods in Natural Language Processing*, 2009. 1

[12] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. 2

[13] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 1, 2, 3, 6

[14] C. Wang, D. M. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009. 1, 2, 3, 5, 8

[15] C. Wang, J. Paisley, and D. Blei. Online variational inference for the Hierarchical Dirichlet Process. In *AISTATS*, 2011. 2, 3, 4, 5, 6

[16] D. Weinshall, G. Levi, and D. Hanukaev. Latent Dirichlet Allocation topic model with soft assignment of descriptors to words. In *ICML*, 2013. 1

[17] B. Xie and R. J. Rassonneau. Supervised HDP using prior knowledge. In *International Conference on Applications of Natural Language Processing and Information Systems*, 2012. 2

[18] C. Zhang, C. H. Ek, A. Damianou, and H. Kjellström. Factorized Topic Models. In *International Conference on Learning Representations*, 2013. 1, 2, 5, 6, 7

[19] C. Zhang, D. Song, and H. Kjellström. Contextual modeling with Labeled Multi-LDA. In *IROS*, 2013. 1, 7

[20] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: Maximum Margin Supervised Topic Models for regression and classification. In *ICML*, 2009. 1

[3] Code available at: http://www.csc.kth.se/ chengz/TopicModelCode.html

A. Appendix

A.1. The expectation of the label part is:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{q}[\log p(y_j|\mathbf{z},\mathbf{c},\mu)] \\
&= \mathop{\mathbb{E}}_{q}\left[\log\left(\frac{\exp\left(\mu_{y_j}^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}})\right)}{\sum_{l=1}^C \exp(\mu_l^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}}))}\right)\right] \\
&= \mathop{\mathbb{E}}_{q}[\mu_{y_j}^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}})] - \mathop{\mathbb{E}}_{q}\left[\log\left(\sum_{l=1}^C \exp(\mu_l^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}}))\right)\right] \\
&= \mu_{y_j}^T(\frac{1}{N}\sum_{n=1}^N \mathop{\mathbb{E}}_{q}[\mathfrak{c}_{jz_{jn}}]) - \mathop{\mathbb{E}}_{q}\left[\log\left(\sum_{l=1}^C \exp(\mu_l^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}}))\right)\right] \\
&= \mu_{y_j}^T(\frac{1}{N}\sum_{n=1}^N \mathop{\mathbb{E}}_{q}[\sum_{t=1}^T \mathfrak{c}_{jt}[z_{jn}=t]]) - \mathop{\mathbb{E}}_{q}\left[\log\left(\sum_{l=1}^C \exp(\mu_l^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}}))\right)\right] \\
&= \mu_{y_j}^T(\frac{1}{N}\sum_{n=1}^N \sum_{t=1}^T \rho_{jt}\zeta_{jnt}) - \mathop{\mathbb{E}}_{q}\left[\log\left(\sum_{l=1}^C \exp(\mu_l^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}}))\right)\right].
\end{aligned}
\tag{15}
$$

The second term is:

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{q}\left[\log\left(\sum_{l=1}^C \exp(\mu_l^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}}))\right)\right] \\
&\geq -\log\left(\mathop{\mathbb{E}}_{q}\left[\sum_{l=1}^C \exp(\mu_l^T(\frac{1}{N}\sum_{n=1}^N \mathfrak{c}_{jz_{jn}}))\right]\right) \\
&= -\log\left(\mathop{\mathbb{E}}_{q}\left[\sum_{l=1}^C \prod_{n=1}^N \exp(\mu_l^T(\frac{1}{N}\mathfrak{c}_{jz_{jn}}))\right]\right)
\end{aligned}
$$

Since $\mathfrak{c}_{jz_{jn}}$ is a vector which a single non-zero element equal to 1,

$$\mu_l^T\mathfrak{c}_{jz_{jn}} = \prod_{e=1}^K \mu_{le}^{[c_{jz_{jn}}=e]}.$$

$$
\begin{aligned}
&= -\log\left(\sum_{l=1}^C \prod_{n=1}^N \mathop{\mathbb{E}}_{q}\left[\exp(\frac{1}{N}\prod_{e=1}^K \mu_{le}^{[c_{jz_{jn}}=e]})\right]\right) \\
&= -\log\left(\sum_{l=1}^C \prod_{n=1}^N \mathop{\mathbb{E}}_{q}\left[\left(\sum_{e=1}^K [c_{jz_{jn}}=e]\exp(\frac{1}{N}\mu_{le})\right)\right]\right)
\end{aligned}
$$

Since $[c_{jz_{jn}}=e] = \sum_{i=1}^T [c_{ji}=e][z_{jn}=i]$,

$$
\begin{aligned}
&= -\log\left(\sum_{l=1}^C \prod_{n=1}^N \mathop{\mathbb{E}}_{q}\left[\left(\sum_{e=1}^K \sum_{i=1}^T [c_{ji}=e][z_{jn}=i]\exp(\frac{1}{N}\mu_{le})\right)\right]\right) \\
&= -\log\left(\sum_{l=1}^C \prod_{n=1}^N \left(\sum_{e=1}^K \sum_{i=1}^T \rho_{jie}\zeta_{jni}\exp(\frac{1}{N}\mu_{le})\right)\right).
\end{aligned}
\tag{16}
$$

Now plugging the result of 16 back into 15, we obtain:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{q}[\log p(y_j|\mathbf{z},\mathbf{c},\mu)] \geq{}& \mu_{y_j}^T(\frac{1}{N}\sum_{n=1}^N \sum_{t=1}^T \rho_{jt}\zeta_{jnt}) \\
&- \log\left(\sum_{l=1}^C \prod_{n=1}^N \left(\sum_{e=1}^K \sum_{i=1}^T \rho_{jie}\zeta_{jni}\exp(\frac{1}{N}\mu_{le})\right)\right).
\end{aligned}
\tag{17}
$$

A.2. Update of ρ

The partial derivative is given below:
Let $\mathfrak{F} = \sum_{l=1}^{C} \prod_{n=1}^{N} \Big(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N}\mu_{le}) \Big)$

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \rho_{jtk}} = & \sum_{n=1}^{N} \zeta_{jnt} \mathop{\mathbb{E}}_{q}[\log p(w_{jn}|\phi_k)] + \mathop{\mathbb{E}}_{q}[\log \beta_k] - 1 - \log \rho_{jtk} \\
& + \mu_{y_j k} (\frac{1}{N} \sum_{n=1}^{N} \zeta_{jnt}) \\
& - \mathfrak{F}^{-1} \Bigg(\sum_{l=1}^{C} \Big(\Big(\prod_{m=1}^{N} (\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jmi} \exp(\frac{1}{N}\mu_{le})) \Big) \\
& \cdot \sum_{n=1}^{N} \big(\frac{\zeta_{jnt} \exp(\frac{1}{N}\mu_{lk})}{\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N}\mu_{le})} \big) \Big) \Bigg).
\end{aligned} \tag{18}$$

A.3. Update of ζ

$$\begin{aligned}
& \mathscr{L}_{\zeta} \\
& = \sum_{j=1}^{M} \sum_{n=1}^{N} \sum_{t=1}^{T} \Bigg(\zeta_{jnt} \Big(\sum_{k=1}^{K} \rho_{jtk} \sum_{i=1}^{V} (\Psi(\lambda_{ki}) - \Psi(\sum_{p} \lambda_{kp}))[w_{jn} = i] \Big) \\
& + \zeta_{jnt} \Big((\Psi(a_{1jt}) - \Psi(a_{1jt} + a_{2jt})) + \sum_{\mathfrak{t}=1}^{t-1} (\Psi(a_{2j\mathfrak{t}}) - \Psi(a_{1j\mathfrak{t}} + a_{2j\mathfrak{t}})) \Big) \\
& - \zeta_{jnt} \log \zeta_{jnt} \Bigg) + \sum_{j=1}^{M} \Bigg(\mu_{y_j}^{T} (\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \rho_{jt} \zeta_{jnt}) \\
& - \log \Big(\sum_{l=1}^{C} \prod_{n=1}^{N} \big(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N}\mu_{le}) \big) \Big) \Bigg),
\end{aligned} \tag{19}$$

where Ψ is the digamma function. For $\mathfrak{i} \in \{1, \ldots, T\}$, we write:

$$h_{\mathfrak{i}} = \sum_{l=1}^{C} \Big(\prod_{\substack{n=1 \\ n \neq n_{now}}}^{N} \big(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N}\mu_{le}) \big) \cdot \big(\sum_{e=1}^{K} \rho_{j\mathfrak{i}e} \exp(\frac{1}{N}\mu_{le}) \big) \Big),$$

which yields:

$$\sum_{l=1}^{C} \prod_{n=1}^{N} \big(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N}\mu_{le}) \big) = \sum_{\mathfrak{i}=1}^{T} h_{\mathfrak{i}} \cdot \zeta_{jn_{now}\mathfrak{i}}.$$

$\mathscr{L}_{\zeta_{jn}}$ can now be rewritten as:

$$\begin{aligned}
\mathscr{L}_{\zeta_{jn}} = & \sum_{t=1}^{T} \Bigg(\zeta_{jnt} \Big(\sum_{k=1}^{K} \rho_{jtk} \sum_{i=1}^{V} (\Psi(\lambda_{ki}) - \Psi(\sum_{p} \lambda_{kp}))[w_{jn} = i] \Big) \\
& + \zeta_{jnt} \Big((\Psi(a_{1jt}) - \Psi(a_{1jt} + a_{2jt})) + \sum_{\mathfrak{t}=1}^{t-1} (\Psi(a_{2j\mathfrak{t}}) - \Psi(a_{1j\mathfrak{t}} + a\ \ _{j\mathfrak{t}})_{2} \Big) \\
& - \zeta_{jnt} \log \zeta_{jnt} \Bigg) + \Bigg(\mu_{y_j}^{T} (\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \rho_{jt} \zeta_{jnt}) - \log(h^{T} \zeta_{jn}) \Bigg).
\end{aligned} \tag{20}$$

We follow the approach of [14] to derive the fixed point update. Suppose we have a previous value ζ_{jn}^{old}. Consider the inequality $\log(x) \leq \mathfrak{x}^{-1} x + log(\mathfrak{x}) - 1$, where equality holds if and only if $x = \mathfrak{x}$. Thus, set $x = h^{T} \zeta_{jn}$ and $\mathfrak{x} = h^{T} \zeta_{jn}^{old}$. The new bound becomes:

$$\begin{aligned}
\mathscr{L}_{\zeta} \geq & \sum_{j=1}^{M} \sum_{n=1}^{N} \sum_{t=1}^{T} \Bigg(\zeta_{jnt} \Big(\sum_{k=1}^{K} \rho_{jtk} \sum_{i=1}^{V} (\Psi(\lambda_{ki}) - \Psi(\sum_{p} \lambda_{kp}))[w_{jn} = i] \Big) \\
& + \zeta_{jnt} \Big((\Psi(a_{1jt}) - \Psi(a_{1jt} + a_{2jt})) \\
& + \sum_{\mathfrak{t}=1}^{t-1} (\Psi(a_{2j\mathfrak{t}}) - \Psi(a_{1j\mathfrak{t}} + a_{2j\mathfrak{t}})) \Big) - \zeta_{jnt} \log \zeta_{jnt} \Bigg) \\
& + \sum_{j=1}^{M} \Bigg(\mu_{y_j}^{T} (\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \rho_{jt} \zeta_{jnt}) - (h^{T} \zeta_{jn}^{old})^{-1} h^{T} \zeta_{jn} \\
& - \log(h^{T} \zeta_{jn}^{old}) + 1 \Bigg) = \mathscr{L}'_{\zeta}.
\end{aligned} \tag{21}$$

We compute the derivative for the new bound:

$$\begin{aligned}
\frac{\partial \mathscr{L}'}{\partial \zeta_{jnt}} = & \sum_{k=1}^{K} \rho_{jtk} \sum_{i=1}^{V} (\Psi(\lambda_{ki}) - \Psi(\sum_{p} \lambda_{kp}))[w_{jn} = i] \\
& + (\Psi(a_{1jt}) - \Psi(a_{1jt} + a_{2jt})) + \sum_{\mathfrak{t}=1}^{t-1} (\Psi(a_{2j\mathfrak{t}}) - \Psi(a_{1j\mathfrak{t}} + a_{2j\mathfrak{t}})) \\
& - 1 - \log \zeta_{jnt} + \frac{1}{N} \mu_{y_j}^{T} \rho_{jt} - (h^{T} \zeta_{jn}^{old})^{-1} h_t.
\end{aligned} \tag{22}$$

Finally, we set the derivative to zero to get the fixed point update: [4]

$$\begin{aligned}
\zeta_{jnt} \propto & \exp \Big(\sum_{k=1}^{K} \rho_{jtk} \sum_{i=1}^{V} (\Psi(\lambda_{ki}) - \Psi(\sum_{p} \lambda_{kp}))[w_{jn} = i] \\
& + (\Psi(a_{1jt}) - \Psi(a_{1jt} + a_{2jt})) + \sum_{\mathfrak{t}=1}^{t-1} (\Psi(a_{2j\mathfrak{t}}) - \Psi(a_{1j\mathfrak{t}} + a_{2j\mathfrak{t}})) \\
& - 1 + \frac{1}{N} \mu_{y_j}^{T} \rho_{jt} - (h^{T} \zeta_{jn}^{old})^{-1} h_t \Big) \\
& \propto \exp \Big(\sum_{k=1}^{K} \rho_{jtk} \mathop{\mathbb{E}}_{q}[\log p(w_{jn}|\phi_k)] + \mathop{\mathbb{E}}_{q}[\log \pi_{jt}] \\
& + \frac{1}{N} \mu_{y_j}^{T} \rho_{jt} - (h^{T} \zeta_{jn}^{old})^{-1} h_t \Big).
\end{aligned} \tag{23}$$

A.4. Update of μ

Let $\mathfrak{F} = \sum_{l=1}^{C} \prod_{n=1}^{N} \Big(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N}\mu_{le}) \Big)$.

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \mu_{yk}} = & \sum_{j=1}^{M} \Bigg([y_j = y] \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \rho_{jtk} \zeta_{jnt} \\
& - \mathfrak{F}^{-1} \prod_{m=1}^{N} \Big(\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jmi} \exp(\frac{1}{N}\mu_{ye}) \Big) \\
& \cdot \sum_{n=1}^{N} \Big(\frac{(\sum_{i=1}^{T} \rho_{jik} \zeta_{jni}) \cdot \frac{1}{N} \cdot \exp(\frac{1}{N}\mu_{yk})}{\sum_{e=1}^{K} \sum_{i=1}^{T} \rho_{jie} \zeta_{jni} \exp(\frac{1}{N}\mu_{ye})} \Big) \Bigg).
\end{aligned} \tag{24}$$

[4] To incorporate the constraint that $\sum_{t=1}^{T} \zeta_{jnt} = 1$, $\propto$ is used here instead of $=$, since the normalizing factor is dropped in the above result.