DEEP UNFOLDING INFERENCE FOR SUPERVISED TOPIC MODEL

Chao-Hsi Lee and Jen-Tzung Chien

Department of Electrical and Computer Engineering
National Chiao Tung University, Taiwan, ROC

ABSTRACT

Conventional supervised topic model for multi-class classification is inferred via the variational inference algorithm where the model parameters are estimated by maximizing the lower bound of the logarithm of marginal likelihood function over input documents and labels. The classification accuracy is constrained by the variational lower bound. In this study, we aim to improve the classification accuracy by relaxing this constraint through directly maximizing the negative cross entropy error function via a deep unfolding inference (DUI). The inference procedure for class posterior is treated as the layer-wise learning in a deep neural network. The classification accuracy in DUI is accordingly increased by using the estimated topic parameters according to the exponentiated updates. Deep learning of supervised topic model is achieved through an error back-propagation algorithm. Experimental results show the superiority of DUI to variational Bayes inference in supervised topic model.

Index Terms— Deep unfolding, variational inference, supervised topic model

1. INTRODUCTION

Probabilistic topic model has been successfully developed for document categorization [1], image annotation [2], text segmentation [3], speech recognition [4, 5], document summarization [6], information retrieval [7], speech separation [8] and many other applications. Using topic model, the latent semantic topics are learned from a bag of words to capture the salient aspects embedded in data collection. The topic proportions is further used as the feature for classification [1]. To improve the classification accuracy, several methods [9, 10, 11, 12] have been studied to develop the supervised topic models to jointly capture the relationship of the documents and class labels.

Traditionally, the those topic models were inferred by the variational Bayes or the Gibbs sampling which are seen as the approximate inference procedures with different degrees of smoothing. It is because that the exact inference does not exist due to the coupling of multiple latent variables including topic assignments and topic proportions. Instead of direct maximization of marginal likelihood, the variational inference is developed by indirectly maximizing the lower bound of marginal likelihood through the variational expectation-maximization (EM) algorithm [13, 14].

However, the end performance is constrained by this lower bound. In general, the approximate inference may lead to errors at both estimation and prediction phases due to two reasons [15]. First, the expression of an underlying model may be deteriorated. Second, the approximation with parameter change may mislead the standard learning algorithm. In [16], the minimization of empirical risk was used instead of maximization of the approximate likelihood of training data. In [17], the errors due to inconsistent parameter estimator were shown to be partially compensated by using an approximate prediction method. In [18, 19, 20, 12], the discriminative training was demonstrated to outperform standard generative training in prediction tasks.

Recently, deep neural networks (DNNs) have been recognized as one of the most successful machine learning approaches to speech-related applications. This study aims to incorporate the power of DNN architecture into the inference of topic model and maximize the negative cross entropy error function directly. A deep unfolding inference is then developed accordingly. We derive the updating rules for parameter estimation which are implemented in a style of error back-propagation [16] similar to that in DNNs. The DNN-inferred topic model is then established. In the experiments, the classification performance is increased by using this deep unfolding inference compared to that using standard variational inference.

2. SUPERVISED TOPIC MODEL

This study investigates the DNN-inferred supervised topic model based on the supervised latent Dirichlet allocation (sLDA) [21] for multi-class classification (sLDAc) [10]. We first survey the model construction and inference based on sLDAc with graphical representation in Figure 1.

2.1. Model construction

sLDAc [10] is the extension of sLDA [21] which characterizes both documents and their labels. Given a set of Documents $\mathbf{w} = \{\mathbf{w}_d\} = \{w_{dn}\}$, the topic $z_{dn} = k$ of a

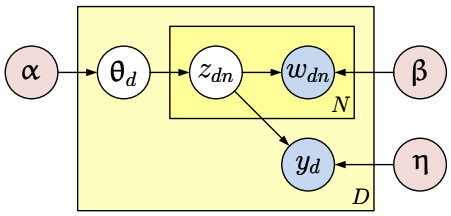


Fig. 1: Graphical representation for supervised topic model

word $w_{dn} = v$ is drawn by a multinomial distribution over V vocabulary words with parameters $\boldsymbol{\beta} = \{\boldsymbol{\beta}_k\} = \{\beta_{kv}\}$. sLDAc introduces a Dirichlet prior with hyper-parameters $\boldsymbol{\alpha}$ for document-dependent topic $\boldsymbol{\theta}_d = \{\theta_{dk}\} \sim \text{Dir}(\boldsymbol{\alpha})$ over K topics with hyper-parameters $\boldsymbol{\alpha} = \{\alpha_k\}$ where $\alpha_k > 0$. Each document is treated as a "random mixture" over different topics. Each word w_{dn} and its topic assignment z_{dn} in a document d are multinomial variables expressed by $w_{dn} = v \sim \text{Mult}(\boldsymbol{\beta}_{z_{dn}})$ and $z_{dn} = k \sim \text{Mult}(\boldsymbol{\theta}_d)$, respectively. Accordingly, the latent variables in sLDAc consist of topic proportions $\boldsymbol{\theta} = \{\theta_{dk}\}$ and topic assignments $\mathbf{z} = \{z_{dn}\}$. The class label outputs $\mathbf{y} = \{y_d\}$ with $y_d = m \sim \text{softmax}(\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)$ have the distributions

$$p(y_d = m \mid \mathbf{z}_d, \boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\eta}_m^\top \bar{\mathbf{z}}_d)}{\sum_{i=1}^{M} \exp(\boldsymbol{\eta}_i^\top \bar{\mathbf{z}}_d)} \tag{1}$$

where $\bar{\mathbf{z}}_d = \{\bar{z}_{dk}\}$ denotes the empirical topic frequency, which satisfies $\bar{z}_{dk} = (1/N_d) \sum_{n=1}^{N_d} \delta(z_{dn}, k)$, $\boldsymbol{\eta} = \{\boldsymbol{\eta}_{mk}\}$ denotes the class label coefficients, and M is the number of class labels.

The model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}\}$ are estimated by maximizing the marginal likelihood of $\{\mathbf{w}, \mathbf{y}\}$

$$\begin{aligned} p(\mathbf{w}, \mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}) = \prod_{d=1}^{D} p(y_d | \mathbf{z}_d, \boldsymbol{\eta}) \int_{\boldsymbol{\theta}_d} p(\boldsymbol{\theta}_d | \boldsymbol{\alpha}) \\ \times \prod_{n=1}^{N_d} \sum_{k=1}^{K} p(z_{dn} = k | \boldsymbol{\theta}_d) p(w_{dn} | z_{dn} = k, \boldsymbol{\beta}) \, d\boldsymbol{\theta}_d \end{aligned} \tag{2}$$

where N_d denotes the number of words in document d. However, the exact solution to model inference based on Eq. (2) does not exist due to the coupling of multiple latent variables $\boldsymbol{\theta}$ and $\mathbf{z}$ in posterior distribution $p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})$.

2.2. Variational Bayes inference

Variational Bayes (VB) is implemented to resolve the intractable posterior distribution $p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})$ [10] by using a factorizable variational distribution

$$q(\boldsymbol{\theta}, \mathbf{z} | \boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{d=1}^{D} q(\boldsymbol{\theta}_d | \boldsymbol{\gamma}_d) \prod_{n=1}^{N_d} q(z_{dn} = k | \phi_{dnk}) \tag{3}$$

through maximizing a lower bound of the logarithm of marginal likelihood $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})$ where $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ denote the variational Dirichlet and multinomial parameters, respectively. We have the relation

$$\begin{aligned} \ln p(\mathbf{w}, \mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}) = \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}) \\ + \text{KL}(q(\boldsymbol{\theta}, \mathbf{z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) \| p(\boldsymbol{\theta}, \mathbf{z} | \mathbf{w}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})). \end{aligned} \tag{4}$$

Therefore, maximizing lower bound $\mathcal{L}$ with respect to $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ is equivalent to estimating the new variational distribution q which is closest to the true posterior p with the smallest Kullback-Leibler divergence $\text{KL}(\cdot \| \cdot)$. The variational EM is solved by [10]. When the label is unknown, lower bound $\mathcal{L}$ with variational parameters $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ degenerates into the expectation step of LDA. The class label output is estimated by

$$\widehat{\mathbf{y}}_d = \mathbb{E}_q[y_d = m | \boldsymbol{\phi}_d, \mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}] = \frac{\boldsymbol{\eta}_m^\top \overline{\boldsymbol{\phi}}_d}{\sum_{i=1}^{M} \boldsymbol{\eta}_i^\top \overline{\boldsymbol{\phi}}_d} \tag{5}$$

where $\overline{\boldsymbol{\phi}}_d = (1/N_d) \sum_{n=1}^{N_d} \boldsymbol{\phi}_{dn}$ denotes the variational empirical topic frequency. We summarize the label prediction rule for sLDAc in Algorithm 1.

Algorithm 1 Prediction rule for supervised topic model

repeat
$\gamma_{dk} = \alpha_k + \sum_v N_{dv} \phi_{dvk}$
$\phi_{dvk} \propto \exp\{\mathbb{E}[\ln \beta_{kv}] + \mathbb{E}[\ln \theta_{dk}]\}$
until γ_{dk} and ϕ_{dvk} converged
return $\widehat{\mathbf{y}}_d = \text{softmax}(\boldsymbol{\eta}^\top \overline{\boldsymbol{\phi}}_d)$

3. DEEP UNFOLDING INFERENCE

We first address a general solution to deep unfolding inference (DUI) and then present how this new algorithm is realized to develop a deep inference for sLDAc.

3.1. General framework

Consider a topic model with parameters $\boldsymbol{\Theta}$. For each instance n, this model specifies an interesting output y_n for an observation x_n. In the inference procedure, we first estimate the data dependent parameters $\boldsymbol{\Psi}_n$ that maximize the objective function $\mathcal{F}$ with respect to $\boldsymbol{\Theta}$. Once the optimal parameters $\widehat{\boldsymbol{\Psi}}_n$ are estimated, we compute the interesting output $\widehat{y}_n$ through an estimator $\mathcal{G}_{\boldsymbol{\Theta}}$ as expressed by [20]

$$\begin{aligned} \widehat{\boldsymbol{\Psi}}(x_n | \boldsymbol{\Theta}) &= \arg\max_{\boldsymbol{\Psi}_n} \mathcal{F}_{\boldsymbol{\Theta}}(x_n, \boldsymbol{\Psi}_n) \\ \widehat{y}(x_n | \boldsymbol{\Theta}) &= \mathcal{G}_{\boldsymbol{\Theta}}(x_n, \widehat{\boldsymbol{\Psi}}(x_n | \boldsymbol{\Theta})). \end{aligned} \tag{6}$$

Typically, the optimal parameters $\widehat{\boldsymbol{\Psi}}_n$ are obtained by running an iterative updating algorithm

$$\boldsymbol{\Psi}^{(l)} = f_{\boldsymbol{\Theta}}(x_n, \boldsymbol{\Psi}^{(l-1)}). \tag{7}$$

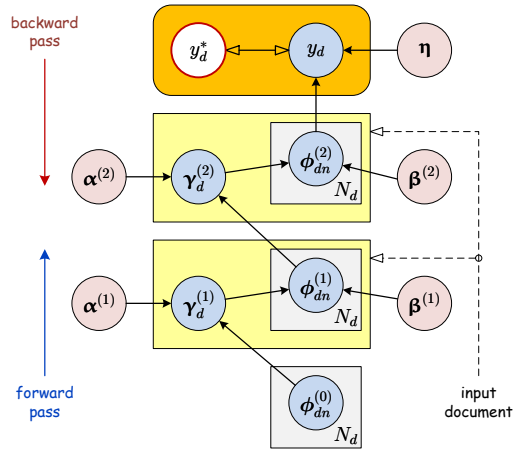


Fig. 2: Deep unfolding inference for supervised topic model

Here, the iteration index l can be treated as the layer in deep unfolding. The performance of a model can be measured by the evaluator $\mathcal{J}_{\boldsymbol{\Theta}}$ with the optimal parameters $\boldsymbol{\Theta}$ for each $\widehat{y}_n$

$$\max_{\boldsymbol{\Theta}} \ \mathcal{J}_{\boldsymbol{\Theta}}(\{\widehat{y}_n, \forall n\}). \tag{8}$$

Accordingly, maximizing the performance of a model is seen as a bi-level optimization problem. Maximizing $\mathcal{J}$ in Eq. (8) with respect to $\boldsymbol{\Theta}$ involves in maximizing $\mathcal{F}$ with respect to parameter $\boldsymbol{\Psi}_n$, and vice versa. Deep unfolding inference provides a way to this general framework where an error back-propagation algorithm is implemented to calculate the differentiations

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\Theta}^{(L)}} = \sum_n \frac{\partial \mathcal{J}}{\partial \widehat{y}_n} \frac{\partial \widehat{y}_n}{\partial \boldsymbol{\Theta}^{(L)}} \qquad \frac{\partial \mathcal{J}}{\partial \boldsymbol{\Psi}_n^{(L)}} = \frac{\partial \mathcal{J}}{\partial \widehat{y}_n} \frac{\partial \widehat{y}_n}{\partial \boldsymbol{\Psi}_n^{(L)}}$$
$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{\Theta}^{(l)}} = \sum_n \frac{\partial \mathcal{J}}{\partial \boldsymbol{\Psi}_n^{(l+1)}} \frac{\partial \boldsymbol{\Psi}_n^{(l+1)}}{\partial \boldsymbol{\Theta}^{(l)}} \qquad \frac{\partial \mathcal{J}}{\partial \boldsymbol{\Psi}_n^{(l)}} = \frac{\partial \mathcal{J}}{\partial \boldsymbol{\Psi}_n^{(l+1)}} \frac{\partial \boldsymbol{\Psi}_n^{(l+1)}}{\partial \boldsymbol{\Psi}_n^{(l)}}. \tag{9}$$

3.2. DUI for supervised topic model

In this study, we would like to carry out deep unfolding inference for supervised latent Dirichlet allocation on classification task. Optimization problem is formulated as

$$\{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}\} = \arg \max_{\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}\}} \ \mathcal{E}(\mathbf{y}^*, \text{softmax}(\boldsymbol{\eta}^\top \overline{\boldsymbol{\phi}})) \tag{10}$$

$$\text{where } \{\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\phi}}\} = \arg \max_{\{\boldsymbol{\phi}, \boldsymbol{\gamma}\}} \ \mathcal{L}(\boldsymbol{\gamma}(\boldsymbol{\alpha}, \boldsymbol{\phi}), \boldsymbol{\phi}(\boldsymbol{\beta}, \boldsymbol{\gamma})) \tag{11}$$

where $\mathbf{y}^* = \{y_d^*\}$ denotes the truth labels and $\mathcal{E}(\cdot)$ is the negative cross entropy error function

$$\mathcal{E}(\mathbf{y}^*, \widehat{\mathbf{y}}) = \sum_{d=1}^{D} \sum_{m=1}^{M} \delta(y_d^*, m) \ln \widehat{y}_{dm}. \tag{12}$$

To maximize $\mathcal{E}$ directly, we resort to DUI method. The model parameters $\boldsymbol{\Theta}$ consist of $\mathbf{A} = \{\boldsymbol{\alpha}^{(l)}\}$, $\mathbf{B} = \{\boldsymbol{\beta}^{(l)}\}$ and $\boldsymbol{\eta}$ while the variational parameters $\boldsymbol{\Psi}$ are composed of $\{\boldsymbol{\gamma}^{(l)}, \boldsymbol{\phi}^{(l)}\}$ with L layers. $\mathcal{F} \triangleq \mathcal{L}$ and $\mathcal{J} \triangleq \mathcal{E}$. These parameters are estimated through unfolding the maximum step in Eq. (10) via

$$\{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\boldsymbol{\eta}}\} = \arg \max_{\{\mathbf{A}, \mathbf{B}, \boldsymbol{\eta}\}} \ \mathcal{E}(\mathbf{y}^*, \text{softmax}(\boldsymbol{\eta}^\top \overline{\boldsymbol{\phi}}(\boldsymbol{\phi}^{(L-1)}))) \tag{13}$$

with L layers

$$\boldsymbol{\phi}_d^{(l)}(\boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}_d^{(l)}(\boldsymbol{\alpha}^{(l)}, \boldsymbol{\phi}_d^{(l-1)}(\dots \boldsymbol{\gamma}_d^{(1)}(\boldsymbol{\alpha}^{(1)}, \boldsymbol{\phi}_d^{(0)}))) \dots) \tag{14}$$

where $\phi_{dnk}^{(0)} = 1/K$. The nested structure is illustrated as shown in Figure 2, which can be realized by a procedure similar to Algorithm 1.

3.3. Error back-propagation procedure

The error back-propagation can be expressed in forward and backward pass. In forward pass, we calculate the variational parameters $\boldsymbol{\Psi}^{(l)}$ for different layers l based on the model parameters $\boldsymbol{\Theta}^{(l)}$ estimated in previous learning epochs. In backward pass, the model parameters are updated layer by layer by directly maximizing the negative cross entropy error function in Eq. (13) through a back-propagation algorithm. Similar to the error back-propagation in standard DNN training.

To deal with back-propagation of variational parameters, we define a set of auxiliary variables $\{a^{(l)}\}$ and re-express the forward pass as

$$\phi_{dvk}^{(l)} = \exp(a_{dvk}^{(l)}) \ / \ \textstyle\sum_j \exp(a_{dvj}^{(l)}) \tag{15}$$

where $l = L-1, ..., 1$ and

$$a_{dvk}^{(L)} = \ln \beta_{vk}^{(L)} + \ln \gamma_{dk}^{(L)} - \ln \textstyle\sum_j \gamma_{dj}^{(L)} \tag{16}$$

$$a_{dvk}^{(l)} = \ln \beta_{vk}^{(l)} + \psi(\gamma_{dk}^{(l)}). \tag{17}$$

The class label output is then estimated by

$$\bar{\phi}_{dk} = (1/N_d) \textstyle\sum_v N_{dv} \phi_{dvk}^{(L-1)} \tag{18}$$

$$a_{dm}^{(L)} = \textstyle\sum_k \eta_{mk} \bar{\phi}_{dk} \tag{19}$$

$$\widehat{y}_d = \exp(a_{dm}) \ / \ \textstyle\sum_{i=1}^{M} \exp(a_{di}^{(L)}). \tag{20}$$

Next, the backward pass can be derived by applying chain rule and find the layer-wise differentiation of negative cross entropy. We summarize the procedure of back-propagation and updating rule in Algorithm 2, where $\psi'(\cdot)$ denotes the trigamma function. Since model parameters $\{\mathbf{A}, \mathbf{B}\}$ are non-negative, we adopt the *exponentiated gradient* method [22] to solve the constrained optimization problem subject to the non-negativity constraints. As a result, we implement the DNN-styled inference for topic model which is seen as an incorporation of the deterministic DNN into the inference of model-based method [20]. The powers of deep learning using DNN and stochastic learning using model-based method are assured in the proposed DUI.

Algorithm 2 Error back-propagation for DUI

/Back-propagation/
for all document d **do**
$\frac{\partial \mathcal{E}}{\partial a_{dm}} = \widehat{y}_{dm} - y^*_{dm}$
$\frac{\partial \mathcal{E}}{\partial \bar{\phi}_{dk}} = \sum_m \frac{\partial \mathcal{E}}{\partial a^{(L)}_{dm}} \eta_{mk}$
$\frac{\partial \mathcal{E}}{\partial \phi^{(L-1)}_{dvk}} = \frac{N_{dv}}{N_d} \left(\frac{\partial \mathcal{E}}{\partial \bar{\phi}_{dk}} - \frac{N_{dv}}{N_d} \sum_j \phi^{(L-1)}_{dvj} \frac{\partial \mathcal{E}}{\partial \bar{\phi}_{dj}} \right)$
for $l = L-1$ **to** 1 **do**
$\frac{\partial \mathcal{E}}{\partial a^{(l)}_{dvk}} = \phi^{(l)}_{dvk} \frac{\partial \mathcal{E}}{\partial \phi^{(l)}_{dvk}} - \phi^{(l)}_{dvk} \sum_j \phi^{(l)}_{dvj} \frac{\partial \mathcal{E}}{\partial \phi^{(l)}_{dvj}}$
$\frac{\partial \mathcal{E}}{\partial \phi^{(l-1)}_{dvk}} = \frac{\partial \mathcal{E}}{\partial a^{(l)}_{dvk}} N_{dv}\, \psi'(\gamma^{(l)}_{dk})$
end for
end for
/Updating rule/
Update $\boldsymbol{\eta}$ with $\frac{\partial \mathcal{E}}{\partial \eta_{km}} = \frac{\partial \mathcal{E}}{\partial a^{(L)}_{dm}} \bar{\phi}_{dk}$
for $l = 1$ **to** L **do**
Update $\boldsymbol{\beta}^{(l)}$ with $\frac{\partial \mathcal{E}}{\partial \beta^{(l)}_{vk}} = \sum_d \frac{\partial \mathcal{E}}{\partial a^{(l)}_{dvk}} \frac{1}{\beta^{(l)}_{vk}}$
Update $\boldsymbol{\alpha}^{(l)}$ with $\frac{\partial \mathcal{E}}{\partial \alpha^{(l)}_{k}} = \sum_{dv} \frac{\partial \mathcal{E}}{\partial \phi^{(l-1)}_{dvk}}$
end for

4. EXPERIMENTS

This section investigates the proposed DUI on document classification and compares the result with VB.

4.1. Experimental setup

For document classification, we evaluated the proposed method by using 20 newsgroups data set. It contains approximately 20,000 newsgroup documents, partitioned across 20 different newsgroups. In this task, we removed a list of English stop words provided by SMART information retrieval system, and the words that document frequencies were less than 5. Stemming was processed to reduce the inflected words into their word stem. Finally, we kept 5000 most frequent words in dictionary for document modeling. The original training set was randomly divided into a smaller training set and test set with 9000 documents and 6000 documents, respectively.

4.2. Experimental result

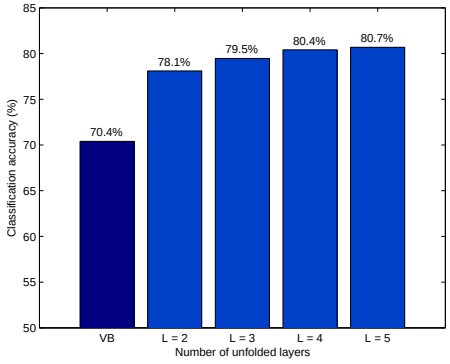
We investigated how the number of unfolded layers affects the accuracy of classification. We train a DUI model with $K = 60$, $\alpha = 0.1$, and vary $L = \{2, 3, 4, 5\}$. Mini-batch size is set to 1500. The model parameters $\mathbf{A}$ and $\mathbf{B}$ are random initialized and tied for all layers. $\boldsymbol{\eta}$ is initialized with a block diagonal matrix and trained with stochastic gradient descent

$$\boldsymbol{\eta} = \{\eta_{mk}\} = \begin{bmatrix} \mathbf{1} & & \\ & \ddots & \\ & & \mathbf{1} \end{bmatrix} \tag{21}$$

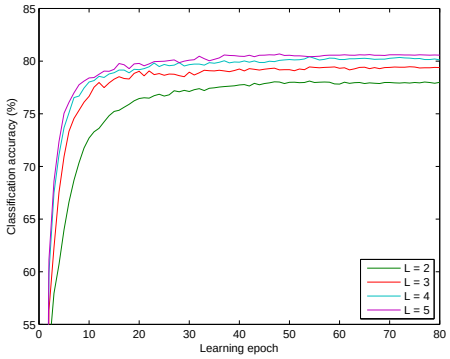
where $\mathbf{1} = (1, \ldots, 1)$ in M/K dimensions, non-diagonal blocks are set to zero. Learning rate is scheduled as

$$\frac{\exp(\text{no. of epochs} - 1)}{200}. \tag{22}$$

Figure 3(a) shows the classification accuracy versus the number of layers using DUI. The result shows that DUI can achieve better classification accuracy around 80.7%, outperforming the result of VB, 70.4%. This implies that DUI can improve the accuracy significantly by this DNN training style. Figure 3(b) shows the convergence of classification accuracy. The deeper the DUI network, the better the accuracy and the convergence rate we can achieve.



(a) Classification accuracy vs. number of unfolded layers



(b) Convergence of classification accuracy

Fig. 3: Evaluation of classification accuracy using DUI

5. CONCLUSIONS

We proposed the DUI for sLDAc by unfolding the prediction rule to obtain a deep model. We derived the error back-propagation algorithm for learning parameters via DUI. Such learning algorithm is able to seek a deep model to improve classification performance. Experiments showed that maximizing the negative cross entropy error function instead of lower bound of marginal likelihood can outperform the conventional VB in terms of classification accuracy for test documents. Such DUI could be extended to the other models.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[2] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65, Nov. 2010.

[3] J.-T. Chien and C.-H. Chueh, "Topic-based hierarchical segmentation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 55–66, 2012.

[4] J.-T. Chien and C.-H. Chueh, "Dirichlet class language models for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 482–495, 2011.

[5] J.-T. Chien, "Hierarchical Pitman-Yor-Dirichlet language model," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 8, pp. 1259–1272, 2015.

[6] Y.-L. Chang and J.-T. Chien, "Latent Dirichlet learning for document summarization," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 1689–1692.

[7] J.-T. Chien and M.-S. Wu, "Adaptive Bayesian latent semantic analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 198–207, Jan 2008.

[8] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in Models for Acoustic Processing Workshop (NIPS)*, 2006.

[9] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: discriminative learning for dimensionality reduction and classification," in *Advances in Neural Information Processing Systems (NIPS)*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2008, pp. 897–904.

[10] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1903–1910.

[11] J. Zhu, A. Ahmed, and E. P. Xing, "MedLDA: maximum margin supervised topic models," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2237–2278, 2012.

[12] J. Chen, J. He, Y. Shen, L. Xiao, X. He, J. Gao, X. Song, and L. Deng, "End-to-end learning of latent Dirichlet allocation by mirror-descent back propagation," *arXiv:1508.03398*, 2015.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, Jan. 1977.

[14] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, University of London, 2003.

[15] A. Kulesza and F. Pereira, "Structured learning with approximate inference," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 785–792.

[16] V. Stoyanov, A. Ropson, and J. Eisner, "Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2011, pp. 725–733.

[17] M. J. Wainwright, "Estimating the "wrong" graphical model: benefits in the computation-limited setting," *Journal of Machine Learning Research*, vol. 7, pp. 1829–1859, 2006.

[18] J. Lasserre, C. M. Bishop, and T. P. Minka, "Principled hybrids of generative and discriminative models," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, vol. 1, pp. 87–94.

[19] A. Holub and P. Perona, "A discriminative framework for modelling object classes," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 664–671.

[20] J. R. Hershey, J. Le Roux, and F. Weninger, "Deep unfolding: model-based inspiration of novel deep architectures," *arXiv:1409.2574*, 2014.

[21] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 121–128.

[22] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, vol. 132, no. 1, pp. 1–63, 1997.