Semi-supervised Max-margin Topic Model with Manifold Posterior Regularization *

Wenbo Hu, Jun Zhu, Hang Su, Jingwei Zhuo, Bo Zhang
Tsinghua National Laboratory for Information Science and Technology (TNList),
State Key Lab for Intelligent Technology and Systems,
Center for Brain-Inspired Computing Research (CBICR),
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
{hwb13@mails., dcszj@, suhangss@, zhuojw10@mails., dcszb@}tsinghua.edu.cn

Abstract

Supervised topic models leverage label information to learn discriminative latent topic representations. As collecting a fully labeled dataset is often time-consuming, semi-supervised learning is of high interest. In this paper, we present an effective semi-supervised max-margin topic model by naturally introducing manifold posterior regularization to a regularized Bayesian topic model, named LapMedLDA. The model jointly learns latent topics and a related classifier with only a small fraction of labeled documents. To perform the approximate inference, we derive an efficient stochastic gradient MCMC method. Unlike the previous semi-supervised topic models, our model adopts a tight coupling between the generative topic model and the discriminative classifier. Extensive experiments demonstrate that such tight coupling brings significant benefits in quantitative and qualitative performance.

1 Introduction

Supervised topic models have been applied in various tasks by discovering latent topic structures and meanwhile predicting interpretable labels, including document classification [Zhu *et al.*, 2012], image classification [Wang *et al.*, 2009], etc. The popular models include supervised LDA (sLDA) [Blei and McAuliffe, 2007], discriminative LDA (DiscLDA) [Lacoste-Julien *et al.*, 2008], Labeled LDA [Ramage *et al.*, 2009] and max-margin LDA (MedLDA) [Zhu *et al.*, 2012]. By incorporating document labels, supervised topic models can capture high-level information and provide guidance to predict meaningful latent topic patterns. However, human-labeling is often costly, especially for a large corpus, which limits the further applicability of the supervised topic models. It is imperative to incorporate a fraction of labeled documents with the massive unlabeled documents to conduct the latent topic modeling and document classification in a semi-supervised learning (SSL) manner [Chapelle *et al.*, 2006].

Various efforts have been made to build semi-supervised topic models (SSTMs). For example, Zhuang et al. [2013] first learn an sLDA model by using labeled data only, then predict labels for the unlabeled documents using the learned sLDA, and finally refine the labels by solving a low-rank graph regularized problem. This method adopts a loose coupling between the classifier and the topic model, which leads to a waste of the unlabeled documents when learning topics. Lu et al. [2013] and Zhang et al. [2014] use both labeled and unlabeled documents in learning the latent topics; discriminative topic modeling (DTM) [Huh and Fienberg, 2012] trains an unsupervised manifold regularized PLSA model [Cai *et al.*, 2009], and then builds a subsequent SVM classifier based on the learned topic representations. However, the unlabeled documents are not taken into consideration for these algorithms when building the classifiers, which restricts the classification performance.

Other semi-supervised topic models have been built based on some miscellaneous document supervision, such as the partially-labeled topic assignments [Ramage *et al.*, 2011] or a complex hierarchical topic structure [Mao *et al.*, 2012]. More recently, in some models with deep architectures, unlabeled documents are used for pre-training before using a supervised CNN [Johnson and Zhang, 2015] or LSTM [Johnson and Zhang, 2016]. However, the performance of deep models is largely depending on plenty of labeled documents.

1.1 Our Proposal

In this paper, we propose a semi-supervised topic model with manifold posterior regularization. Specifically, we introduce the manifold regularization to the posterior of a supervised topic model under the generic regularized Bayesian inference (RegBayes) [Zhu *et al.*, 2014b] framework, such that the close samples in the bag-of-words feature domain would have the similar topic representation and predicted labels. We build a tight-coupling model that learns latent topics and classifiers with both labeled and unlabeled documents. With the tight coupling, our semi-supervised model jointly discovers discriminative topic representations and builds powerful classifiers.

However, the algorithms based on graph Laplacian regular-

*The work is supported by National 973 Project (2013CB329403), NSFC Projects (Nos. 61620106010, 61621136008, 61332007, 61571261), the Youth Top-notch Talent Support Program and the Collaborative Project from MSRA.

Table 1: A summary of (semi-)supervised document classification methods.

methods	Max-margin	Topic Modeling	Manifold Regularization	Stochastic training
TSVM [Joachims, 1999]	√	×	×	√
LapSVM [Belkin *et al.*, 2006]	√	×	√	×
MedLDA [Zhu *et al.*, 2012]	√	√	×	√
DTM [Huh and Fienberg, 2012]	×	√	√	×
Laplacian MedLDA(Proposed)	√	√	√	√

ization generally scale poorly with the data size. To address this issue, we present an efficient stochastic gradient Markov chain Monte Carlo (MCMC) to infer the model. In each iteration, only a small subset of the data is used, reducing the computational cost significantly.

In summary, our contributions are as follows

- We introduce manifold posterior regularization for latent topic models, which can incorporate both the labeled and unlabeled data via manifold learning;
- We build a Laplacian MedLDA model, a semi-supervised max-margin topic model with manifold posterior regularization, under the *RegBayes* framework, such that the induced model has a tight coupling between the latent topic modeling and the max-margin classifier;
- We develop a stochastic gradient MCMC method for the efficient inference of our model.

We summarize the merits of our model in Table. 1, and compare it with the representative (topic) models.

2 Preliminaries

We review the manifold regularization method for semi-supervised learning as well as the max-margin supervised topic model.

2.1 Manifold Regularization

The manifold regularization methodology was popularized in machine learning by Laplacian Eigenmap [Belkin and Niyogi, 2001], a representative dimensionality reduction method. Later on, manifold regularization has been used for semi-supervised learning (SSL) and served as a regularization term in existing empirical loss minimization schemes [Belkin *et al.*, 2006]. Manifold regularization assumes that the learned manifolds should be smooth, which means nearby data pairs have similar prediction scores.

Specifically, for a set of training instances $\mathbf{X} = \{x_i\}_{i=1}^{l+u}$, where $x_i \in \mathbb{R}^d$, we only have a part of the instance labels $\mathbf{Y} = \{y_i\}_{i=1}^{l}$, where $y_i \in \{-1, 1\}$. The goal of SSL is to learn an unknown function f that can fit well to the observed labels while having nice properties on the unlabeled instances. In particular, the manifold regularization SSL solves the problem

$$\min_{f \in \mathcal{H}_\kappa} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + c_1 \Omega(f) + c_2 \mathbf{f}^\top L \mathbf{f}, \qquad (1)$$

where V is some loss function (e.g., squared loss or hinge loss), $\Omega(f)$ is a structural penalty of the classifier f, c_1 and c_2 are the regularization parameters, and the manifold regularization term is defined as $\sum_{i,j=1}^{l+u}(f(x_i) - f(x_j))^2 w_{ij} = \mathbf{f}^\top L \mathbf{f}$, which represents graph smoothness. The weight w_{ij} is a pairwise similarity between samples i and j; and L is the graph Laplacian which is analogous to the Laplace-Beltrami operator on manifolds. Particularly, this method is *Laplacian SVM* (LapSVM) [Belkin *et al.*, 2006], when V is set as the hinge loss function $\max[0, 1 - y_i f(x_i)]$.

2.2 MedLDA

Max-margin topic model [Zhu *et al.*, 2012] is a popular supervised topic model, which can be formulated as a regularized Bayesian (RegBayes) method [Zhu *et al.*, 2014b] with a max-margin posterior regularization. Specifically, MedLDA consists of two parts: 1) a latent Dirichlet allocation (LDA) model for modeling the latent topic structures of the corpus and 2) a max-margin classifier for predicting document labels.

LDA [Blei *et al.*, 2003] is a hierarchical Bayesian model that uses an admixture of topics as a latent document representation. Let $\mathbf{\Phi} = \{\mathbf{\Phi}_k\}_{k=1}^{K}$ be a set of K topics, in which each topic $\mathbf{\Phi}_k$ is a multinomial distribution with a symmetric Dirichlet prior $\mathrm{Dir}(\beta)$ over a V-word vocabulary. For a single document i with N_i words, the generative process of the vanilla LDA is:

1. Draw a topic proportion $\theta_i \sim \mathrm{Dir}(\alpha)$,
2. For each word n $(1 \le n \le N_i)$:
   (a) Draw a topic assignment $z_{in} \sim \mathrm{Multinomial}(\theta_i)$,
   (b) Draw a word $w_{in} \sim \mathrm{Multinomial}(\Phi_{z_{in}})$.

Given a set of documents $\mathbf{X} = \{x_i\}_{i=1}^{l}$, we denote the collection of latent topic proportions as $\mathbf{\Theta} = \{\theta_i\}_{i=1}^{l}$ and topic assignments as $\mathbf{Z} = \{z_i\}_{i=1}^{l}$ where $x_i = \{w_{in}\}_{n=1}^{N_i}$ and $z_i = \{z_{in}\}_{n=1}^{N_i}$.

MedLDA is a supervised model that builds a classifier on the latent topic representations and in this paper we use the Gibbs classifier formulation to define our model, which is proven to have good generalization performance [McAllester, 2003; Germain *et al.*, 2009]. With the posterior samples of $\mathbf{Z}$, we can get the average topic assignments of the words in document i as $\bar{z}_i$, with element $\bar{z}_{ik} = \frac{1}{N_i}\sum_{n=1}^{N_i} \mathbb{I}(z_{in} = k)$[1]. Then we predict the document label y_i,

$$\hat{y}_i = \mathrm{sgn}(f(\bar{z}_i, \eta)), \quad f(\bar{z}_i, \eta) = \eta^\top \bar{z}_i, \qquad (2)$$

where η is the vector of classifier weights; and $\mathrm{sgn}(\cdot)$ is the sign function. We define the corresponding *expected hinge loss*

$$\mathcal{R}_1(q) = \sum_{i=1}^{l} \mathbb{E}_q\left[\max\left(0, \ell - y_i f(\bar{z}_i, \eta)\right)\right], \qquad (3)$$

[1] $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the predicate holds otherwise 0.

where we take expectation over $q(\eta, \mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi})$ to consider the uncertainty of latent variables and $\ell(\geq 1)$ is the cost of making a wrong prediction. By optimizing a variational-reformulated objective, the MedLDA model solves the Reg-Bayes problem:

$$\min_{q(\eta,\mathbf{\Theta},\mathbf{Z},\mathbf{\Phi})} \mathcal{L}\left(q(\eta, \mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi})\right) + c_1 \cdot \mathcal{R}_1\left(q(\eta, \mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi})\right), \quad (4)$$

where q is the target posterior distribution; $\mathcal{L}(\cdot)$ is the variational objective of the vanilla topic model; c_1 is the positive regularization parameter. In this way, we define the MedLDA model under the Gibbs classifier formulation, named Gibbs MedLDA. We refer readers to [Zhu *et al.*, 2014a] for more details of Gibbs MedLDA.

By solving problem (4), we get the unnormalized posterior distribution as:

$$p(\eta, \mathbf{\Theta}, \mathbf{\Phi}, \mathbf{Z}) \propto p_0(\eta, \mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi}) p(\mathbf{X}|\mathbf{Z}, \Phi) \psi_1(\mathbf{Y}|\mathbf{Z}, \eta), \quad (5)$$

where ψ_1 is the unnormalized likelihood of the supervised signal,

$$\psi_1(\mathbf{Y}|\mathbf{Z}, \eta) = \prod_{i=1}^{l} \exp\left[-c_1 \max\left(0, \ell - y_i f(\bar{z}_i, \eta)\right)\right]. \quad (6)$$

Inference can be implemented by Gibbs sampling with data augmentation [Zhu *et al.*, 2014a], a fast Gibbs sampler [Zheng *et al.*, 2015] and stochastic gradient MCMC method [Hu *et al.*, 2017].

3 Laplacian MedLDA

We now formalize our Laplacian MedLDA (LapMedLDA) to learn latent topics and document classifiers with labeled and unlabeled documents. We also present a stochastic gradient MCMC method for fast inference.

3.1 Formulation of Laplacian MedLDA

For semi-supervised document classification, we have a set of partially-labeled documents $\left\{\mathbf{X} = \{x_i\}_{i=1}^{l+u}, \mathbf{Y} = \{y_i\}_{i=1}^{l}\right\}$ and aim to infer the labels for unlabeled data. To incorporate the useful information of the unlabeled documents, we directly use the discriminant function score to build a regularization term $\mathcal{R}_2(q)$. In the Gibbs classifier formulation, we take expectation over the target posterior q and define the manifold regularization loss as,

$$\mathcal{R}_2 = \mathbb{E}_q\left[\sum_{i,j=1}^{l+u} (\eta^\top \bar{z}_i - \eta^\top \bar{z}_j)^2 w_{ij}\right] = \mathbb{E}_q\left[\mathbf{f}^\top L \mathbf{f}\right], \quad (7)$$

where w_{ij} denotes the pairwise similarity between the bag-of-words vectors of documents i and j; L is the Laplacian matrix of the neighborhood graph derived from $W = [w_{ij}]$; $\mathbf{f} = \bar{Z}^\top \eta$ is the prediction score vector; and $\bar{Z}$ is the latent topic assignment matrix with the i-the column being $\bar{z}_i$. This regularization encourages that if two documents are close to each other on the graph, they should have similar latent representations and hence similar prediction labels.

We use the manifold regularization in Eqn. (7) as a posterior regularization term, yielding the following regularized Bayesian model:

$$\min_{q(\eta,\mathbf{\Theta},\mathbf{Z},\mathbf{\Phi})} \mathcal{L}\left(q(\eta, \mathbf{\Theta}, \mathbf{Z}, \mathbf{\Phi})\right) + c_1 \cdot \mathcal{R}_1\left(q\right) + c_2 \cdot \mathcal{R}_2(q),$$

where f is the discriminant function as in MedLDA; $\mathcal{L}$ is the variational objective for the vanilla topic model with both labeled and unlabeled documents; c_2 is another positive regularization parameter; $\mathcal{R}_1$ and $\mathcal{R}_2$ are corresponding to the max-margin regularization and manifold regularization, respectively.

By solving the problem (8), we get the unnormalized posterior distribution of LapMedLDA as

$$\begin{aligned} &p(\eta, \mathbf{\Theta}, \mathbf{\Phi}, \mathbf{Z}) \\ \propto &p_0(\eta, \mathbf{\Theta}, \mathbf{\Phi}, \mathbf{Z}) p(\mathbf{X}|\mathbf{Z}, \Theta) \psi_1(\mathbf{Y}|\mathbf{Z}, \eta) \psi_2(L|\mathbf{Z}, \eta), \end{aligned} \quad (8)$$

where ψ_2 is the unmormalized likelihood of the graph Laplacian:

$$\psi_2(\mathbf{Z}, \eta|L) = \exp\left(-c_2 \mathbf{f}^\top L \mathbf{f}\right). \quad (9)$$

Note the normalizing term for the above posterior distribution can be omitted in the Gibbs sampling and the stochastic gradient MCMC steps. LapMedLDA captures not only the max-margin supervision of the document labels but also the manifold structure of the documents. Compared with the previous work on the likelihood manifold regularized topic models [Cai *et al.*, 2009; Huh and Fienberg, 2012], our manifold posterior regularization is more direct and has a tight coupling between the unlabeled data and the classifiers, which brings the better latent topic representation and classification performance.

Note that the latent modeling of $\mathbf{Z}$ can be regarded as a non-linear feature extractor from the raw bag-of-words feature, and it allows us to use the linear kernel for the RKHS hypothesis space. In this case, our model does not include any sensitive parameter, such as the polynomial order for the polynomial kernel and the scale parameter for the RBF kernel.

In particular, when $c_2 = 0$, the manifold regularization is actually unused and unlabeled data is only used in the generative topic modeling. Our model degenerates to a naive version of semi-supervised max-margin topic model, which is called SS-MedLDA.

3.2 Stochastic Gradient MCMC for LapMedLDA

Due to the tight coupling of the max-margin loss and manifold regularization, previous Gibbs sampling methods with data augmentation [Zhu *et al.*, 2014a] cannot be directly applied. Moreover, it is hard to store and compute the pairwise similarities of the neighborhood graph for large-scale datasets. Therefore, it is imperative to develop a fast inference method. In this section, we present a stochastic gradient MCMC method for LapMedLDA.

In each iteration, we first calculate the expectation of the latent topic assignments from the conditional distribution $q(\mathbf{Z}|\mathbf{W}, \mathbf{\Phi}, \eta, L, \alpha)$ and then use them to update the posterior parameter η and Φ via the stochastic gradient Riemannian MCMC.

To speed up the inference procedure, we collapse out Θ and get the collapsed generative distribution of our model as

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{\Phi}, \mathbf{Y}|\alpha, \beta, L) \tag{10}$$
$$\propto p_0(\eta)p(\mathbf{\Phi}|\beta)p(\mathbf{X}, \mathbf{Z}|\alpha, \mathbf{\Phi})\psi_2(\mathbf{Z}, \eta|L)\prod_{i=1}^{l}\psi_1(y_i|z_i, \eta),$$

where

$$p(\mathbf{X}, \mathbf{Z}|\alpha, \mathbf{\Phi}) = \prod_{i=1}^{l+u}\prod_{k=1}^{K}\frac{\Gamma(\alpha + C_{ik\cdot})}{\Gamma(\alpha)}\prod_{i=1}^{V}\Phi_{kv}^{C_{ikv}}. \tag{11}$$

$C_{ik\cdot}$ is the number of words in document i that are assigned to topic k and C_{ikv} is the number of word v in document i that are assigned to topic k. We draw posterior samples from the collapsed posterior distribution $q(\mathbf{\Phi}, \eta|\mathbf{X}, \mathbf{Y}, \alpha, \beta, L)$ and update η and Φ by turns. Using a randomly-drawn document subset $\tilde{\mathbf{X}}$, we get the unbiased noisy estimate of the gradients of the log posterior of Φ with respect to Φ as

$$\frac{\partial}{\partial \Phi_{kv}}\log q(\mathbf{\Phi}, \eta|\mathbf{X}, \alpha, \beta, \mathbf{Y}, L) \tag{12}$$
$$\approx \frac{\beta - 1}{\Phi_{kv}} - 1 + \frac{l+u}{|\tilde{\mathbf{X}}|}\sum_{i\in\tilde{X}}\mathbb{E}_{z_i|\mathbf{x}_i,\theta,\alpha}\left[\frac{C_{ikv}}{\Phi_k} - \frac{C_{ik\cdot}}{\Phi_{k\cdot}}\right],$$

and then we update the Φ using stochastic gradient Riemannian Langevin dynamics [Patterson and Teh, 2013].

Similarly, we also get the unbiased stochastic estimate of the log posterior gradient of η with respect to η as

$$\frac{\partial}{\partial \eta}\log q(\mathbf{\Phi}, \eta|\mathbf{X}, \alpha, \beta, \mathbf{Y}, L) \approx -\eta + \frac{\partial}{\partial \eta}\log\psi_2 \tag{13}$$
$$+ \frac{l}{|\tilde{\mathbf{X}} \cap \{x_i\}_{i=1}^{l}|}\sum_{x_i\in\tilde{X}, i\le l}\frac{\partial}{\partial \eta}\log\psi_1(y_i|z_i, \eta),$$

where the subgradient $\frac{\partial}{\partial \eta}\log\psi_1(y_i|z_i, \eta)$ equals $-c_1\bar{z}_i$ if $\psi_1(y_i|z_i, \eta) < 1$ and 0 otherwise; we also calculate the stochastic gradient of ψ_2 with respect to η, yielding $\frac{\partial}{\partial \eta}\log\psi_2$ as

$$\frac{\partial}{\partial \eta}\log\psi_2 \approx \frac{l+u}{|\tilde{\mathbf{X}}|}\left(-2c_2\bar{Z}_{\#}L_{\#\#}\bar{Z}_{\#}^{\top}\eta\right), \tag{14}$$

where the matrix subscript(s) $\#$ represent the matrix slicing with the indexes of the minibatch $\tilde{X}$, either the one-dimensional for Z or two-dimensional for L. Then, we use SGLD steps to update η [Welling and Teh, 2011].

Note that the matrix slice for L indicates that in each iteration, we only use part of the documents and their neighborhood information, which means that we use a subgraph in each iteration. This process will be detailed in Sec. 3.4.

To calculate the expectation of $\bar{z}$ to get the observed counts in the above posterior gradients, the Gibbs sampling iterations for the topic assignments of word n in document i is as follows:

$$p(z_{in} = k|z_{i,-n}, \Phi, \eta) \propto (\alpha + C_{ik\cdot}^{-n})\Phi_{kn}$$
$$\psi_2(\bar{Z}^*, \eta|L,)\psi_1(y_i|\bar{z}_i^*, \eta), \tag{15}$$

where $z_{i,-n}$ is the topic assignments of other documents, $\bar{z}_i^*$ is the average topic assignments $\bar{z}_i$ after setting topic z_{in} as k and $C_{ik\cdot}^{-n}$ is the number of words assignment as topic k in document i after removing word n.

3.3 Multi-Class Extension

In many applications, document labels are multi-class and we present a multi-class extension of our model. The multi-class hinge-loss $\rho(y_i, \bar{z}_i, \eta)$ is defined as [Crammer and Singer, 2001],

$$\rho = \max\left[0, \ell + \max_{y\neq y_i} f(y, \bar{z}_i|\eta) - f(y_i, \bar{z}_i|\eta)\right], \tag{16}$$

where $f(y, \bar{z}_i) = \eta_y^{\top}\bar{z}_i$, $y \in \{1, 2, \cdots, \gamma\}$, η is a $\gamma \times K$ matrix and η_y is the y-th row of the matrix η. We now retain the LDA model part $\mathcal{L}(q)$ in the LapMedLDA model (Eqn. (8)), and give the multi-class regularizations (γ-class) as,

$$\mathcal{R}_1(q) = \sum_{i=1}^{l}\mathbb{E}_q\rho(y_i, \bar{z}_i, \eta),\ \mathcal{R}_2(q) = \mathbb{E}_q\left[\mathrm{Tr}(\mathbf{f}^{\top}L\mathbf{f})\right].$$

With the regularizations for the multi-class setting, we get the similar collapsed posterior as Eqn. (10) with the unnormalized pseudo likelihoods, $\psi_1(y_i|z_i, \eta)$ and ψ_2 as,

$$\begin{aligned}\psi_1(y_i|z_i, \eta) &= \exp\rho(y_i, \bar{z}_i, \eta),\\ \psi_2(L|\mathbf{Z}, \eta) &= \exp\mathrm{Tr}(-c_2\mathbf{f}^{\top}L\mathbf{f}).\end{aligned} \tag{17}$$

The log-posterior gradient with respect to Φ are identical to that of the binary case. Now we give the subgradients of the unnormalized likelihood ψ_1 with respect to η as,

$$\begin{cases}\partial_{\eta}\log\psi_1 = 0; & \text{if } \psi_1(y_i|z_i, \eta) \ge 1,\\ \partial_{\eta_y}\log\psi_1 = -c\bar{z}_i; & \text{if } \psi_1(y_i|z_i, \eta) < 1, y = y^*\\ \partial_{\eta_y}\log\psi_1 = c\bar{z}_i; & \text{if } \psi_1(y_i|z_i, \eta) < 1, y \neq y^*\end{cases}$$

where $y^* = \mathrm{argmax}_{y\neq y}\,\eta^{\top}\bar{z}_i$. The subgradient of $\log\psi_2$ with respect to η is identical to Eqn. (14) with η being a $\gamma \times K$ matrix. With the posterior stochastic gradients, we can use SGRLD and SGLD to sample Φ and η.

3.4 Graph Construction

For every data point pair, we first construct the pair-wise similarities using the cosine distances of the bag-of-word vectors and remain the r nearest neighbors. The similarities between two documents are defined as:

$$w_{ij} = \begin{cases}1, & x_i \in \Delta_r(x_j) \text{ or } x_j \in \Delta_r(x_i)\\ 0, & \text{otherwise}\end{cases}, \tag{18}$$

where $\Delta_r(x)$ is the set of r nearest neighbors of x. Moreover, we further utilize the label information by adding the edges of documents in the same categories and removing the edges of documents in the different categories. For the practical useage, we choose the symmetric normalized Laplacian matrix to normalize the node degrees. Similiar graph construction methods were used in [Huh and Fienberg, 2012; Cai *et al.*, 2009].

In the stochastic gradient MCMC method, we use one tiny minibatch in each iteration. For each minibatch, we extracted the corresponding subgraph that consists of the pair-wise similarities between the data instances in the minibatch. The merit of this method is that we do not need to build and store a large pair-wise graph in advance. Our subgraph construction is related to the online manifold regularization in the streaming setting [Goldberg *et al.*, 2008].

Table 2: Classification Accuracy(%) on binary Dataset

Dataset	Label Ratio	0.07	0.09	0.1	0.3	0.5	0.7	1
20Newsgroup (atheism vs. religion)	SVM	0.59±0.12	0.61±0.12	0.60±0.11	0.70±0.09	0.74±0.05	0.76±0.02	0.79±0.01
	LDA+SVM	0.56±0.10	0.56±0.12	0.56±0.12	0.60±0.11	0.65±0.05	0.65±0.02	0.66±0.02
	MedLDA	0.55±0.00	0.55±0.00	0.60±0.09	0.61±0.10	0.71±0.06	0.77±0.02	0.79±0.02
	TSVM	0.61±0.12	0.63±0.12	0.65±0.11	0.73±0.09	0.75±0.05	0.77±0.02	0.79±0.02
	LapSVM	0.58±0.12	0.62±0.11	0.62±0.10	0.72±0.07	0.72±0.05	0.76±0.03	0.80±0.03
	DTM	**0.64±0.06**	0.66±0.05	0.69±0.04	0.70±0.03	0.71±0.02	0.72±0.02	0.77±0.02
	SS-MedLDA	0.59±0.12	0.66±0.11	0.72±0.10	0.74±0.04	0.75±0.04	0.79±0.02	0.79±0.02
	LapMedLDA	0.62±0.11	**0.68±0.10**	**0.74±0.05**	**0.76±0.02**	**0.77±0.02**	0.79±0.02	0.79±0.02

4 Experiments

We now present the empirical results of our semi-supervised topic model. We first present the document classification performance and the efficiency analysis of our model. Then we show the qualitative analysis .

4.1 Datasets and Experimental Setup

We compare our model with several state-of-the-art document classification models. The involved supervised models are linear-SVM (SVM), MedLDA and LDA+SVM. These three methods use labeled documents and for LDA+SVM, unlabeled documents are also used in the generative topic model part. The involved semi-supervised models are transductive SVM (TSVM), Laplacian SVM (LapSVM), discriminative topic modeling (DTM), SS-MedLDA and LapMedLDA. The SVM, TSVM and LapSVM methods directly use the bag-of-word feature as the input features.

We consider a binary document set which consists of two subgroups of the 20Newsgroups data[2], *alt.atheism* and *talk.religion.misc*. This sub-dataset consists of 856 training documents and 569 testing documents. Next we test the multi-class model on the Yahoo! News K-series[3] and the whole 20Newsgroups data. The 20Newsgroups and Yahoo! News K-series datasets have 18,274 and 2,340 documents respectively.

We use random label discards to construct a partially labeled corpus. At every single test, we randomly discard part of the training document labels, which follows a binomial distribution with a parameter called *label ratio*. For example, when label ratio is 0.1, we randomly discard every existing document with probability 0.9. All the testing accuracy performances are averaged over 100 repetitions.

Since the classification performance of MedLDA is insensitive to the hyper-parameters in a wide range [Zhu *et al.*, 2014a] and we use the same hyper-parameter setting of MedLDA. For LDA-based models, we set $\alpha = 1$, $\beta = 1$ and topic number $K = 20$. For MedLDA-based models, we set $\ell = 164$ and $c_1 = 1$. The parameter c_2 is the regularization parameter for the manifold regularization which is chosen from $\{0.1, 0.01, 0.001\}$ via 5-fold cross validation. The expectation of the topic assignments $\bar{\mathbf{Z}}$ are calculated with 5 samples and for graph construction, we set the nearest neighbor number as 10 for 20Newsgroups dataset and 5 for Yahoo news dataset. As we shall see in the sensitivity analysis, the performances of our model is insensitive to K, $zloop$ and nearest neighborhood number used in the graph construction. For the stochastic gradient MCMC, the stepsizes for classifier weights η are AdaGrad stepsize [Duchi *et al.*, 2011] and stepsizes for topic-word parameter Φ are set as $10 * (1 + t/100)^{-0.6}$ at iteration t.

[2] http://qwone.com/∼jason/20Newsgroups

[3] http://www-users.cs.umn.edu/∼boley/ftp/PDDPdata/README.html

4.2 Classification Performance

We report the classification performance in Table. 2 for binary setting and Table. 3 for multi-class. We mark the best classification results as bold when the label ratio is lower than 0.5. We can see that our LapMedLDA generally performs best when there are a few labels (label ratio ≤ 0.5), indicating that the manifold regularization term captures the global manifold information. When the label ratio is larger than 0.5, semi-supervised models generally perform as good as supervised models.

It is worth stressing that in all the semi-supervised models, we use the inductive setting and all testing documents are not available during the training process. For semi-supervised models with latent topic variables, the latent topic assignments of the testing documents are predicted only with the topic-word parameter inferred from the training process. The prediction process uses either collapsed Gibbs sampling for LDA or EM-based optimization method for PLSA [Zhu *et al.*, 2014a]. Please note that Huh et al. [2012] reported the results in a transductive setting, where the testing data features are seen during training.

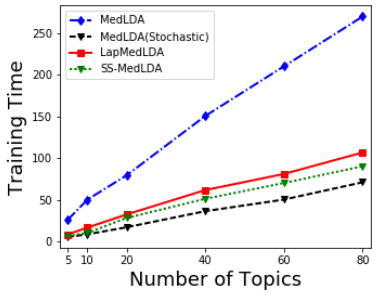

Figure 1: Training time efficiency of three MedLDA models

4.3 Time Efficiency

To compare the efficiency of the algorithms, we show the training time of related models on the multi-class 20newsgroup dataset in Figure. 1. The red solid line (LapMedLDA) shows the training time of the stochastic sampler for LapMedLDA, the blue dotted line (MedLDA)

Table 3: Classification Accuracy on multi-class Datasets

Dataset	Label Ratio	0.07	0.09	0.1	0.3	0.5	0.7	1
Yahoo k-series	SVM	0.59±0.01	0.65±0.01	0.66±0.03	0.77±0.01	0.79±0.01	0.83±0.00	0.83±0.00
	LDA+SVM	0.30±0.10	0.42±0.08	0.56±0.07	0.68±0.10	0.70±0.03	0.74±0.02	0.75±0.01
	MedLDA	0.50±0.09	0.56±0.04	0.67±0.02	0.75±0.02	0.77±0.02	0.83±0.01	0.83±0.01
	TSVM	0.64±0.04	0.68±0.03	0.70±0.03	**0.78±0.01**	0.80±0.01	0.83±0.00	0.83±0.00
	LapSVM	0.66±0.05	0.70±0.05	0.72±0.02	0.78±0.02	0.81±0.02	0.83±0.01	0.83±0.01
	DTM	0.54±0.04	0.56±0.04	0.60±0.03	0.62±0.03	0.63±0.03	0.63±0.02	0.63±0.02
	SS-MedLDA	0.63±0.06	0.67±0.04	0.69±0.03	0.73±0.02	**0.81±0.01**	0.83±0.00	0.83±0.00
	LapMedLDA	**0.67±0.05**	**0.72±0.03**	**0.74±0.01**	**0.78±0.01**	**0.81±0.01**	0.83±0.00	0.83±0.00
20Newsgroup	SVM	0.63±0.01	0.65±0.01	0.66±0.00	0.72±0.01	0.75±0.00	0.78±0.00	0.78±0.00
	LDA+SVM	0.20±0.08	0.34±0.13	0.47±0.11	0.53±0.10	0.56±0.08	0.61±0.07	0.62±0.07
	MedLDA	0.19±0.03	0.22±0.04	0.23±0.08	0.50±0.03	0.77±0.01	0.78±0.01	0.78±0.01
	TSVM	0.61±0.03	0.67±0.02	0.69±0.03	0.73±0.03	0.75±0.02	0.77±0.01	0.77±0.01
	LapSVM	0.63±0.05	0.69±0.03	0.70±0.03	0.73±0.02	0.75±0.02	0.76±0.02	0.76±0.02
	DTM	0.50±0.05	0.53±0.05	0.54±0.03	0.56±0.03	0.56±0.02	0.58±0.02	0.58±0.02
	SS-MedLDA	0.48±0.10	0.49±0.10	0.50±0.08	0.77±0.02	**0.79±0.01**	0.80±0.00	0.80±0.00
	LapMedLDA	**0.65±0.06**	**0.69±0.07**	**0.72±0.04**	**0.78±0.01**	**0.79±0.01**	0.79±0.01	0.79±0.01

shows the training time of Gibbs sampler for MedLDA [Zhu *et al.*, 2014a], the green dotted line shows the training time of SS-MedLDA and the black dash-dotted line(MedLDA(Stochastic)) shows the training time of the stochastic gradient MCMC for MedLDA [Hu *et al.*, 2017]. It can be seen that, the stochastic samplers are faster than the Gibbs sampler. Comparing the results of the stochastic samplers, we can see that the adding manifold regularization only has a limited effect on the efficiency.

4.4 Qualitative Analysis

For the binary classification setting of 20news dataset, we show some latent topic patterns of the unlabeled documents in Table. 4. The left column shows the true category labels of the documents, 70 percent of which are actually discarded. We choose the salient topics of the two categories and show the average topic proportions and the representative words in the middle and the right columns. As can be seen that, with only a fraction of document labels, we still can learn salient latent topic(s) for each specific category.

We use the 2-dimensional t-SNE method [Maaten and Hinton, 2008] to get the visualization of the learned latent topics for multi-class 20news Dataset, where the 90 percent of labels are discarded. As can be seen in Figure. 2, by adding unlabeled data in the generative topic modeling part (from left to middle), the labeled data tend to be more concentrated. By adding the manifold regularization (from middle to right), unlabeled data tend to be more concentrated and tend to cluster around the labeled documents. This concentration indicates that with the posterior manifold regularization, our model learns more discriminative latent topics of the unlabeled documents and this is beneficial for the coupled classifiers.

Table 4: Learnt Latent Topic Representation

Category	Salient Topics	Top words
atheism	T37=0.6744	god, don, atheism
	T12=0.0452	question, kill, people
	T6=0.0457	christian, bible, church
religion	T6=0.7216	christian, bible, church
	T31=0.0063	read, belief, mean
	T3=0.0045	mormon, kill, word

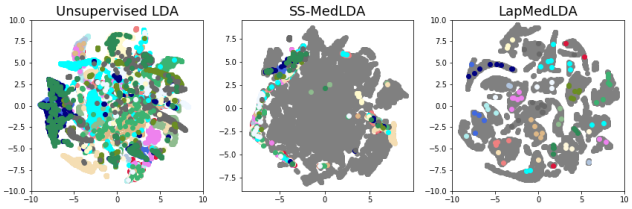

Figure 2: T-SNE embeddings of learned document representations (best viewed in color). Different colors mean different categories and the gray color means the unlabeled data in the semi-supervised setting.

4.5 Sensitivity Analysis

Figure. 3 shows that LapMedLDA is insensitive to the variation in the three parameters, the number of the nearest neighbors, the topic number and the number of collecting topic samples for calculating the expectation of $\bar{z}$ (Eqn. (15)).

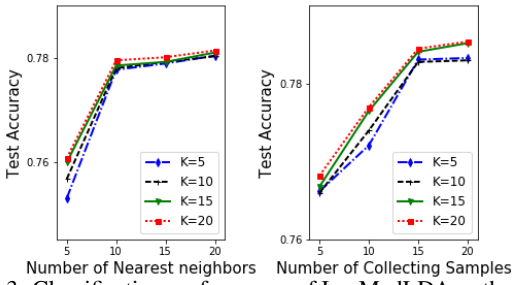

Figure 3: Classification performance of LapMedLDA as the parameters are varied.

5 Conclusions

We present a semi-supervised max-margin topic model with manifold regularization, named Laplacian MedLDA (LapMedLDA). By adopting a manifold posterior regularization term, our model jointly learns the latent topics and a related max-margin classifier for semi-supervised document classification. Under a tight coupling between topic modeling and the semi-supervised max-margin classifier, we learn discriminative topic representations and a powerful semi-supervised classifier via an efficient stochastic gradient MCMC method. Extensive experimental results show the effectiveness in the semi-supervised setting.

References

[Belkin and Niyogi, 2001] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.

[Belkin *et al.*, 2006] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[Blei and McAuliffe, 2007] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[Blei *et al.*, 2003] D. Blei, A. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[Cai *et al.*, 2009] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In *ICML*, 2009.

[Chapelle *et al.*, 2006] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised learning*. MIT Press, 2006.

[Crammer and Singer, 2001] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001.

[Duchi *et al.*, 2011] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159, 2011.

[Germain *et al.*, 2009] P. Germain, A. Lacasse, F. Laviolette, and Marchand M. PAC-Bayesian learning of linear classifiers. In *ICML*, 2009.

[Goldberg *et al.*, 2008] AB. Goldberg, M. Li, and X. Zhu. Online manifold regularization: A new learning setting and empirical study. In *ECML/PKDD*, pages 393–407. Springer, 2008.

[Hu *et al.*, 2017] W. Hu, J. Zhu, and B. Zhang. Fast sampling for Bayesian max-margin models. *Expert Systems with Applications*, 69:277–287, 2017.

[Huh and Fienberg, 2012] Seungil Huh and Stephen E Fienberg. Discriminative topic modeling based on manifold learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):20, 2012.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.

[Johnson and Zhang, 2015] R. Johnson and T. Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In *NIPS*, 2015.

[Johnson and Zhang, 2016] R. Johnson and T. Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. In *ICML*, 2016.

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

[Lu *et al.*, 2013] Y. Lu, S. Okada, and K. Nitta. Semi-supervised latent dirichlet allocation for multi-label text classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 351–360. Springer, 2013.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[Mao *et al.*, 2012] X. Mao, Z. Ming, T. Chua, S. Li, H. Yan, and X. Li. SSHLDA: a semi-supervised hierarchical topic model. In *EMNLP*, 2012.

[McAllester, 2003] D. McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

[Patterson and Teh, 2013] S. Patterson and Y. W. Teh. Stochastic gradient Riemannian langevin dynamics on the probability simplex. In *NIPS*, 2013.

[Ramage *et al.*, 2009] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

[Ramage *et al.*, 2011] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *KDD*, 2011.

[Wang *et al.*, 2009] C. Wang, D. Blei, and F. Li. Simultaneous image classification and annotation. In *CVPR*, 2009.

[Welling and Teh, 2011] M Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 2011.

[Zhang and Wei, 2014] Y. Zhang and W. Wei. A jointly distributed semi-supervised topic model. *Neurocomputing*, 134:38–45, 2014.

[Zheng *et al.*, 2015] X. Zheng, Y. Yu, and E. P. Xing. Linear time samplers for supervised topic models using compositional proposals. In *KDD*, 2015.

[Zhu *et al.*, 2012] J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: maximum margin supervised topic models. *JMLR*, 13(1):2237–2278, 2012.

[Zhu *et al.*, 2014a] J. Zhu, N. Chen, H. Perkins, and E. P. Xing. Gibbs max-margin topic models with data augmentation. *JMLR*, 15:1073–1110, 2014.

[Zhu *et al.*, 2014b] J. Zhu, N. Chen, and E. P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *JMLR*, 15:1799–1847, 2014.

[Zhuang *et al.*, 2013] L. Zhuang, H. Gao, J. Luo, and Z. Lin. Regularized semi-supervised latent dirichlet allocation for visual concept learning. *Neurocomputing*, 119:26–32, 2013.