Gibbs Max-margin Topic Models with Data Augmentation

Jun Zhu   DCSZJ@MAIL.TSINGHUA.EDU.CN
Ning Chen   NINGCHEN@MAIL.TSINGHUA.EDU.CN
Hugh Perkins   NGLS11@MAILS.TSINGHUA.EDU.CN
Bo Zhang   DCSZB@MAIL.TSINGHUA.EDU.CN
State Key Lab of Intelligent Technology and Systems
Tsinghua National Lab for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University
Beijing, 100084, China

Editor: David Blei

Abstract

Max-margin learning is a powerful approach to building classifiers and structured output predictors. Recent work on max-margin supervised topic models has successfully integrated it with Bayesian topic models to discover discriminative latent semantic structures and make accurate predictions for unseen testing data. However, the resulting learning problems are usually hard to solve because of the non-smoothness of the margin loss. Existing approaches to building max-margin supervised topic models rely on an iterative procedure to solve multiple latent SVM subproblems with additional mean-field assumptions on the desired posterior distributions. This paper presents an alternative approach by defining a new max-margin loss. Namely, we present Gibbs max-margin supervised topic models, a latent variable Gibbs classifier to discover hidden topic representations for various tasks, including classification, regression and multi-task learning. Gibbs max-margin supervised topic models minimize an expected margin loss, which is an upper bound of the existing margin loss derived from an expected prediction rule. By introducing augmented variables and integrating out the Dirichlet variables analytically by conjugacy, we develop simple Gibbs sampling algorithms with no restrictive assumptions and no need to solve SVM subproblems. Furthermore, each step of the "augment-and-collapse" Gibbs sampling algorithms has an analytical conditional distribution, from which samples can be easily drawn. Experimental results on several medium-sized and large-scale data sets demonstrate significant improvements on time efficiency. The classification performance is also improved over competitors on binary, multi-class and multi-label classification tasks.

Keywords: supervised topic models, max-margin learning, Gibbs classifiers, regularized Bayesian inference, support vector machines

1. Introduction

As the availability and scope of complex data increase, developing statistical tools to discover latent structures and reveal hidden explanatory factors has become a major theme in statistics and machine learning. Topic models represent one type of such useful tools to discover latent semantic structures that are organized in an automatically learned latent topic space, where each topic (i.e., a coordinate of the latent space) is a unigram distribution over

the terms in a vocabulary. Due to its nice interpretability and extensibility, the Bayesian formulation of topic models (Blei et al., 2003) has motivated substantially broad extensions and applications to various fields, such as document analysis, image categorization (Fei-Fei and Perona, 2005), and network data analysis (Airoldi et al., 2008). Besides discovering latent topic representations, many models usually have a goal to make good predictions, such as relational topic models (Chang and Blei, 2009; Chen et al., 2013) whose major goal is to make accurate predictions on the link structures of a document network. Another example is supervised topic models, our focus in this paper, which learn a prediction model for regression and classification tasks. As supervising information (e.g., user-input rating scores for product reviews) gets easier to obtain on the Web, developing supervised latent topic models has attracted a lot of attention. Both maximum likelihood estimation (MLE) and max-margin learning have been applied to learn supervised topic models. Different from the MLE-based approaches (Blei and McAuliffe, 2007), which define a normalized likelihood model for response variables, max-margin supervised topic models, such as maximum entropy discrimination LDA (MedLDA) (Zhu et al., 2012), directly minimize a margin-based loss derived from an expected (or averaging) prediction rule.

By performing discriminative learning, max-margin supervised topic models can discover predictive latent topic representations and have shown promising performance in various prediction tasks, such as text document categorization (Zhu et al., 2012) and image annotation (Yang et al., 2010). However, their learning problems are generally hard to solve due to the non-smoothness of the margin-based loss function. Most existing solvers rely on a variational approximation scheme with strict mean-field assumptions on posterior distributions, and they normally need to solve multiple latent SVM subproblems in an EM-type iterative procedure. By showing a new interpretation of MedLDA as a regularized Bayesian inference method, the recent work (Jiang et al., 2012) successfully developed Monte Carlo methods for such max-margin topic models, with a weaker mean-field assumption. Though the prediction performance is improved because of more accurate inference, the Monte Carlo methods still need to solve multiple SVM subproblems. Thus, their efficiency could be limited as learning SVMs is normally computationally demanding. Furthermore, due to the dependence on SVM solvers, it is not easy to parallelize these algorithms for large-scale data analysis tasks, although substantial efforts have been made to develop parallel Monte Carlo methods for unsupervised topic models (Newman et al., 2009; Smola and Narayanamurthy, 2010; Ahmed et al., 2012).

This paper presents Gibbs MedLDA, an alternative formulation of max-margin supervised topic models, for which we can develop simple and efficient inference algorithms. Technically, instead of minimizing the margin loss of an expected (averaging) prediction rule as adopted in existing max-margin topic models, Gibbs MedLDA minimizes the expected margin loss of many latent prediction rules, of which each rule corresponds to a configuration of topic assignments and the prediction model, drawn from a post-data posterior distribution. Theoretically, the expected margin loss is an upper bound of the existing margin loss of an expected prediction rule. Computationally, although the expected margin loss can still be hard to optimize using variational algorithms, we successfully develop simple and fast Gibbs sampling algorithms without any restrictive assumptions on the posterior distribution and without solving multiple latent SVM subproblems. By introducing a set of auxiliary variables and integrating out the Dirichlet variables by conjugacy, each of the

sampling substeps has a closed-form conditional distribution, from which samples can be efficiently drawn.

Our algorithms represent an extension of the classical ideas of data augmentation (Dempster et al., 1977; Tanner and Wong, 1987; van Dyk and Meng, 2001) and its recent developments in learning fully observed max-margin classifiers (Polson and Scott, 2011) to learn the sophisticated latent topic models. On the other hand, Gibbs MedLDA represents a generalization of Gibbs (or stochastic) classifiers (McAllester, 2003; Catoni, 2007; Germain et al., 2009) to incorporate a hierarchy of latent variables for discovering latent topic representations. We further generalize the ideas to develop a Gibbs MedLDA regression model and a multi-task Gibbs MedLDA model, for which we also develop efficient "augment-and-collapse" Gibbs sampling algorithms by exploring the same ideas of data augmentation. Empirical results on real data sets demonstrate significant improvements in time efficiency. The classification performance is also significantly improved in binary, multi-class, and multi-label classification tasks.

The rest of the paper is structured as follows. Section 2 summarizes some related work. Section 3 reviews MedLDA and its EM-type algorithms. Section 4 presents Gibbs MedLDA and its sampling algorithms for classification. Section 5 presents two extensions of Gibbs MedLDA for regression and multi-task learning. Section 6 presents empirical results. Finally, Section 7 concludes and discusses future directions.

2. Related Work

Max-margin learning has been very successful in building classifiers (Vapnik, 1995) and structured output prediction models (Taskar et al., 2003) in the last decade. Recently, research on learning max-margin models in the presence of latent variable models has received increasing attention because of the promise of using latent variables to capture the underlying structures of the complex problems. Deterministic approaches (Yu and Joachims, 2009) fill in the unknown values of the hidden structures by using some estimates (e.g., MAP estimates), and then a max-margin loss function is defined with the filled-in hidden structures, while probabilistic approaches aim to infer an entire distribution profile of the hidden structures given evidence and some prior distribution, following the Bayes' way of thinking. Though the former works well in practice, we focus on Bayesian approaches, which can naturally incorporate prior beliefs, maintain the entire distribution profile of latent structures, and be extensible to nonparametric methods. One representative work along this line is maximum entropy discrimination (MED) (Jaakkola et al., 1999; Jebara, 2001), which learns a distribution of model parameters given a set of labeled training data.

MedLDA (Zhu et al., 2012) is one extension of MED to infer hidden topical structures from data and MMH (max-margin Harmoniums) (Chen et al., 2012) is another extension that infers the hidden semantic features from multi-view data. Along similar lines, recent work has also successfully developed nonparametric Bayesian max-margin models, such as infinite SVMs (iSVM) (Zhu et al., 2011) for discovering clustering structures when building SVM classifiers and infinite latent SVMs (iLSVM) (Zhu et al., 2014) for automatically learning predictive features for SVM classifiers. Both iSVM and iLSVM can automatically resolve the model complexity, for example, the number of components in a mixture model or the number of latent features in a factor analysis model. The nonparametric Bayesian max-

margin ideas have been proven to be effective in dealing with more challenging problems, such as link prediction in social networks (Zhu, 2012) and low-rank matrix factorization for collaborative recommendation (Xu et al., 2012, 2013).

One common challenge of these Bayesian max-margin latent variable models is on the posterior inference, which is normally intractable. Almost all the existing work adopts a variational approximation scheme, with some mean-field assumptions. Very little research has been done on developing Monte Carlo methods, except the work (Jiang et al., 2012) which still makes mean-field assumptions. The work in the present paper provides a novel way to formulate Bayesian max-margin models and we show that these new formulations can have very simple and efficient Monte Carlo inference algorithms without making restrictive assumptions. The key step to deriving our algorithms is a data augmentation formulation of the expected margin-based loss. Other work on inferring the posterior distributions of latent variables includes max-margin min-entropy models (Miller et al., 2012) which learn a single set of model parameters, different from our focus of inferring the model posterior distribution.

Data augmentation refers to methods of augmenting the observed data so as to make it easy to analyze with an iterative optimization or sampling algorithm. For deterministic algorithms, the technique has been popularized in the statistics community by the seminal expectation-maximization (EM) algorithm (Dempster et al., 1977) for maximum likelihood estimation (MLE) with missing values. For stochastic algorithms, the technique has been popularized in statistics by Tanner and Wong's data augmentation algorithm for posterior sampling (Tanner and Wong, 1987) and in physics by Swendsen and Wang's sampling algorithms for Ising and Potts models (Swendsen and Wang, 1987). When using the idea to solve estimation or posterior inference problems, the key step is to find a set of augmented variables, conditioned on which the distribution of our models can be easily sampled. The speed of mixing or convergence is another important concern when designing a data augmentation method. While the conflict between simplicity and speed is a common phenomenon with many standard augmentation schemes, some work has demonstrated that with more creative augmentation schemes it is possible to construct EM-type algorithms (Meng and van Dyk, 1997) or Markov Chain Monte Carlo methods (known as slice sampling) (Neal, 1997) that are both fast and simple. We refer the readers to van Dyk and Meng (2001) for an excellent review of the broad literature of data augmentation and an effective search strategy for selecting good augmentation schemes.

For our focus on max-margin classifiers, the recent work (Polson and Scott, 2011) provides an elegant data augmentation formulation for support vector machines (SVM) with fully observed input data, which leads to analytical conditional distributions that are easy to sample from and fast to mix. Our work in the present paper builds on the method of Polson et al. and presents a successful implementation of data augmentation to deal with the challenging posterior inference problems of Bayesian max-margin latent topic models. Our approach can be generalized to deal with other Bayesian max-margin latent variable models, for example, max-margin matrix factorization (Xu et al., 2013), as reviewed above.

Finally, some preliminary results were presented in a conference paper (Zhu et al., 2013a). This paper presents a full extension.

3. MedLDA

We begin with a brief overview of MedLDA and its learning algorithms, which motivate our developments of Gibbs MedLDA.

3.1 MedLDA: A Regularized Bayesian Model

We consider binary classification with a labeled training set $\mathcal{D} = \{(\mathbf{w}_d, y_d)\}_{d=1}^{D}$, where $\mathbf{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ denotes the bag-of-words appearing in document d and the response variable Y takes values from the output space $\mathcal{Y} = \{-1, +1\}$. D is the data set size. Basically, MedLDA consists of two parts—an LDA model for describing input documents $\mathbf{W} = \{\mathbf{w}_d\}_{d=1}^{D}$, and an expected classifier for considering the supervising signal $\mathbf{y} = \{y_d\}_{d=1}^{D}$. Below, we introduce each of them in turn.

LDA: Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a hierarchical Bayesian model that posits each document as an admixture of K topics, where each topic $\boldsymbol{\Phi}_k$ is a multinomial distribution over a V-word vocabulary. For document d, the generating process can be described as

1. draw a topic proportion $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$

2. for each word n ($1 \leq n \leq N_d$):

    (a) draw a topic assignment[1] $z_{dn}|\boldsymbol{\theta}_d \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$

    (b) draw the observed word $w_{dn}|z_{dn}, \boldsymbol{\Phi} \sim \mathrm{Mult}(\boldsymbol{\Phi}_{z_{dn}})$

where $\mathrm{Dir}(\cdot)$ is a Dirichlet distribution; $\mathrm{Mult}(\cdot)$ is multinomial; and $\boldsymbol{\Phi}_{z_{dn}}$ denotes the topic selected by the non-zero entry of z_{dn}. For a fully-Bayesian LDA, the topics are random samples drawn from a prior, for example, $\boldsymbol{\Phi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$.

Given a set of documents $\mathbf{W}$, we let $\mathbf{z}_d = \{z_{dn}\}_{n=1}^{N_d}$ denote the set of topic assignments for document d and let $\mathbf{Z} = \{\mathbf{z}_d\}_{d=1}^{D}$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}_{d=1}^{D}$ denote all the topic assignments and mixing proportions for the whole corpus, respectively. Then, LDA infers the posterior distribution using Bayes' rule

$$p(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}|\mathbf{W}) = \frac{p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})}{p(\mathbf{W})},$$

where $p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) = \prod_{k=1}^{K} p_0(\boldsymbol{\Phi}_k|\boldsymbol{\beta}) \prod_{d=1}^{D} p_0(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(z_{dn}|\boldsymbol{\theta}_d)$ according to the generating process of LDA; and $p(\mathbf{W})$ is the marginal evidence. We can show that the posterior distribution by Bayes' rule is the solution of an information theoretical optimization problem

$$\min_{q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathrm{KL}\left[q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\|p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right] - \mathbb{E}_q\left[\log p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})\right]$$
$$\text{s.t.} : q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \in \mathcal{P}, \tag{1}$$

where $\mathrm{KL}(q\|p)$ is the Kullback-Leibler divergence and $\mathcal{P}$ is the space of probability distributions with an appropriate dimension. In fact, if we add the constant $\log p(\mathbf{W})$ to the objective, the problem is the minimization of the KL-divergence $\mathrm{KL}(q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\|p(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}|\mathbf{W}))$,

---

1. z_{dn} is a K-dimensional binary vector with only one nonzero entry.

whose solution is the desired posterior distribution by Bayes' rule. One advantage of this variational formulation of Bayesian inference is that it can be naturally extended to include some regularization terms on the desired post-data posterior distribution q. This insight has been taken to develop regularized Bayesian inference (RegBayes) (Zhu et al., 2014), a computational framework for doing Bayesian inference with posterior regularization.[2] As shown in Jiang et al. (2012) and detailed below, MedLDA is one example of RegBayes models. Moreover, as we shall see in Section 4, our Gibbs max-margin topic models follow this similar idea too.

Expected Classifier: Given a training set $\mathcal{D}$, an expected (or averaging) classifier chooses a posterior distribution $q(h|\mathcal{D})$ over a hypothesis space $\mathcal{H}$ of classifiers such that the q-weighted (expected) classifier

$$h_q(\mathbf{w}) = \text{sign } \mathbb{E}_q[h(\mathbf{w})]$$

will have the smallest possible risk. MedLDA follows this principle to learn a posterior distribution $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}|\mathcal{D})$ such that the expected classifier

$$\hat{y} = \text{sign } F(\mathbf{w}) \tag{2}$$

has the smallest possible risk, approximated by the training error $\mathcal{R}_{\mathcal{D}}(q) = \sum_{d=1}^{D} \mathbb{I}(\hat{y}_d \neq y_d)$. The discriminant function is defined as

$$F(\mathbf{w}) = \mathbb{E}_{q(\boldsymbol{\eta}, \mathbf{z}|\mathcal{D})}[F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w})],$$
$$F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w}) = \boldsymbol{\eta}^{\top}\bar{\mathbf{z}},$$

where $\bar{\mathbf{z}}$ is the average topic assignment associated with the words $\mathbf{w}$, a vector with element $\bar{z}_k = \frac{1}{N}\sum_{n=1}^{N} z_n^k$, and $\boldsymbol{\eta}$ is the classifier weights. Note that the expected classifier and the LDA likelihood are coupled via the latent topic assignments $\mathbf{Z}$. The strong coupling makes it possible for MedLDA to learn a posterior distribution that can describe the observed words well and make accurate predictions.

Regularized Bayesian Inference: To integrate the above two components for hybrid learning, MedLDA regularizes the properties of the topic representations by imposing the following max-margin constraints derived from the classifier (2) to a standard LDA inference problem (1)

$$y_d F(\mathbf{w}_d) \geq \ell - \xi_d, \ \forall d,$$

where ℓ (≥ 1) is the cost of making a wrong prediction; and $\boldsymbol{\xi} = \{\xi_d\}_{d=1}^{D}$ are non-negative slack variables for inseparable cases. Let $\mathcal{L}(q) = \text{KL}(q||p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})]$ be the objective for doing standard Bayesian inference with the classifier $\boldsymbol{\eta}$ and $p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) = p_0(\boldsymbol{\eta})p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$. MedLDA solves the regularized Bayesian inference (Zhu et al., 2014) problem

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \in \mathcal{P}, \boldsymbol{\xi}} \mathcal{L}\left(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right) + 2c\sum_{d=1}^{D} \xi_d \tag{3}$$

$$\text{s.t.:} \ \ y_d F(\mathbf{w}_d) \geq \ell - \xi_d, \ \xi_d \geq 0, \forall d,$$

---

2. Posterior regularization was first used in Ganchev et al. (2010) for maximum likelihood estimation and was later extended in Zhu et al. (2014) to Bayesian and nonparametric Bayesian methods.

where the margin constraints directly regularize the properties of the post-data distribution and c is the positive regularization parameter. Equivalently, MedLDA solves the unconstrained problem[3]

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})} \mathcal{L}\left(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\right) + 2c\mathcal{R}\left(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\right), \tag{4}$$

where $\mathcal{R}(q) = \sum_{d=1}^{D} \max(0, \ell - y_d F(\mathbf{w}_d))$ is the hinge loss that upper-bounds the training error $\mathcal{R}_{\mathcal{D}}(q)$ of the expected classifier (2). Note that the constant 2 is included simply for convenience.

3.2 Existing Iterative Algorithms

Since it is difficult to solve problem (3) or (4) directly because of the non-conjugacy (between priors and likelihood) and the max-margin constraints, corresponding to a non-smooth posterior regularization term in (4), both variational and Monte Carlo methods have been developed for approximate solutions. It can be shown that the variational method (Zhu et al., 2012) is a coordinate descent algorithm to solve problem (4) with the fully-factorized assumption that

$$q(\boldsymbol{\eta},\Theta,\mathbf{Z},\boldsymbol{\Phi}) = q(\boldsymbol{\eta})\left(\prod_{d=1}^{D} q(\boldsymbol{\theta}_d)\prod_{n=1}^{N_d} q(z_{dn})\right)\prod_{k=1}^{K} q(\boldsymbol{\Phi}_k);$$

while the Monte Carlo methods (Jiang et al., 2012) make a weaker assumption that

$$q(\boldsymbol{\eta},\Theta,\mathbf{Z},\boldsymbol{\Phi}) = q(\boldsymbol{\eta})q(\Theta,\mathbf{Z},\boldsymbol{\Phi}).$$

All these methods have a similar EM-type iterative procedure, which solves many latent SVM subproblems, as outlined below.

Estimate $q(\boldsymbol{\eta})$: Given $q(\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$, we solve problem (4) with respect to $q(\boldsymbol{\eta})$. In the equivalent constrained form, this step solves

$$\min_{q(\boldsymbol{\eta}),\boldsymbol{\xi}} \quad \mathrm{KL}\left(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})\right) + 2c\sum_{d=1}^{D}\xi_d \tag{5}$$
$$\text{s.t.}: \quad y_d\mathbb{E}_q[\boldsymbol{\eta}]^{\top}\mathbb{E}_q[\bar{\mathbf{z}}_d] \geq \ell - \xi_d, \ \xi_d \geq 0, \forall d.$$

This problem is convex and can be solved with Lagrangian methods. Specifically, let μ_d be the Lagrange multipliers, one per constraint. When the prior $p_0(\boldsymbol{\eta})$ is the commonly used standard normal distribution, we have the optimum solution $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\kappa}, I)$, where $\boldsymbol{\kappa} = \sum_{d=1}^{D} y_d\mu_d\mathbb{E}_q[\bar{\mathbf{z}}_d]$. It can be shown that the dual problem of (5) is the dual of a standard binary linear SVM and we can solve it or its primal form efficiently using existing high-performance SVM learners. We denote the optimum solution of this problem by $(q^*(\boldsymbol{\eta}), \boldsymbol{\kappa}^*, \boldsymbol{\xi}^*, \boldsymbol{\mu}^*)$.

---

3. If not specified, q is subject to the constraint $q \in \mathcal{P}$.

Estimate $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$: Given $q(\boldsymbol{\eta})$, we solve problem (4) with respect to $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$. In the constrained form, this step solves

$$\min_{q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}), \boldsymbol{\xi}} \quad \mathcal{L}\left(q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right) + 2c \sum_{d=1}^{D} \xi_d \tag{6}$$
$$\text{s.t.}: \ y_d (\boldsymbol{\kappa}^*)^\top \mathbb{E}_q[\bar{\mathbf{z}}_d] \geq \ell - \xi_d, \ \xi_d \geq 0, \forall d.$$

Although we can solve this problem using Lagrangian methods, it would be hard to derive the dual objective. An effective approximation strategy was used in Zhu et al. (2012) and Jiang et al. (2012), which updates $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ for only one step with $\boldsymbol{\xi}$ fixed at $\boldsymbol{\xi}^*$. By fixing $\boldsymbol{\xi}$ at $\boldsymbol{\xi}^*$, we have the solution

$$q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \propto p(\mathbf{W}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \exp \left\{ (\boldsymbol{\kappa}^*)^\top \sum_d \mu_d^* \bar{\mathbf{z}}_d \right\},$$

where the second term indicates the regularization effects due to the max-margin posterior constraints. For those data with non-zero Lagrange multipliers (i.e., support vectors), the second term will bias MedLDA towards a new posterior distribution that favors more discriminative representations on these "hard" data points. The Monte Carlo methods (Jiang et al., 2012) directly draw samples from the posterior distribution $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ or its collapsed form using Gibbs sampling to estimate $\mathbb{E}_q[\bar{\mathbf{z}}_d]$, the expectations required to learn $q(\boldsymbol{\eta})$. In contrast, the variational methods (Zhu et al., 2012) solve problem (6) using coordinate descent to estimate $\mathbb{E}_q[\bar{\mathbf{z}}_d]$ with a fully factorized assumption.

4. Gibbs MedLDA

Now, we present Gibbs max-margin topic models for binary classification and their "augment-and-collapse" sampling algorithms. We will discuss further extensions in the next section.

4.1 Learning with an Expected Margin Loss

As stated above, MedLDA chooses the strategy to minimize the hinge loss of an expected classifier. In learning theory, an alternative approach to building classifiers with a posterior distribution of models is to minimize an expected loss, under the framework known as Gibbs classifiers (or stochastic classifiers) (McAllester, 2003; Catoni, 2007; Germain et al., 2009) which have nice theoretical properties on generalization performance.

For our case of inferring the distribution of latent topic assignments $\mathbf{Z} = \{\mathbf{z}_d\}_{d=1}^{D}$ and the classification model $\boldsymbol{\eta}$, the expected margin loss is defined as follows. If we have drawn a sample of the topic assignments $\mathbf{Z}$ and the prediction model $\boldsymbol{\eta}$ from a posterior distribution $q(\boldsymbol{\eta}, \mathbf{Z})$, we can define the linear discriminant function

$$F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w}) = \boldsymbol{\eta}^\top \bar{\mathbf{z}}$$

as before and make prediction using the *latent prediction rule*

$$\hat{y}(\boldsymbol{\eta}, \mathbf{z}) = \text{sign} \ F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w}). \tag{7}$$

Note that the prediction is a function of the configuration $(\boldsymbol{\eta}, \mathbf{z})$. Let $\zeta_d = \ell - y_d \boldsymbol{\eta}^\top \bar{\mathbf{z}}_d$, where ℓ is a cost parameter as defined before. The hinge loss of the stochastic classifier is

$$\mathcal{R}(\boldsymbol{\eta}, \mathbf{Z}) = \sum_{d=1}^{D} \max(0, \zeta_d),$$

a function of the latent variables $(\boldsymbol{\eta}, \mathbf{Z})$, and the expected hinge loss is

$$\mathcal{R}'(q) = \mathbb{E}_q[\mathcal{R}(\boldsymbol{\eta}, \mathbf{Z})] = \sum_{d=1}^{D} \mathbb{E}_q \left[\max(0, \zeta_d) \right],$$

a functional of the posterior distribution $q(\boldsymbol{\eta}, \mathbf{Z})$. Since for any $(\boldsymbol{\eta}, \mathbf{Z})$, the hinge loss $\mathcal{R}(\boldsymbol{\eta}, \mathbf{Z})$ is an upper bound of the training error of the latent Gibbs classifier (7), that is,

$$\mathcal{R}(\boldsymbol{\eta}, \mathbf{Z}) \geq \sum_{d=1}^{D} \mathbb{I} \left(y_d \neq \hat{y}_d(\boldsymbol{\eta}, \mathbf{z}_d) \right),$$

we have

$$\mathcal{R}'(q) \geq \sum_{d=1}^{D} \mathbb{E}_q \left[\mathbb{I}(y_d \neq \hat{y}_d(\boldsymbol{\eta}, \mathbf{z}_d)) \right],$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals to 1 if the predicate holds otherwise 0. In other words, the expected hinge loss $\mathcal{R}'(q)$ is an upper bound of the expected training error of the Gibbs classifier (7). Thus, it is a good surrogate loss for learning a posterior distribution which could lead to a low training error in expectation.

Then, with the same goal as MedLDA of finding a posterior distribution $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ that on one hand describes the observed data and on the other hand predicts as well as possible on training data, we define Gibbs MedLDA as solving the new regularized Bayesian inference problem

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathcal{L} \left(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \right) + 2c\mathcal{R}' \left(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \right). \tag{8}$$

Note that we have written the expected margin loss $\mathcal{R}'$ as a function of the complete distribution $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$. This does not conflict with our definition of $\mathcal{R}'$ as a function of the marginal distribution $q(\boldsymbol{\eta}, \mathbf{Z})$ because the other irrelevant variables (i.e., $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$) are integrated out when we compute the expectation.

Comparing to MedLDA in problem (4), we have the following lemma by applying Jensen's inequality.

Lemma 1 *The expected hinge loss $\mathcal{R}'$ is an upper bound of the hinge loss of the expected classifier (2):*

$$\mathcal{R}'(q) \geq \mathcal{R}(q) = \sum_{d=1}^{D} \max \left(0, \mathbb{E}_q[\zeta_d] \right);$$

and thus the objective in (8) is an upper bound of that in (4) when c values are the same.

4.2 Formulation with Data Augmentation

If we directly solve problem (8), the expected hinge loss $\mathcal{R}'$ is hard to deal with because of the non-differentiable max function. Fortunately, we can develop a simple collapsed Gibbs sampling algorithm with analytical forms of local conditional distributions, based on a data augmentation formulation of the expected hinge-loss.

Let $\phi(y_d|\mathbf{z}_d, \boldsymbol{\eta}) = \exp\{-2c\max(0, \zeta_d)\}$ be the unnormalized likelihood of the response variable for document d. Then, problem (8) can be written as

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathcal{L}\left(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right) - \mathbb{E}_q\left[\log \phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta})\right], \tag{9}$$

where $\phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta}) = \prod_{d=1}^{D} \phi(y_d|\mathbf{z}_d, \boldsymbol{\eta})$. Solving problem (9) with the constraint that $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \in \mathcal{P}$, we can get the normalized posterior distribution

$$q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) = \frac{p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})\phi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta})}{\psi(\mathbf{y}, \mathbf{W})},$$

where $\psi(\mathbf{y}, \mathbf{W})$ is the normalization constant that depends on the observed data only. Due to the complicated form of ϕ, it will not have simple conditional distributions if we want to derive a Gibbs sampling algorithm for $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ directly. This motivates our exploration of data augmentation techniques. Specifically, using the ideas of data augmentation (Tanner and Wong, 1987; Polson and Scott, 2011), we have Lemma 2.

Lemma 2 (Scale Mixture Representation) *The unnormalized likelihood can be expressed as*

$$\phi(y_d|\mathbf{z}_d, \boldsymbol{\eta}) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right) \mathrm{d}\lambda_d.$$

Proof Due to the fact that $a\max(0, x) = \max(0, ax)$ if $a \geq 0$, we have $-2c\max(0, \zeta_d) = -2\max(0, c\zeta_d)$. Then, we can follow the proof in Polson and Scott (2011) to get the results. ∎

Lemma 2 indicates that the posterior distribution of Gibbs MedLDA can be expressed as the marginal of a higher-dimensional distribution that includes the augmented variables $\boldsymbol{\lambda} = \{\lambda_d\}_{d=1}^{D}$, that is,

$$q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) = \int_0^\infty \cdots \int_0^\infty q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_D = \int_{\mathbb{R}_+^D} q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\mathrm{d}\boldsymbol{\lambda},$$

where $\mathbb{R}_+ = \{x : x \in \mathbb{R}, \ x > 0\}$ is the set of positive real numbers; the complete posterior distribution is

$$q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) = \frac{p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})\phi(\mathbf{y}, \boldsymbol{\lambda}|\mathbf{Z}, \boldsymbol{\eta})}{\psi(\mathbf{y}, \mathbf{W})};$$

and the unnormalized joint distribution of $\mathbf{y}$ and $\boldsymbol{\lambda}$ is

$$\phi(\mathbf{y}, \boldsymbol{\lambda}|\mathbf{Z}, \boldsymbol{\eta}) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right).$$

In fact, we can show that the complete posterior distribution is the solution of the data augmentation problem of Gibbs MedLDA

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathcal{L}\left(q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right) - \mathbb{E}_q\left[\log \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta})\right],$$

which is again subject to the normalization constraint that $q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \in \mathcal{P}$. The first term in the objective is $\mathcal{L}(q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) = \mathrm{KL}(q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) || p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) p_0(\boldsymbol{\lambda})) - \mathbb{E}_q[\log p(\mathbf{W} | \mathbf{Z}, \boldsymbol{\Phi})]$, where a prior distribution is imposed on the augmented variables $\boldsymbol{\lambda}$. One good choice of $p_0(\boldsymbol{\lambda})$ is the noninformative uniform prior.

Remark 3 *The objective of this augmented problem is an upper bound of the objective in (9) (thus, also an upper bound of MedLDA's objective due to Lemma 1). This is because by using the data augmentation we can show that*

$$\mathbb{E}_{q(\mathbf{V})}[\log \phi(\mathbf{y} | \mathbf{Z}, \boldsymbol{\eta})] = \mathbb{E}_{q(\mathbf{V})} \left[\log \int_{\mathbb{R}_+^D} \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta}) \mathrm{d}\boldsymbol{\lambda} \right]$$

$$= \mathbb{E}_{q(\mathbf{V})} \left[\log \int_{\mathbb{R}_+^D} \frac{q(\boldsymbol{\lambda} | \mathbf{V})}{q(\boldsymbol{\lambda} | \mathbf{V})} \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta}) \mathrm{d}\boldsymbol{\lambda} \right]$$

$$\geq \mathbb{E}_{q(\mathbf{V})} \left[\mathbb{E}_{q(\boldsymbol{\lambda} | \mathbf{V})} [\log \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta})] - \mathbb{E}_{q(\boldsymbol{\lambda} | \mathbf{V})} [\log q(\boldsymbol{\lambda} | \mathbf{V})] \right]$$

$$= \mathbb{E}_{q(\mathbf{V}, \boldsymbol{\lambda})} [\log \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta})] - \mathbb{E}_{q(\mathbf{V}, \boldsymbol{\lambda})} [\log q(\boldsymbol{\lambda} | \mathbf{V})],$$

where $\mathbf{V} = \{\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}\}$ *denotes all the random variables in MedLDA. Therefore, we have*

$$\mathcal{L}(q(\mathbf{V})) - \mathbb{E}_{q(\mathbf{V})}[\log \phi(\mathbf{y} | \mathbf{Z}, \boldsymbol{\eta})] \leq \mathcal{L}(q(\mathbf{V})) - \mathbb{E}_{q(\mathbf{V}, \boldsymbol{\lambda})} [\log \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta})] + \mathbb{E}_{q(\mathbf{V}, \boldsymbol{\lambda})} [\log q(\boldsymbol{\lambda} | \mathbf{V})]$$

$$= \mathcal{L}(q(\mathbf{V}, \boldsymbol{\lambda})) - \mathbb{E}_q[\log \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta})].$$

4.3 Inference with Collapsed Gibbs Sampling

Although with the above data augmentation formulation we can do Gibbs sampling to infer the complete posterior distribution $q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ and thus $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ by ignoring $\boldsymbol{\lambda}$, the mixing speed would be slow because of the large sample space of the latent variables. One way to effectively reduce the sample space and improve mixing rates is to integrate out the intermediate Dirichlet variables $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ and build a Markov chain whose equilibrium distribution is the resulting marginal distribution $q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \mathbf{Z})$. We propose to use collapsed Gibbs sampling, which has been successfully used in LDA (Griffiths and Steyvers, 2004). With the data augmentation representation, this leads to our "augment-and-collapse" sampling algorithm for Gibbs MedLDA, as detailed below.

For the data augmented formulation of Gibbs MedLDA, by integrating out the Dirichlet variables $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$, we get the collapsed posterior distribution:

$$q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \mathbf{Z}) \propto p_0(\boldsymbol{\eta}) p(\mathbf{W}, \mathbf{Z} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \phi(\mathbf{y}, \boldsymbol{\lambda} | \mathbf{Z}, \boldsymbol{\eta})$$

$$= p_0(\boldsymbol{\eta}) \left[\prod_{d=1}^{D} \frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \right] \prod_{k=1}^{K} \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d} \right),$$

where

$$\delta(\mathbf{x}) = \frac{\prod_{i=1}^{\dim(\mathbf{x})} \Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim(\mathbf{x})} x_i)};$$

$\Gamma(\cdot)$ is the Gamma function; C_k^t is the number of times that the term t is assigned to topic k over the whole corpus; $\mathbf{C}_k = \{C_k^t\}_{t=1}^V$ is the set of word counts associated with topic k; C_d^k is the number of times that terms are associated with topic k within the d-th document; and $\mathbf{C}_d = \{C_d^k\}_{k=1}^K$ is the set of topic counts for document d. Then, the conditional distributions used in collapsed Gibbs sampling are as follows.

For $\boldsymbol{\eta}$: Let us assume its prior is the commonly used isotropic Gaussian distribution $p_0(\boldsymbol{\eta}) = \prod_{k=1}^K \mathcal{N}(\eta_k; 0, \nu^2)$, where ν is a non-zero parameter. Then, we have the conditional distribution of $\boldsymbol{\eta}$ given the other variables:

$$q(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}) \propto p_0(\boldsymbol{\eta}) \prod_{d=1}^D \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right)$$

$$\propto \exp\left(-\sum_{k=1}^K \frac{\eta_k^2}{2\nu^2} - \sum_{d=1}^D \frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right)$$

$$= \exp\left(-\frac{1}{2}\boldsymbol{\eta}^\top \left(\frac{1}{\nu^2}I + c^2 \sum_{d=1}^D \frac{\bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top}{\lambda_d}\right)\boldsymbol{\eta} + \left(c \sum_{d=1}^D y_d \frac{\lambda_d + c\ell}{\lambda_d} \bar{\mathbf{z}}_d\right)^\top \boldsymbol{\eta}\right)$$

$$= \mathcal{N}(\boldsymbol{\eta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{10}$$

a K-dimensional Gaussian distribution, where the posterior mean and the covariance matrix are

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(c \sum_{d=1}^D y_d \frac{\lambda_d + c\ell}{\lambda_d} \bar{\mathbf{z}}_d\right), \text{ and } \boldsymbol{\Sigma} = \left(\frac{1}{\nu^2}I + c^2 \sum_{d=1}^D \frac{\bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top}{\lambda_d}\right)^{-1}.$$

Therefore, we can easily draw a sample from this multivariate Gaussian distribution. The inverse can be robustly done using Cholesky decomposition, an $O(K^3)$ procedure. Since K is normally not large, the inversion can be done efficiently, especially in applications where the number of documents is much larger than the number of topics.

For $\mathbf{Z}$: The conditional distribution of $\mathbf{Z}$ given the other variables is

$$q(\mathbf{Z}|\boldsymbol{\eta}, \boldsymbol{\lambda}) \propto \prod_{d=1}^D \frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right) \prod_{k=1}^K \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})}.$$

By canceling common factors, we can derive the conditional distribution of one variable z_{dn} given others $\mathbf{Z}_\neg$ as:

$$q(z_{dn}^k = 1|\mathbf{Z}_\neg, \boldsymbol{\eta}, \boldsymbol{\lambda}, w_{dn} = t) \propto \frac{(C_{k,\neg n}^t + \beta_t)(C_{d,\neg n}^k + \alpha_k)}{\sum_{t=1}^V C_{k,\neg n}^t + \sum_{t=1}^V \beta_t} \exp\left(\frac{\gamma y_d(c\ell + \lambda_d)\eta_k}{\lambda_d}\right.$$

$$\left. -c^2 \frac{\gamma^2 \eta_k^2 + 2\gamma(1-\gamma)\eta_k \Lambda_{dn}^k}{2\lambda_d}\right), \tag{11}$$

where $C_{\cdot,\neg n}$ indicates that term n is excluded from the corresponding document or topic; $\gamma = \frac{1}{N_d}$; and

$$\Lambda_{dn}^k = \frac{1}{N_d - 1} \sum_{k'=1}^K \eta_{k'} C_{d,\neg n}^{k'}$$

Algorithm 1 collapsed Gibbs sampling algorithm for Gibbs MedLDA classification models

1: **Initialization:** set $\boldsymbol{\lambda} = 1$ and randomly draw z_{dk} from a uniform distribution.
2: **for** $m = 1$ **to** M **do**
3:     draw the classifier from the normal distribution (10)
4:     **for** $d = 1$ **to** D **do**
5:        **for** each word n in document d **do**
6:           draw a topic from the multinomial distribution (11)
7:        **end for**
8:        draw λ_d^{-1} (and thus λ_d) from the inverse Gaussian distribution (12).
9:     **end for**
10: **end for**

---

is the discriminant function value without word n. We can see that the first term is from the LDA model for observed word counts and the second term is from the supervised signal **y**.

For $\boldsymbol{\lambda}$: Finally, the conditional distribution of the augmented variables $\boldsymbol{\lambda}$ given the other variables is factorized and we can derive the conditional distribution for each λ_d as

$$
\begin{aligned}
q(\lambda_d | \mathbf{Z}, \boldsymbol{\eta}) &\propto \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right) \\
&\propto \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{c^2\zeta_d^2}{2\lambda_d} - \frac{\lambda_d}{2}\right) \\
&= \mathcal{GIG}\left(\lambda_d; \frac{1}{2}, 1, c^2\zeta_d^2\right),
\end{aligned}
$$

where

$$
\mathcal{GIG}(x; p, a, b) = C(p, a, b) x^{p-1} \exp\left(-\frac{1}{2}\left(\frac{b}{x} + ax\right)\right)
$$

is a generalized inverse Gaussian distribution (Devroye, 1986) and $C(p, a, b)$ is a normalization constant. Therefore, we can derive that λ_d^{-1} follows an inverse Gaussian distribution

$$
p(\lambda_d^{-1} | \mathbf{Z}, \boldsymbol{\eta}) = \mathcal{IG}\left(\lambda_d^{-1}; \frac{1}{c|\zeta_d|}, 1\right), \tag{12}
$$

where

$$
\mathcal{IG}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}} \exp\left(-\frac{b(x-a)^2}{2a^2 x}\right)
$$

for $a > 0$ and $b > 0$.

With the above conditional distributions, we can construct a Markov chain which iteratively draws samples of the classifier weights $\boldsymbol{\eta}$ using Equation (10), the topic assignments $\mathbf{Z}$ using Equation (11) and the augmented variables $\boldsymbol{\lambda}$ using Equation (12), with an initial condition. To sample from an inverse Gaussian distribution, we apply the transformation method with multiple roots (Michael et al., 1976) which is very efficient with a constant time complexity. Overall, the per-iteration time complexity is $\mathcal{O}(K^3 + N_{total}K)$, where $N_{total} = \sum_{d=1}^{D} N_d$ is the total number of words in all documents. If K is not very large

(e.g., $K \ll \sqrt{N_{total}}$), which is the common case in practice as N_{total} is often very large, the per-iteration time complexity is $\mathcal{O}(N_{total}K)$; if K is large (e.g., $K \gg \sqrt{N_{total}}$), which is not common in practice, drawing the global classifier weights will dominate and the per-iteration time complexity is $\mathcal{O}(K^3)$. In our experiments, we initially set $\lambda_d = 1$, $\forall d$ and randomly draw $\mathbf{Z}$ from a uniform distribution. In training, we run this Markov chain to finish the burn-in stage with M iterations, as outlined in Algorithm 1. Then, we draw a sample $\hat{\boldsymbol{\eta}}$ as the Gibbs classifier to make predictions on testing data.

In general, there is no theoretical guarantee that a Markov chain constructed using data augmentation can converge to the target distribution; see Hobert (2011) for a failure example. However, for our algorithms, we can justify that the Markov transition distribution of the chain satisfies the condition $\mathcal{K}$ from Hobert (2011), that is, the transition probability from one state to any other state is larger than 0. Condition $\mathcal{K}$ implies that the Markov chain is Harris ergodic (Tan, 2009, Lemma 1). Therefore, no matter how the chain is started, our sampling algorithms can be employed to effectively explore the intractable posterior distribution. In practice, the sampling algorithm as well as the ones to be presented require only a few iterations to get stable prediction performance, as we shall see in Section 6.5.1. More theoretical analysis such as convergence rates requires a good bit of technical Markov chain theory and is our future work.

4.4 Prediction

To apply the Gibbs classifier $\hat{\boldsymbol{\eta}}$, we need to infer the topic assignments for testing document, denoted by $\mathbf{w}$. A fully Bayesian treatment needs to compute an integral in order to get the posterior distribution of the topic assignment given the training data $\mathcal{D}$ and the testing document content $\mathbf{w}$:

$$p(\mathbf{z}|\mathbf{w}, \mathcal{D}) \propto \int_{\mathcal{P}_V^K} p(\mathbf{z}, \mathbf{w}, \Phi|\mathcal{D})\mathrm{d}\Phi = \int_{\mathcal{P}_V^K} p(\mathbf{z}, \mathbf{w}|\Phi)p(\Phi|\mathcal{D})\mathrm{d}\Phi,$$

where $\mathcal{P}_V$ is the $V-1$ dimensional simplex; and the second equality holds due to the conditional independence assumption of the documents given the topics. Various approximation methods can be applied to compute the integral. Here, we take the approach applied in Zhu et al. (2012), which uses a point estimate of topics $\boldsymbol{\Phi}$ from training data and makes predictions based on them. Specifically, we use a point estimate $\hat{\boldsymbol{\Phi}}$ (a Dirac measure) to approximate the probability distribution $p(\boldsymbol{\Phi}|\mathcal{D})$. For the collapsed Gibbs sampler, an estimate of $\hat{\boldsymbol{\Phi}}$ using the samples is the posterior mean

$$\hat{\phi}_{kt} \propto C_k^t + \beta_t.$$

Then, given a testing document $\mathbf{w}$, we infer its latent components $\mathbf{z}$ using $\hat{\boldsymbol{\Phi}}$ by drawing samples from the local conditional distribution

$$p(z_n^k = 1|\mathbf{z}_{\neg n}, \mathbf{w}, \mathcal{D}) \propto \hat{\phi}_{kw_n} \left(C_{\neg n}^k + \alpha_k \right), \tag{13}$$

where $C_{\neg n}^k$ is the number of times that the terms in this document $\mathbf{w}$ assigned to topic k with the n-th term excluded. To start the sampler, we randomly set each word to one topic. Then, we run the Gibbs sampler for a few iterations until some stop criterion is satisfied,

for example, after a few burn-in steps or the relative change of data likelihood is lower than some threshold. Here, we adopt the latter, the same as in Jiang et al. (2012). After this burn-in stage, we keep one sample of $\mathbf{z}$ for prediction using the stochastic classifier. Empirically, using the average of a few (e.g., 10) samples of $\mathbf{z}$ could lead to slightly more robust predictions, as we shall see in Section 6.5.4.

5. Extensions to Regression and Multi-task Learning

The above ideas can be naturally generalized to develop Gibbs max-margin supervised topic models for various prediction tasks. In this section, we present two examples for regression and multi-task learning, respectively.

5.1 Gibbs MedLDA Regression Model

We first discuss how to generalize the above ideas to develop a regression model, where the response variable Y takes real values. Formally, the Gibbs MedLDA regression model also has two components—an LDA model to describe input bag-of-words documents and a Gibbs regression model for the response variables. Since the LDA component is the same as in the classification model, we focus on presenting the Gibbs regression model.

5.1.1 The Models with Data Augmentation

If a sample of the topic assignments $\mathbf{Z}$ and the prediction model $\boldsymbol{\eta}$ are drawn from the posterior distribution $q(\boldsymbol{\eta}, \mathbf{Z})$, we define the latent regression rule as

$$\hat{y}(\boldsymbol{\eta}, \mathbf{z}) = \boldsymbol{\eta}^{\top} \bar{\mathbf{z}}. \tag{14}$$

To measure the goodness of the prediction rule (14), we adopt the widely used ϵ-insensitive loss

$$\mathcal{R}_{\epsilon}(\boldsymbol{\eta}, \mathbf{Z}) = \sum_{d=1}^{D} \max\left(0, |\Delta_d| - \epsilon\right),$$

where $\Delta_d = y_d - \boldsymbol{\eta}^{\top} \bar{\mathbf{z}}_d$ is the margin between the true score and the predicted score. The ϵ-insensitive loss has been successfully used in learning fully observed support vector regression (Smola and Scholkopf, 2003). In our case, the loss is a function of predictive model $\boldsymbol{\eta}$ as well as the topic assignments $\mathbf{Z}$ which are hidden from the input data. To resolve this uncertainty, we define the expected ϵ-insensitive loss

$$\mathcal{R}_{\epsilon}(q) = \mathbb{E}_q\left[\mathcal{R}_{\epsilon}(\boldsymbol{\eta}, \mathbf{Z})\right] = \sum_{d=1}^{D} \mathbb{E}_q\left[\max(0, |\Delta_d| - \epsilon)\right],$$

a function of the desired posterior distribution $q(\boldsymbol{\eta}, \mathbf{Z})$.

With the above definitions, we can follow the same principle as Gibbs MedLDA to define the Gibbs MedLDA regression model as solving the regularized Bayesian inference problem

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathcal{L}\left(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right) + 2c\mathcal{R}_{\epsilon}\left(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right). \tag{15}$$

Note that as in the classification model, we have put the complete distribution $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ as the argument of the expected loss $\mathcal{R}_\epsilon$, which only depends on the marginal distribution $q(\boldsymbol{\eta}, \mathbf{Z})$. This does not affect the results because we are taking the expectation to compute $\mathcal{R}_\epsilon$ and any irrelevant variables will be marginalized out.

As in the Gibbs MedLDA classification model, we can show that $\mathcal{R}_\epsilon$ is an upper bound of the ϵ-insensitive loss of MedLDA's expected prediction rule, by applying Jensen's inequality to the convex function $h(x) = \max(0, |x| - \epsilon)$.

Lemma 4 *We have $\mathcal{R}_\epsilon \geq \sum_{d=1}^D \max(0, |\mathbb{E}_q[\Delta_d]| - \epsilon)$.*

We can reformulate problem (15) in the same form as problem (9), with the unnormalized likelihood

$$\phi(y_d | \boldsymbol{\eta}, \mathbf{z}_d) = \exp\left(-2c \max(0, |\Delta_d| - \epsilon)\right).$$

Then, we have the dual scale of mixture representation, by noting that

$$\max(0, |x| - \epsilon) = \max(0, x - \epsilon) + \max(0, -x - \epsilon). \tag{16}$$

Lemma 5 (Dual Scale Mixture Representation) *For regression, the unnormalized likelihood can be expressed as*

$$\phi(y_d | \boldsymbol{\eta}, \mathbf{z}_d) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c(\Delta_d - \epsilon))^2}{2\lambda_d}\right) \mathrm{d}\lambda_d$$
$$\times \int_0^\infty \frac{1}{\sqrt{2\pi\omega_d}} \exp\left(-\frac{(\omega_d - c(\Delta_d + \epsilon))^2}{2\omega_d}\right) \mathrm{d}\omega_d.$$

Proof By the equality (16), we have $\phi(y_d | \boldsymbol{\eta}, \mathbf{z}_d) = \exp\{-2c \max(0, \Delta_d - \epsilon)\}$ $\exp\{-2c \max(0, -\Delta_d - \epsilon)\}$. Each of the exponential terms can be formulated as a scale mixture of Gaussians due to Lemma 2. ∎

Then, the data augmented learning problem of the Gibbs MedLDA regression model is

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathcal{L}\left(q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})\right) - \mathbb{E}_q\left[\log \phi(\mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\omega} | \mathbf{Z}, \boldsymbol{\eta})\right],$$

where $\phi(\mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\omega} | \mathbf{Z}, \boldsymbol{\eta}) = \prod_{d=1}^D \phi(y_d, \lambda_d, \omega_d | \mathbf{Z}, \boldsymbol{\omega})$ and

$$\phi(y_d, \lambda_d, \omega_d | \mathbf{Z}, \boldsymbol{\eta}) = \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c(\Delta_d - \epsilon))^2}{2\lambda_d}\right) \frac{1}{\sqrt{2\pi\omega_d}} \exp\left(-\frac{(\omega_d - c(\Delta_d + \epsilon))^2}{2\omega_d}\right).$$

Solving the augmented problem and integrating out $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$, we can get the collapsed posterior distribution

$$q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{Z}) \propto p_0(\boldsymbol{\eta}) p(\mathbf{W}, \mathbf{Z} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \phi(\mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\omega} | \mathbf{Z}, \boldsymbol{\eta}).$$

Algorithm 2 collapsed Gibbs sampling algorithm for Gibbs MedLDA regression models

---
1: **Initialization:** set $\boldsymbol{\lambda} = 1$ and randomly draw z_{dk} from a uniform distribution.
2: **for** $m = 1$ **to** M **do**
3:     draw the classifier from the normal distribution (17)
4:     **for** $d = 1$ **to** D **do**
5:         **for** each word n in document d **do**
6:             draw a topic from the multinomial distribution (18)
7:         **end for**
8:         draw λ_d^{-1} (and thus λ_d) from the inverse Gaussian distribution (19).
9:         draw ω_d^{-1} (and thus ω_d) from the inverse Gaussian distribution (20).
10:     **end for**
11: **end for**

5.1.2 A Collapsed Gibbs Sampling Algorithm

Following similar derivations as in the classification model, the Gibbs sampling algorithm to infer the posterior has the following conditional distributions, with an outline in Algorithm 2.

For $\boldsymbol{\eta}$: Again, with the isotropic Gaussian prior $p_0(\boldsymbol{\eta}) = \prod_{k=1}^{K} \mathcal{N}(\eta_k; 0, \nu^2)$, we have

$$
q(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\omega}) \propto p_0(\boldsymbol{\eta}) \prod_{d=1}^{D} \exp\left(-\frac{(\lambda_d + c(\Delta_d - \epsilon))^2}{2\lambda_d}\right) \exp\left(-\frac{(\omega_d - c(\Delta_d + \epsilon))^2}{2\omega_d}\right)
$$

$$
\propto \exp\left(-\sum_{k=1}^{K} \frac{\eta_k^2}{2\nu^2} - \sum_{d=1}^{D}\left(\frac{(\lambda_d + c(\Delta_d - \epsilon))^2}{2\lambda_d} + \frac{(\omega_d - c(\Delta_d + \epsilon))^2}{2\omega_d}\right)\right)
$$

$$
= \exp\left(-\frac{1}{2}\boldsymbol{\eta}^\top \left(\frac{1}{\nu^2}I + c^2 \sum_{d=1}^{D} \rho_d \bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top\right)\boldsymbol{\eta} + c\left(\sum_{d=1}^{D} \psi_d \bar{\mathbf{z}}_d\right)^\top \boldsymbol{\eta}\right)
$$

$$
= \mathcal{N}(\boldsymbol{\eta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{17}
$$

where the posterior covariance matrix and the posterior mean are

$$
\boldsymbol{\Sigma} = \left(\frac{1}{\nu^2}I + c^2 \sum_{d=1}^{D} \rho_d \bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top\right)^{-1}, \quad \boldsymbol{\mu} = c\boldsymbol{\Sigma}\left(\sum_{d=1}^{D} \psi_d \bar{\mathbf{z}}_d\right),
$$

and $\rho_d = \frac{1}{\lambda_d} + \frac{1}{\omega_d}$ and $\psi_d = \frac{y_d - \epsilon}{\lambda_d} + \frac{y_d + \epsilon}{\omega_d}$ are two parameters. We can easily draw a sample from a K-dimensional multivariate Gaussian distribution. The inverse can be robustly done using Cholesky decomposition.

For $\mathbf{Z}$: We can derive the conditional distribution of one variable z_{dn} given others $\mathbf{Z}_{\neg}$ as:

$$
q(z_{dn}^k = 1|\mathbf{Z}_{\neg}, \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, w_{dn} = t) \propto \frac{(C_{k,\neg n}^t + \beta_t)(C_{d,\neg n}^k + \alpha_k)}{\sum_{t=1}^{V} C_{k,\neg n}^t + \sum_{t=1}^{V} \beta_t} \exp\left(c\gamma\psi_d\eta_k\right.
$$

$$
\left. -c^2\left(\frac{\gamma^2\rho_d\eta_k^2}{2} + \gamma(1-\gamma)\rho_d\eta_k\Upsilon_{dn}^k\right)\right), \tag{18}
$$

where $\gamma = \frac{1}{N_d}$; and $\Upsilon_{dn}^k = \frac{1}{N_d - 1} \sum_{k'=1}^{K} \eta_{k'} C_{d,\neg n}^{k'}$ is the discriminant function value without word n. The first term is from the LDA model for observed word counts. The second term is from the supervised signal $\mathbf{y}$.

For $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$: Finally, we can derive that λ_d^{-1} and ω_d^{-1} follow the inverse Gaussian distributions:

$$q(\lambda_d^{-1}|\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\omega}) = \mathcal{IG}\left(\lambda_d^{-1}; \frac{1}{c|\Delta_d - \epsilon|}, 1\right), \tag{19}$$

$$q(\omega_d^{-1}|\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\lambda}) = \mathcal{IG}\left(\omega_d^{-1}; \frac{1}{c|\Delta_d + \epsilon|}, 1\right). \tag{20}$$

The per-iteration time complexity of this algorithm is similar to that of the binary Gibbs MedLDA model, that is, linear to the number of documents and the number of topics if K is not too large.

5.2 Multi-task Gibbs MedLDA

The second extension is a multi-task Gibbs MedLDA. Multi-task learning is a scenario where multiple potentially related tasks are learned jointly with the hope that their performance can be boosted by sharing some statistic information among these tasks, and it has attracted a lot of research attention. In particular, learning a common latent representation shared by all the related tasks has proven to be an effective way to capture task relationships (Ando and Zhang, 2005; Argyriou et al., 2007; Zhu et al., 2014). Here, we take the similar approach to learning multiple predictive models which share the common latent topic representations. As we shall see in Section 6.3.2, one natural application of our approach is to do multi-label classification (Tsoumakas et al., 2010), where each document can belong to multiple categories, by defining each task as a binary classifier to determine whether a data point belongs to a particular category; and it can also be applied to multi-class classification, where each document belongs to only one of the many categories, by defining a single output prediction rule (See Section 6.3.2 for details).

5.2.1 THE MODEL WITH DATA AUGMENTATION

We consider L binary classification tasks and each task i is associated with a classifier with weights $\boldsymbol{\eta}_i$. We assume that all the tasks work on the same set of input data $\mathbf{W} = \{\mathbf{w}_d\}_{d=1}^D$, but each data d has different binary labels $\{y_d^i\}_{i=1}^L$ in different tasks. A multi-task Gibbs MedLDA model has two components—an LDA model to describe input words (the same as in Gibbs MedLDA); and multiple Gibbs classifiers sharing the same topic representations. When we have the classifier weights $\boldsymbol{\eta}$ and the topic assignments $\mathbf{Z}$, drawn from a posterior distribution $q(\boldsymbol{\eta}, \mathbf{Z})$, we follow the same principle as in Gibbs MedLDA and define the latent Gibbs rule for each task as

$$\forall i = 1, \ldots L : \quad \hat{y}^i(\boldsymbol{\eta}_i, \mathbf{z}) = \text{sign}\, F(\boldsymbol{\eta}_i, \mathbf{z}; \mathbf{w}) = \text{sign}(\boldsymbol{\eta}_i^\top \bar{\mathbf{z}}). \tag{21}$$

Let $\zeta_d^i = \ell - y_d^i \boldsymbol{\eta}_i^\top \bar{\mathbf{z}}_d$. The hinge loss of the stochastic classifier i is

$$\mathcal{R}_i(\boldsymbol{\eta}_i, \mathbf{Z}) = \sum_{d=1}^D \max(0, \zeta_d^i)$$

and the expected hinge loss is

$$\mathcal{R}_i'(q) = \mathbb{E}_q[\mathcal{R}_i(\boldsymbol{\eta}_i, \mathbf{Z})] = \sum_{d=1}^{D} \mathbb{E}_q\left[\max(0, \zeta_d^i)\right].$$

For each task i, we can follow the argument as in Gibbs MedLDA to show that the expected loss $\mathcal{R}_i'(q)$ is an upper bound of the expected training error $\sum_{d=1}^{D} \mathbb{E}_q[\mathbb{I}(y_d^i \neq \hat{y}_d^i(\boldsymbol{\eta}_i, \mathbf{z}_d))]$ of the Gibbs classifier (21). Thus, it is a good surrogate loss for learning a posterior distribution which could lead to a low expected training error.

Then, following a similar procedure of defining the binary GibbsMedLDA classifier, we define the multi-task GibbsMedLDA model as solving the following RegBayes problem:

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})} \mathcal{L}\left(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\right) + 2c\mathcal{R}_{MT}'\left(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\right),$$

where the multi-task expected hinge loss is defined as a summation of the expected hinge loss of all the tasks:

$$\mathcal{R}_{MT}'\left(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\right) = \sum_{i=1}^{L} \mathcal{R}_i'\left(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})\right).$$

Due to the separability of the multi-task expected hinge loss, we can apply Lemma 2 to reformulate each task-specific expected hinge loss $\mathcal{R}_i'$ as a scale mixture by introducing a set of augmented variables $\{\lambda_d^i\}_{d=1}^{D}$. More specifically, let $\phi_i(y_d^i|\mathbf{z}_d, \boldsymbol{\eta}) = \exp\{-2c\max(0, \zeta_d^i)\}$ be the unnormalized likelihood of the response variable for document d in task i. Then, we have

$$\phi_i(y_d^i|\mathbf{z}_d, \boldsymbol{\eta}) = \int_0^{\infty} \frac{1}{\sqrt{2\pi\lambda_d^i}} \exp\left(-\frac{(\lambda_d^i + c\zeta_d^i)^2}{2\lambda_d^i}\right) d\lambda_d^i.$$

5.2.2 A Collapsed Gibbs Sampling Algorithm

Similar to the binary Gibbs MedLDA classification model, we can derive the collapsed Gibbs sampling algorithm, as outlined in Algorithm 3. Specifically, let

$$\phi_i(\mathbf{y}^i, \boldsymbol{\lambda}^i|\mathbf{Z}, \boldsymbol{\eta}) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d^i}} \exp\left(-\frac{(\lambda_d^i + c\zeta_d^i)^2}{2\lambda_d^i}\right)$$

be the joint unnormalized likelihood of the class labels $\mathbf{y}^i = \{y_d^i\}_{d=1}^{D}$ and the augmentation variables $\boldsymbol{\lambda}^i = \{\lambda_d^i\}_{d=1}^{D}$. Then, for the multi-task Gibbs MedLDA, we can integrate out the Dirichlet variables $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ and get the collapsed posterior distribution

$$q(\boldsymbol{\eta},\boldsymbol{\lambda},\mathbf{Z}) \propto p_0(\boldsymbol{\eta})p(\mathbf{W},\mathbf{Z}|\boldsymbol{\alpha},\boldsymbol{\beta}) \prod_{i=1}^{L} \phi_i(\mathbf{y}^i, \boldsymbol{\lambda}^i|\mathbf{Z}, \boldsymbol{\eta})$$

$$= p_0(\boldsymbol{\eta}) \left[\prod_{d=1}^{D} \frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})}\right] \prod_{k=1}^{K} \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \prod_{i=1}^{L}\prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\lambda_d^i}} \exp\left(-\frac{(\lambda_d^i + c\zeta_d^i)^2}{2\lambda_d^i}\right).$$

Algorithm 3 collapsed Gibbs sampling algorithm for multi-task Gibbs MedLDA

1: **Initialization:** set $\boldsymbol{\lambda} = 1$ and randomly draw z_{dk} from a uniform distribution.
2: **for** $m = 1$ **to** M **do**
3:     **for** $i = 1$ **to** L **do**
4:         draw the classifier $\boldsymbol{\eta}_i$ from the normal distribution (22)
5:     **end for**
6:     **for** $d = 1$ **to** D **do**
7:         **for** each word n in document d **do**
8:             draw a topic from the multinomial distribution (23)
9:         **end for**
10:         **for** $i = 1$ **to** L **do**
11:             draw $(\lambda_d^i)^{-1}$ (and thus λ_d^i) from the inverse Gaussian distribution (24).
12:         **end for**
13:     **end for**
14: **end for**

Then, we can derive the conditional distributions used in collapsed Gibbs sampling as follows.

For $\boldsymbol{\eta}$: We also assume its prior is an isotropic Gaussian $p_0(\boldsymbol{\eta}) = \prod_{i=1}^{L} \prod_{k=1}^{K} \mathcal{N}(\eta_{ik}; 0, \nu^2)$. Then, we have the factorized form $q(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}) = \prod_{i=1}^{L} q(\boldsymbol{\eta}_i|\mathbf{Z}, \boldsymbol{\lambda})$, and each individual distribution is

$$q(\boldsymbol{\eta}_i|\mathbf{Z}, \boldsymbol{\lambda}) \propto p_0(\boldsymbol{\eta}_i) \prod_{d=1}^{D} \exp\left(-\frac{(\lambda_d^i + c\zeta_d^i)^2}{2\lambda_d^i} \right) = \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{22}$$

where the posterior covariance matrix and posterior mean are

$$\boldsymbol{\Sigma}_i = \left(\frac{1}{\nu^2} I + c^2 \sum_{d=1}^{D} \frac{\bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top}{\lambda_d^i} \right)^{-1}, \text{ and } \boldsymbol{\mu}_i = \boldsymbol{\Sigma}_i \left(c \sum_{d=1}^{D} y_d^i \frac{\lambda_d^i + c\ell}{\lambda_d^i} \bar{\mathbf{z}}_d \right).$$

Similarly, the inverse can be robustly done using Cholesky decomposition, an $O(K^3)$ procedure. Since K is normally not large, the inversion can be done efficiently.

For $\mathbf{Z}$: The conditional distribution of $\mathbf{Z}$ is

$$q(\mathbf{Z}|\boldsymbol{\eta}, \boldsymbol{\lambda}) \propto \prod_{d=1}^{D} \frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \left[\prod_{i=1}^{L} \exp\left(-\frac{(\lambda_d^i + c\zeta_d^i)^2}{2\lambda_d^i} \right) \right] \prod_{k=1}^{K} \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})}.$$

By canceling common factors, we can derive the conditional distribution of one variable z_{dn} given others $\mathbf{Z}_\neg$ as:

$$
\begin{aligned}
q(z_{dn}^k = 1|\mathbf{Z}_\neg, \boldsymbol{\eta}, \boldsymbol{\lambda}, w_{dn} = t) \propto &\frac{(C_{k,\neg n}^t + \beta_t)(C_{d,\neg n}^k + \alpha_k)}{\sum_{t=1}^{V} C_{k,\neg n}^t + \sum_{t=1}^{V} \beta_t} \prod_{i=1}^{L} \exp\left(\frac{\gamma y_d^i (c\ell + \lambda_d^i)\eta_{ik}}{\lambda_d^i} \right. \\
&\left. - c^2 \frac{\gamma^2 \eta_{ik}^2 + 2\gamma(1-\gamma)\eta_{ik}\Lambda_{dn}^i}{2\lambda_d^i} \right),
\end{aligned}
\tag{23}
$$

where $\Lambda_{dn}^i = \frac{1}{N_d - 1} \sum_{k'=1}^{K} \eta_{ik'} C_{d, \neg n}^{k'}$ is the discriminant function value without word n. We can see that the first term is from the LDA model for observed word counts and the second term is from the supervised signal $\{y_d^i\}$ from all the multiple tasks.

For $\boldsymbol{\lambda}$: Finally, the conditional distribution of the augmented variables $\boldsymbol{\lambda}$ is fully factorized, $q(\boldsymbol{\lambda}|\mathbf{Z}, \boldsymbol{\eta}) = \prod_{i=1}^{L} \prod_{d=1}^{D} q(\lambda_d^i|\mathbf{Z}, \boldsymbol{\eta})$, and each variable follows a generalized inverse Gaussian distribution

$$q(\lambda_d^i|\mathbf{Z}, \boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi\lambda_d^i}} \exp\left(-\frac{(\lambda_d^i + c\zeta_d^i)^2}{2\lambda_d^i}\right) = \mathcal{GIG}\left(\lambda_d^i; \frac{1}{2}, 1, c^2(\zeta_d^i)^2\right).$$

Therefore, we can derive that $(\lambda_d^i)^{-1}$ follows an inverse Gaussian distribution

$$p((\lambda_d^i)^{-1}|\mathbf{Z}, \boldsymbol{\eta}) = \mathcal{IG}\left((\lambda_d^i)^{-1}; \frac{1}{c|\zeta_d^i|}, 1\right), \tag{24}$$

from which a sample can be efficiently drawn with a constant time complexity.

The per-iteration time complexity of the algorithm is $\mathcal{O}(LK^3 + N_{total}K + DL)$. For common large-scale applications where K and L are not too large while D (thus N_{total}) is very large, the step of sampling latent topic assignments takes most of the time. If L is very large, for example, in the PASCAL large-scale text/image categorization challenge tasks which have tens of thousands of categories,[4] the step of drawing global classifier weights may dominate. A nice property of the algorithm is that we can easily parallelize this step since there is no coupling among these classifiers once the topic assignments are given. A preliminary investigation of the parallel algorithm is presented in Zhu et al. (2013b).

6. Experiments

We present empirical results to demonstrate the efficiency and prediction performance of Gibbs MedLDA (denoted by GibbsMedLDA) on the 20Newsgroups data set for classification, a hotel review data set for regression, and a Wikipedia data set with more than 1 million documents for multi-label classification. We also analyze its sensitivity to key parameters and examine the learned latent topic representations qualitatively. The 20Newsgroups data set contains about 20K postings within 20 groups. We follow the same setting as in Zhu et al. (2012) and remove a standard list of stop words for both binary and multi-class classification. For all the experiments, we use the standard normal prior $p_0(\boldsymbol{\eta})$ (i.e., $\nu^2 = 1$) and the symmetric Dirichlet priors $\boldsymbol{\alpha} = \frac{\alpha}{K}\mathbf{1}$, $\boldsymbol{\beta} = 0.01 \times \mathbf{1}$, where $\mathbf{1}$ is a vector with all entries being 1. For each setting, we report the average performance and standard deviation with five randomly initialized runs. All the experiments, except the those on the large Wikipedia data set, are done on a standard desktop computer.

6.1 Binary Classification

The binary classification task is to distinguish postings of the newsgroup *alt.atheism* and postings of the newsgroup *talk.religion.misc*. The training set contains 856 documents, and

---

4. See the websites: `http://lshtc.iit.demokritos.gr/`;
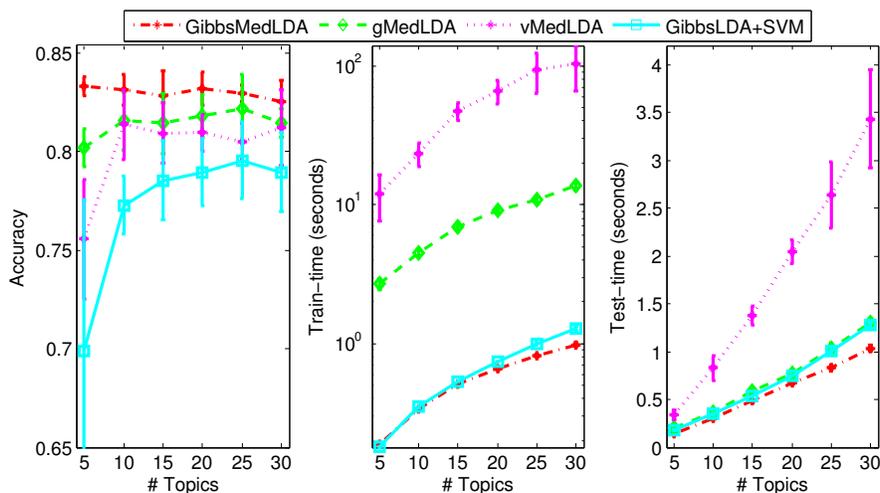   and `http://www.image-net.org/challenges/LSVRC/2012/index`.

Figure 1: Classification accuracy, training time (in log-scale) and testing time (in linear scale) on the 20Newsgroups binary classification data set.

the test set contains 569 documents. We compare Gibbs MedLDA with the MedLDA model that uses variational methods (denoted by vMedLDA) (Zhu et al., 2012) and the MedLDA that uses collapsed Gibbs sampling algorithms (denoted by gMedLDA) (Jiang et al., 2012). We also include unsupervised LDA using collapsed Gibbs sampling as a baseline, denoted by GibbsLDA. For GibbsLDA, we learn a binary linear SVM on its topic representations using SVMLight (Joachims, 1999). The results of other supervised topic models, such as sLDA and DiscLDA (Lacoste-Jullien et al., 2009), were reported in Zhu et al. (2012). For Gibbs MedLDA, we set $\alpha = 1$, $\ell = 164$ and $M = 10$. As we shall see in Section 6.5, Gibbs MedLDA is insensitive to α, ℓ and M in a wide range. Although tuning c (e.g., via cross-validation) can produce slightly better results, we fix $c = 1$ for simplicity.

Figure 1 shows the accuracy, training time, and testing time of different methods with various numbers of topics. We can see that by minimizing an expected hinge-loss and making no restrictive assumptions on the posterior distributions, GibbsMedLDA achieves higher accuracy than other max-margin topic models, which make some restrictive mean-field assumptions. Similarly, as gMedLDA makes a weaker mean-field assumption, it achieves slightly higher accuracy than vMedLDA, which assumes that the posterior distribution is fully factorized. For the training time, GibbsMedLDA is about two orders of magnitudes faster than vMedLDA, and about one order of magnitude faster than gMedLDA. This is partly because both vMedLDA and gMedLDA need to solve multiple SVM problems. For the testing time, GibbsMedLDA is comparable with gMedLDA and the unsupervised GibbsLDA, but faster than the variational algorithm used by vMedLDA, especially when the number of topics K is large. There are several possible reasons for the faster testing than vMedLDA, though they use the same stopping criterion. For example, vMedLDA performs mean-field inference in a full space which leads to a low convergence speed, while GibbsMedLDA carries out Gibbs sampling in a collapsed space. Also, the sparsity of the
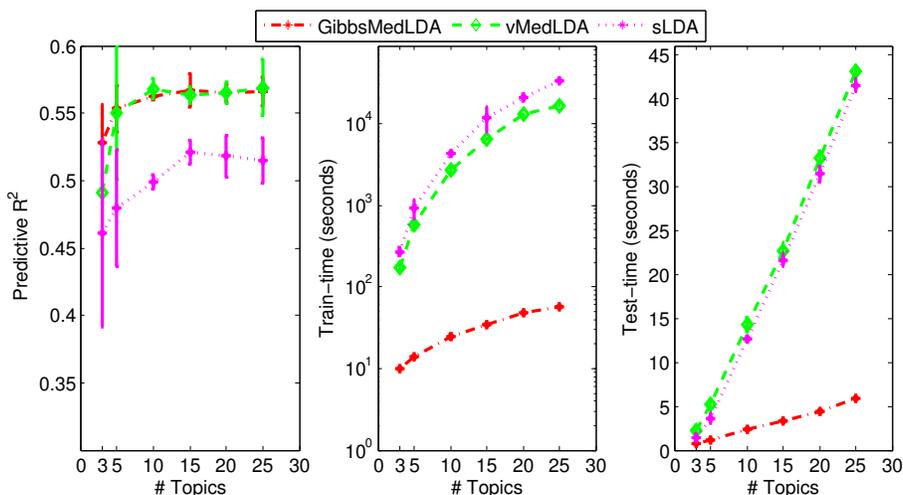
Figure 2: Predictive R2, training time and testing time on the hotel review data set.

sampled topics in GibbsMedLDA could save time, while vMedLDA needs to carry out computation for each dimension of the variational parameters.

6.2 Regression

We use the hotel review data set (Zhu and Xing, 2010) built by randomly crawling hotel reviews from the TripAdvisor website,[5] where each review is associated with a global rating score ranging from 1 to 5. In these experiments, we focus on predicting the global rating scores for reviews using the bag-of-words features only, with a vocabulary of 12,000 terms, though the other manually extracted features (e.g., part-of-speech tags) are provided. All the reviews have character lengths between 1,500 and 6,000. The data set consists of 5,000 reviews, with 1,000 reviews per rating. The data set is uniformly partitioned into training and testing sets. We compare the Gibbs MedLDA regression model with the MedLDA regression model that uses variational inference and supervised LDA (sLDA) which also uses variational inference. For Gibbs MedLDA and vMedLDA, the precision is set at $\epsilon = 1e^{-3}$ and c is selected via 5 fold cross-validation during training. Again, we set the Dirichlet parameter $\alpha = 1$ and the number of burn-in $M = 10$.

Figure 2 shows the predictive R^2 (Blei and McAuliffe, 2007) of different methods. We can see that GibbsMedLDA achieves comparable prediction performance with vMedLDA, which is better than sLDA. Note that vMedLDA uses a full likelihood model for both input words and response variables, while GibbsMedLDA uses a simpler likelihood model for words only.[6] For training time, GibbsMedLDA is about two orders of magnitudes faster than vMedLDA (as well as sLDA), again due to the fact that GibbsMedLDA does not need to solve multiple SVM problems. For testing time, GibbsMedLDA is also much faster than vMedLDA and sLDA, especially when the number of topics is large, due to the same reasons as stated in Section 6.1.

---

5. See the website: `http://www.tripadvisor.com/`.

6. The MedLDA with a simple likelihood on words only doesn't perform well for regression.

6.3 Multi-class Classification

We perform multi-class classification on 20Newsgroups with all 20 categories. The data set has a balanced distribution over the categories. The test set consists of 7,505 documents, in which the smallest category has 251 documents and the largest category has 399 documents. The training set consists of 11,269 documents, in which the smallest and the largest categories contain 376 and 599 documents, respectively. We consider two approaches to doing multi-class classification—one is to build multiple independent binary Gibbs MedLDA models, one for each category, and the other one is to build multiple dependent binary Gibbs MedLDA models under the framework of multi-task learning, as presented in Section 5.2.

6.3.1 Multiple One-vs-All Classifiers

Various methods exist to apply binary classifiers to do multi-class classification, including the popular "one-vs-all" and "one-vs-one" strategies. Here we choose the "one-vs-all" strategy, which has shown effective (Rifkin and Klautau, 2004), to provide some preliminary analysis. Let $\hat{\boldsymbol{\eta}}_i$ be the sampled classifier weights of the 20 "one-vs-all" binary classifiers after the burn-in stage. For a test document $\mathbf{w}$, we need to infer the latent topic assignments $\mathbf{z}_i$ under each "one-vs-all" binary classifier using a Gibbs sampler with the conditional distribution (13). Then, we predict the document as belonging to the single category which has the largest discriminant function value, that is,

$$\hat{y} = \operatorname*{argmax}_{i=1,\dots,L} (\hat{\boldsymbol{\eta}}_i^\top \bar{\mathbf{z}}_i),$$

where L is the number of categories (i.e., 20 in this experiment). Again, since GibbsMedLDA is insensitive to α and ℓ, we set $\alpha = 1$ and $\ell = 64$. We also fix $c = 1$ for simplicity. The number of burn-in iterations is set as $M = 20$, which is sufficiently large as will be shown in Figure 7.

Figure 3 shows the classification accuracy and training time, where GibbsMedLDA builds 20 binary Gibbs MedLDA classifiers. Note that for GibbsMedLDA the horizontal axis denotes the number of topics used by each single binary classifier. Since there is no coupling among these 20 binary classifiers, we can learn them in parallel, which we denote by pGibbsMedLDA. We can see that GibbsMedLDA clearly improves over other competitors on the classification accuracy, which may be due to the different strategies on building the multi-class classifiers.[7] However, given the performance gain on the binary classification task, we believe that the Gibbs sampling algorithm without any restrictive factorization assumptions is another factor leading to the improved performance. For training time, GibbsMedLDA takes slightly less time than the variational MedLDA as well as gMedLDA. But if we train the 20 binary GibbsMedLDA classifiers in parallel, we can save a lot of training time. These results are promising since it is now not uncommon to have a desktop computer with multiple processors or a cluster with tens or hundreds of computing nodes.

6.3.2 Multi-class Classification as a Multi-task Learning Problem

The second approach to performing multi-class classification is to formulate it as a multiple task learning problem, with a single output prediction rule. Specifically, let the label space

---

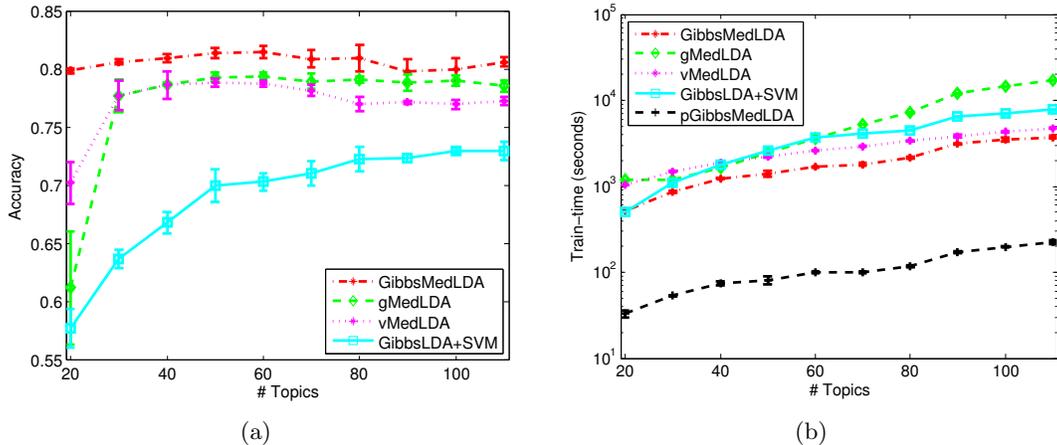7. MedLDA learns multi-class SVM (Zhu et al., 2012).

Figure 3: (a) classification accuracy and (b) training time of the one-vs-all Gibbs MedLDA classifiers for multi-class classification on the whole 20Newsgroups data set.

be $\mathcal{Y} = \{1, \ldots, L\}$. We can define one binary classification task for each category i and the task is to distinguish whether a data example belongs to the class i (with binary label $+1$) or not (with binary label -1). All the binary tasks share the same topic representations. To apply the model as we have presented in Section 5.2, we need to determine the true binary label of each document in a task. Given the multi-class label y_d of document d, this can be easily done by defining

$$\forall i = 1, \ldots, L: \ y_d^i = \begin{cases} +1 \text{ if } y_d = i \\ -1 \text{ otherwise} \end{cases}.$$

Then, we can learn a multi-task Gibbs MedLDA model using the data with transferred multiple labels. Let $\hat{\boldsymbol{\eta}}_i$ be the sampled classifier weights of task i after the burn-in stage. For a test document $\mathbf{w}$, once we have inferred the latent topic assignments $\mathbf{z}$ using a Gibbs sampler with the conditional distribution (13), we compute the discriminant function value $\hat{\boldsymbol{\eta}}_i^\top \bar{\mathbf{z}}$ for each task i, and predict the document as belonging to the single category which has the largest discriminant function value, that is,

$$\hat{y} = \underset{i=1,\ldots,L}{\operatorname{argmax}} (\hat{\boldsymbol{\eta}}_i^\top \bar{\mathbf{z}}).$$

Figure 4 shows the performance of the multi-task Gibbs MedLDA with comparison to the high-performance methods of the one-vs-all GibbsMedLDA and gMedLDA. Note again that for the one-vs-all GibbsMedLDA the horizontal axis denotes the number of topics used by each single binary classifier. We can see that although the multi-task GibbsMedLDA uses 20 times fewer topics than the one-vs-all GibbsMedLDA, their prediction accuracy scores are comparable when the multi-task GibbsMedLDA uses a reasonable number of topics (e.g., larger than 40). Both implementations of Gibbs MedLDA yield higher performance than gMedLDA. Looking at training time, when there is only a single processor core available, the multi-task GibbsMedLDA is about 3 times faster than the one-vs-all GibbsMedLDA. When there are multiple processor cores available, the naive parallel one-vs-all Gibbs MedLDA is
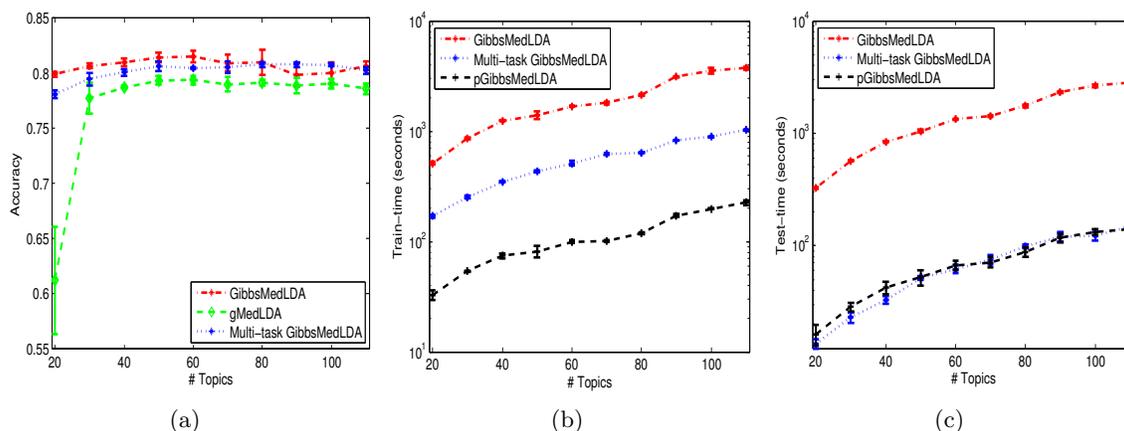
Figure 4: (a) classification accuracy; (b) training time; and (c) testing time of the multi-task Gibbs MedLDA classifiers for multi-class classification on the whole 20News-groups data set.

faster. In this case, using 20 processor cores, the parallel one-vs-all GibbsMedLDA is about 7 times faster than the multi-task GibbsMedLDA. In some scenarios, the testing time is significant. We can see that using a single core, the multi-task GibbsMedLDA is about 20 times faster than the one-vs-all GibbsMedLDA. Again however, in the presence of multiple processor cores, in this case 20, the parallel one-vs-all GibbsMedLDA tests at least as fast, at the expense of using more processor resources. So, depending on the processor cores available, both the parallel one-vs-all GibbsMedLDA and the multi-task GibbsMedLDA can be excellent choices. Where high efficiency single-core processing is key, then the multi-task GibbsMedLDA is a great choice. When there are many processor cores available, then the parallel one-vs-all GibbsMedLDA might be an appropriate choice.

6.4 Multi-label Classification

We also present some results on a multi-label classification task. We use the Wiki data set which is built from the large Wikipedia set used in the PASCAL LSHC challenge 2012, and where each document has multiple labels. The original data set is extremely imbalanced.[8] We built our data set by selecting the 20 categories that have the largest numbers of documents and keeping all the documents that are labeled by at least one of these 20 categories. The training set consists of 1.1 millions of documents and the testing set consists of 5,000 documents. The vocabulary has 917,683 terms in total. To examine the effectiveness of Gibbs MedLDA, which performs topic discovery and classifier learning jointly, we compare it with a linear SVM classifier built on the raw bag-of-words features and a two-step approach denoted by GibbsLDA+SVM. The GibbsLDA+SVM method first uses LDA with collapsed Gibbs sampling to discover latent topic representations for all the documents and then builds 20 separate binary SVM classifiers using the training documents with their

---

8. The data set is available at: `http://lshtc.iit.demokritos.gr/`.
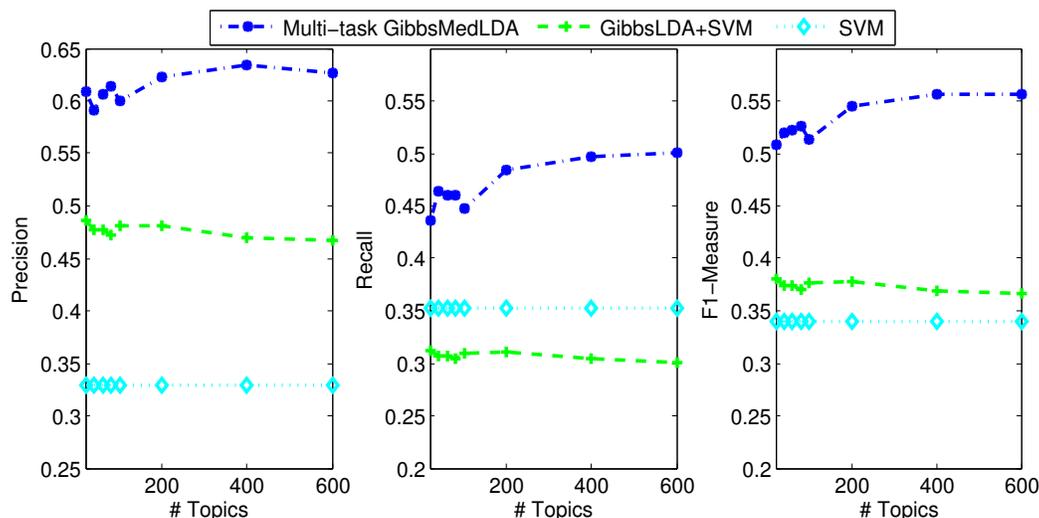
Figure 5: Precision, recall and F1-measure of the multi-label classification using different models on the Wiki data set.

discovered topic representations. For multi-task Gibbs MedLDA, we use 40 burn-in steps, which is sufficiently large. The model is insensitive to other parameters, similar to the multi-class classification task.

Figure 5 shows the precision, recall and F1 measure (i.e., the harmonic mean of precision and recall) of various models running on a distributed cluster with 20 nodes (each node is equipped with two 6-core CPUs).[9] We can see that the multi-task Gibbs MedLDA performs much better than other competitors. There are several reasons for the improvements. Since the vocabulary has about 1 million terms, the raw features are in a high-dimensional space and each document gives rise to a sparse feature vector, that is, only a few elements are nonzero. Thus, learning SVM classifiers on the raw data leads not just to over-fitting but a wider failure to generalize. For example, two documents from the same category might contain non-intersecting sets of words, yet contain similar latent topics. Using LDA to discover latent topic representations can produce dense features. Building SVM classifiers using the latent topic features improves the overall F1 measure, by improving the ability to generalize, and reducing overfitting. But, due to its two-step procedure, the discovered topic representations may not be very predictive. By doing max-margin learning and topic discovery jointly, the multi-task GibbsMedLDA can discover more discriminative topic features, thus improving significantly over the two-step GibbsLDA+SVM algorithm.

6.5 Sensitivity Analysis

We now provide a more careful analysis of the various Gibbs MedLDA models on their sensitivity to some key parameters in the classification tasks. Specifically, we will look at

---

9. For GibbsLDA, we use the parallel implementation in Yahoo-LDA, which is publicly available at: `https://github.com/shravanmn/Yahoo_LDA`. For Gibbs MedLDA, the parallel implementation of our Gibbs sampler is presented in Zhu et al. (2013b).
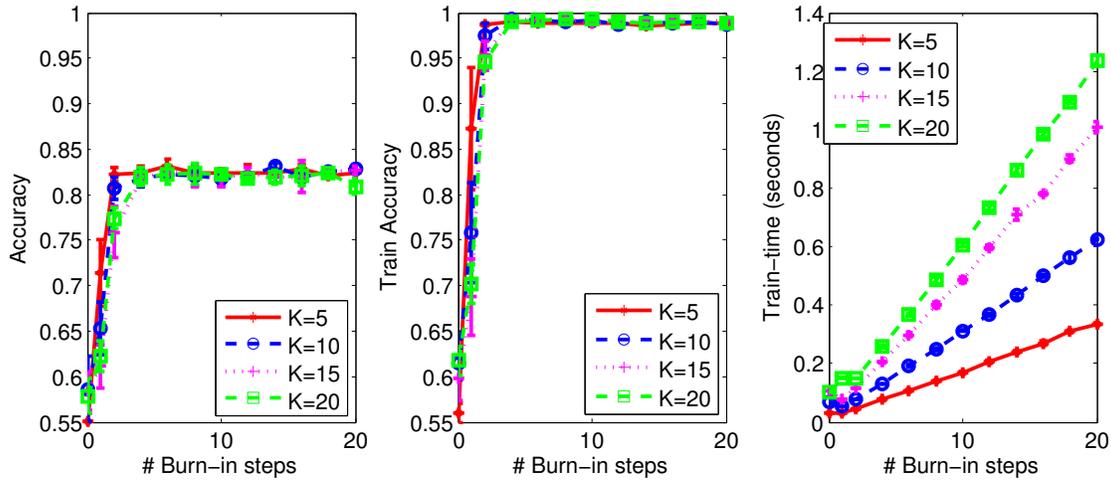
Figure 6: (Left) testing accuracy, (Middle) training accuracy, and (Right) training time of GibbsMedLDA with different numbers of burn-in steps for binary classification.

the effects of the number of burn-in steps, the Dirichlet prior α, the loss penalty ℓ, and the number of testing samples.

6.5.1 BURN-IN STEPS

Figure 6 shows the classification accuracy, training accuracy and training time of GibbsMedLDA with different numbers of burn-in samples in the binary classification task. When $M = 0$, the model is essentially random, for which we draw a classifier with the randomly initialized topic assignments for training data. We can see that both the training accuracy and testing accuracy increase very quickly and converge to their stable values with 5 to 10 burn-in steps. As expected, the training time increases about linearly in general when using more burn-in steps. Moreover, the training time increases linearly as K increases. In the previous experiments, we have chosen $M = 10$, which is sufficiently large.

Figure 7 shows the performance of GibbsMedLDA for multi-class classification with different numbers of burn-in steps when using the one-vs-all strategy. We show the total training time as well as the training time of the naive parallel implementation of pGibbsMedLDA. We can see that when the number of burn-in steps is larger than 20, the performance is quite stable, especially when K is large. Again, the training time grows about linearly as the number of burn-in steps increases. Even if we use 40 or 60 steps of burn-in, the training time is still competitive, compared with the variational MedLDA, especially considering that GibbsMedLDA can be naively parallelized by learning different binary classifiers simultaneously.

Figure 8 shows the testing classification accuracy, training accuracy and training time of the multi-task Gibbs MedLDA for multi-class classification with different numbers of burn-in steps. We can see that again both the training accuracy and testing accuracy increase fast and converge to their stable scores after about 30 burn-in steps. Also, the training time increases about linearly as the number of burn-in steps increases.
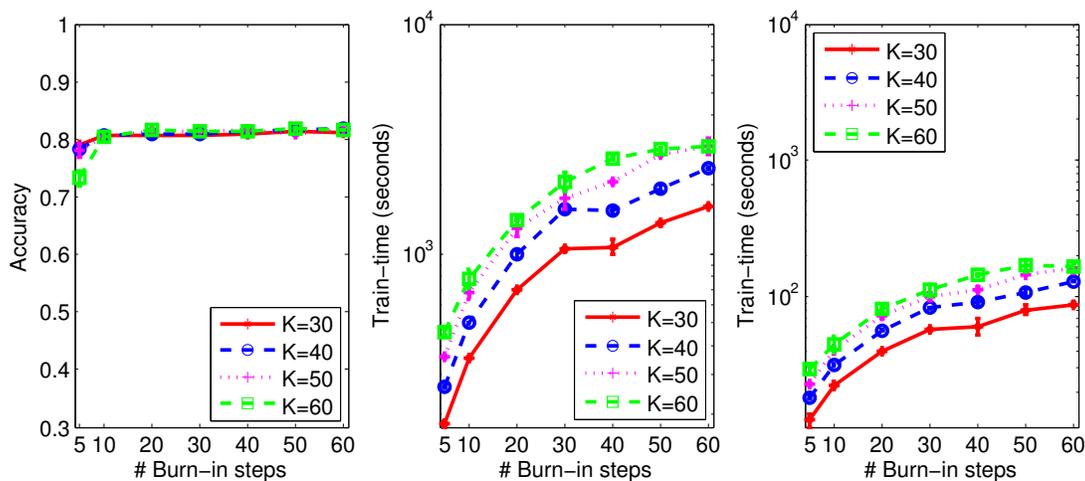
Figure 7: (Left) classification accuracy of GibbsMedLDA, (Middle) training time of GibbsMedLDA and (Right) training time of the parallel pGibbsMedLDA with different numbers of burn-in steps for multi-class classification.
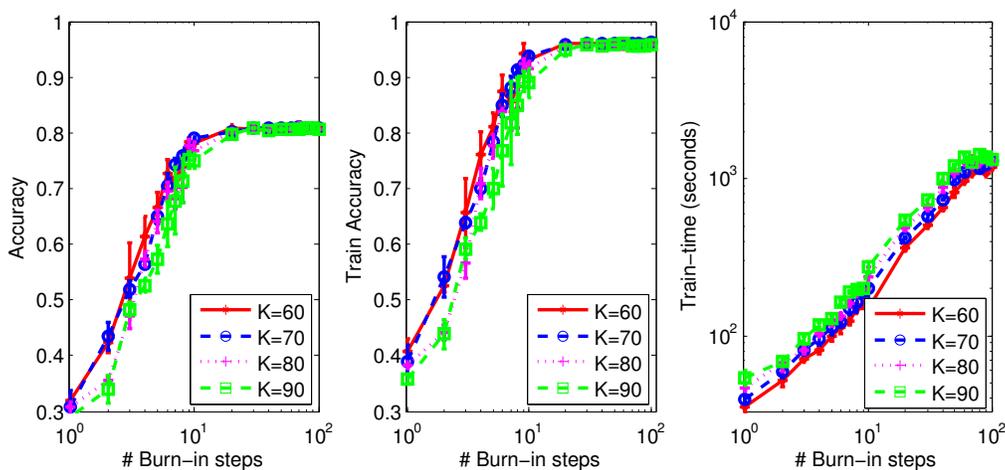

Figure 8: (Left) test accuracy, (Middle) training accuracy, and (Right) training time of the multi-task GibbsMedLDA with different numbers of burn-in steps for multi-class classification.

6.5.2 DIRICHLET PRIOR

For topic models with a Dirichlet prior, the Dirichlet hyper-parameter can be automatically estimated, such as using the Newton-Raphson method (Blei et al., 2003). Here, we analyze its effects on the performance by setting different values. Figure 9 shows the classification performance of GibbsMedLDA on the binary task with different α values for the symmetric Dirichlet prior $\boldsymbol{\alpha} = \frac{\alpha}{K}\mathbf{1}$. For the three different topic numbers, we can see that the performance is quite stable in a wide range of α values, for example, from 0.1 to 10. We can also
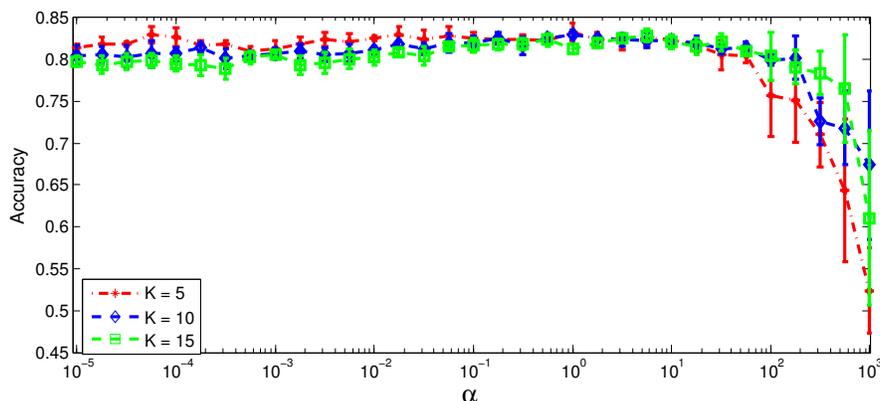
Figure 9: Classification accuracy of GibbsMedLDA on the binary classification data set with different α values.
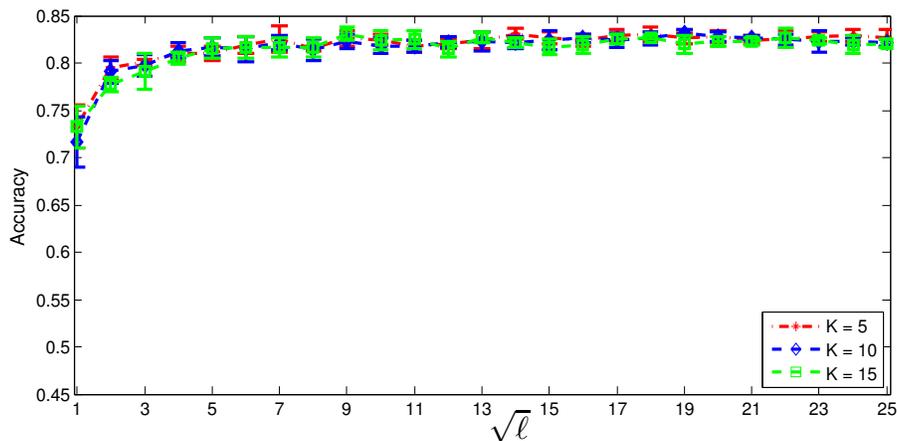

Figure 10: Classification accuracy of GibbsMedLDA on the binary classification data set with different ℓ values.

see that it generally needs a larger α in order to get the best results when K becomes larger. For example, when $\alpha < 0.1$, using fewer topics results in slightly higher performance. This is mainly because a large K tends to produce sparse topic representations and an appropriately large α is needed to smooth the representations, as the effective Dirichlet prior is $\alpha_k = \alpha/K$.

6.5.3 Loss Penalty ℓ

Figure 10 shows the classification performance of GibbsMedLDA on the binary classification task with different ℓ values. Again, we can see that in a wide range, for example, from 25 to 625, the performance is quite stable for all the three different K values. In the
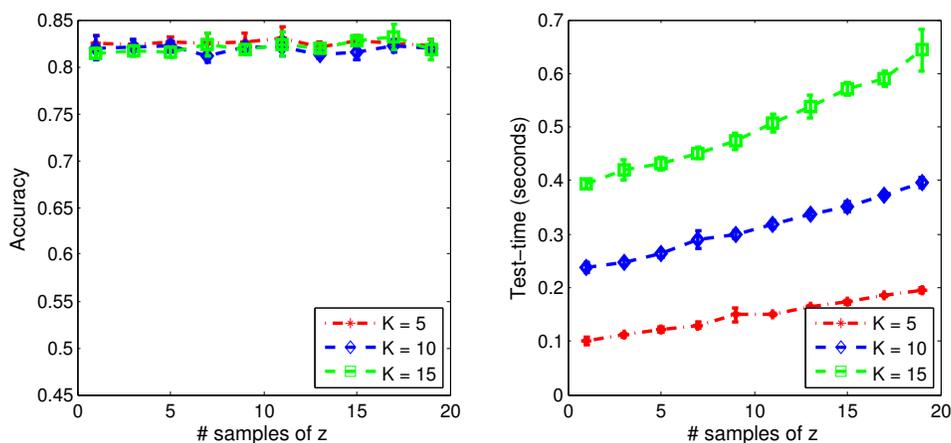
Figure 11: (Left) classification accuracy and (Right) testing time of GibbsMedLDA on the binary classification data set with different numbers of **z** samples in making predictions.

above experiments, we set $\ell = 164$. For the multi-class classification task, we have similar observations, and we set $\ell = 64$ in the previous experiments.

6.5.4 The Number of Testing Samples

Figure 11 shows the classification performance and testing time of GibbsMedLDA in the binary classification task with different numbers of **z** samples when making predictions, as stated in Section 4.4. We can see that in a wide range, for example, from 1 to 19, the classification performance is quite stable for all the three different K values we have tested; and the testing time increases about linearly as the number of **z** samples increases. For the multi-class classification task, we have similar observations.

6.6 Topic Representations

Finally, we also visualize the discovered latent topic representations of Gibbs MedLDA on the 20Newsgroup data set. We choose the multi-task Gibbs MedLDA, since it learns a single common topic space shared by multiple classifiers. We set the number of topics at 40. Figure 12 shows the average topic representations of the documents from each category, and Table 1 presents the 10 most probable words in each topic. We can see that for different categories, the average representations are quite different, indicating that the topic representations are good at distinguishing documents from different classes. We can also see that on average the documents in each category have very few salient topics, that is, the topics with a high probability of describing the documents. For example, the first two most salient topics for describing the documents in the category *alt.atheism* are topic 20 and topic 29, whose top-ranked words (see Table 1) reflect the semantic meaning of the category. For *graphics* category, the documents have the most salient topic 23, which
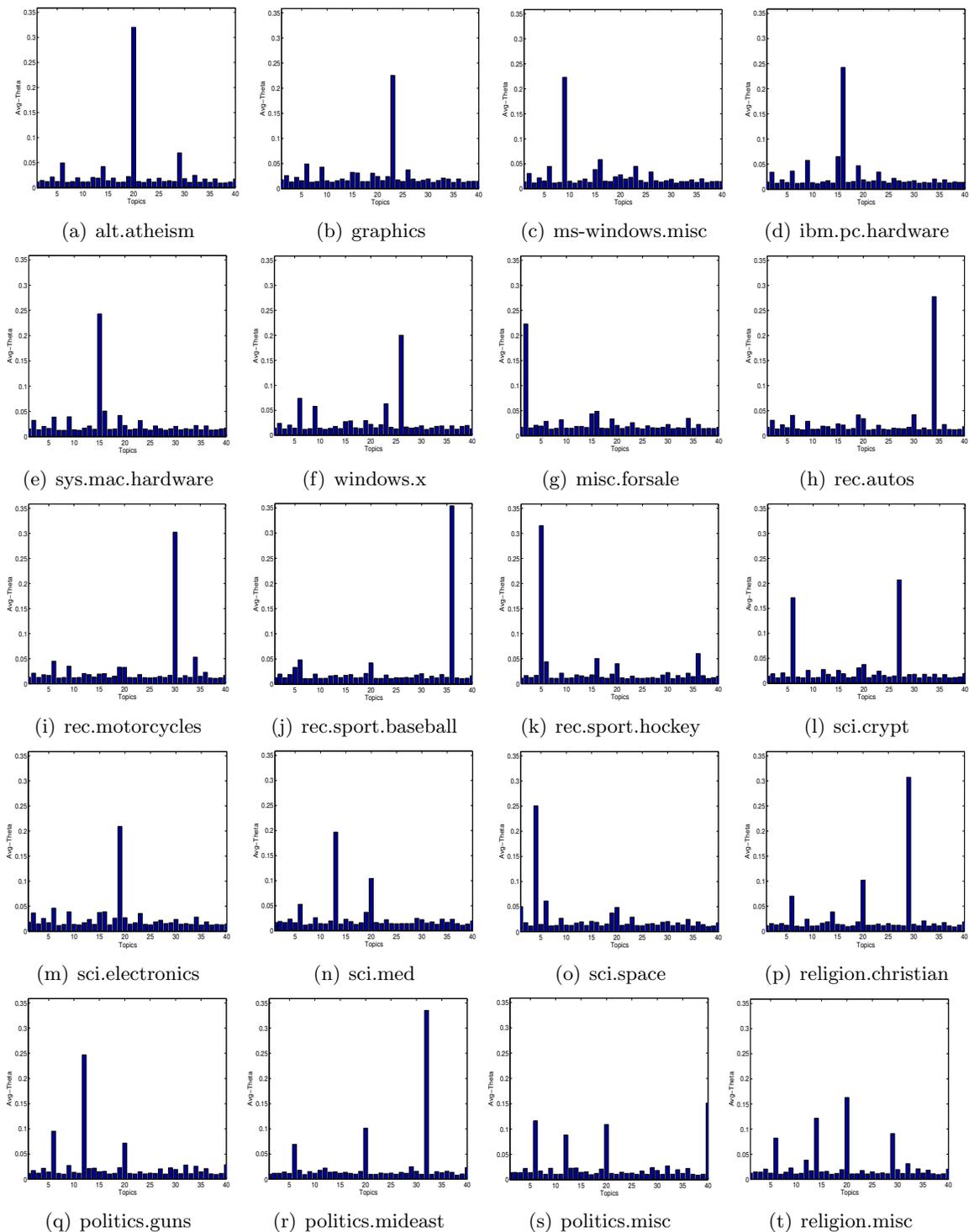
Figure 12: (a-t) per-class average topic representations on the 20Newsgroups data set.

has topic words *image*, *graphics*, *file*, *jpeg*, and etc., all of which are closely related to the semantic of graphics. For other categories, we have similar observations.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
data	sale	woman	space	team	writes	mr	db	windows	file
mission	offer	told	nasa	game	don	president	cs	writes	congress
center	shipping	afraid	launch	hockey	time	stephanopoulos	mov	article	january
sci	dos	building	writes	play	article	jobs	bh	file	centers
jpl	mail	couldn	earth	season	information	russian	si	files	bill
planetary	price	floor	article	nhl	number	administration	al	problem	quotes
mass	condition	beat	orbit	ca	people	meeting	byte	dos	hr
probe	good	standing	moon	games	make	george	theory	don	states
ames	interested	immediately	shuttle	players	work	russia	bits	run	march
atmosphere	sell	crowd	gov	year	part	working	larson	win	included
Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
organizations	gun	msg	jesus	mac	drive	ma	wiring	writes	writes
began	people	health	god	apple	scsi	nazis	supply	power	people
security	guns	medical	people	writes	mb	mu	boxes	article	article
terrible	weapons	food	christian	drive	card	conflict	bnr	don	don
association	firearms	article	bible	problem	system	ql	plants	ground	god
sy	writes	disease	sandvik	mb	controller	te	reduce	good	life
publication	article	patients	christians	article	bus	ne	corp	current	things
helped	government	writes	objective	system	hard	eu	relay	work	apr
organized	fire	doctor	ra	bit	ide	cx	onur	circuit	evidence
bullock	law	research	christ	don	disk	qu	damage	ve	time
Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**	**Topic 26**	**Topic 27**	**Topic 28**	**Topic 29**	**Topic 30**
ms	pub	image	internet	fallacy	window	key	unit	god	bike
myers	ftp	graphics	anonymous	conclusion	server	encryption	heads	jesus	dod
os	mail	file	privacy	rule	motif	chip	master	people	writes
vote	anonymous	files	posting	innocent	widget	clipper	multi	church	article
votes	archive	software	email	perfect	sun	government	led	christians	ride
santa	electronic	jpeg	anonymity	assertion	display	keys	vpic	christ	don
fee	server	images	users	true	application	security	dual	christian	apr
impression	faq	version	postings	ad	mit	escrow	ut	bible	ca
issued	eff	program	service	consistent	file	secure	ratio	faith	motorcycle
voting	directory	data	usenet	perspective	xterm	nsa	protected	truth	good
Topic 31	**Topic 32**	**Topic 33**	**Topic 34**	**Topic 35**	**Topic 36**	**Topic 37**	**Topic 38**	**Topic 39**	**Topic 40**
matthew	israel	courtesy	car	didn	year	south	entry	open	people
wire	people	announced	writes	apartment	writes	war	output	return	government
neutral	turkish	sequence	cars	sumgait	article	tb	file	filename	article
judas	armenian	length	article	started	game	nuclear	program	read	tax
reported	jews	rates	don	mamma	team	rockefeller	build	input	make
acts	israeli	molecular	good	father	baseball	georgia	um	char	health
island	armenians	sets	engine	karina	good	military	entries	write	state
safety	article	pm	speed	ll	don	bay	ei	enter	don
hanging	writes	automatic	apr	guy	games	acquired	eof	judges	money
outlets	turkey	barbara	oil	knew	runs	ships	printf	year	cramer

Table 1: The ten most probable words in the topics discovered by Multi-task Gibbs MedLDA ($K = 40$) on the 20Newsgroups data set.

7. Conclusions and Discussions

We have presented Gibbs MedLDA, an alternative approach to learning max-margin supervised topic models by minimizing an expected margin loss. We have applied Gibbs MedLDA to various tasks including text categorization, regression, and multi-task learning. By using data augmentation techniques, we have presented simple and highly efficient "augment-and-collapse" Gibbs sampling algorithms, without making any restricting assumptions on posterior distributions. Empirical results on the medium-sized 20Newsgroups and hotel re-

view data sets and a large-scale Wikipedia data set demonstrate significant improvements on time efficiency and classification accuracy over existing max-margin topic models. Our approaches are applicable to building other max-margin latent variable models, such as the max-margin nonparametric latent feature models for link prediction (Zhu, 2012) and matrix factorization (Xu et al., 2012). Finally, we release the code for public use.[10]

The new data augmentation formulation without any need to solve constrained sub-problems has shown great promise on improving the time efficiency of max-margin topic models. For future work, we are interested in developing highly scalable sampling algorithms (e.g., using a distributed architecture) (Newman et al., 2009; Smola and Narayanamurthy, 2010; Ahmed et al., 2012) to deal with large scale data sets. One nice property of the sampling algorithms is that the augmented variables are local to each document. Therefore, they can be effectively handled in a distributed architecture. But, the global prediction model weights bring in new challenges. Some preliminary work has been investigated in Zhu et al. (2013b). Another interesting topic is to apply the data augmentation technique to deal with the multiclass max-margin formulation, which was proposed by Crammer and Singer (2001) and used in MedLDA for learning multi-class max-margin topic models. Intuitively, it can be solved following an iterative procedure that infers the classifier weights associated with each category by fixing the others, similar as in polychomotous logistic regression (Holmes and Held, 2006), in which each substep may involve solving a binary hinge loss and thus our data augmentation techniques can be applied. A systematical investigation is our future work.

Acknowledgments

References

Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan Narayanamurthy, and Alex Smola. Scalable inference in latent variable models. In *International Conference on Web Search and Data Mining (WSDM)*, pages 123–132, 2012.

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research (JMLR)*, 9:1981–2014, 2008.

Rie K. Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research (JMLR)*, (6):1817–1853, 2005.

---

10. The code is available at: `http://www.ml-thu.net/~jun/gibbs-medlda.shtml`.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 41–48, 2007.

David M. Blei and John D. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 121–128, 2007.

David M. Blei, Andrew Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *Monograph Series of the Institute of Mathematical Statistics*, 2007.

Jonathan Chang and David Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics (AISTATS)*, pages 81–88, 2009.

Ning Chen, Jun Zhu, Fuchun Sun, and Eric P. Xing. Large-margin predictive latent subspace learning for multiview data analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(12):2365–2378, 2012.

Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. Generalized relational topic models with data augmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1273–1279, 2013.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research (JMLR)*, (2):265–292, 2001.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, (39):1–38, 1977.

Luc Devroye. *Non-uniform Random Variate Generation*. Springer-Verlag, 1986.

Li Fei-Fei and Pietro Perona. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, 2005.

Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research (JMLR)*, 11:2001–2049, 2010.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning (ICML)*, pages 353–360, 2009.

Thomas Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of National Academy of Science (PNAS)*, pages 5228–5235, 2004.

James P. Hobert. *The Data Augmentation Algorithm: Theory and Methodology*. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds.). Chapman & Hall/CRC Press, Boca Raton, FL., 2011.

Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems (NIPS)*, 1999.

Tony Jebara. *Discriminative, Generative and Imitative Learning*. PhD thesis, Media Laboratory, MIT, Dec 2001.

Qixia Jiang, Jun Zhu, Maosong Sun, and Eric P. Xing. Monte Carlo methods for maximum margin supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1609, 2012.

Thorsten Joachims. *Making Large-scale SVM Learning Practical*. MIT press, 1999.

Simons Lacoste-Jullien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 897–904, 2009.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Xiao-Li Meng and David A. van Dyk. The EM algorithm – an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society, Ser. B*, (59):511–567, 1997.

John R. Michael, William R. Schucany, and Roy W. Haas. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.

Kevin Miller, M. Pawan Kumar, Ben Packer, Danny Goodman, and Daphne Koller. Max-margin min-entropy models. In *Artificial Intelligence and Statistics (AISTATS)*, pages 779–787, 2012.

Radford M. Neal. Markov chain Monte Carlo methods based on 'slicing' the density function. *Technical Report No. 9722, Department of Statistics, University of Toronto*, 1997.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *Journal of Machine Learning Research (JMLR)*, (10):1801–1828, 2009.

Nicholas G. Polson and Steven L. Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–24, 2011.

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research (JMLR)*, (5):101–141, 2004.

Alex Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Very Large Data Base (VLDB)*, 3(1-2):703–710, 2010.

Alex Smola and Bernhard Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2003.

Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, (58):86–88, 1987.

Aixin Tan. *Convergence Rates and Regeneration of the Block Gibbs Sampler for Bayesian Random Effects Models.* PhD thesis, Department of Statistics, University of Florida, 2009.

Martin A. Tanner and Wing-Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the Americal Statistical Association (JASA)*, 82(398):528–540, 1987.

Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook, 2nd ed.*, pages 667–685, 2010.

David van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics (JCGS)*, 10(1):1–50, 2001.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

Minjie Xu, Jun Zhu, and Bo Zhang. Bayesian nonparametric maximum margin matrix factorization for collaborative prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 64–72, 2012.

Minjie Xu, Jun Zhu, and Bo Zhang. Fast max-margin matrix factorization with data augmentation. In *International Conference on Machine Learning (ICML)*, pages 978–986, 2013.

Shuanghong Yang, Jiang Bian, and Hongyuan Zha. Hybrid generative/discriminative learning for automatic image annotation. In *Uncertainty in Artificial Intelligence (UAI)*, pages 683–690, 2010.

Chun-Nam Yu and Thorsten Joachims. Learning structural SVMs with latent variables. In *International Conference on Machine Learning (ICML)*, pages 1169–1176, 2009.

Jun Zhu. Max-margin nonparametric latent feature models for link prediction. In *International Conference on Machine Learning (ICML)*, pages 719–726, 2012.

Jun Zhu and Eric P. Xing. Conditional topic random fields. In *International Conference on Machine Learning (ICML)*, pages 1239–1246, 2010.

Jun Zhu, Ning Chen, and Eric P. Xing. Infinite SVM: Dirichlet process mixtures of large-margin kernel machines. In *International Conference on Machine Learning (ICML)*, pages 617–624, 2011.

Jun Zhu, Amr Ahmed, and Eric. P. Xing. MedLDA: maximum margin supervised topic models. *Journal of Machine Learning Research (JMLR)*, (13):2237–2278, 2012.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with fast inference algorithms. In *International Conference on Machine Learning (ICML)*, pages 124–132, 2013a.

Jun Zhu, Xun Zheng, Li Zhou, and Bo Zhang. Scalable inference in max-margin topic models. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 964–972, 2013b.

Jun Zhu, Ning Chen, and Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *Journal of Machine Learning Research (JMLR, in press) (Technical Report, arXiv:1210.1766v3)*, 2014.