Improved Bayesian Logistic Supervised Topic Models with Data Augmentation

Jun Zhu, Xun Zheng, Bo Zhang
Department of Computer Science and Technology
TNLIST Lab and State Key Lab of Intelligent Technology and Systems
Tsinghua University, Beijing, China
{dcszj, dcszb}@tsinghua.edu.cn; vforveri.zheng@gmail.com

Abstract

Supervised topic models with a logistic likelihood have two issues that potentially limit their practical use: 1) response variables are usually over-weighted by document word counts; and 2) existing variational inference methods make strict mean-field assumptions. We address these issues by: 1) introducing a regularization constant to better balance the two parts based on an optimization formulation of Bayesian inference; and 2) developing a simple Gibbs sampling algorithm by introducing auxiliary Polya-Gamma variables and collapsing out Dirichlet variables. Our augment-and-collapse sampling algorithm has analytical forms of each conditional distribution without making any restricting assumptions and can be easily parallelized. Empirical results demonstrate significant improvements on prediction performance and time efficiency.

1 Introduction

As widely adopted in supervised latent Dirichlet allocation (sLDA) models (Blei and McAuliffe, 2010; Wang et al., 2009), one way to improve the predictive power of LDA is to define a likelihood model for the widely available document-level response variables, in addition to the likelihood model for document words. For example, the logistic likelihood model is commonly used for binary or multinomial responses. By imposing some priors, posterior inference is done with the Bayes' rule. Though powerful, one issue that could limit the use of existing logistic supervised LDA models is that they treat the document-level response variable as one additional word via a normalized likelihood model. Although some special treatment is carried out on defining the likelihood of the single response variable, it is normally of a much smaller scale than the likelihood of the usually tens or hundreds of words in each document. As noted by (Halpern et al., 2012) and observed in our experiments, this model imbalance could result in a weak influence of response variables on the topic representations and thus non-satisfactory prediction performance. Another difficulty arises when dealing with categorical response variables is that the commonly used normal priors are no longer conjugate to the logistic likelihood and thus lead to hard inference problems. Existing approaches rely on variational approximation techniques which normally make strict mean-field assumptions.

To address the above issues, we present two improvements. First, we present a general framework of Bayesian logistic supervised topic models with a regularization parameter to better balance response variables and words. Technically, instead of doing standard Bayesian inference via Bayes' rule, which requires a normalized likelihood model, we propose to do regularized Bayesian inference (Zhu et al., 2011; Zhu et al., 2013b) via solving an optimization problem, where the posterior regularization is defined as an expectation of a logistic loss, a surrogate loss of the expected misclassification error; and a regularization parameter is introduced to balance the surrogate classification loss (i.e., the response log-likelihood) and the word likelihood. The general formulation subsumes standard sLDA as a special case.

Second, to solve the intractable posterior inference problem of the generalized Bayesian logistic supervised topic models, we present a simple Gibbs sampling algorithm by exploring the ideas of data augmentation (Tanner and Wong, 1987; van Dyk and Meng, 2001; Holmes and Held, 2006). More specifically, we extend Polson's method for Bayesian logistic regression (Polson et al., 2012) to the generalized logistic supervised topic models, which are much more challeng-

ing due to the presence of non-trivial latent variables. Technically, we introduce a set of Polya-Gamma variables, one per document, to reformulate the generalized logistic pseudo-likelihood model (with the regularization parameter) as a scale mixture, where the mixture component is conditionally normal for classifier parameters. Then, we develop a simple and efficient Gibbs sampling algorithms with analytic conditional distributions without Metropolis-Hastings accept/reject steps. For Bayesian LDA models, we can also explore the conjugacy of the Dirichlet-Multinomial prior-likelihood pairs to collapse out the Dirichlet variables (i.e., topics and mixing proportions) to do collapsed Gibbs sampling, which can have better mixing rates (Griffiths and Steyvers, 2004). Finally, our empirical results on real data sets demonstrate significant improvements on time efficiency. The classification performance is also significantly improved by using appropriate regularization parameters. We also provide a parallel implementation with GraphLab (Gonzalez et al., 2012), which shows great promise in our preliminary studies.

The paper is structured as follows. Sec. 2 introduces logistic supervised topic models as a general optimization problem. Sec. 3 presents Gibbs sampling algorithms with data augmentation. Sec. 4 presents experiments. Sec. 5 concludes.

2 Logistic Supervised Topic Models

We now present the generalized Bayesian logistic supervised topic models.

2.1 The Generalized Models

We consider binary classification with a training set $\mathcal{D} = \{(\mathbf{w}_d, y_d)\}_{d=1}^D$, where the response variable Y takes values from the output space $\mathcal{Y} = \{0, 1\}$. A logistic supervised topic model consists of two parts — an LDA model (Blei et al., 2003) for describing the words $\mathbf{W} = \{\mathbf{w}_d\}_{d=1}^D$, where $\mathbf{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ denote the words within document d, and a logistic classifier for considering the supervising signal $\mathbf{y} = \{y_d\}_{d=1}^D$. Below, we introduce each of them in turn.

LDA: LDA 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 is

1. draw a topic proportion $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
2. for each word $n = 1, 2, \ldots, N_d$:
   (a) draw a topic[1] $z_{dn} \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$
   (b) draw the word $w_{dn} \sim \mathrm{Mult}(\boldsymbol{\Phi}_{z_{dn}})$

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

Let $\mathbf{z}_d = \{z_{dn}\}_{n=1}^{N_d}$ denote the set of topic assignments for document d. 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 entire corpus. LDA infers the posterior distribution $p(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}|\mathbf{W}) \propto p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})$, where $p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) = \big(\prod_d p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_n p(z_{dn}|\theta_d)\big) \prod_k p(\Phi_k|\boldsymbol{\beta})$ is the joint distribution defined by the model. As noticed in (Jiang et al., 2012), the posterior distribution by Bayes' rule is equivalent to the solution of an information theoretical optimization problem

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

where $\mathrm{KL}(q||p)$ is the Kullback-Leibler divergence from q to p and $\mathcal{P}$ is the space of probability distributions.

Logistic classifier: To consider binary supervising information, a logistic supervised topic model (e.g., sLDA) builds a logistic classifier using the topic representations as input features

$$p(y = 1|\boldsymbol{\eta}, \mathbf{z}) = \frac{\exp(\boldsymbol{\eta}^\top \bar{\mathbf{z}})}{1 + \exp(\boldsymbol{\eta}^\top \bar{\mathbf{z}})}, \quad (2)$$

where $\bar{\mathbf{z}}$ is a K-vector with $\bar{z}_k = \frac{1}{N} \sum_{n=1}^N \mathbb{I}(z_n^k = 1)$, and $\mathbb{I}(\cdot)$ is an indicator function that equals to 1 if predicate holds otherwise 0. If the classifier weights $\boldsymbol{\eta}$ and topic assignments $\mathbf{z}$ are given, the prediction rule is

$$\hat{y}|_{\boldsymbol{\eta}, \mathbf{z}} = \mathbb{I}(p(y = 1|\boldsymbol{\eta}, \mathbf{z}) > 0.5) = \mathbb{I}(\boldsymbol{\eta}^\top \bar{\mathbf{z}} > 0). \quad (3)$$

Since both $\boldsymbol{\eta}$ and $\mathbf{Z}$ are hidden variables, we propose to infer a posterior distribution $q(\boldsymbol{\eta}, \mathbf{Z})$ that has the minimal expected log-logistic loss

$$\mathcal{R}(q(\boldsymbol{\eta}, \mathbf{Z})) = -\sum_d \mathbb{E}_q[\log p(y_d|\boldsymbol{\eta}, \mathbf{z}_d)], \quad (4)$$

which is a good surrogate loss for the expected misclassification loss, $\sum_d \mathbb{E}_q[\mathbb{I}(\hat{y}|_{\boldsymbol{\eta}, \mathbf{z}_d} \neq y_d)]$, of a Gibbs classifier that randomly draws a model $\boldsymbol{\eta}$ from the posterior distribution and makes predictions (McAllester, 2003; Germain et al., 2009). In fact, this choice is motivated from the observation that logistic loss has been widely used as a convex surrogate loss for the misclassification

[1] A K-binary vector with only one entry equaling to 1.

loss (Rosasco et al., 2004) in the task of fully observed binary classification. Also, note that the logistic classifier and the LDA likelihood are coupled by sharing the latent topic assignments $\mathbf{z}$. The strong coupling makes it possible 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, a logistic supervised topic model solves the joint Bayesian inference problem

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})} \mathcal{L}(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})) + c\mathcal{R}(q(\boldsymbol{\eta},\mathbf{Z})) \quad (5)$$
$$\text{s.t.: } q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) \in \mathcal{P},$$

where $\mathcal{L}(q) = \mathrm{KL}(q||p_0(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})) - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})]$ is the objective for doing standard Bayesian inference with the classifier weights $\boldsymbol{\eta}$; $p_0(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = p_0(\boldsymbol{\eta})p_0(\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$; and c is a regularization parameter balancing the influence from response variables and words.

In general, we define the pseudo-likelihood for the supervision information

$$\psi(y_d|\mathbf{z}_d,\boldsymbol{\eta}) = p^c(y_d|\boldsymbol{\eta},\mathbf{z}_d) = \frac{\{\exp(\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d)\}^{cy_d}}{(1+\exp(\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d))^c}, \quad (6)$$

which is un-normalized if $c \neq 1$. But, as we shall see this un-normalization does not affect our subsequent inference. Then, the generalized inference problem (5) of logistic supervised topic models can be written in the "standard" Bayesian inference form (1)

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})} \mathcal{L}(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})) - \mathbb{E}_q[\log \psi(\mathbf{y}|\mathbf{Z},\boldsymbol{\eta})] \quad (7)$$
$$\text{s.t.: } q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) \in \mathcal{P},$$

where $\psi(\mathbf{y}|\mathbf{Z},\boldsymbol{\eta}) = \prod_d \psi(y_d|\mathbf{z}_d,\boldsymbol{\eta})$. It is easy to show that the optimum solution of problem (5) or the equivalent problem (7) is the posterior distribution with supervising information, i.e.,

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

where $\phi(\mathbf{y},\mathbf{W})$ is the normalization constant to make q a distribution. We can see that when $c = 1$, the model reduces to the standard sLDA, which in practice has the imbalance issue that the response variable (can be viewed as one additional word) is usually dominated by the words. This imbalance was noticed in (Halpern et al., 2012). We will see that c can make a big difference later.

Comparison with MedLDA: The above formulation of logistic supervised topic models as an instance of regularized Bayesian inference provides a direct comparison with the max-margin supervised topic model (MedLDA) (Jiang et al., 2012), which has the same form of the optimization problems. The difference lies in the posterior regularization, for which MedLDA uses a hinge loss of an expected classifier while the logistic supervised topic model uses an expected log-logistic loss. Gibbs MedLDA (Zhu et al., 2013a) is another max-margin model that adopts the expected hinge loss as posterior regularization. As we shall see in the experiments, by using appropriate regularization constants, logistic supervised topic models achieve comparable performance as max-margin methods. We note that the relationship between a logistic loss and a hinge loss has been discussed extensively in various settings (Rosasco et al., 2004; Globerson et al., 2007). But the presence of latent variables poses additional challenges in carrying out a formal theoretical analysis of these surrogate losses (Lin, 2001) in the topic model setting.

2.2 Variational Approximation Algorithms

The commonly used normal prior for $\boldsymbol{\eta}$ is non-conjugate to the logistic likelihood, which makes the posterior inference hard. Moreover, the latent variables $\mathbf{Z}$ make the inference problem harder than that of Bayesian logistic regression models (Chen et al., 1999; Meyer and Laud, 2002; Polson et al., 2012). Previous algorithms to solve problem (5) rely on variational approximation techniques. It is easy to show that the variational method (Wang et al., 2009) is a coordinate descent algorithm to solve problem (5) with the additional fully-factorized constraint $q(\boldsymbol{\eta},\Theta,\mathbf{Z},\boldsymbol{\Phi}) = q(\boldsymbol{\eta})(\prod_d q(\boldsymbol{\theta}_d)\prod_n q(z_{dn}))\prod_k q(\boldsymbol{\Phi}_k)$ and a variational approximation to the expectation of the log-logistic likelihood, which is intractable to compute directly. Note that the non-Bayesian treatment of $\boldsymbol{\eta}$ as unknown parameters in (Wang et al., 2009) results in an EM algorithm, which still needs to make strict mean-field assumptions together with a variational bound of the expectation of the log-logistic likelihood. In this paper, we consider the full Bayesian treatment, which can principally consider prior distributions and infer the posterior covariance.

3 A Gibbs Sampling Algorithm

Now, we present a simple and efficient Gibbs sampling algorithm for the generalized Bayesian logistic supervised topic models.

3.1 Formulation with Data Augmentation

Since the logistic pseudo-likelihood $\psi(\mathbf{y}|\mathbf{Z}, \boldsymbol{\eta})$ is not conjugate with normal priors, it is not easy to derive the sampling algorithms directly. Instead, we develop our algorithms by introducing auxiliary variables, which lead to a scale mixture of Gaussian components and analytic conditional distributions for automatical Bayesian inference without an accept/reject ratio. Our algorithm represents a first attempt to extend Polson's approach (Polson et al., 2012) to deal with highly non-trivial Bayesian latent variable models. Let us first introduce the Polya-Gamma variables.

Definition 1 *(Polson et al., 2012) A random variable X has a Polya-Gamma distribution, denoted by $X \sim \mathcal{PG}(a, b)$, if*

$$X = \frac{1}{2\pi^2} \sum_{i=1}^{\infty} \frac{g_k}{(i-1)^2/2 + b^2/(4\pi^2)},$$

where $a, b > 0$ and each $g_i \sim \mathcal{G}(a, 1)$ is an independent Gamma random variable.

Let $\omega_d = \boldsymbol{\eta}^\top \bar{\mathbf{z}}_d$. Then, using the ideas of data augmentation (Tanner and Wong, 1987; Polson et al., 2012), we can show that the generalized pseudo-likelihood can be expressed as

$$\psi(y_d|\mathbf{z}_d, \boldsymbol{\eta}) = \frac{1}{2^c} e^{\kappa_d \omega_d} \int_0^\infty \exp\Big(-\frac{\lambda_d \omega_d^2}{2}\Big) p(\lambda_d|c, 0) d\lambda_d,$$

where $\kappa_d = c(y_d - 1/2)$ and λ_d is a Polya-Gamma variable with parameters $a = c$ and $b = 0$. This result indicates that the posterior distribution of the generalized Bayesian logistic supervised topic models, i.e., $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$, can be expressed as the marginal of a higher dimensional distribution that includes the augmented variables $\boldsymbol{\lambda}$. The complete posterior distribution is

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

where the pseudo-joint distribution of $\mathbf{y}$ and $\boldsymbol{\lambda}$ is

$$\phi(\mathbf{y}, \boldsymbol{\lambda}|\mathbf{Z}, \boldsymbol{\eta}) = \prod_d \exp\Big(\kappa_d \omega_d - \frac{\lambda_d \omega_d^2}{2}\Big) p(\lambda_d|c, 0).$$

3.2 Inference with Collapsed Gibbs Sampling

Although 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 rate would be slow due to the large sample space. One way to effectively improve mixing rates is to integrate out the intermediate variables $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ and build a Markov chain whose equilibrium distribution is the 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). For our model, the collapsed posterior distribution is

$$\begin{aligned} q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \mathbf{Z}) &\propto p_0(\boldsymbol{\eta}) p(\mathbf{W}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \phi(\mathbf{y}, \boldsymbol{\lambda}|\mathbf{Z}, \boldsymbol{\eta}) \\ &= p_0(\boldsymbol{\eta}) \prod_{k=1}^K \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \prod_{d=1}^D \Big[\frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \\ &\quad \times \exp\Big(\kappa_d \omega_d - \frac{\lambda_d \omega_d^2}{2}\Big) p(\lambda_d|c, 0)\Big], \end{aligned}$$

where $\delta(\mathbf{x}) = \frac{\prod_{i=1}^{\dim(\mathbf{x})} \Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim(\mathbf{x})} x_i)}$, C_k^t is the number of times the term t being assigned to topic k over the whole corpus and $\mathbf{C}_k = \{C_k^t\}_{t=1}^V$; C_d^k is the number of times that terms being associated with topic k within the d-th document and $\mathbf{C}_d = \{C_d^k\}_{k=1}^K$. Then, the conditional distributions used in collapsed Gibbs sampling are as follows.

For $\boldsymbol{\eta}$: for the commonly used isotropic Gaussian prior $p_0(\boldsymbol{\eta}) = \prod_k \mathcal{N}(\eta_k; 0, \nu^2)$, we have

$$\begin{aligned} q(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}) &\propto p_0(\boldsymbol{\eta}) \prod_d \exp\Big(\kappa_d \omega_d - \frac{\lambda_d \omega_d^2}{2}\Big) \\ &= \mathcal{N}(\boldsymbol{\eta}; \boldsymbol{\mu}, \Sigma), \end{aligned} \tag{8}$$

where the posterior mean is $\boldsymbol{\mu} = \Sigma(\sum_d \kappa_d \bar{\mathbf{z}}_d)$ and the covariance is $\Sigma = (\frac{1}{\nu^2} I + \sum_d \lambda_d \bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top)^{-1}$. We can easily draw a sample from a K-dimensional 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.

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

$$\begin{aligned} q(\mathbf{Z}|\boldsymbol{\eta}, \boldsymbol{\lambda}) &\propto \prod_{k=1}^K \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \prod_{d=1}^D \Big[\frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \\ &\quad \times \exp\Big(\kappa_d \omega_d - \frac{\lambda_d \omega_d^2}{2}\Big)\Big]. \end{aligned}$$

By canceling common factors, we can derive the local conditional of one variable z_{dn} as:

$$\begin{aligned} &q(z_{dn}^k = 1 \mid \mathbf{Z}_\neg, \boldsymbol{\eta}, \boldsymbol{\lambda}, w_{dn} = t) \\ &\quad \propto \frac{(C_{k,\neg n}^t + \beta_t)(C_{d,\neg n}^k + \alpha_k)}{\sum_t C_{k,\neg n}^t + \sum_{t=1}^V \beta_t} \exp\Big(\gamma \kappa_d \eta_k \\ &\quad\quad - \lambda_d \frac{\gamma^2 \eta_k^2 + 2\gamma(1-\gamma)\eta_k \Lambda_{dn}^k}{2}\Big), \end{aligned} \tag{9}$$

where $C_{\cdot,\neg n}^{\cdot}$ indicates that term n is excluded from the corresponding document or topic; $\gamma = \frac{1}{N_d}$; and $\Lambda_{dn}^k = \frac{1}{N_d - 1} \sum_{k'} \eta_{k'} C_{d,\neg n}^{k'}$ 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 supervising signal $\mathbf{y}$.

For $\boldsymbol{\lambda}$: Finally, the conditional distribution of the augmented variables $\boldsymbol{\lambda}$ is

$$\begin{aligned} q(\lambda_d|\mathbf{Z}, \boldsymbol{\eta}) &\propto \exp\Big(-\frac{\lambda_d \omega_d^2}{2}\Big) p(\lambda_d|c, 0) \\ &= \mathcal{PG}(\lambda_d; c, \omega_d), \end{aligned} \tag{10}$$

Algorithm 1 for collapsed Gibbs sampling

Initialization: set $\boldsymbol{\lambda} = 1$ and randomly draw z_{dn} from a uniform distribution.
for $m = 1$ **to** M **do**
  draw a classifier from the distribution (8)
  for $d = 1$ **to** D **do**
    for each word n in document d **do**
      draw the topic using distribution (9)
    end for
    draw λ_d from distribution (10).
  end for
end for

which is a Polya-Gamma distribution. The equality has been achieved by using the construction definition of the general $\mathcal{PG}(a, b)$ class through an exponential tilting of the $\mathcal{PG}(a, 0)$ density (Polson et al., 2012). To draw samples from the Polya-Gamma distribution, we adopt the efficient method[2] proposed in (Polson et al., 2012), which draws the samples through drawing samples from the closely related exponentially tilted Jacobi distribution.

With the above conditional distributions, we can construct a Markov chain which iteratively draws samples of $\boldsymbol{\eta}$ using Eq. (8), $\mathbf{Z}$ using Eq. (9) and $\boldsymbol{\lambda}$ using Eq. (10), with an initial condition. In our experiments, we initially set $\boldsymbol{\lambda} = 1$ and randomly draw $\mathbf{Z}$ from a uniform distribution. In training, we run the Markov chain for M iterations (i.e., the burn-in stage), as outlined in Algorithm 1. Then, we draw a sample $\hat{\boldsymbol{\eta}}$ as the final classifier to make predictions on testing data. As we shall see, the Markov chain converges to stable prediction performance with a few burn-in iterations.

3.3 Prediction

To apply the classifier $\hat{\boldsymbol{\eta}}$ on testing data, we need to infer their topic assignments. We take the approach in (Zhu et al., 2012; Jiang et al., 2012), which uses a point estimate of topics $\boldsymbol{\Phi}$ from training data and makes prediction based on them. Specifically, we use the MAP estimate $\hat{\boldsymbol{\Phi}}$ to replace the probability distribution $p(\boldsymbol{\Phi})$. For the Gibbs sampler, an estimate of $\hat{\boldsymbol{\Phi}}$ using the samples is $\hat{\phi}_{kt} \propto C_k^t + \beta_t$. Then, given a testing document $\mathbf{w}$, we infer its latent components $\mathbf{z}$ using $\hat{\boldsymbol{\Phi}}$ as $p(z_n = k|\mathbf{z}_{\neg n}) \propto \hat{\phi}_{kw_n}(C_{\neg n}^k + \alpha_k)$, where $C_{\neg n}^k$ is the times that the terms in this document $\mathbf{w}$ assigned to topic k with the n-th term excluded.

[2]The basic sampler was implemented in the R package BayesLogit. We implemented the sampling algorithm in C++ together with our topic model sampler.

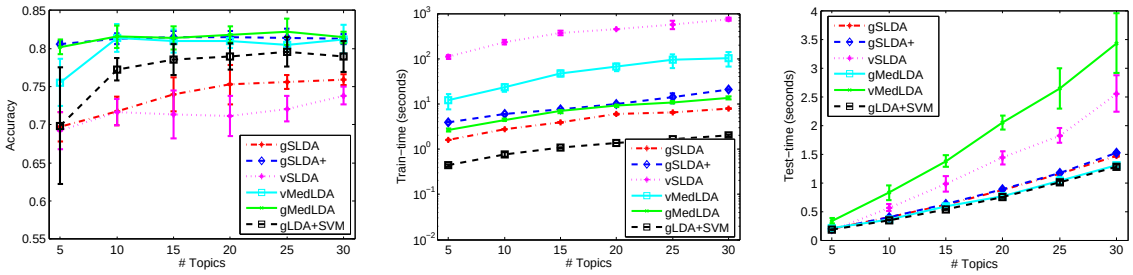
4 Experiments

We present empirical results and sensitivity analysis to demonstrate the efficiency and prediction performance[3] of the generalized logistic supervised topic models on the 20Newsgroups (20NG) data set, which contains about 20,000 postings within 20 news 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 the standard deviation with five randomly initialized runs.

4.1 Binary classification

Following the same setting in (Lacoste-Jullien et al., 2009; Zhu et al., 2012), the task is to distinguish postings of the newsgroup *alt.atheism* and those of the group *talk.religion.misc*. The training set contains 856 documents and the test set contains 569 documents. We compare the generalized logistic supervised LDA using Gibbs sampling (denoted by gSLDA) with various competitors, including the standard sLDA using variational mean-field methods (denoted by vSLDA) (Wang et al., 2009), the MedLDA model using variational mean-field methods (denoted by vMedLDA) (Zhu et al., 2012), and the MedLDA model using collapsed Gibbs sampling algorithms (denoted by gMedLDA) (Jiang et al., 2012). We also include the unsupervised LDA using collapsed Gibbs sampling as a baseline, denoted by gLDA. For gLDA, we learn a binary linear SVM on its topic representations using SVMLight (Joachims, 1999). The results of DiscLDA (Lacoste-Jullien et al., 2009) and linear SVM on raw bag-of-words features were reported in (Zhu et al., 2012). For gSLDA, we compare two versions – the standard sLDA with $c = 1$ and the sLDA with a well-tuned c value. To distinguish, we denote the latter by gSLDA+. We set $c = 25$ for gSLDA+, and set $\alpha = 1$ and $M = 100$ for both gSLDA and gSLDA+. As we shall see, gSLDA is insensitive to α,

[3]Due to space limit, the topic visualization (similar to that of MedLDA) is deferred to a longer version.



(a) accuracy (b) training time (c) testing time

Figure 1: Accuracy, training time (in log-scale) and testing time on the 20NG binary data set.

c and M in a wide range.

Fig. 1 shows the performance of different methods with various numbers of topics. For accuracy, we can draw two conclusions: 1) without making restricting assumptions on the posterior distributions, gSLDA achieves higher accuracy than vSLDA that uses strict variational mean-field approximation; and 2) by using the regularization constant c to improve the influence of supervision information, gSLDA+ achieves much better classification results, in fact comparable with those of MedLDA models since they have the similar mechanism to improve the influence of supervision by tuning a regularization constant. The fact that gLDA+SVM performs better than the standard gSLDA is due to the same reason, since the SVM part of gLDA+SVM can well capture the supervision information to learn a classifier for good prediction, while standard sLDA can't well-balance the influence of supervision. In contrast, the well-balanced gSLDA+ model successfully outperforms the two-stage approach, gLDA+SVM, by performing topic discovery and prediction jointly[4].

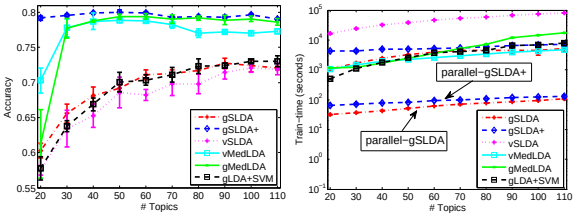
For training time, both gSLDA and gSLDA+ are very efficient, e.g., about 2 orders of magnitudes faster than vSLDA and about 1 order of magnitude faster than vMedLDA. For testing time, gSLDA and gSLDA+ are comparable with gMedLDA and the unsupervised gLDA, but faster than the variational vMedLDA and vSLDA, especially when K is large.

4.2 Multi-class classification

We perform multi-class classification on the 20NG data set with all the 20 categories. For multi-class classification, one possible extension is to use a multinomial logistic regression model for categorical variables Y by using topic representations $\bar{\mathbf{z}}$ as input features. However, it is non-trivial to develop a Gibbs sampling algorithm using the similar data augmentation idea, due to the presence of latent variables and the nonlinearity of the soft-max function. In fact, this is harder than the multinomial Bayesian logistic regression, which can be done via a coordinate strategy (Polson et al., 2012). Here, we apply the binary gSLDA to do the multi-class classification, following the "one-vs-all" strategy, which has been shown effective (Rifkin and Klautau, 2004), to provide some preliminary analysis. Namely, we learn 20 binary gSLDA models and aggregate their predictions by taking the most likely ones as the final predictions. We again evaluate two versions of gSLDA – the standard gSLDA with $c = 1$ and the improved gSLDA+ with a well-tuned c value. Since gSLDA is also insensitive to α and c for the multi-class task, we set $\alpha = 5.6$ for both gSLDA and gSLDA+, and set $c = 256$ for gSLDA+. The number of burn-in is set as $M = 40$, which is sufficiently large to get stable results, as we shall see.

Fig. 2 shows the accuracy and training time. We can see that: 1) by using Gibbs sampling without restricting assumptions, gSLDA performs better than the variational vSLDA that uses strict mean-field approximation; 2) due to the imbalance between the single supervision and a large set of word counts, gSLDA doesn't outperform the decoupled approach, gLDA+SVM; and 3) if we increase the value of the regularization constant c, supervision information can be better captured to infer predictive topic representations, and gSLDA+ performs much better than gSLDA. In fact, gSLDA+ is even better than the MedLDA that uses mean-field approximation, while is comparable with the MedLDA using collapsed Gibbs sampling. Finally, we should note that the improvement on the accuracy might be due to the different strategies on building the multi-class classifiers. But given the performance gain in the binary task, we believe that the Gibbs sampling algorith-

[4]The variational sLDA with a well-tuned c is significantly better than the standard sLDA, but a bit inferior to gSLDA+.



(a) accuracy (b) training time

Figure 2: Multi-class classification.

Table 1: Split of training time over various steps.

	SAMPLE $\boldsymbol{\lambda}$	SAMPLE $\boldsymbol{\eta}$	SAMPLE $\mathbf{Z}$
K=20	2841.67 (65.80%)	7.70 (0.18%)	1455.25 (34.02%)
K=30	2417.95 (56.10%)	10.34 (0.24%)	1888.78 (43.66%)
K=40	2393.77 (49.00%)	14.66 (0.30%)	2476.82 (50.70%)
K=50	2161.09 (43.67%)	16.33 (0.33%)	2771.26 (56.00%)

m without factorization assumptions is the main factor for the improved performance.

For training time, gSLDA models are about 10 times faster than variational vSLDA. Table 1 shows in detail the percentages of the training time (see the numbers in brackets) spent at each sampling step for gSLDA+. We can see that: 1) sampling the global variables $\boldsymbol{\eta}$ is very efficient, while sampling local variables $(\boldsymbol{\lambda}, \mathbf{Z})$ are much more expensive; and 2) sampling $\boldsymbol{\lambda}$ is relatively stable as K increases, while sampling $\mathbf{Z}$ takes more time as K becomes larger. But, the good news is that our Gibbs sampling algorithm can be easily parallelized to speedup the sampling of local variables, following the similar architectures as in LDA.

A Parallel Implementation: GraphLab is a graph-based programming framework for parallel computing (Gonzalez et al., 2012). It provides a high-level abstraction of parallel tasks by expressing data dependencies with a distributed graph. GraphLab implements a GAS (gather, apply, scatter) model, where the data required to compute a vertex (edge) are gathered along its neighboring components, and modification of a vertex (edge) will trigger its adjacent components to recompute their values. Since GAS has been successfully applied to several machine learning algorithms[5] including Gibbs sampling of LDA, we choose it as a preliminary attempt to parallelize our Gibbs sampling algorithm. A systematical investigation of the parallel computation with various architectures in interesting, but beyond the scope of this paper.

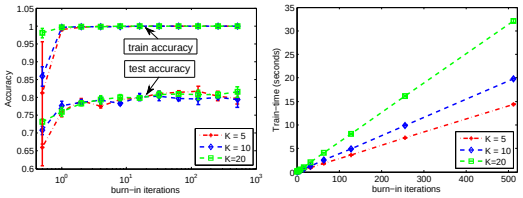
For our task, since there is no coupling among the 20 binary gSLDA classifiers, we can learn them in parallel. This suggests an efficient hybrid multi-core/multi-machine implementation, which can avoid the time consumption of IPC (i.e., inter-process communication). Namely, we run our experiments on a cluster with 20 nodes where each node is equipped with two 6-core CPUs (2.93GHz). Each node is responsible for learning one binary gSLDA classifier with a parallel implementation on its 12-cores. For each binary gSLDA model, we construct a bipartite graph connecting train documents with corresponding terms. The graph works as follows: 1) the edges contain the token counts and topic assignments; 2) the vertices contain individual topic counts and the augmented variables $\boldsymbol{\lambda}$; 3) the global topic counts and $\boldsymbol{\eta}$ are aggregated from the vertices periodically, and the topic assignments and $\boldsymbol{\lambda}$ are sampled asynchronously during the GAS phases. Once started, sampling and signaling will propagate over the graph. One thing to note is that since we cannot directly measure the number of iterations of an asynchronous model, here we estimate it with the total number of topic samplings, which is again aggregated periodically, divided by the number of tokens. We denote the parallel models by parallel-gSLDA ($c = 1$) and parallel-gSLDA+ ($c = 256$). From Fig. 2 (b), we can see that the parallel gSLDA models are about 2 orders of magnitudes faster than their sequential counterpart models, which is very promising. Also, the prediction performance is not sacrificed as we shall see in Fig. 4.

4.3 Sensitivity analysis

Burn-In: Fig. 3 shows the performance of gSLDA+ with different burn-in steps for binary classification. When $M = 0$ (see the most left points), the models are built on random topic assignments. We can see that the classification performance increases fast and converges to the stable optimum with about 20 burn-in steps. 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 set $M = 100$.

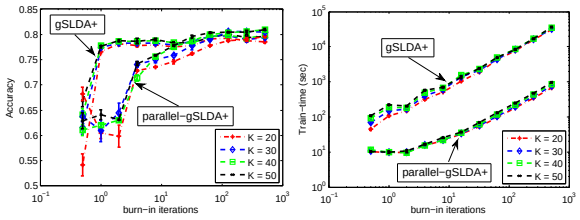
Fig. 4 shows the performance of gSLDA+ and its parallel implementation (i.e., parallel-gSLDA+) for the multi-class classification with different burn-in steps. We can see when the number of burn-in steps is larger than 20, the performance of gSLDA+ is quite stable. Again, in the log-log scale, since the slopes of the lines in Fig. 4 (b) are close to the constant 1, the training time grows about linearly as the number of

[5] http://docs.graphlab.org/toolkits.html



(a) accuracy (b) training time

Figure 3: Performance of gSLDA+ with different burn-in steps for binary classification. The most left points are for the settings with no burn in.



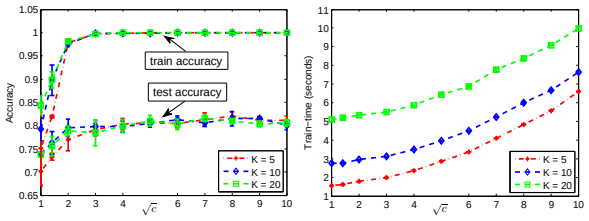
(a) accuracy (b) training time

Figure 4: Performance of gSLDA+ and parallel-gSLDA+ with different burn-in steps for multi-class classification. The most left points are for the settings with no burn in.

burn-in steps increases. Even when we use 40 or 60 burn-in steps, the training time is still competitive, compared with the variational vSLDA. For parallel-gSLDA+ using GraphLab, the training is consistently about 2 orders of magnitudes faster. Meanwhile, the classification performance is also comparable with that of gSLDA+, when the number of burn-in steps is larger than 40. In the previous experiments, we have set $M = 40$ for both gSLDA+ and parallel-gSLDA+.

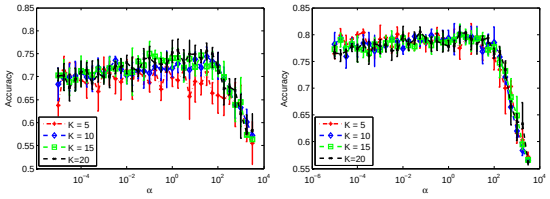
Regularization constant c: Fig. 5 shows the performance of gSLDA in the binary classification task with different c values. We can see that in a wide range, e.g., from 9 to 100, the performance is quite stable for all the three K values. But for the standard sLDA model, i.e., $c = 1$, both the training accuracy and test accuracy are low, which indicates that sLDA doesn't fit the supervision data well. When c becomes larger, the training accuracy gets higher, but it doesn't seem to over-fit and the generalization performance is stable. In the above experiments, we set $c = 25$. For multi-class classification, we have similar observations and set $c = 256$ in the previous experiments.

Dirichlet prior α: Fig. 6 shows the performance of gSLDA on the binary task with different α values. We report two cases with $c = 1$ and $c = 9$. We can see that the performance is quite stable in a wide range of α values, e.g., from 0.1 to 10. We also noted that the change of α does not affect the training time much.



(a) accuracy (b) training time

Figure 5: Performance of gSLDA for binary classification with different c values.



(a) $c = 1$ (b) $c = 9$

Figure 6: Accuracy of gSLDA for binary classification with different α values in two settings with $c = 1$ and $c = 9$.

5 Conclusions and Discussions

We present two improvements to Bayesian logistic supervised topic models, namely, a general formulation by introducing a regularization parameter to avoid model imbalance and a highly efficient Gibbs sampling algorithm without restricting assumptions on the posterior distributions by exploring the idea of data augmentation. The algorithm can also be parallelized. Empirical results for both binary and multi-class classification demonstrate significant improvements over the existing logistic supervised topic models. Our preliminary results with GraphLab have shown promise on parallelizing the Gibbs sampling algorithm.

For future work, we plan to carry out more careful investigations, e.g., using various distributed architectures (Ahmed et al., 2012; Newman et al., 2009; Smola and Narayanamurthy, 2010), to make the sampling algorithm highly scalable to deal with massive data corpora. Moreover, the data augmentation technique can be applied to deal with other types of response variables, such as count data with a negative-binomial likelihood (Polson et al., 2012).

Acknowledgments

This work is supported by National Key Foundation R&D Projects (No.s 2013CB329403,

2012CB316301), Tsinghua Initiative Scientific Research Program No.20121088071, Tsinghua National Laboratory for Information Science and Technology, and the 221 Basic Research Plan for Young Faculties at Tsinghua University.

References

A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. Smola. 2012. Scalable inference in latent variable models. In *International Conference on Web Search and Data Mining (WSDM)*.

D.M. Blei and J.D. McAuliffe. 2010. Supervised topic models. *arXiv:1003.0783v1*.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3:993–1022.

M. Chen, J. Ibrahim, and C. Yiannoutsos. 1999. Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of Royal Statistical Society, Ser. B*, (61):223–242.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. 2009. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning (ICML)*, pages 353–360.

A. Globerson, T. Koo, X. Carreras, and M. Collins. 2007. Exponentiated gradient algorithms for log-linear structured prediction. In *ICML*, pages 305–312.

J.E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. 2012. Powergraph: Distributed graph-parallel computation on natural graphs. In *the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.

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

Y. Halpern, S. Horng, L. Nathanson, N. Shapiro, and D. Sontag. 2012. A comparison of dimensionality reduction techniques for unstructured clinical text. In *ICML 2012 Workshop on Clinical Data Analysis*.

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

Q. Jiang, J. Zhu, M. Sun, and E.P. Xing. 2012. Monte Carlo methods for maximum margin supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*.

T. Joachims. 1999. *Making large-scale SVM learning practical*. MIT press.

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

Y. Lin. 2001. A note on margin-based loss functions in classification. *Technical Report No. 1044. University of Wisconsin*.

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

M. Meyer and P. Laud. 2002. Predictive variable selection in generalized linear models. *Journal of American Statistical Association*, 97(459):859–871.

D. Newman, A. Asuncion, P. Smyth, and M. Welling. 2009. Distributed algorithms for topic models. *Journal of Machine Learning Research (JMLR)*, (10):1801–1828.

N.G. Polson, J.G. Scott, and J. Windle. 2012. Bayesian inference for logistic models using Polya-Gamma latent variables. *arXiv:1205.0310v1*.

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

L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. 2004. Are loss functions all the same? *Neural Computation*, (16):1063–1076.

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

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

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

C. Wang, D.M. Blei, and Li F.F. 2009. Simultaneous image classification and annotation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

J. Zhu, N. Chen, and E.P. Xing. 2011. Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1620–1628.

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

J. Zhu, N. Chen, H. Perkins, and B. Zhang. 2013a. Gibbs max-margin topic models with fast sampling algorithms. In *International Conference on Machine Learning (ICML)*.

J. Zhu, N. Chen, and E.P. Xing. 2013b. Bayesian inference with posterior regularization and applications to infinite latent svms. *arXiv:1210.1766v2*.