Journal of Machine Learning Research 13 (2012) xx-xx Submitted 6/10; Revised 9/11; Published xx/12

MedLDA: Maximum Margin Supervised Topic Models

Jun Zhu DCSZJ@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
Amr Ahmed AMAHMED@CS.CMU.EDU
Eric P. Xing EPXING@CS.CMU.EDU
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: David Blei

Abstract

A supervised topic model can utilize side information such as ratings or labels associated with documents or images to discover more predictive low dimensional topical representations of the data. However, existing supervised topic models predominantly employ likelihood-driven objective functions for learning and inference, leaving the popular and potentially powerful max-margin principle unexploited for seeking predictive representations of data and more discriminative topic bases for the corpus. In this paper, we propose the maximum entropy discrimination latent Dirichlet allocation (MedLDA) model, which integrates the mechanism behind the max-margin prediction models (e.g., SVMs) with the mechanism behind the hierarchical Bayesian topic models (e.g., LDA) under a unified constrained optimization framework, and yields latent topical representations that are more discriminative and more suitable for prediction tasks such as document classification or regression. The principle underlying the MedLDA formalism is quite general and can be applied for jointly max-margin and maximum likelihood learning of directed or undirected topic models when supervising side information is available. Efficient variational methods for posterior inference and parameter estimation are derived and extensive empirical studies on several real data sets are also provided. Our experimental results demonstrate qualitatively and quantitatively that MedLDA could: 1) discover sparse and highly discriminative topical representations; 2) achieve state of the art prediction performance; and 3) be more efficient than existing supervised topic models, especially for classification.

Keywords: supervised topic models, max-margin learning, maximum entropy discrimination, latent Dirichlet allocation, support vector machines.

1. Introduction

Probabilistic latent aspect models such as the latent Dirichlet allocation (LDA) model (Blei et al., 2003) have recently gained much popularity for stratifying a large collection of documents by projecting every document into a low dimensional space spanned by a

set of bases that capture the semantic aspects, also known as *topics*, of the collection. An LDA model posits that each document is an admixture of latent topics, of which each topic is represented as a unique unigram distribution over a given vocabulary. The document-specific admixture proportion vector $\boldsymbol{\theta}$, also known as the *topic vector*, is modeled as a latent Dirichlet random variable, and can be regarded as a low dimensional representation of the document in a topical space. This low dimensional representation can be used for downstream tasks such as classification, clustering, or merely as a tool for structurally visualizing the otherwise unstructured document collection.

The original LDA is an unsupervised model and is typically built on a discrete bag-of-words representation of input contents, which can be text documents (Blei et al., 2003), images (Fei-Fei and Perona, 2005), or even network entities (Airoldi et al., 2008). However, in many practical applications, we can easily obtain useful side information besides the document or image contents. For example, when online users post their reviews for products or restaurants, they usually associate each review with a rating score or a thumb-up/thumb-down opinion; web sites or pages in the public Yahoo! Directory [1] can have their categorical labels; and images in the LabelMe (Russell et al., 2008) database are organized by a visual ontology and additionally each image is associated with a set of annotation tags. Furthermore, there is an increasing trend towards using online crowdsourcing services (such as Amazon Mechanical Turk [2]) to collect large collections of labeled data with a reasonably low price (Snow et al., 2008). Such side information often provides useful high-level or direct summarization of the content, but it is not directly utilized in the original LDA or models alike to influence topic inference. One would expect that incorporating such information into latent aspect modeling could guide a topic model towards discovering secondary or non-dominant, albeit semantically more salient statistical patterns (Chechik and Tishby, 2002) that may be more interesting or relevant to the user's goal, such as prediction on unlabeled data.

To explore this potential, developing new topic models that appropriately capture side information mentioned above has recently gained increasing attention. Representative attempts include supervised topic model (sLDA) (Blei and McAuliffe, 2007), which captures real-valued document rating as a regression response; multi-class sLDA (Wang et al., 2009), which directly captures discrete labels of documents as a classification response; and discriminative LDA (DiscLDA) (Lacoste-Julien et al., 2008), which also performs classification, but with a mechanism different from that of sLDA. All these models focus on the document-level side information such as document categories or review rating scores to supervise model learning. More variants of supervised topic models can be found in a number of applied domains, such as the aspect rating model (Titov and McDonald, 2008) for predicting ratings for each aspect of a hotel and the credit attribution model (Ramage et al., 2009) that associates each word with a label. In computer vision, several supervised topic models have been designed for understanding complex scene images (Sudderth et al., 2005; Fei-Fei and Perona, 2005; Li et al., 2009). Mimno and McCallum (2008) also proposed a topic model for considering document-level meta-data, e.g., publication date and venue of a paper.

It is worth pointing out that among existing supervised topic models for incorporating side information, there are two classes of approaches, namely, *downstream supervised topic*

1. http://dir.yahoo.com/
2. https://www.mturk.com/

model (DSTM) and *upstream supervised topic model* (USTM). In a DSTM the response variable is predicted based on the latent representation of the document, whereas in an USTM the response variable is being conditioned on to generate the latent representation of the document. Examples of USTM [3] include DiscLDA and the scene understanding models (Sudderth et al., 2005; Li et al., 2009), whereas sLDA is an example of DSTM. Another distinction between existing supervised topic models is the training criterion, or more precisely, the choice of objective function in the optimization-based learning. The sLDA model is trained by maximizing the *joint* likelihood of the content data (e.g., text or image) and the responses (e.g., labeling or rating), whereas DiscLDA is trained by maximizing the *conditional* likelihood of the responses given contents. To the best of our knowledge, all the existing supervised topic models are trained by optimizing a likelihood-based objective; the highly successful margin-based objectives such as the hinge loss commonly used in discriminative models such as SVMs have never been employed.

In this paper, we propose *maximum entropy discrimination latent Dirichlet allocation* (MedLDA), a supervised topic model leveraging the maximum margin principle for making more effective use of side information during estimation of latent topical representations. Unlike existing supervised topic models mentioned above, MedLDA employs an arguably more discriminative max-margin learning technique within a probabilistic framework; and unlike the commonly adopted two-stage heuristic which first estimates a latent topic vector for each document using a topic model and then feeds them to another downstream prediction model, MedLDA integrates the mechanism behind the max-margin prediction models (e.g., SVMs) with the mechanism behind the hierarchical Bayesian topic models (e.g., LDA) under a unified constrained optimization framework. It employs a composite objective motivated by a tradeoff between two components – the negative log-likelihood of an underlying topic model which measures the goodness of fit for document contents, and a measure of prediction error on training data. It then seeks a regularized posterior distribution of the predictive function in a feasible space defined by a set of *expected* margin constraints generalized from the SVM-style margin constraints. The resultant inference problem is intractable; to circumvent this, we relax the original objective by using a variational upper bound of the negative log-likelihood and a surrogate convex loss function that upper bounds the training error. Our proposed approach builds on earlier developments in maximum entropy discrimination (MED) (Jaakkola et al., 1999; Jebara, 2001) and partially observed maximum entropy discrimination Markov network (PoMEN) (Zhu et al., 2008), but is significantly different and more powerful. In MedLDA, because of the influence of both the likelihood function over content data (e.g., text or image) and margin constraints induced by the side information, the discovery of latent topics is therefore coupled with the max-margin estimation of model parameters. This interplay can yield latent topical representations that are more discriminative and more suitable for supervised prediction tasks, as we demonstrate in the experimental section.

In fact, the methodology we develop in this paper generalizes beyond learning topic models; it can be applied to perform max-margin learning for various types of graphical models, including directed Bayesian networks, e.g., LDA, sLDA and topic models with different priors such as the correlated topic models (Blei and Lafferty, 2005), and undirected

3. The model presented by (Mimno and McCallum, 2008) is also an upstream model for incorporating document meta-features.

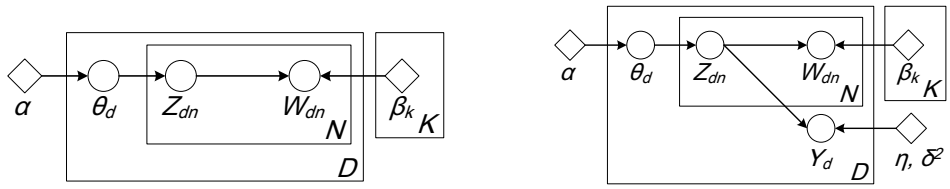


Figure 1: Graphical illustration of (Left) unsupervised LDA (Blei et al., 2003); and (Right) supervised LDA (Blei and McAuliffe, 2007).

Markov networks, e.g., exponential family harmoniums (Welling et al., 2004) and replicated softmax (Salakhutdinov and Hinton, 2009) (See Section 4 for an extensive discussion). In this paper, we focus on the scenario of downstream supervised topic models, and we present several concrete examples of MedLDA that build on the original LDA to learn "discriminative topics" that allow more salient topic proportion vector $\boldsymbol{\theta}$ to be inferred for every document, evidenced by a significant improvement of accuracy of both regression and classification of documents based on the $\boldsymbol{\theta}$ resulted from MedLDA, over the $\boldsymbol{\theta}$ resulted from either the vanilla unsupervised LDA or even sLDA and alike. We also present an efficient and easy-to-implement variational approach for inference under MedLDA, with a running time comparable to that of an unsupervised LDA and lower than other likelihood-based supervised LDAs. This advantage stems from the fact that MedLDA can directly optimize a margin-based loss instead of a likelihood-based one, and thereby avoids dealing with the normalization factor resultant from a full probabilistic generative formulation (e.g., sLDA), which generally makes learning harder.

The rest of this paper is structured as follows. Section 2 introduces the preliminaries that are needed to present MedLDA. Section 3 presents MedLDA models for both regression and classification, together with efficient variational algorithms. Section 4 discusses the generalization of MedLDA to other topic models. Section 5 presents empirical studies of MedLDA. Finally, Section 6 concludes this paper with future research directions discussed. Part of the materials of this paper build on conference proceedings presented earlier in (Zhu et al., 2009; Zhu and Xing, 2010).

2. Preliminaries

We begin with a brief overview of the fundamentals of topic models, support vector machines, and the maximum entropy discrimination formulism (Jaakkola et al., 1999), which constitute the major building blocks of the proposed MedLDA model.

2.1 Unsupervised and Supervised Topic Models

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a hierarchical Bayesian model that projects a text document into a latent low dimensional space spanned by a set of automatically learned topical bases. Each topic is a multinomial distribution over M words in a given vocabulary. Let $\mathbf{w} = (w_1, \ldots, w_N)$ denote the vector of words appearing in a document (for notation simplicity, we suppress the indexing subscript of N and assume that all documents have the same length N); assume the number of topics to be an integer K, where

K can be manually specified by a user or via cross-validation; and let $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$ denote the $M \times K$ matrix of topic distribution parameters, of which each $\boldsymbol{\beta}_k$ parameterizes a topic-specific multinomial word distribution. Under an LDA, the likelihood of a document d corresponds to the following generative process:

1. Draw a topic mixing proportion vector $\boldsymbol{\theta}_d$ according to a K-dimensional Dirichlet prior: $\boldsymbol{\theta}_d|\boldsymbol{\alpha} \sim \text{Dir}(\boldsymbol{\alpha})$;
2. For the n-th word in document d, where $1 \leq n \leq N$,
   (a) draw a topic assignment z_{dn} according to $\boldsymbol{\theta}_d$: $z_{dn}|\boldsymbol{\theta}_d \sim \text{Mult}(\boldsymbol{\theta}_d)$;
   (b) draw the word instance w_{dn} according to z_{dn}: $w_{dn}|z_{dn}, \boldsymbol{\beta} \sim \text{Mult}(\boldsymbol{\beta}_{z_{dn}})$,

where z_{dn} is a K-dimensional indicator vector (i.e., only one element is 1; all others are 0), an instance of the topic assignment random variable Z_{dn}. With a little abuse of notations, we use $\boldsymbol{\beta}_{z_{dn}}$ to denote the topic that is selected by the non-zero element of z_{dn}.

According to the above generative process, an *unsupervised* LDA defines the following joint distribution for a corpus $\mathcal{D}$ that contains D documents:

$$p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \Big(\prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \boldsymbol{\beta}) \Big),$$

where $\mathbf{W} \triangleq \{\mathbf{w}_1; \cdots; \mathbf{w}_D\}$ denotes all the words in $\mathcal{D}$, and $\mathbf{z}_d \triangleq \{z_{d1}; \cdots; z_{dN}\}$. To estimate the unknown parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and to infer the posterior distributions of latent variables $\{\boldsymbol{\theta}_d, \mathbf{z}_d\}$, an EM procedure is developed to maximize the marginal data likelihood $p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ [4]. As we have stated, $\boldsymbol{\theta}_d$ represents the mixing proportion over K topics for document d, which can be treated as a low-dimensional representation of the document. Moreover, since the posterior of z_{dn} represents the probability distribution that word n is assigned to one of the K topics; the average topic assignment $\bar{\mathbf{z}}_d \triangleq \frac{1}{N}\sum_n z_{dn}$ can also be treated as a representation of the document, as commonly done in downstream supervised topic models (Blei and McAuliffe, 2007; Wang et al., 2009).

Due to intractability of the likelihood $p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})$, approximate inference algorithms based on variational (Blei et al., 2003) or Markov Chain Monte Carlo (MCMC) (Griffiths and Steyvers, 2004) methods have been widely used for parameter estimation and posterior inference under LDA. We focus on variational inference in this paper. The following variational bound for unsupervised LDA will be used later. Let $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$ represent a variational distribution that approximates the true model posterior $p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$, one can derive a variational bound $\mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta})$ for the likelihood under unsupervised LDA:

$$\begin{aligned} \mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta}) &\triangleq -\mathbb{E}_q[\log p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})] - \mathcal{H}(q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})) \\ &\geq -\log p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}), \end{aligned} \tag{1}$$

4. We restrict ourselves to treat $\boldsymbol{\beta}$ as unknown parameters, as done in (Blei and McAuliffe, 2007; Wang et al., 2009). Extension to a Bayesian treatment of $\boldsymbol{\beta}$ (i.e., by putting a prior over $\boldsymbol{\beta}$ and inferring its posterior) can be easily done both in LDA as shown in the literature (Blei et al., 2003) and in the MedLDA proposed here based on the regularized Bayesian inference framework (Zhu et al., 2011a). But a systematical discussion is beyond the scope of this paper.

where $\mathcal{H}(q) \triangleq -\mathbb{E}_q[\log q]$ is the entropy of q. By making some independence assumption (e.g., mean field) about q, $\mathcal{L}^u(q)$ can be efficiently optimized (Blei et al., 2003).

As we have stated, the unsupervised LDA described above does not utilize side information for learning topics and inferring topic vectors $\boldsymbol{\theta}$. In order to consider side information appropriately for discovering more predictive representations, supervised topic models (sLDA) (Blei and McAuliffe, 2007) introduce a response variable Y to LDA for each document, as shown in Figure 1. For regression, where $y \in \mathbb{R}$, the generative process of sLDA is similar to LDA, but with an additional step – *draw a response variable:* $y|\mathbf{z}_d, \boldsymbol{\eta}, \delta^2 \sim \mathcal{N}(\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d, \delta^2)$ *for each document* d, where $\boldsymbol{\eta}$ is the regression weight vector and δ^2 is a noise variance parameter. Then, the joint distribution of sLDA is:

$$p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \Big(\prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \boldsymbol{\beta}) \Big) p(y_d|\boldsymbol{\eta}^\top \bar{\mathbf{z}}_d, \delta^2), \quad (2)$$

where $\mathbf{y} \triangleq \{y_1; \cdots; y_D\}$. In this case, the joint likelihood is $p(\mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2)$. Given a new document, the prediction is the expected response value

$$\hat{y} \triangleq \mathbb{E}[Y|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2] = \boldsymbol{\eta}^\top \mathbb{E}[\bar{Z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2], \quad (3)$$

where the average topic assignment random variable $\bar{Z} \triangleq \frac{1}{N}\sum_n Z_n$ ($\bar{\mathbf{z}}$ is an instance of $\bar{Z}$), and the expectation is taken with respect to the posterior distribution of $\mathbf{Z} \triangleq \{Z_1; \cdots; Z_N\}$. However, exact inference is again intractable, and one can use the following variational upper bound $\mathcal{L}^s(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2)$ for *supervised* sLDA for approximate inference:

$$\begin{aligned} \mathcal{L}^s(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2) &\triangleq -\mathbb{E}_q[\log p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2)] - \mathcal{H}(q(\{\theta_d, \mathbf{z}_d\})) \quad (4) \\ &\geq -\log p(\mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2). \end{aligned}$$

By changing the model of generating Y, sLDA can deal with other types of response variables, such as discrete ones for classification (Wang et al., 2009) using the multi-class logistic regression

$$p(y|\boldsymbol{\eta}, \mathbf{z}) = \frac{\exp(\boldsymbol{\eta}_y^\top \bar{\mathbf{z}})}{\sum_{y'} \exp(\boldsymbol{\eta}_{y'}^\top \bar{\mathbf{z}})}, \quad (5)$$

where $\boldsymbol{\eta}_y$ is the parameter vector associated with class label y. However, posterior inference in an sLDA classification model can be more challenging than that in the sLDA regression model. This is because the non-Gaussian probability distribution in Eq. (5) is highly nonlinear of $\boldsymbol{\eta}$ and $\mathbf{z}$ and its normalization factor can make the topic assignments of different words in the same document strongly coupled. Variational methods were successfully used to approximate the normalization factor (Wang et al., 2009), but they can be computationally expensive as we shall demonstrate in the experimental section.

DiscLDA (Lacoste-Julien et al., 2008) is yet another supervised topic model for classification. DiscLDA is an upstream supervised topic model and as such the unknown parameter is the transformation matrix that is used to generate the document latent representations conditioned on the class label; and this transformation matrix is learned by maximizing the conditional marginal likelihood of the text given class labels.

This progress notwithstanding, to the best of our knowledge, current developments of supervised topic models have been solely built on a likelihood-driven probabilistic inference paradigm. The arguably sometimes more powerful max-margin based techniques widely used in learning discriminative models have not been exploited to learn supervised topic models. The main goal of this paper is to systematically investigate how the max-margin principe can be exploited inside a topic model to learn topics that are better at discriminating documents than current likelihood-driven learning achieves while retaining semantic interpretability as the later allows. For this purpose, below we briefly review the max-margin principle underlying a major technique built on this principle, the support vector machines.

2.2 Support Vector Machines

Max-margin methods, such as support vector machines (SVMs) (Vapnik, 1998) and max-margin Markov networks (M^3N) (Taskar et al., 2003), have been successfully applied to a wide range of discriminative problems such as document categorization and handwritten character recognition. It has been shown that such methods enjoy strong generalization guarantees (Vapnik, 1998; Taskar et al., 2003). Depending on the nature of the response variable, the max-margin principle can be exploited in both classification and regression. Below we use document rating prediction as an example to recapitulate the ideas behind support vector regression (SVR) (Smola and Schölkopf, 2003), which we will shortly leverage to build our first instance of max-margin topic model.

Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_D, y_D)\}$ be a training set, where $\mathbf{x} \in \mathcal{X}$ are inputs such as document-feature vectors, and $y \in \mathbb{R}$ are response values such as user ratings. Using SVR, one obtains a function $h(\mathbf{x}) \in \mathcal{F}$ that makes at most ϵ deviation from the true response value y for each training example, and at the same time is as flat as possible. One common choice of the function family $\mathcal{F}$ is linear functions, that is, $h(\mathbf{x}; \boldsymbol{\eta}) = \boldsymbol{\eta}^\top \mathbf{f}(\mathbf{x})$, where $\mathbf{f} = \{f_1, \cdots, f_I\}$ is a vector of feature functions $f_i : \mathcal{X} \to \mathbb{R}$, and $\boldsymbol{\eta}$ is the corresponding weight vector. Formally, the linear SVR finds an optimal linear function by solving the following constrained optimization problem:

$$\text{P0(SVR)}: \quad \min_{\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \quad \frac{1}{2}\|\boldsymbol{\eta}\|_2^2 + C\sum_{d=1}^{D}(\xi_d + \xi_d^*) \tag{6}$$

$$\forall d, \text{ s.t.}: \quad \begin{cases} y_d - \boldsymbol{\eta}^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d \\ -y_d + \boldsymbol{\eta}^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \geq 0 \end{cases},$$

where $\|\boldsymbol{\eta}\|_2 \triangleq \sqrt{\boldsymbol{\eta}^\top \boldsymbol{\eta}}$ is the ℓ_2-norm; $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$ are slack variables that tolerate some errors in the training data; ϵ is a precision parameter; and C is a positive regularization constant. Problem P0 can be equivalently formulated as a regularized empirical loss minimization, where the loss is the so-called ϵ-insensitive loss (Smola and Schölkopf, 2003).

Under a standard SVR, P0 is a quadratic programming (QP) problem and can be easily solved in a Lagrangian dual formulation. Samples with non-zero lagrange multipliers are called support vectors, as in the SVM classification model. There exist several free packages for solving standard SVR, such as SVM-light (Joachims, 1999). We will use these methods as a sub-routine in our proposed approach, as we will detail in the sequel.

2.3 Maximum Entropy Discrimination

To unite the principles behind topic models and SVR, namely, Bayesian inference and max-margin learning, we employ a formalism known as *maximum entropy discrimination* (MED) (Jaakkola et al., 1999; Jebara, 2001), which learns a distribution of all possible regression/classification models that belong to a particular parametric family, subject to a set of margin-based constraints. For instance, the MED regression model, or simply MED^r, learns a distribution $q(\boldsymbol{\eta})$ through solving the following optimization problem:

$$\text{P1(MED}^{\text{r}}\text{)}: \quad \min_{q(\boldsymbol{\eta}),\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + C\sum_{d=1}^{D}(\xi_d + \xi_d^*) \tag{7}$$
$$\forall d, \text{ s.t.}: \quad \begin{cases} y_d - \mathbb{E}[\boldsymbol{\eta}]^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d \\ -y_d + \mathbb{E}[\boldsymbol{\eta}]^\top \mathbf{f}(\mathbf{x}_d) \leq \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \geq 0 \end{cases},$$

where $p_0(\boldsymbol{\eta})$ is a prior distribution over the parameters and $KL(p\|q) \triangleq \mathbb{E}_p[\log(p/q)]$ is the Kullback-Leibler (KL) divergence.

As studied in (Jebara, 2001), this MED problem leads to an entropic-regularized posterior distribution of the SVR coefficients, $q(\boldsymbol{\eta})$; and the resultant predictor $\hat{y} = \mathbb{E}_{q(\boldsymbol{\eta})}[h(\mathbf{x};\boldsymbol{\eta})]$ enjoys several nice properties and subsumes the standard SVR as special cases when the prior $p_0(\eta)$ is standard normal (Jebara, 2001). Moreover, as shown in (Zhu and Xing, 2009; Zhu et al., 2011b), with different choices of the prior over $\boldsymbol{\eta}$, such as a sparsity-inducing Laplace or a nonparametric Dirichlet process, the resultant $q(\boldsymbol{\eta})$ can exhibit a wide variety of characteristics and are suitable for diverse utilities such as feature selection or learning complex non-linear discriminating functions. Finally, the recent developments of the maximum entropy discrimination Markov network (MaxEnDNet) (Zhu and Xing, 2009) and partially observed MaxEnDNet (PoMEN) (Zhu et al., 2008) have extended the basic MED to the much broader scenarios of learning structured prediction functions with or without latent variables.

To apply the MED idea to learn a supervised topic model, a major difficulty is the presence of heterogeneous latent variables in the topic models, such as the topic vector $\boldsymbol{\theta}$ and topic indicator Z. In the sequel, we present a novel formalism called *maximum entropy discrimination LDA* (MedLDA) that extends the basic MED to make this possible, and at the same time discovers latent discriminating topics present in the study corpus based on available discriminant side information.

3. MedLDA: Maximum Margin Supervised Topic Models

Now we present a new class of supervised topic models that explicitly employ labeling information in the context of document classification or regression, under a unified statistical framework that jointly optimizes over the cross entropy between a user supplied model prior and the aimed model posterior, and over the margin of ensuing predictive tasks based on the learned model. This is to contrast conventional heuristics that first learn a topic model, and then independently train a classifier such as SVM using the per-document topic vectors resultant from the first step as inputs. In such a heuristic, the document labels are never

able to influence the way topics can be learned, and the per-document topic vectors are often found to be not strongly predictive (Xing et al., 2005).

3.1 Regressional MedLDA

We first consider the scenario where the numerical-valued rating of documents in the corpus is available, and our goal is to learn a supervised topic model specialized at predicting the rating of new documents through a regression function. We call this model a Regressional MedLDA, or simply, *MedLDA*^{r}.

Instead of learning a point estimate of regression coefficient $\boldsymbol{\eta}$ as in sLDA or SVR, we take the more general Bayesian-style (i.e., an averaging model) approach as in MED and learn a joint distribution [5] $q(\boldsymbol{\eta}, \mathbf{z})$ in a max-margin manner. For prediction, we take a weighted average over all the possible models (represented by $\boldsymbol{\eta}$) and latent topical representations $\mathbf{z}$, or more precisely, an expectation of the prediction over $q(\boldsymbol{\eta}, \mathbf{z})$, which is similar to that in Eq. (3), but now over both $\boldsymbol{\eta}$ and $\mathbf{Z}$, rather than only over $\mathbf{Z}$:

$$\hat{y} \triangleq \mathbb{E}[Y|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2] = \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2]. \tag{8}$$

Now, the question underlying the prediction rule (8) is how we can devise an appropriate objective function as well as constraints to learn a $q(\cdot)$ that leverages both the max-margin principle (for strong predictivity) and the topic model architecture (for topic discovery). Below we begin with a simple reformulation of the sLDA that makes this possible.

3.1.1 Max-Margin Training of sLDA

Without loss of generality, we let $q(\boldsymbol{\eta}, \mathbf{z}) = \int_{\boldsymbol{\theta}} q(\boldsymbol{\eta}) q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\eta})$, where $q(\boldsymbol{\eta})$ is the learned distribution of the predictive regression coefficient, and $q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\eta})$ is the learned distribution of the topic elements of the documents analogous to an sLDA-style topic model, but estimated from a different learning paradigm that leverages margin-based supervised training. As reviewed in Section 2.1, two good templates for $q(\mathbf{z}, \boldsymbol{\theta}|\boldsymbol{\eta})$ can be the original LDA or sLDA. For brevity, here we present a regressional MedLDA that uses the supervised sLDA as the underlying topic model. As we shall see in Section 3.2 and Appendix B, the underlying topic model can also be an unsupervised LDA.

Let $p_0(\boldsymbol{\eta})$ denote a prior distribution of $\boldsymbol{\eta}$, then MedLDA^r defines a joint distribution

$$p(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) = p_0(\boldsymbol{\eta}) p(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \delta^2),$$

where the second factor has the same form as Eq. (2) for sLDA, except that now $\boldsymbol{\eta}$ is a random variable and follows a prior $p_0(\boldsymbol{\eta})$. Accordingly, the likelihood $p(\mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)$ is an expectation of the likelihood of sLDA under $p_0(\boldsymbol{\eta})$, which makes it even harder than in sLDA to directly optimize. Therefore, we choose to optimize a variational upper bound of the log-likelihood. We will discuss other approximation methods in Section 4.

Let $q(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\})$ be a variational approximation to the posterior $p(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2, \mathbf{y}, \mathbf{W})$. Then, an upper bound $\mathcal{L}^{bs}(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)$ [6] of the negative log-likelihood is

$$\mathcal{L}^{bs}(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2) \triangleq -\mathbb{E}_q[\log p(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\}, \mathbf{y}, \mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)] - \mathcal{H}(q(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\}))$$

5. In principle, we can perform Bayesian-style estimation for other parameters, like δ^2. For simplicity, we only consider $\boldsymbol{\eta}$ as a random variable in this paper.
6. "bs" stands for "Bayesian Supervised".

$$= KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + \mathbb{E}_{q(\boldsymbol{\eta})}[\mathcal{L}^s]. \quad (9)$$

We can see that the bound is also an expectation of sLDA's variational bound $\mathcal{L}^s$ in Eq. (4). To derive Eq. (9), we should note that the variational distribution for sLDA is "conditioned on" its model parameters, which include $\boldsymbol{\eta}$. Similarly, the distribution q in $\mathcal{L}^{bs}$ depends on the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)$. For notation clarity, we have omitted the explicit dependence on parameters in variational distributions.

Based on the MED principle and the variational bound in Eq. (9), we define the learning problem of MedLDA^r as follows:

$$\text{P2(MedLDA}^r\text{)}: \min_{q,\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2,\boldsymbol{\xi},\boldsymbol{\xi}^*} \mathbb{E}_{q(\boldsymbol{\eta})}[\mathcal{L}^s(q;\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2)] + KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + C\sum_{d=1}^{D}(\xi_d + \xi_d^*) \quad (10)$$

$$\forall d, \text{ s.t.}: \begin{cases} y_d - \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \leq \epsilon + \xi_d \\ -y_d + \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \leq \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \geq 0, \end{cases}$$

where $\boldsymbol{\xi}, \boldsymbol{\xi}^*$ are slack variables, and ϵ is a precision parameter as in SVR. The margin constraints in P2 are of the same form as those in P0, but in an expectation version because both the topic assignments Z and parameters $\boldsymbol{\eta}$ are latent random variables in MedLDA^r.

It is easy to verify that at the optimum, at most one of ξ_d and ξ_d^* can be non-zero and $\xi_d + \xi_d^* = \max(0, |y_d - \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d]| - \epsilon)$, which is known as ϵ-insensitive loss (Smola and Schölkopf, 2003), that is, if the current prediction $\hat{y}$ as in Eq. (8) does not deviate from the true response value too much (i.e., less than ϵ), there is no loss; otherwise, a linear loss will be penalized. Mathematically, problem P2 can be equivalently written as a loss minimization problem without using slack variables:

$$\min_{q,\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2} \mathcal{L}^{bs}(q;\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2) + C\sum_{d=1}^{D}\max(0, |y_d - \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d]| - \epsilon), \quad (11)$$

where the variational bound $\mathcal{L}^{bs}$ plays two roles – *regularization* and *maximum likelihood estimation*. Specifically, as shown in Eq. (9), $\mathcal{L}^{bs}$ decomposes into two parts. The first part of KL-divergence is an entropic regularizer for $q(\boldsymbol{\eta})$; and the second term is an expected bound of the data likelihood, as we have discussed. Therefore, problem P2 is a joint maximum margin learning and maximum likelihood estimation (with appropriate regularization), and the two components are coupled by sharing latent topic assignments Z and parameters $\boldsymbol{\eta}$.

The rationale underlying MedLDA^r is that: by minimizing an integrated objective function, we aim to find a latent topical representation and a document-rating prediction function which, on one hand, can predict accurately on unseen data with a sufficient margin, and on the other hand, can explain the data well (i.e., minimizing a variational bound of the negative log-likelihood). The max-margin learning and topic discovery procedure are coupled together via the constraints, which are defined on the expectations of model parameters $\boldsymbol{\eta}$ and latent topical assignments Z. This interplay will yield a topical representation that could be more suitable for prediction tasks, as explained below and verified in experiments.

3.1.2 Variational Approximation Algorithm for MedLDA^r

Minimizing $\mathcal{L}^{bs}$ is intractable. Here, we use mean field methods (Jordan et al., 1999) widely employed in fitting LDA and sLDA to efficiently obtain an approximate q for problem P2.

Algorithm 1 Variational MedLDA^r

Input: corpus $\mathcal{D} = \{(\mathbf{y}, \mathbf{W})\}$, constants C and ϵ, and topic number K.
Output: Dirichlet parameters $\boldsymbol{\gamma}$, posterior distribution $q(\boldsymbol{\eta})$, parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and δ^2.
repeat
  for $d = 1$ **to** D **do**
    Update $\boldsymbol{\gamma}_d$ as in Eq. (18).
    for $n = 1$ **to** N **do**
      Update $\boldsymbol{\phi}_{dn}$ as in Eq. (19).
    end for
  end for
  Solve the dual problem D2 to get $q(\boldsymbol{\eta})$, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}^*$.
  Update $\boldsymbol{\beta}$ using Eq. (15), and update δ^2 using Eq. (16). Optimize $\boldsymbol{\alpha}$ with gradient descent or fix $\boldsymbol{\alpha}$ as $1/K$ times the ones vector.
until convergence

Specifically, we assume that q is a fully factorized mean-field approximation to p:

$$q(\boldsymbol{\eta}, \{\boldsymbol{\theta}_d, \mathbf{z}_d\}) = q(\boldsymbol{\eta}) \prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \prod_{n=1}^{N} q(z_{dn}|\boldsymbol{\phi}_{dn}), \tag{12}$$

where $\boldsymbol{\gamma}_d$ is a K-dimensional vector of Dirichlet parameters and each $\boldsymbol{\phi}_{dn}$ parameterizes a multinomial distribution over K topics. It is easy to verify that:

$$\mathbb{E}[Z_{dn}] = \boldsymbol{\phi}_{dn}, \text{ and } \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] = \mathbb{E}[\boldsymbol{\eta}]^\top (\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\phi}_{dn}). \tag{13}$$

Now, we develop a coordinate descent algorithm to solve the equivalent "unconstrained" formulation (11). The algorithm is outlined in Alg. 1 and detailed below.

(1) **Solve for** $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)$ **and** $q(\boldsymbol{\eta})$: When $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$ is fixed, this substep (in an equivalent constrained form) is to solve

$$\min_{q(\boldsymbol{\eta}), \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \mathbb{E}_{q(\boldsymbol{\eta})}[\mathcal{L}^s(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)] + KL(q(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})) + C \sum_{d=1}^{D} (\xi_d + \xi_d^*) \tag{14}$$

$$\forall d, \text{ s.t. : } \begin{cases} y_d - \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \leq \epsilon + \xi_d, & (\mu_d) \\ -y_d + \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \leq \epsilon + \xi_d^*, & (\mu_d^*) \\ \xi_d \geq 0, & (v_d) \\ \xi_d^* \geq 0, & (v_d^*), \end{cases}$$

where $\{\mu_d, \mu_d^*, v_d, v_d^*\}$ are lagrange multipliers. Since the margin constraints are not dependent on $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)$, we can solve for them using the same procedure as in sLDA, when $q(\boldsymbol{\eta})$ and $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$ are given. Specifically, for $\boldsymbol{\alpha}$, the same gradient descent method as in (Blei et al., 2003) can be applied; for $\boldsymbol{\beta}$, the update equations are the same as for sLDA:

$$\beta_{kw} \propto \sum_{d=1}^{D} \sum_{n=1}^{N} \mathbb{I}(w_{dn} = w) \phi_{dn}^k, \tag{15}$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals to 1 if the condition holds; otherwise 0; and for δ^2, the update rule is similar as that of sLDA but in an expected version, because $\boldsymbol{\eta}$ is a random variable:

$$\delta^2 = \frac{1}{D}\Big(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbb{E}[A]\mathbb{E}[\boldsymbol{\eta}] + \mathbb{E}[\boldsymbol{\eta}^\top\mathbb{E}[A^\top A]\boldsymbol{\eta}]\Big), \tag{16}$$

where $\mathbb{E}[\boldsymbol{\eta}^\top\mathbb{E}[A^\top A]\boldsymbol{\eta}] = \text{tr}(\mathbb{E}[A^\top A]\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^\top])$, and A is a $D \times K$ matrix whose rows are the vectors $\bar{Z}_d^\top$.

Solving for $q(\boldsymbol{\eta})$ can be done using Lagrangian methods, but it is a bit more delicate. For brevity, we postpone the details of this step after we have finished presenting the overall procedure. We denote the optimum lagrange multipliers by $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^*)$ and the optimum slack variables by $(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\xi}}^*)$.

(2) **Solve for** $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$: By fixing $q(\boldsymbol{\eta})$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)$, this substep (in an equivalent constrained form) is to solve

$$\min_{q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}), \boldsymbol{\xi}, \boldsymbol{\xi}^*} \mathbb{E}_{q(\boldsymbol{\eta})}[\mathcal{L}^s(q; \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta^2)] + C\sum_{d=1}^{D}(\xi_d + \xi_d^*) \tag{17}$$

$$\forall d, \text{ s.t. : } \begin{cases} y_d - \mathbb{E}[\boldsymbol{\eta}^\top\bar{Z}_d] \leq \epsilon + \xi_d \\ -y_d + \mathbb{E}[\boldsymbol{\eta}^\top\bar{Z}_d] \leq \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \geq 0, \end{cases}$$

Since the constraints are not dependent on $\boldsymbol{\gamma}_d$ and $q(\boldsymbol{\eta})$ is also not directly connected with $\boldsymbol{\theta}_d$, we get the same update rule for $\boldsymbol{\gamma}_d$ as in sLDA:

$$\boldsymbol{\gamma}_d = \boldsymbol{\alpha} + \sum_{n=1}^{N}\boldsymbol{\phi}_{dn}. \tag{18}$$

For $q(\mathbf{z}_d)$, in theory, we can do the optimization to get the optimal solution of $\boldsymbol{\phi}$ and the corresponding optimal lagrange multipliers. But the full optimization would be expensive, especially considering that this sub-step is within the most inner iteration loop and it would be performed for many times. Here, we adopt an approximation strategy, which performs a single step update of $\boldsymbol{\phi}$, rather than a full optimization. Note that this one-step approximation could lead to a slight increase of the objective function during the iterations. Our empirical studies show that this increase is usually within an acceptable range. More specifically, we fix $(\boldsymbol{\xi}, \boldsymbol{\xi}^*)$ at $(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\xi}}^*)$ (the optimum solution of the previous step) and set the lagrange multipliers to be $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^*)$ [7]. Then, we have the closed-form update equation

$$\boldsymbol{\phi}_{dn} \propto \exp\Big(\mathbb{E}[\log\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d] + \log p(w_{dn}|\boldsymbol{\beta}) + \frac{y_d}{N\delta^2}\mathbb{E}[\boldsymbol{\eta}] - \frac{2\mathbb{E}[\boldsymbol{\eta}^\top\boldsymbol{\phi}_{d,-n}\boldsymbol{\eta}] + \mathbb{E}[\boldsymbol{\eta}\circ\boldsymbol{\eta}]}{2N^2\delta^2} + \frac{\mathbb{E}[\boldsymbol{\eta}]}{N}(\hat{\mu}_d - \hat{\mu}_d^*)\Big), \tag{19}$$

7. Before we update $\boldsymbol{\phi}$, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^*)$ and $(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\xi}}^*)$ satisfy the optimal conditions (e.g., KKT conditions) of problem (17). So, they are the initially optimal solutions. But after we have updated $\boldsymbol{\phi}$, the KKT conditions do not hold. This is the reason why our strategy of not updating $(\boldsymbol{\mu}, \boldsymbol{\mu}^*)$ and $(\boldsymbol{\xi}, \boldsymbol{\xi}^*)$ could lead to a slight increase of the objective function.

where $\boldsymbol{\phi}_{d,-n} \triangleq \sum_{i \neq n} \boldsymbol{\phi}_{di}$; $\boldsymbol{\eta} \circ \boldsymbol{\eta}$ is the element-wise product; and the result of exponentiating a vector is a vector of the exponentials of its corresponding components. Note that the first two terms in the exponential are the same as those in LDA.

Remark 1 *From the update rule of $\boldsymbol{\phi}$ in Eq. (19), we can see that the essential differences between MedLDA^r and sLDA lie in the last three terms in the exponential of $\boldsymbol{\phi}_{dn}$. Firstly, the third and fourth terms are similar to those of sLDA, but in an expected version since we are learning the distribution $q(\boldsymbol{\eta})$ instead of a point estimate of $\boldsymbol{\eta}$. The second-order expectations $\mathbb{E}[\boldsymbol{\eta}^\top \boldsymbol{\phi}_{d,-n} \boldsymbol{\eta}]$ and $\mathbb{E}[\boldsymbol{\eta} \circ \boldsymbol{\eta}]$ mean that the co-variances of $\boldsymbol{\eta}$ (See Corollary 3 for an example) affect the distribution over topics. This makes our approach significantly different from a point estimation method, like sLDA, where no expectations or co-variances are involved in updating $\boldsymbol{\phi}_{dn}$. Secondly, the last term is from the max-margin regression formulation. For a document d, which lies on the decision boundary, i.e., a support vector, either μ_d or μ_d^* is non-zero, and the last term biases $\boldsymbol{\phi}_{dn}$ towards a distribution that favors a more accurate prediction on the document. Moreover, the last term is fixed for words in the document and thus will directly affect the latent representation of the document, i.e., $\boldsymbol{\gamma}_d$. Therefore, the latent representation $\boldsymbol{\theta}_d$ inferred under MedLDA^r can be more suitable for supervised prediction tasks. Our empirical studies further verify this, as we shall see in Section 5.*

Now, we turn to the sub-step of solving for $q(\boldsymbol{\eta})$, as well as the slack variables and lagrange multipliers. Specifically, we have the following result.

Proposition 2 *For MedLDA^r, the optimum solution of $q(\boldsymbol{\eta})$ has the form:*

$$q(\boldsymbol{\eta}) = \frac{p_0(\boldsymbol{\eta})}{Z} \exp\Big(\boldsymbol{\eta}^\top \sum_{d=1}^{D} (\hat{\mu}_d - \hat{\mu}_d^* + \frac{y_d}{\delta^2}) \mathbb{E}[\bar{Z}_d] - \boldsymbol{\eta}^\top \frac{\mathbb{E}[A^\top A]}{2\delta^2} \boldsymbol{\eta}\Big), \tag{20}$$

where $\mathbb{E}[A^\top A] = \sum_{d=1}^{D} \mathbb{E}[\bar{Z}_d \bar{Z}_d^\top]$, and $\mathbb{E}[\bar{Z}_d \bar{Z}_d^\top] = \frac{1}{N^2}(\sum_{n=1}^{N} \sum_{m \neq n} \boldsymbol{\phi}_{dn} \boldsymbol{\phi}_{dm}^\top + \sum_{n=1}^{N} \mathrm{diag}\{\boldsymbol{\phi}_{dn}\})$. The lagrange multipliers $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^*)$ are the solution of the dual problem of (14):

$$\begin{aligned} \text{D2}: \quad & \max_{\boldsymbol{\mu}, \boldsymbol{\mu}^*} \quad -\log Z - \epsilon \sum_{d=1}^{D} (\mu_d + \mu_d^*) + \sum_{d=1}^{D} y_d (\mu_d - \mu_d^*) \\ & \forall d, \text{ s.t.}: \quad \mu_d, \mu_d^* \in [0, C]. \end{aligned} \tag{21}$$

Proof (sketch) By setting the partial derivative of the Lagrangian functional over $q(\boldsymbol{\eta})$ equal to zero, we can get the solution of $q(\boldsymbol{\eta})$. Plugging $q(\boldsymbol{\eta})$ into the Lagrangian functional and solving for the optimal (v_d, v_d^*) and (ξ_d, ξ_d^*) as in the standard SVR to get the box constraints, we get the dual problem. ■

In MedLDA^r, we can choose different priors to introduce some regularization effects. For the standard normal prior, we have the following corollary:

Corollary 3 *Assume the prior $p_0(\boldsymbol{\eta}) = \mathcal{N}(0, I)$, where I is the identity matrix, then the optimum solution of $q(\boldsymbol{\eta})$ is*

$$q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\lambda}, \Sigma), \tag{22}$$

where $\boldsymbol{\lambda} = \Sigma(\sum_{d=1}^{D}(\hat{\mu}_d - \hat{\mu}_d^* + \frac{y_d}{\delta^2})\mathbb{E}[\bar{Z}_d])$ *is the mean and* $\Sigma = (I + 1/\delta^2\mathbb{E}[A^\top A])^{-1}$ *is a* $K \times K$ *co-variance matrix. The dual problem D2 is now:*

$$\begin{aligned} \max_{\boldsymbol{\mu},\boldsymbol{\mu}^*} \quad & -\frac{1}{2}\boldsymbol{\omega}^\top\Sigma\boldsymbol{\omega} - \epsilon\sum_{d=1}^{D}(\mu_d + \mu_d^*) + \sum_{d=1}^{D} y_d(\mu_d - \mu_d^*) \\ \forall d,\ \text{s.t.}: \quad & \mu_d, \mu_d^* \in [0, C], \end{aligned} \tag{23}$$

where $\boldsymbol{\omega} = \sum_{d=1}^{D}(\mu_d - \mu_d^* + \frac{y_d}{\delta^2})\mathbb{E}[\bar{Z}_d]$.

In the above Corollary, computation of Σ can be done robustly through Cholesky decomposition of $\delta^2 I + \mathbb{E}[A^\top A]$, an $O(K^3)$ procedure. Another example is the Laplace prior, which can lead to a shrinkage effect (Zhu and Xing, 2009) that is useful in sparse problems. In this paper, we focus on the normal prior and extension to the Laplace prior can be done similarly as in (Zhu and Xing, 2009). For the standard normal prior, the dual optimization problem is a QP problem and can be solved with any standard QP solvers, although they may not be so efficient. To leverage recent developments in learning support vector regression models, we first prove the following corollary:

Corollary 4 *Assume the prior* $p_0(\boldsymbol{\eta}) = \mathcal{N}(0, I)$*, then the mean* $\boldsymbol{\lambda}$ *of* $q(\boldsymbol{\eta})$ *in problem (14) is the optimum solution of the following problem:*

$$\begin{aligned} \min_{\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad & \frac{1}{2}\boldsymbol{\lambda}^\top\Sigma^{-1}\boldsymbol{\lambda} - \boldsymbol{\lambda}^\top(\sum_{d=1}^{D}\frac{y_d}{\delta^2}\mathbb{E}[\bar{Z}_d]) + C\sum_{d=1}^{D}(\xi_d + \xi_d^*) \\ \forall d,\ \text{s.t.}: \quad & \begin{cases} y_d - \boldsymbol{\lambda}^\top\mathbb{E}[\bar{Z}_d] \le \epsilon + \xi_d \\ -y_d + \boldsymbol{\lambda}^\top\mathbb{E}[\bar{Z}_d] \le \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \ge 0 \end{cases} \end{aligned} \tag{24}$$

Proof See Appendix A for details. ■

The above primal form can be re-formulated as a standard SVR problem. Specifically, we do Cholesky decomposition $\Sigma^{-1} = U^\top U$, where U is an upper triangular matrix with strict positive diagonal entries. Let $\boldsymbol{\nu} = \sum_{d=1}^{D}\frac{y_d}{\delta^2}\mathbb{E}[\bar{Z}_d]$, and we define $\boldsymbol{\lambda}' = U(\boldsymbol{\lambda} - \Sigma\boldsymbol{\nu})$; $y'_d = y_d - \boldsymbol{\nu}^\top\Sigma\mathbb{E}[\bar{Z}_d]$; and $\mathbf{x}_d = (U^{-1})^\top\mathbb{E}[\bar{Z}_d]$. Then, the above primal problem in Corollary 4 can be re-formulated as the following standard form:

$$\begin{aligned} \min_{\boldsymbol{\lambda}',\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad & \frac{1}{2}\|\boldsymbol{\lambda}'\|_2^2 + C\sum_{d=1}^{D}(\xi_d + \xi_d^*) \\ \forall d,\ \text{s.t.}: \quad & \begin{cases} y'_d - (\boldsymbol{\lambda}')^\top\mathbf{x}_d \le \epsilon + \xi_d \\ -y'_d + (\boldsymbol{\lambda}')^\top\mathbf{x}_d \le \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \ge 0 \end{cases}. \end{aligned} \tag{25}$$

Then, we can solve the standard SVR problem using existing algorithms, such as the working set selection algorithm implemented in SVM-light (Joachims, 1999), to get the dual parameters [8] $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}^*$ (as well as slack variables $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\xi}}^*$), which are needed to

8. Not all existing solvers return the dual parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}^*$. SVM-light is one nice package that provides both primal parameters $\boldsymbol{\lambda}'$ and the dual parameters. Note that the above transformation from (24) to (25) is done in the primal form and does not affect the solution of dual parameters of (23).

infer $\boldsymbol{\phi}$ as defined in (19), and the primal parameters $\boldsymbol{\lambda}'$ which we use to get $\boldsymbol{\lambda}$ by doing a reverse transformation since $\boldsymbol{\lambda}' = U(\boldsymbol{\lambda} - \Sigma\boldsymbol{\nu})$ as defined above. The other lagrange multipliers, which are not explicitly involved in topic inference and estimation of $q(\boldsymbol{\eta})$, are solved according to KKT conditions.

3.2 Classificational MedLDA

Now, we present the MedLDA classification model, of which the discrete labels of the documents are available, and our goal is to learn a supervised topic model specialized at predicting the labels of new documents through a discriminant function. We call this model a Classificational MedLDA, or simply, *MedLDA*^{c}.

Denoting the discrete response variable by Y, for brevity, we only consider the multi-class classification, where y takes values from a finite set $\mathcal{C} \triangleq \{1, 2, \cdots, J\}$. The binary case, where $\mathcal{C} \triangleq \{+1, -1\}$, can be easily defined based on a binary SVM and the optimization problem can be solved similarly. For classification, if the latent topic assignments $\mathbf{z} \triangleq \{z_1; \cdots; z_N\}$ of all the words in a document are given, we define the *latent* linear discriminant function

$$F(y, \mathbf{z}, \boldsymbol{\eta}; \mathbf{w}) = \boldsymbol{\eta}_y^\top \bar{\mathbf{z}}, \tag{26}$$

where $\bar{\mathbf{z}} \triangleq 1/N \sum_n z_n$, the same as in the case of MedLDA regression model; $\boldsymbol{\eta}_y$ is a class-specific K-dimensional parameter vector associated with class y; and $\boldsymbol{\eta}$ is a $|\mathcal{C}|K$-dimensional vector by stacking the elements of $\boldsymbol{\eta}_y$. Equivalently, F can be written as $F(y, \mathbf{z}, \boldsymbol{\eta}; \mathbf{w}) = \boldsymbol{\eta}^\top \mathbf{f}(y, \bar{\mathbf{z}})$, where $\mathbf{f}(y, \bar{\mathbf{z}})$ is a feature vector whose components from $(y-1)K+1$ to yK are those of the vector $\bar{\mathbf{z}}$ and all the others are 0.

However, we cannot directly use the latent function $F(y, \mathbf{z}, \boldsymbol{\eta}; \mathbf{w})$ to make prediction for an observed input $\mathbf{w}$ of a document because the topic assignments $\mathbf{z}$ are hidden variables. Here, we also treat $\boldsymbol{\eta}$ as a random vector and consider the general case to learn a distribution of $q(\boldsymbol{\eta})$. In order to deal with the uncertainty of $\mathbf{z}$ and $\boldsymbol{\eta}$, similar to MedLDA^r, we take the expectation over $q(\boldsymbol{\eta}, \mathbf{z})$ and define the *effective* discriminant function

$$F(y; \mathbf{w}) = \mathbb{E}[F(y, \mathbf{Z}, \boldsymbol{\eta}; \mathbf{w})] = \mathbb{E}[\boldsymbol{\eta}^\top \mathbf{f}(y, \bar{Z}) | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}], \tag{27}$$

where $\mathbf{Z} \triangleq \{Z_1; \cdots; Z_N\}$ is the set of topic assignment random variables and $\bar{Z} \triangleq 1/N \sum_n Z_n$ is the average topic assignment random variable as defined before. Then, the prediction rule for multi-class classification is naturally

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{C}} F(y; \mathbf{w}) = \operatorname*{argmax}_{y \in \mathcal{C}} \mathbb{E}[\boldsymbol{\eta}^\top \mathbf{f}(y, \bar{Z}) | \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}]. \tag{28}$$

Our goal here is to learn an optimal set of parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and distribution $q(\boldsymbol{\eta})$. As in MedLDA^r, we have the option of using either a supervised sLDA (Wang et al., 2009) or an unsupervised LDA as a building block of MedLDA^c to discover latent topical representations. However, as we have discussed in Section 2.1 and shown in (Wang et al., 2009) as well as Section 5.3.1, inference under sLDA can be harder and slower because the probability model of discrete Y in Eq. (5) is highly nonlinear over $\boldsymbol{\eta}$ and Z, both of which are latent variables in our case, and its normalization factor strongly couples the topic assignments of different words in the same document. Therefore, in this paper we

focus on the case of using an LDA that only models the likelihood of document contents $\mathbf{W}$ but not document label Y as the underlying topic model to discover latent representations Z. Even with this likelihood model, document labels can still influence topic learning and inference because they induce margin constraints pertinent to the topical distributions. As we shall see, the resultant MedLDA classification model can be easily and efficiently learned by utilizing existing high-performance SVM solvers. Moreover, since the goal of max-margin learning is to directly minimize a hinge loss (i.e., an upper bound of the empirical loss), we do not need a normalized distribution model for response variables Y.

3.2.1 Max-Margin Learning of LDA for Classification

The LDA component inside the MedLDA^c defines a likelihood function $p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ over the corpus $\mathcal{D}$, which is known to be intractable. Therefore, we choose to optimize its variational bound $\mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta})$ in Eq. (1), which facilitates efficient approximation algorithms. The integrated problem of discovering latent topical representations and learning a distribution of classifiers is defined as follows:

$$\text{P3(MedLDA}^{\text{c}}): \quad \min_{q, q(\boldsymbol{\eta}), \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}} \quad \mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta}) + KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta})) + \frac{C}{D}\sum_{d=1}^{D}\xi_d \tag{29}$$

$$\forall d,\ y \in \mathcal{C},\ \text{s.t.}: \quad \begin{cases} \mathbb{E}[\boldsymbol{\eta}^\top \Delta\mathbf{f}_d(y)] \geq \Delta\ell_d(y) - \xi_d \\ \xi_d \geq 0, \end{cases}$$

where q denotes the variational distribution $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$; $\Delta\ell_d(y)$ is a non-negative cost function (e.g., 0/1 cost as typically used in SVMs) that measures how different the prediction y is from the true class label y_d; $\Delta\mathbf{f}_d(y) \triangleq \mathbf{f}(y_d, \bar{Z}_d) - \mathbf{f}(y, \bar{Z}_d)$ [9]; and $\boldsymbol{\xi}$ are slack variables. It is typically assumed that $\Delta\ell_d(y_d) = 0$, i.e., no cost for a correct prediction. Finally,

$$\mathbb{E}[\boldsymbol{\eta}^\top \Delta\mathbf{f}_d(y)] = F(y_d; \mathbf{w}_d) - F(y; \mathbf{w}_d) \tag{30}$$

is the "*expected* margin" by which the true label y_d is favored over a prediction y.

Note that we have taken a full expectation to define $F(y; \mathbf{w})$, instead of taking the mode as done in latent SVMs (Felzenszwalb et al., 2010; Yu and Joachims, 2009), because expectation is a nice linear functional of the distributions under which it is taken, whereas taking the mode involves the highly nonlinear *argmax* function for discrete Z, which could lead to a harder inference task. Furthermore, due to the same reason to avoid dealing with a highly nonlinear discriminant function, we did not adopt the method in (Jebara, 2001) either, which uses log-likelihood ratio to define the discriminant function when considering latent variables in MED. Specifically, in our case, the max-margin constraints of the standard MED would be

$$\forall d,\ \forall y \in \mathcal{C},\ \log\frac{p(y_d|\mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(y|\mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta})} \geq \Delta\ell_d(y) - \xi_d, \tag{31}$$

which are highly nonlinear due to the complex form of the marginal likelihood $p(y|\mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int_{\boldsymbol{\theta}_d}\sum_{\mathbf{z}_d} p(y, \boldsymbol{\theta}_d, \mathbf{z}_d|\mathbf{w}_d, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Our linear expectation operator is an effective tool to deal with

9. Since multi-class SVM is a special case of max-margin Markov networks, we follow the common conventions and use the same notations as in structured max-margin methods (Taskar et al., 2003; Joachims et al., 2009).

latent variables in the context of maximum margin learning. In fact, besides the present work, we have successfully applied this operator to other challenging settings of learning latent variable structured prediction models with nontrivial dependence structures among output variables (Zhu et al., 2008) and learning nonparametric Bayesian models (Zhu et al., 2011a,b). These expected margin constraints also make MedLDA^c fundamentally different from the mixture of conditional max-entropy models (Pavlov et al., 2003), where constraints are based on moment matching, i.e., empirical expectations of features equal to their model expectations.

By setting $\boldsymbol{\xi}$ to their optimum solutions, i.e., $\xi_d = \max_y(\Delta\ell_d(y) - \mathbb{E}[\boldsymbol{\eta}^\top \Delta\mathbf{f}_d(y)])$, we can rewrite problem P3 in the form of regularized empirical loss minimization

$$\min_{q, q(\boldsymbol{\eta}), \boldsymbol{\alpha}, \boldsymbol{\beta}} \mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta}) + KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta})) + C\mathcal{R}(q, q(\boldsymbol{\eta})), \tag{32}$$

where

$$\mathcal{R}(q, q(\boldsymbol{\eta})) \triangleq \frac{1}{D}\sum_{d=1}^{D} \max_{y \in \mathcal{C}}(\Delta\ell_d(y) - \mathbb{E}[\boldsymbol{\eta}^\top \Delta\mathbf{f}_d(y)]) \tag{33}$$

is an upper bound of the training error of the prediction rule in Eq. (28) and C is again the regularization constant. However, different from MedLDA^r, which uses a Bayesian supervised sLDA as the underlying likelihood model, here the variational bound $\mathcal{L}^u$ does not contain a cross-entropy term on $q(\boldsymbol{\eta})$ for its regularization (as in $\mathcal{L}^{bs}$ in Eq. (9)). Therefore, we include the KL-divergence in problem P3 as an explicit entropic regularizer for the distribution $q(\boldsymbol{\eta})$.

The rationale underlying MedLDA^c is similar to that of MedLDA^r, that is, we want to find latent topical representations $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$ and a model parameter distribution $q(\boldsymbol{\eta})$ which on one hand tend to predict as accurate as possible on training data, while on the other hand tend to explain the data well. The two parts are closely coupled by the expected margin constraints.

3.2.2 Variational Algorithm for MedLDA^c

As in MedLDA^r, we make the fully-factorized mean field assumption that

$$q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}) = \prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \prod_{n=1}^{N} q(z_{dn}|\boldsymbol{\phi}_{dn}), \tag{34}$$

where $\boldsymbol{\gamma}_d$ and $\boldsymbol{\phi}_{dn}$ are variational parameters, having the same meaning as in MedLDA^r. Then, we have $\mathbb{E}[\boldsymbol{\eta}^\top \mathbf{f}(y, \bar{Z}_d)] = \mathbb{E}[\boldsymbol{\eta}]^\top \mathbf{f}(y, 1/N \sum_{n=1}^{N} \boldsymbol{\phi}_{dn})$. We develop a similar coordinate descent algorithm to solve the "unconstrained" formulation in (32). Since the constraints in P3 are not on $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, their update rules are the same as in the case of MedLDA^r and we omit the details here. Below, we explain the optimization over $q(\{\mathbf{z}_d\})$ and $q(\boldsymbol{\eta})$ and show the insights of the max-margin topic model.

Optimize over $q(\boldsymbol{\eta})$: As in the case of regression, we have the following solution:

Corollary 5 *When $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$ are fixed, the optimum solution $q(\boldsymbol{\eta})$ of MedLDA^c in problem P3 has the form:*

$$q(\boldsymbol{\eta}) = \frac{1}{Z} p_0(\boldsymbol{\eta}) \exp\Big(\boldsymbol{\eta}^\top (\sum_{d=1}^{D} \sum_{y \in \mathcal{C}} \hat{\mu}_d^y \mathbb{E}[\Delta \mathbf{f}_d(y)])\Big), \tag{35}$$

where the lagrange multipliers $\hat{\boldsymbol{\mu}}$ are the optimum solution of the dual problem:

$$\begin{aligned} \text{D3}: \quad & \max_{\boldsymbol{\mu}} - \log Z + \sum_{d=1}^{D} \sum_{y \in \mathcal{C}} \mu_d^y \Delta \ell_d(y) \\ & \forall d, \text{ s.t. }: \sum_{y \in \mathcal{C}} \mu_d^y \in [0, \frac{C}{D}], \end{aligned} \tag{36}$$

Again, we can choose different priors in MedLDA^c for different regularization effects. We consider the normal prior in this paper. For the standard normal prior $p_0(\boldsymbol{\eta}) = \mathcal{N}(0, I)$, we can get: $q(\boldsymbol{\eta})$ is a normal with a shifted mean, i.e., $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\lambda}, I)$, where $\boldsymbol{\lambda} = \sum_{d=1}^{D} \sum_{y \in \mathcal{C}} \mu_d^y \mathbb{E}[\Delta \mathbf{f}_d(y)]$, and the dual problem D3 thus becomes the same as the dual problem of a standard multi-class SVM (Crammer and Singer, 2001):

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & -\frac{1}{2} \| \sum_{d=1}^{D} \sum_{y \in \mathcal{C}} \mu_d^y \mathbb{E}[\Delta \mathbf{f}_d(y)] \|_2^2 + \sum_{d=1}^{D} \sum_{y \in \mathcal{C}} \mu_d^y \Delta \ell_d(y) \\ \forall d, \text{ s.t. }: \quad & \sum_{y \in \mathcal{C}} \mu_d^y \in [0, \frac{C}{D}]. \end{aligned} \tag{37}$$

The primal form of problem (37) is

$$\begin{aligned} \min_{\boldsymbol{\lambda}, \boldsymbol{\xi}} \quad & \frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 + \frac{C}{D} \sum_{d=1}^{D} \xi_d \\ \forall d, \ \forall y \in \mathcal{C}, \text{ s.t. }: \quad & \begin{cases} \boldsymbol{\lambda}^\top \mathbb{E}[\Delta \mathbf{f}_d(y)] \geq \Delta \ell_d(y) - \xi_d \\ \xi_d \geq 0. \end{cases} \end{aligned} \tag{38}$$

Optimize over $q(\{\mathbf{z}_d\})$: again, since q is fully factorized, we can perform the optimization on each document separately. We have

$$\boldsymbol{\phi}_{dn} \propto \exp\Big(\mathbb{E}[\log \boldsymbol{\theta}_d | \boldsymbol{\gamma}_d] + \log p(w_{dn} | \boldsymbol{\beta}) + \frac{1}{N} \sum_{y \in \mathcal{C}} \hat{\mu}_d^y \mathbb{E}[\boldsymbol{\eta}_{y_d} - \boldsymbol{\eta}_y]\Big), \tag{39}$$

where we can see that the first two terms in Eq. (39) are the same as in unsupervised LDA (Blei et al., 2003), and the last term is due to the max-margin formulation of P3 and reflects our intuition that the discovered latent topical representation is influenced by the margin constraints. For those examples that are on the decision boundary, i.e., support vectors, their associated lagrange multipliers are non-zero and thus the last term acts as a regularizer that biases the model towards discovering latent representations that tend to make more accurate prediction on these difficult examples. Moreover, this term is fixed

for words in the document and thus will directly affect the latent representation of the document (i.e., $\boldsymbol{\gamma}_d$) and therefore leads to a discriminative latent representation. As we shall see in Section 5, such an estimate is more suitable for the classification task: for instance, MedLDA^c needs much fewer support vectors than the max-margin classifiers that are built on raw text or the topical representations discovered by LDA.

The above formulation of MedLDA^c has a slack variable associated with each document. This is known as the *n-slack* formulation (Joachims et al., 2009). Another equivalent formulation, which can be more efficiently solved, is the so called *1-slack* formulation. The 1-slack MedLDA^c can be written as follows

$$\text{P4(1-slack MedLDA}^c\text{)}: \min_{q,q(\boldsymbol{\eta}),\boldsymbol{\alpha},\boldsymbol{\beta},\xi} \quad \mathcal{L}^u(q) + KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta})) + C\xi \tag{40}$$

$$\forall(\bar{y}_1,\cdots,\bar{y}_D),\ \text{s.t.}: \begin{cases} \frac{1}{D}\sum_{d=1}^{D}\mathbb{E}[\boldsymbol{\eta}^\top\Delta\mathbf{f}_d(\bar{y}_d)] \geq \frac{1}{D}\sum_{d=1}^{D}\Delta\ell_d(\bar{y}_d) - \xi \\ \xi \geq 0. \end{cases}$$

By using the above developed variational algorithm and the cutting plane algorithm for solving the 1-slack as well as n-slack multi-class SVMs (Joachims et al., 2009), which is implemented in the SVM^{struct} package [10], we can solve the 1-slack or n-slack MedLDA^c model efficiently, as we shall see in Section 5.3.1. SVM^{struct} provides the solutions of the primal parameters $\boldsymbol{\lambda}$ as well as the dual parameters $\boldsymbol{\mu}$, which are needed to do inference.

4. MedTM: a general framework

We have presented two variants of MedLDA for discovering predictive latent topical representations of documents, as well as learning discriminating topics from the corpus; and we have shown that the underlying topic model that defines data likelihood can be either a supervised or an unsupervised LDA. In fact, the likelihood component of MedLDA can be any other form of generative topic model, such as correlated topic models (Blei and Lafferty, 2005), or latent space Markov random fields, such as exponential family harmoniums (Welling et al., 2004; Xing et al., 2005; Chen et al., 2010). The same principle can also be applied to upstream latent topic models, which have been widely used in computer vision applications (Sudderth et al., 2005; Fei-Fei and Perona, 2005; Zhu et al., 2010). In this section, we formulate a general framework of applying the max-margin principle to learn discriminative latent topic models when supervising side information is available, and we discuss more insights on developing approximate inference algorithms.

Formally, a *maximum entropy discrimination topic model* (MedTM) consists of two components – an underlying topic model that fits observed data and a MED max-margin model that performs prediction. In an MedTM, we distinguish two types of latent variables – we use Υ to denote the parameters of the model pertaining to the prediction task (e.g., $\boldsymbol{\eta}$ in sLDA), and H to denote the topic assignment and mixing variables (e.g., $\mathbf{z}$ and $\boldsymbol{\theta}$). Let Ψ denote the parameters of the underlying topic model (e.g., the Dirichlet parameter $\boldsymbol{\alpha}$ and topics $\boldsymbol{\beta}$). Then, $p(\mathcal{D}|\Psi)$ is the marginal data likelihood of the corpus $\mathcal{D}$, which may or may not include the supervising side information depending on choice of specific form of the underlying topic model.

10. http://svmlight.joachims.org/svm_multiclass.html

As discussed before, for a general topic model, $p(\mathcal{D}|\Psi)$ is intractable, therefore a generic variational method can be employed. Let $q(\Upsilon, H)$ be a variational distribution to approximate the posterior $p(\Upsilon, H|\mathcal{D}, \Psi)$. By the properties of KL-divergence, the following equality holds if we do not make any restricting assumption of $q(\Upsilon, H)$

$$\begin{aligned} -\log p(\mathcal{D}|\Psi) &= \min_{q(\Upsilon,H)} \Big(-\mathbb{E}_{q(\Upsilon,H)}[\log p(\Upsilon, H, \mathcal{D}|\Psi)] - \mathcal{H}(q(\Upsilon, H))\Big) \qquad (41) \\ &= \min_{q(\Upsilon,H)} \Big(\mathbb{E}_{q(\Upsilon)}\Big[-\mathbb{E}_{q(H|\Upsilon)}[\log p(H, \mathcal{D}|\Psi, \Upsilon)] - \mathcal{H}(q(H|\Upsilon))\Big] + KL(q(\Upsilon)\|p_0(\Upsilon))\Big), \end{aligned}$$

where $p_0(\Upsilon)$ is the prior distribution of Υ. Let us define

$$\mathcal{L}^t(q(H|\Upsilon); \Psi, \Upsilon) \triangleq -\mathbb{E}_{q(H|\Upsilon)}[\log p(H, \mathcal{D}|\Psi, \Upsilon)] - \mathcal{H}(q(H|\Upsilon)).$$

Then, $\mathcal{L}^t(q(H|\Upsilon); \Psi, \Upsilon)$ is the variational bound of the data likelihood associated with the underlying topic model. For instance, when the underlying topic model is supervised sLDA, $\mathcal{L}^t$ reduces to $\mathcal{L}^s$, as we discussed in Eq. (9). When the underlying topic model is unsupervised LDA, the corpus $\mathcal{D}$ only contains document contents, and $p(H, \mathcal{D}|\Psi, \Upsilon) = p(H, \mathcal{D}|\Psi)$. The reduction of $\mathcal{L}^t$ to $\mathcal{L}^u$ needs a simplifying assumption that $q(\Upsilon, H) = q(\Upsilon)q(H)$ (in fact, much stricter assumptions on q are usually needed to make the learning of MedLDA^c tractable).

Mathematically, we define MedTM as solving the following entropic-regularized problem:

$$\begin{aligned} \text{P5(MedTM)}: \min_{q(\Upsilon,H),\Psi,\boldsymbol{\xi}} \mathbb{E}_{q(\Upsilon)}\Big[\mathcal{L}^t(q(H|\Upsilon); \Psi, \Upsilon)\Big] + KL(q(\Upsilon)\|p_0(\Upsilon)) + U(\boldsymbol{\xi}) \qquad (42) \\ \text{s.t.}: \quad q(\Upsilon, H) \text{ satisfies the expected margin constraints.} \end{aligned}$$

where U is a convex function over slack variables, such as $U(\boldsymbol{\xi}) = \frac{C}{D}\sum_d \xi_d$ in MedLDA^c. As we have discussed in Section 3.2.1, by using the linear expectation operator, our expected margin constraints are different from and simpler than those derived using a log-likelihood ratio function in the standard MED with latent variables (Jebara, 2001).

This formulation allows efficient approximate inference to be developed. In general, the difficulty of solving the optimization problem of MedTM lies in two aspects. First, the data likelihood or its equivalent variational form as involved in the objective function is generally intractable to compute if we do not make any restricting assumption about $q(\Upsilon, H)$. Second, the posterior inference (e.g., in LDA) as required in evaluating the margin constraints is generally intractable. Based on recent developments on learning latent topic models, two commonly used approaches can be applied to get an approximate solution to P5(MedTM), namely, Markov Chain Monte Carlo (MCMC) (Griffiths and Steyvers, 2004) and variational (Blei et al., 2003; Teh et al., 2006) methods. For variational methods, which are our focus in this paper, we need to make some additional restricting assumptions, such as the commonly used mean field assumption, about the distribution $q(\Upsilon, H)$. Then, P5 can be efficiently solved with a coordinate descent procedure, similar to what we have done for MedLDA^r and MedLDA^c. For MCMC methods, the difference lies in sampling from the distribution $q(\Upsilon, H)$ under margin constraints – evaluating the expected margin constraints is easy once we obtain samples from the posterior. Several approaches were proposed to deal

with the problem of sampling from a distribution under some constraints such as (Schofield, 2007; Griffiths, 2002; Rodriguez-Yam et al., 2004; Damien and Walker, 2001) to name a few, and we plan to investigate their suitability to our case in the future.

Finally, based on the recent extensions of MED to the structured prediction setting (Zhu and Xing, 2009; Zhu et al., 2008), the basic principle of MedLDA can be similarly extended to perform structured prediction, where multiple response variables are predicted simultaneously and thus their mutual dependencies can be exploited to achieve globally consistent and optimal predictions. Likelihood based structured prediction latent topic models have been developed in different scenarios, such as image annotation (He and Zemel, 2008) and statistical machine translation (Zhao and Xing, 2006). Extension of MedLDA to the structured prediction setting could provide a promising alternative for such problems.

5. Experiments

In this section, we provide qualitative as well as quantitative evaluation of MedLDA on topic estimation, document classification and regression. For MedLDA and other topic models (except DiscLDA whose implementation details are explained in footnote 14), we optimize the K-dimensional Dirichlet parameters $\boldsymbol{\alpha}$ using the Newton-Raphson method (Blei et al., 2003). For initialization, we set $\boldsymbol{\phi}$ to be uniform and each topic $\boldsymbol{\beta}_k$ to be a uniform distribution plus a very small random noise, and the posterior mean of $\boldsymbol{\eta}$ to be zero. We have published our implementation on the website: *http://www.ml-thu.net/∼jun/software.html*. In all the experimental results, by default, we also report the standard deviation for a topic model with five randomly initialized runs.

5.1 Topic Estimation

We begin with an empirical assessment of topic estimation by MedLDA on the 20 Newsgroups data set with a standard list of stop words [11] removed. The data set contains about 20,000 postings in 20 related categories. We compare with unsupervised LDA [12]. We fit the data set to a 110-topic MedLDA^c model, which exploits the supervising category information, and a 110-topic unsupervised LDA, which ignores category information.

Figure 2 shows the 2D embedding of the inferred topic proportions $\boldsymbol{\theta}$ (approximated by the inferred variational posterior means) by MedLDA^c and LDA using the t-SNE stochastic neighborhood embedding (van der Maaten and Hinton, 2008) method, where each dot represents a document and each color-shape pair represents a category. Visually, the max-margin based MedLDA^c produces a better grouping and separation of the documents in different categories. In contrast, unsupervised LDA does not produce a well separated embedding, and documents in different categories tend to mix together. Intuitively, a well-separated representation is more discriminative for document categorization. This is further empirically supported in Section 5.2. Note that a similar embedding was presented in (Lacoste-Julien et al., 2008), where the transformation matrix in their model is predesigned. The results of MedLDA^c in Figure 2 are *automatically* learned.

11. http://mallet.cs.umass.edu/
12. We implemented LDA based on the public variational inference code by Dr. David Blei, using same data structures as MedLDA for fair comparison.

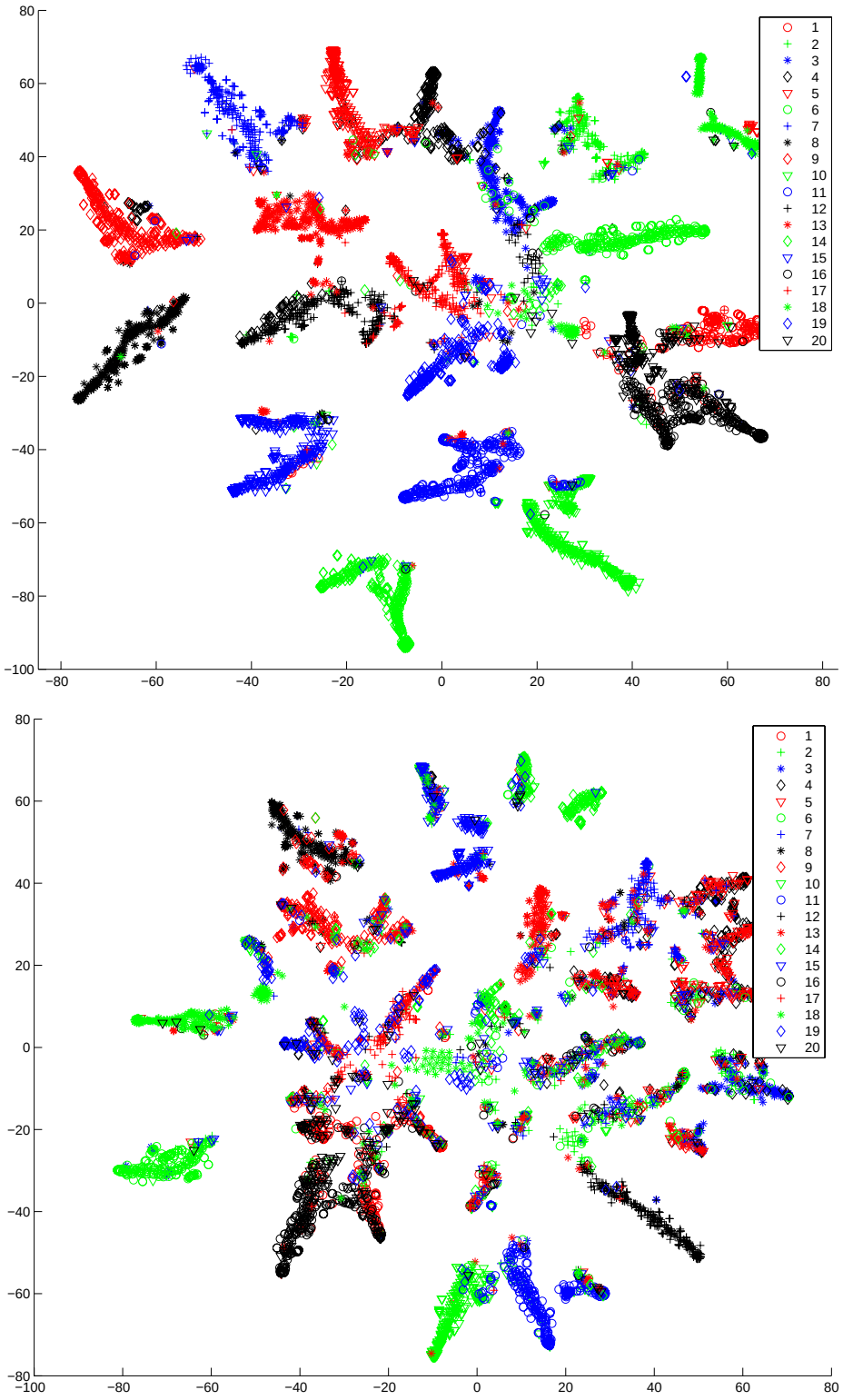


Figure 2: t-SNE 2D embedding of the topical representation by: MedLDA^c (above) and unsupervised LDA (below). The mapping between each index and category name can be found in: http://people.csail.mit.edu/jrennie/20Newsgroups/.

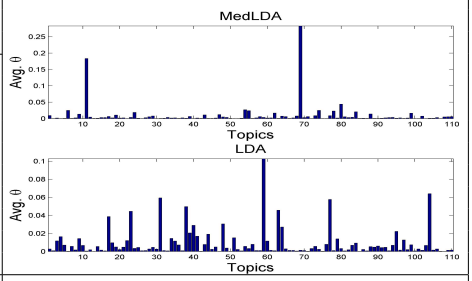
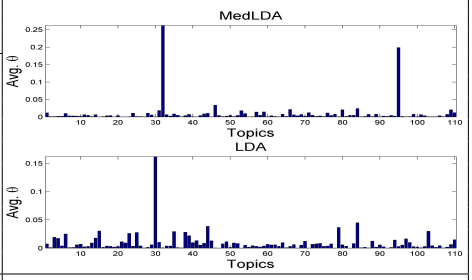
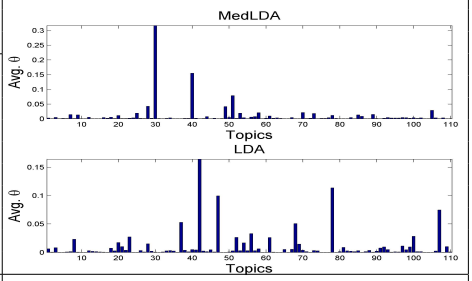
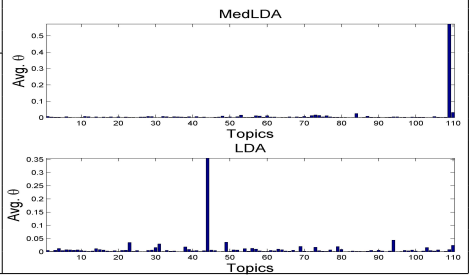
Class	MedLDA			LDA			Average θ per class
comp.graphics	T 69	T 11	T 80	T 59	T 104	T 31	
	image	graphics	db	image	ftp	card	
	jpeg	image	key	jpeg	pub	monitor	
	gif	data	chip	color	graphics	dos	
	file	ftp	encryption	file	mail	video	
	color	software	clipper	gif	version	apple	
	files	pub	system	images	tar	windows	
	bit	mail	government	format	file	drivers	
	images	package	keys	bit	information	vga	
	format	fax	law	files	send	cards	
	program	images	escrow	display	server	graphics	
sci.electronics	T 32	T 95	T 46	T 30	T 84	T 44	
	ground	audio	source	power	water	sale	
	wire	output	rs	ground	energy	price	
	power	input	time	wire	air	offer	
	wiring	signal	john	circuit	nuclear	shipping	
	don	chip	cycle	supply	loop	sell	
	current	high	low	voltage	hot	interested	
	circuit	data	dixie	current	cold	mail	
	neutral	mhz	dog	wiring	cooling	condition	
	writes	time	weeks	signal	heat	email	
	work	good	face	cable	temperature	cd	
politics.mideast	T 30	T 40	T 51	T 42	T 78	T 47	
	israel	turkish	israel	israel	jews	armenian	
	israeli	armenian	lebanese	israeli	jewish	turkish	
	jews	armenians	israeli	peace	israel	armenians	
	arab	armenia	lebanon	writes	israeli	armenia	
	writes	people	people	article	arab	turks	
	people	turks	attacks	arab	people	genocide	
	article	greek	soldiers	war	arabs	russian	
	jewish	turkey	villages	lebanese	center	soviet	
	state	government	peace	lebanon	jew	people	
	rights	soviet	writes	people	nazi	muslim	
misc.forsale	T 109	T 110	T 84	T 44	T 94	T 49	
	sale	drive	mac	sale	don	drive	
	price	scsi	apple	price	mail	scsi	
	shipping	mb	monitor	offer	call	disk	
	offer	drives	bit	shipping	package	hard	
	mail	controller	mhz	sell	writes	mb	
	condition	disk	card	interested	send	drives	
	interested	ide	video	mail	number	ide	
	sell	hard	speed	condition	ve	controller	
	email	bus	memory	email	hotel	floppy	
	dos	system	system	cd	credit	system	

Figure 3: Top topics under each class as discovered by the MedLDA and LDA models.

It is also interesting to examine the discovered topics and their relevance to class labels. In Figure 3 we show the top topics in four example categories as discovered by both MedLDA^c and LDA. Here, the semantic meaning of each topic is represented by the first 10 high probability words.

To visually illustrate the discriminative power of the latent representations, i.e., the topic proportion vector $\boldsymbol{\theta}$ of documents, we illustrate and compare the per-class distribution over topics for each model at the right side of Figure 3. This distribution is computed by averaging the expected topic vector of the documents in each class. We can see that MedLDA^c yields sharper, sparser and fast decaying per-class distributions over topics. For the documents in different categories, we can see that their per-class average distributions over

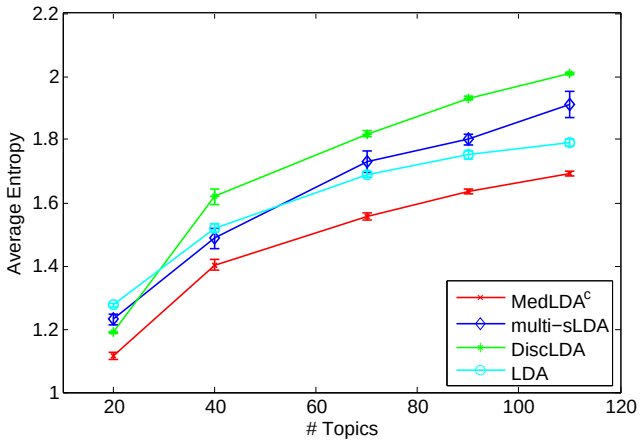


Figure 4: The average entropy of $\boldsymbol{\theta}$ over documents of different topic models on 20 Newsgroups data.

topics are very different, which suggests that the topical representations by MedLDA^c have a good discrimination power. Also, the sharper and sparser representations by MedLDA^c can result in a simpler max-margin classifier (e.g., with fewer support vectors), as we shall see in Section 5.2.1. All these observations suggest that the topical representations discovered by MedLDA^c have a better discriminative power and are more suitable for prediction tasks (Please see Section 5.2 for prediction performance). This behavior of MedLDA^c is in fact due to the regularization effect enforced over $\boldsymbol{\phi}$ as shown in Eq. (39). On the other hand, LDA seems to discover topics that model the fine details of documents, possibly at the cost of achieving weaker discrimination power (i.e., it discovers different variations of the same topic which results in a flat per-class distribution over topics). For instance, in the class *comp.graphics*, MedLDA^c mainly models documents in this class using two salient, discriminative topics (T69 and T11) whereas LDA results in a much flatter distribution. Moreover, in the cases where LDA and MedLDA^c discover comparably the same set of topics in a given class (like *politics.mideast* and *misc.forsale*), MedLDA^c results in a sharper low dimensional representation.

A quantitative measure for the sparsity or sharpness of the distributions over topics is the entropy. We compute the entropy of the inferred topic proportion for each document and take the average over the corpus. Here, we compare MedLDA^c with unsupervised LDA, supervised sLDA for multi-class classification (multi-sLDA) [13] (Wang et al., 2009), and DiscLDA [14] (Lacoste-Julien et al., 2008). For DiscLDA, as in the original paper, we

13. We thank the authors for providing their implementation, on which we made necessary slight modifications, e.g., improving the time efficiency and optimizing $\boldsymbol{\alpha}$.

14. DiscLDA is a conditional model that uses class-specific topics and shared topics. Since the code is not publicly available, we implemented an in-house version by following the same strategy in the original paper and share K_1 topics across classes and allocate K_0 topics to each class, where $K_1 = 2K_0$, and we varied $K_0 = \{1, 2, \cdots\}$. We should note here that (Lacoste-Julien et al., 2008; Lacoste-Julien, 2009) gave an optimization algorithm for learning the topic structure (i.e., a transformation matrix), however since the code is not available, we resorted to one of the fixed splitting strategies mentioned in the paper. Moreover, for the multi-class case, the authors only reported results using the same fixed splitting strategy we mentioned above. For the number of iterations for training and inference, we followed (Lacoste-Julien, 2009). Moreover, following (Lacoste-Julien, 2009) and personal communication with the first author, we

fix the transformation matrix and set it to be diagonally sparse. We use the standard training/testing split [15] to fit the models on training data and infer the topic distributions on testing documents. Figure 4 shows the average entropy of different models on testing documents when different topic numbers are chosen. For DiscLDA, we set the class-specific topic number $K_0 = 1, 2, 3, 4, 5$ and correspondingly $K = 22, 44, 66, 88, 110$. We can see that MedLDA^c yields the smallest entropy, which indicates that the probability mass is concentrated on quite a few topics, consistent with the observations in Figure 3. In contrast, for unsupervised LDA, the probability mass is more uniformly distributed on many topics (again consistent with Figure 3), which results in a higher entropy. For DiscLDA, although the transformation matrix is designed to be diagonally sparse, the distributions over the class-specific topics and shared topics are flat. Therefore, the entropy is also high. Using automatically learned transition matrices might improve the sparsity of DiscLDA.

5.2 Prediction Accuracy

In this subsection, we provide a quantitative evaluation of MedLDA on prediction performance for both document classification and regression.

5.2.1 Classification

We perform binary and multi-class classification on the 20 Newsgroup data set. To obtain a baseline, we first fit all the data to an LDA model, and then use the latent representation of the training [16] documents as features to build a binary or multi-class SVM classifier. We denote this baseline by *LDA+SVM*.

Binary Classification: As in (Lacoste-Julien et al., 2008), the binary classification is to distinguish postings of the newsgroup *alt.atheism* and the postings of the group *talk.religion.misc*. The training set contains 856 documents with a split of 480/376 over the two categories, and the test set contains 569 documents with a split of 318/251 over the two categories. Therefore, the *naïve baseline* that predicts the most frequent category for all test documents has accuracy 0.672.

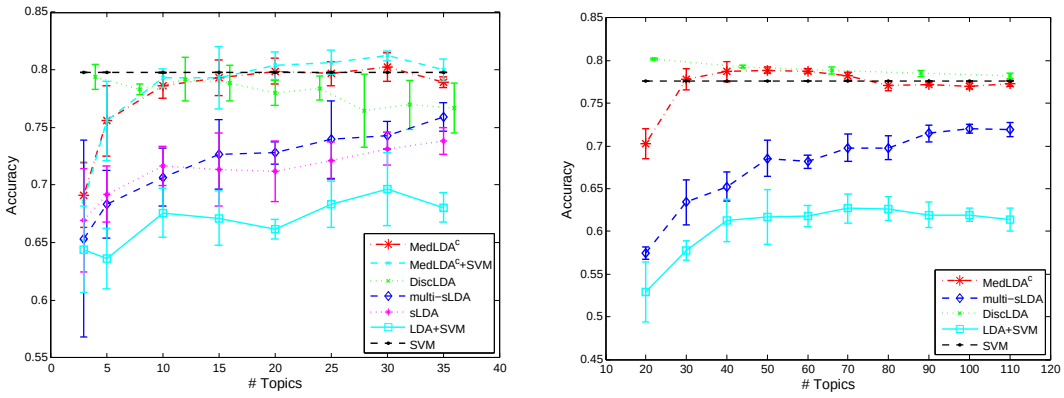
We compare the binary MedLDA^c with supervised LDA, DiscLDA, LDA+SVM, and the standard binary SVM built on raw text features. For supervised LDA, we use both the regression model (sLDA) (Blei and McAuliffe, 2007) and the multi-class classification model (multi-sLDA) (Wang et al., 2009). For the sLDA regression model, we fit it using the binary representation (0/1) of the classes, and use a threshold 0.5 to make prediction. For MedLDA^c, to see whether a second-stage max-margin classifier can improve the performance, we also build a method of *MedLDA^c+SVM*, similar to LDA+SVM. For DiscLDA, we fix the transition matrix. Automatically learning the transition matrix can yield slightly better results, as reported in (Lacoste-Julien, 2009). For all the above methods that utilize the class label information, they are fit *ONLY* on the training data.

We use the SVM-light (Joachims, 1999), which provides both primal and dual parameters, to build SVM classifiers and to estimate the posterior mean of $\boldsymbol{\eta}$ in MedLDA^c. The

used symmetric Dirichlet priors on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and set the Dirichlet parameters at 0.01 and $0.1/(K_0 + K_1)$, respectively.

15. http://people.csail.mit.edu/jrennie/20Newsgroups/

16. We use the training/testing split in: http://people.csail.mit.edu/jrennie/20Newsgroups/



(a) (b)

Figure 5: Classification accuracy of different models for: (a) binary and (b) multi-class classification on the 20 Newsgroup data.

parameter C is chosen via 5 fold cross-validation during training from $\{k^2 : k = 1, \cdots, 8\}$. For each model, we run the experiments for 5 times and take the average as the final results. The prediction accuracy of different models with respect to the number of topics is shown in Figure 5(a). For DiscLDA, we follow (Lacoste-Julien et al., 2008) to set $K = 2K_0 + K_1$, where K_0 is the number of class-specific topics and K_1 is the number of shared topics, and $K_1 = 2K_0$. Here, we set $K_0 = 1, \cdots, 8, 10$.

We can see that the max-margin MedLDA^c performs better than the likelihood-based downstream models, include multi-sLDA, sLDA, and the baseline LDA+SVM. The best performances of the two discriminative models (i.e., MedLDA^c and DiscLDA) are comparable. However, MedLDA^c is easier to learn and faster in testing, as we shall see in Section 5.3.2. Moreover, the different approximate inference algorithms used in MedLDA^c (i.e., variational approximation) and DiscLDA (i.e., Monte Carlo sampling methods) can also make the performance different. In our alternative implementation using collapsed variational inference (Teh et al., 2006) method for MedLDA^c (preliminary results in preparation for submission), we were able to achieve slightly better results. However, the collapsed variational method is much more expensive. Finally, since MedLDA^c already integrates the max-margin principle into its training, our conjecture is that the combination of MedLDA^c and SVM does not further improve the performance much on this task. We believe that the slight differences between MedLDA^c and MedLDA^c+SVM are due to the tuning of regularization parameters. For efficiency, we do not change the regularization constant C during training MedLDA^c. The performance of MedLDA^c would be improved if we select a good C in different iterations because the data representation is changing.

Multi-class Classification: We perform multi-class classification on 20 Newsgroups with all the 20 categories. The data set has a balanced distribution over the categories. For the test set, which contains 7505 documents in total, the smallest category has 251 documents and the largest category has 399 documents. For the training set, which contains 11269 documents, the smallest and the largest categories contain 376 and 599 documents,

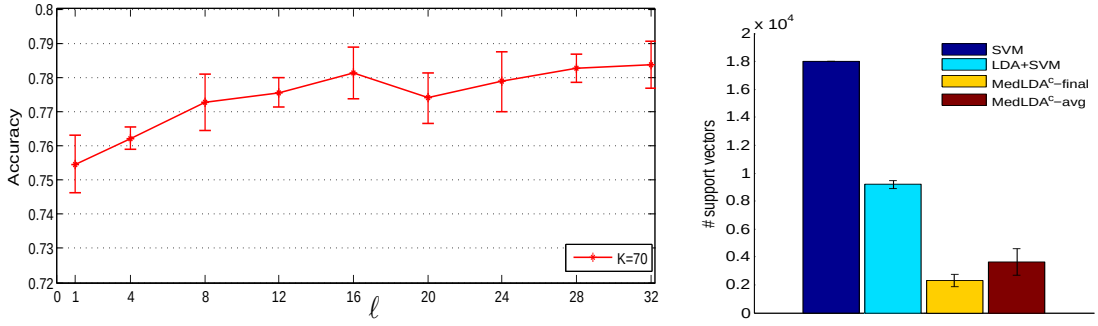
respectively. Therefore, the *naïve baseline* that predicts the most frequent category for all the test documents has the classification accuracy 0.0532.

We compare MedLDA^c with LDA+SVM, multi-sLDA, DiscLDA, and the standard multi-class SVM built on raw text. We use the SVM^{struct} package with a cost function as $\Delta\ell_d(y) \triangleq \ell\mathbb{I}(y \neq y_d)$ to solve the sub-step of learning $q(\boldsymbol{\eta})$ and build the SVM classifiers for LDA+SVM. The parameter ℓ is selected with 5 fold cross-validation [17]. The average results as well as standard deviations over 5 randomly initialized runs are shown in Figure 5(b). For DiscLDA, we use the same equation as in (Lacoste-Julien et al., 2008) to set the number of topics and set $K_0 = 1, \cdots, 5$. We can see that all the supervised topic models discover more predictive topical representations for classification, and the discriminative max-margin MedLDA^c and DiscLDA perform comparably, slightly better than the standard multi-class SVM (about 0.013 ± 0.003 improvement in accuracy). However, as we have stated and will show in Section 5.3.2, MedLDA^c is faster in testing than DiscLDA. As we shall see shortly, MedLDA^c needs much fewer support vectors than standard SVM.

Figure 6(a) shows the multi-class classification accuracy on the 20 Newsgroups data set for MedLDA^c with 70 topics. We show the results with ℓ manually set at $1, 4, 8, 12, \cdots, 32$. We can see that although the default 0/1-cost works well for MedLDA^c, we can get better accuracy if we use a larger cost for penalizing wrong predictions. The performance is quite stable when ℓ is set to be larger than 8. The reason why ℓ affects the performance is that ℓ as well as C control: 1) the scale of the posterior mean of $\boldsymbol{\eta}$ and the Lagrangian multipliers $\boldsymbol{\mu}$, whose dot-product regularizes the topic mixing proportions in Eq. (39); and 2) the goodness of fit of the MED large-margin classifier on the data (Please see (Joachims et al., 2009) for another practical example that uses $0/\ell$-cost, where ℓ is set at 100). For practical reasons, we only try a small subset of candidate C values in parameter search, which can also influence the difference on performance in Figure 6(a). Performing very careful parameter search on C could possibly shrink the difference. Finally, for a small ℓ (e.g., 1 for the standard 0/1-cost), we usually need a large C in order to obtain good performance. But our empirical experience with SVM^{struct} shows that the multi-class SVM with a larger C (and smaller ℓ) is typically more expensive to train than the SVM with a larger ℓ (and smaller C). That is one reason why we choose to use a large ℓ.

Figure 6(b) shows the number of support vectors for MedLDA^c, LDA+SVM, and the multi-class SVM built on raw text features, which are high-dimensional ($\sim$60,000 dimension for 20 Newsgroup data) and sparse. Here we consider the traditional n-slack formulation of multi-class SVM and n-slack MedLDA^c using the SVM^{struct} package, where a support vector corresponds to a document-label pair. For MedLDA^c and LDA+SVM, we set $K = 70$. For MedLDA^c, we report both the number of support vectors at the final iteration and the average number of support vectors over all iterations. We can see that both MedLDA^c and LDA+SVM generally need much fewer support vectors than the standard SVM on raw text. The major reason is that both MedLDA^c and LDA+SVM uses a much lower dimensional and more compact representation for each document. Moreover, MedLDA^c needs (about 4 times) fewer support vectors than LDA+SVM. This could be because MedLDA^c make use of both text contents and the supervising class labels in the training data and its estimated topics tend to be more discriminative when being used to infer the latent topical

17. The traditional 0/1 cost does not yield the best results. In most cases, the selected ℓ's are around 16.



(a) (b)

Figure 6: (a) Sensitivity to the cost parameter ℓ for the MedLDA^c; and (b) the number of support vectors for n-slack multi-class SVM, LDA+SVM, and n-slack MedLDA^c. For MedLDA^c, we show both the number of support vectors at the final iteration and the average number during training.

representations of documents, that is, using these latent representations by MedLDA^c, the documents in different categories are more likely to be well-separated, and therefore the max-margin classifier is simpler (i.e., needs fewer support vectors). This observation is consistent with what we have observed on the per-class distributions over topics in Figure 3. Finally, we observed that about 32% of the support vectors in MedLDA^c are also the support vectors in multi-class SVM on the raw features.

5.2.2 Regression

We first evaluate MedLDA^r on the movie review data set used in (Blei and McAuliffe, 2007), which contains 5006 documents and comprises 1.6M words, with a 5000-term vocabulary chosen by tf-idf. The data set was compiled from the one provided in (Pang and Lee, 2005). As in (Blei and McAuliffe, 2007), we take logs of the response values to make them approximately normal. We compare MedLDA^r with unsupervised LDA, supervised sLDA, MedLDA^r_p – a MedLDA regression model which uses unsupervised LDA as the underlying topic model (Please see Appendix B for details), and the linear SVR that uses the empirical word frequency as input features. For LDA, we use its low dimensional representation of documents as input features to a linear SVR and denote this method by *LDA+SVR*. The evaluation criterion is predictive R^2 (pR^2), which is defined as one minus the mean squared error divided by the data variance (Blei and McAuliffe, 2007), specificlly,

$$\mathrm{pR}^2 = 1 - \frac{\sum_{d=1}^{D}(y_d - \hat{y}_d)^2}{\sum_{d=1}^{D}(y_d - \bar{y})^2},$$

where y_d and $\hat{y}_d$ are the true and estimated response values of document d, respectively; and $\bar{y}$ is the mean of true response values on the whole data set. When we report pR^2, by default it is computed on the testing data set. Note that the *naïve baseline* that predicts the mean response value for all documents (i.e., $\forall d,\ \hat{y}_d = \bar{y}$) will have 0 on pR^2. Any method that have a positive pR^2 performs better than the naïve baseline.

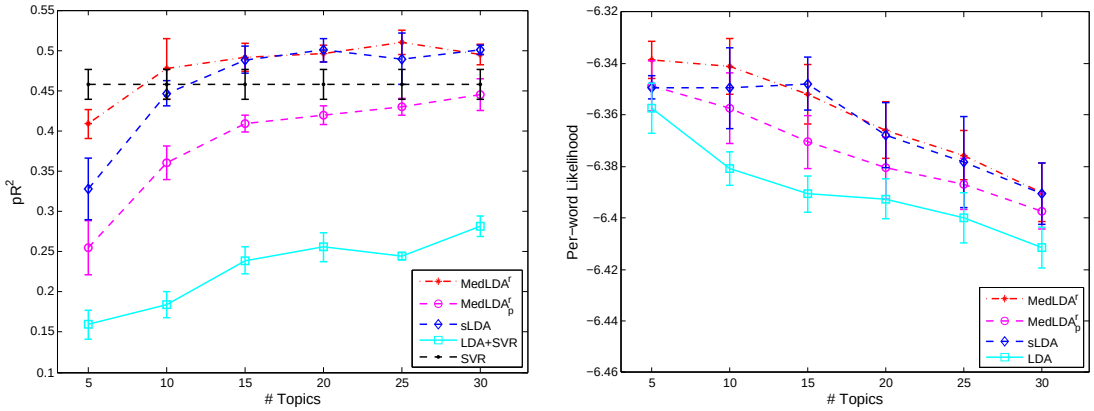


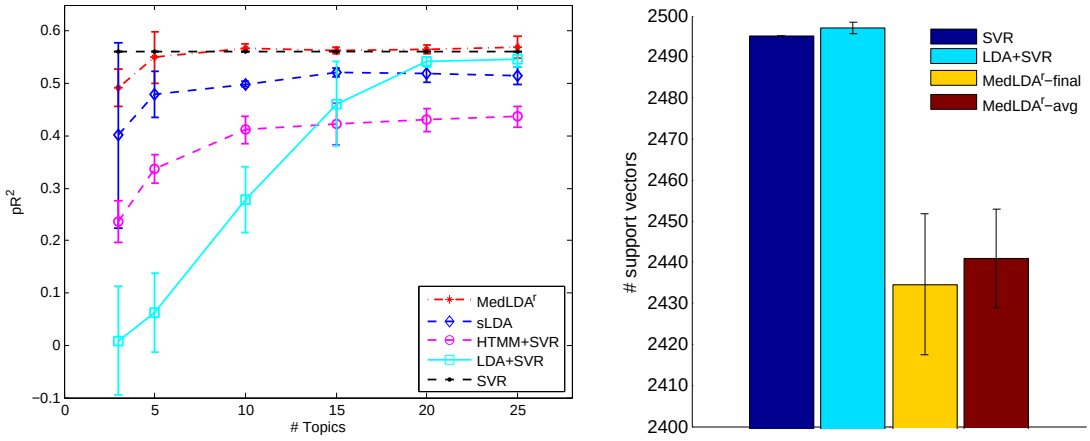
Figure 7: Predictive R^2 (left) and per-word likelihood (right) of different models on the movie review data set.

Figure 7 shows the average results as well as standard deviations over 5 randomly initialized runs, together with the per-word likelihood. For MedLDA and SVR, we fix the precision $\epsilon = 1e^{-3}$ and select C via cross-validation during training. We can see that the supervised MedLDA and sLDA can get better results than unsupervised LDA, which ignores supervised responses during discovering topical representations, and the linear SVR regression model. By using max-margin learning, MedLDA^r can get slightly better results than the likelihood-based sLDA, especially when the number of topics is small (e.g., ≤ 15). Indeed, when the number of topics is small, the latent representation of sLDA alone does not result in a highly separable problem, thus the integration of max-margin training helps in discovering a more discriminative latent representation using the same number of topics. In fact, the number of support vectors (i.e., documents that have at least one non-zero lagrange multiplier) decreases dramatically at $T = 15$ and stays nearly the same for $T > 15$, which with reference to Eq. (19) explains why the relative improvement over sLDA decreased as T increases. This behavior suggests that MedLDA^r can discover more predictive latent structures for *difficult*, non-separable regression problems.

For the two variants of MedLDA regression models, we can see an obvious improvement of MedLDA^r over MedLDA^r_p. This is because for MedLDA^r_p, the update rule of ϕ does not have the third and fourth terms of Eq. (19). Those terms make the max-margin estimation and latent topic discovery attached more tightly.

We also build another real data set of hotel review rating [18] by randomly crawling hotel reviews from TripAdvisor [19], where each review is associated with a global rating score and five aspect rating scores for the aspects [20]–*Value*, *Rooms*, *Location*, *Cleanliness*, and *Service*. This data set is very interesting and can be used for many data mining tasks, for example, extracting the textual mentions of each aspect. Also, the rich features in reviews can be exploited to discover interesting latent structures with a conditional topic model (Zhu and Xing, 2010). In these experiments, we focus on predicting the global rating

18. The data set is available at: http://www.cs.cmu.edu/∼junzhu/ReviewData.htm.
19. http://www.tripadvisor.com/
20. The website is subject to change. Our data set was built in December, 2009.



(a) (b)

Figure 8: (a) Predictive R^2 of different models on the hotel review data set; and (b) the number of support vectors for SVR, LDA+SVR, and MedLDA^r. For MedLDA^r, we show both the number of support vectors at the final iteration and the average number during training.

scores for reviews. To avoid too short and too long reviews, we only keep those reviews whose character length is between 1500 and 6000. On TripAdvisor, the global ratings rank from 1 to 5. We randomly select 1000 reviews for each rating and the data set consists of 5000 reviews in total. We uniformly partition it into training and testing sets. By removing a standard list of stopping words and those terms whose count frequency is less than 5, we build a dictionary with 12000 terms. Similarly, we take logarithm to make the response approximately normal. Figure 8(a) shows the predictive R^2 of different methods. Here, we also compare with the hidden topic Markov model (HTMM) (Gruber et al., 2007), which assumes the words in the same sentence have the same topic assignment. We use HTMM to discover latent representations of documents and use SVR to do regression. On this data set, we see a clear improvement of the supervised MedLDA^r compared to sLDA. The performance of unsupervised LDA (with a combination with SVR) is generally very unstable. The HTMM is more robust but its performance is worse than those of the supervised topic models. Finally, a linear SVR on empirical word frequency achieves a pR^2 of about 0.56, comparable to the best performance that can be achieved by MedLDA^r.

Figure 8(b) shows the number of support vectors for MedLDA^r, the standard SVR built on empirical word frequency, and the two-stage approach LDA+SVR. For MedLDA^r, we report both the number of support vectors at the last iteration and the average number of support vectors during training. Here, we set $K = 10$ for LDA and MedLDA^r. Again, we can see that MedLDA^r needs fewer support vectors than SVR and LDA+SVR. In contrast, LDA+SVR needs about the same number of support vectors as SVR. This observation suggests that the topical representations by the supervised MedLDA^r are more suitable for

learning a simple max-margin predictor, which is consistent with what we have observed in the classification case.

5.2.3 When and Why Should MedLDA be Preferred to SVM? A Discussion and Simulation Study

The above results show that the MedLDA classification model works comparably or slightly better than the SVM classifiers built on raw input features; and for the two regression problems, MedLDA outperforms the support vector regression model (i.e., SVR) on one data set while they are comparable on the other data set. These results raise the question "when should we choose MedLDA?" Our answers are as follows.

First of all, MedLDA is a topic model. Besides making prediction on unseen data, one major function of MedLDA is that it can discover semantic patterns underlying complex data, and facilitate dimensionality reduction (and compression) of data. In contrast, SVM models are more like black box machines which take raw input features and find good decision boundaries or regression curves; but they are incapable of discovering or considering hidden structures of complex data, and performing dimensionality reduction [21]. Our main goal of including SVM/SVR into our comparison of predictive accuracy is indeed to demonstrate that dimensionality reduction and information extraction from raw data via MedLDA does not cause serious loss (if at all) predictive information, which is not the case for many alternative probabilistic or non-probabilistic information extractors (e.g., LDA or LSI). As an integration of SVM with LDA, MedLDA performs both predictive and exploratory tasks simultaneously. So, the first selection rule is: *if we want to disclose some underlying patterns and extract a lower dimensional semantic-preserving representation of raw data besides doing prediction, MedLDA should be preferred to SVM.*

Second, even if our goal is focusing on prediction performance, MedLDA should also be considered as one competitive alternative. As shown in the above experiments, our simulation experiments below, as well as the follow-up works (Yang et al., 2010; Wang and Mori, 2011; Li et al., 2011), depending on the data and problems, max-margin supervised topic models can outperform SVM models, or they are comparable if no gains on predictive performance are obtained. There are several possible reasons for the comparable (not dramatically superior) classification performance we obtained on the 20 Newsgroups data:

(1) The fully factorized mean field inference method could potentially lead to inaccurate estimates. We have tried more sophisticated inference methods such as collapsed variational inference and collapsed Gibbs sampling [22], both of which could lead to superior prediction performance (e.g., about 4 percent improvement over SVM on multi-class classification accuracy);

(2) The much lower dimensional topical representations could be too compact, compared to the original high-dimensional inputs. A clever combination (e.g., concatenation with appropriate re-scaling of different features) of the discovered latent topical represen-

21. Some strategies like sparse feature selection can be incorporated to make an SVM more interpretable in the original feature space. But this is beyond the scope of this paper.
22. Sampling methods for MedLDA can be developed by using Lagrangian dual methods. But a full discussion on this topic is beyond the scope.

tations and the original input features could potentially improve the performance, as demonstrated in (Wang and Mori, 2011) for image classification.

To further substantiate the claimed advantages of MedLDA over SVM for admixed (i.e., multi-topical) data such as text and image, we conduct some simulation experiments to empirically study when MedLDA can perform well. We generate the observed word counts from an LDA model with K topics. The Dirichlet parameters are $\boldsymbol{\alpha} = (1, \ldots, 1)$. For the topics, we randomly draw $\beta_{kn} \propto \text{Beta}(1, 1)$, where $\propto$ means that we need to normalize $\boldsymbol{\beta}_k$ to be a distribution over the terms in a given vocabulary. We consider three different settings of binary classification with a vocabulary of 500 terms. The document lengths for each setting are randomly draw from a Poisson distribution, whose mean parameter is L, that is,

$$\forall d,\ N_d \sim \text{Poisson}(L).$$

(1) **Setting 1:** We set $K = 40$. We randomly draw the class label for document d from a distribution model

$$p(y_d = 1|\boldsymbol{\theta}_d) = \frac{1}{1 + \exp\{-\boldsymbol{\eta}^\top \boldsymbol{\theta}_d\}},\ \text{where } \boldsymbol{\eta}_k \sim \mathcal{N}(0, 0.1).$$

In other words, the class labels are solely influenced by the latent topic representations. Therefore, the true model that generates the labeled data follows the assumptions of sLDA and MedLDA. We set $L = 25, 50, 150, 300, 500$.

(2) **Setting 2:** We set $K = 150$. We randomly draw the class label for document d from a distribution model

$$p(y_d = 1|\boldsymbol{\theta}_d) = \frac{1}{1 + \exp\{-(\boldsymbol{\eta}_1^\top \boldsymbol{\theta}_d + \boldsymbol{\eta}_2^\top \mathbf{w}_d)\}},\ \text{where } \boldsymbol{\eta}_{ij} \sim \mathcal{N}(0, 0.1),\ i = 1, 2.$$

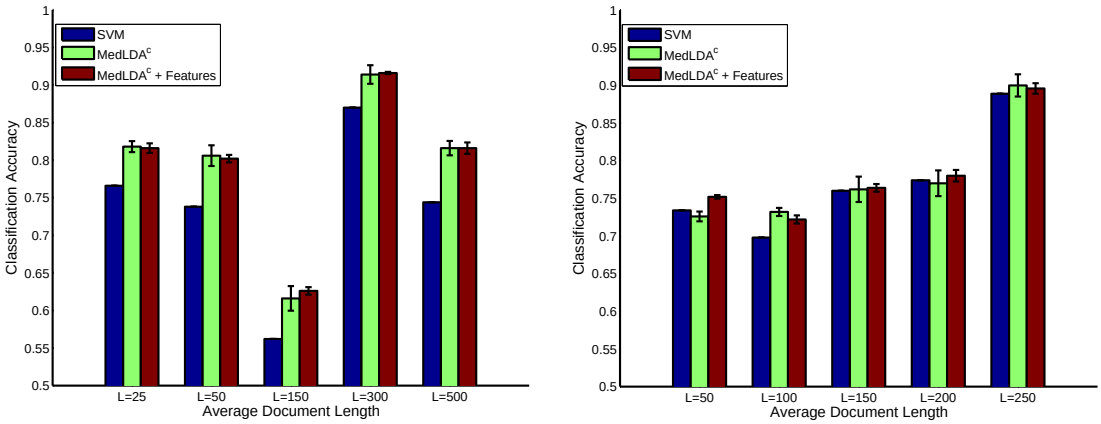
In other words, the true model that generates the labeled data does not follow the assumptions of sLDA. The class labels are influenced by the observed word counts. In fact, due to the law of conservation of belief (i.e., the total probability mass of a distribution must sum to one), the influence of $\boldsymbol{\theta}$ would be generally weaker than that of $\mathbf{w}$ in determining the true class labels. We set $L = 50, 100, 150, 200, 250$.

(3) **Setting 3:** Similar as in setting 2, but we improve the influence of $\boldsymbol{\theta}$ on class labels by using larger weights $\boldsymbol{\eta}_1$. Specifically, we sample the weights

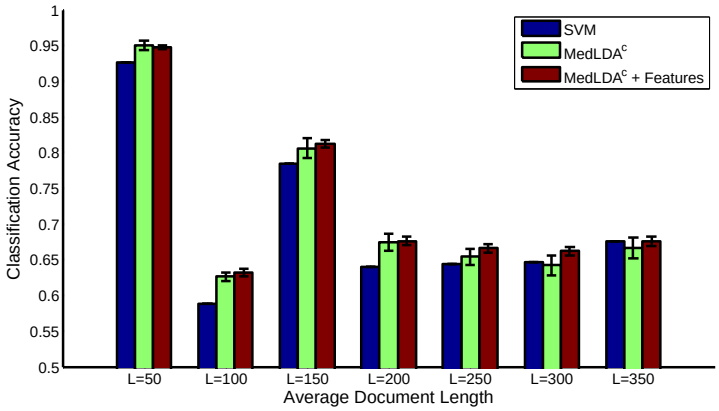
$$\boldsymbol{\eta}_{1j} \sim K \times \mathcal{N}(0, 0.1) \text{ and } \boldsymbol{\eta}_{2j} \sim \mathcal{N}(0, 0.1).$$

We set $L = 50, 100, 150, 200, 250, 300, 350$.

In summary, the first two settings generally represent two extremes where the true model matches the assumptions of MedLDA or SVM, while Setting 3 is somewhat in the middle place between Setting 1 and Setting 2. Since the synthetic words do not have real meanings, below we focus on presenting the prediction performance, rather than visualizing the discovered topic representations.



(a) (b)



(c)

Figure 9: Classification accuracy of different methods in (a) Setting 1; (b) Setting 2; and (c) Setting 3.

Figure 9 shows the classification accuracy of MedLDA^c, the SVM classifiers built on word counts, and the MedLDA^c models using both $\boldsymbol{\theta}$ and word counts to learn classifiers [23] at each iteration step of solving for $q(\boldsymbol{\eta})$. We can see that for Setting 1, where the true model that generates the data matches the assumptions of MedLDA (and sLDA models too) well, we can achieve significant improvements compared to the SVM classifiers built on raw input word counts for all settings with various average document lengths. In contrast, for Setting 2, where the true model largely violates the assumptions of MedLDA (in fact, it matches the assumptions of SVM well), we generally do not have much improvements. But still, we can have comparable performance. For the middle ground in Setting 3, we have mixed results. When the average document length is small (e.g., $\leq$ 250), which

23. We simply concatenate the two types of features without considering the scale difference.

means the influence of word counts on class labels is weak, MedLDA^c can improve a lot over SVM. But when the influence of word counts gets bigger (e.g., $L \geq 300$), using the low dimensional topic representations tends to be insufficient to get good performance. Translating to empirical text analysis, MedLDA will be particularly helpful when analyzing short texts, such as abstracts, reviews, users comments, and user status updates, which are nowadays the dominant forms of user texts on social media.

In all the three settings, we can see that a naïve combination of both latent topic representations and input word counts could improve the performance in some cases, or at least it will produce comparable performance with the better model between MedLDA^c and SVM. Finally, comparing the three settings, we can see that for Setting 2, since the true class labels heavily depend on the input word counts, increasing the average document length L generally improves the classification performance of all models. In other words, the classification problems become easier because of more discriminant information is provided as L increases. In contrast, we do not have the similar observations in the other two settings because the true labels are heavily (or solely in Setting 1) determined by $\boldsymbol{\theta}$, whose dimensionality is fixed.

The last reason that we think MedLDA should be considered as an important novel development with one root being from SVM because it presents one of the first successful attempts, in the particular context of Bayesian topic models, towards pushing forward the interface between max-margin learning and Bayesian generative modeling. As further demonstrated in others' work (Yang et al., 2010; Wang and Mori, 2011; Li et al., 2011) as well as our recent work on regularized Bayesian inference (Zhu et al., 2011a,b), the max-margin principle can be a fruitful addition to "regularize" the desired posterior distributions of Bayesian models for performing better prediction in a broad range of scenarios, such as image annotation, classification, multi-task learning, etc.

5.3 Time Efficiency

In this section, we report empirical results on time efficiency in training and testing. All the following results are achieved on a standard desktop with a 2.66GHz Intel processor. We implement all the models in C++ language, without any special optimization of the code.

5.3.1 Training Time

Figure 10 shows the average training time of different models together with standard deviations on both binary and multi-class classification tasks with 5 randomly initialized runs. Here, we do not compare with DiscLDA because learning the transition matrix is not fully implemented in (Lacoste-Julien, 2009), but we will compare the testing time with it. From the results, we can see that for binary classification, MedLDA^c is more efficient than multi-class sLDA and is comparable with LDA+SVM. The slowness of multi-class sLDA is because the normalization factor in the distribution model of y strongly couples the topic assignments of different words in the same document. Therefore, the posterior inference is slower than that of unsupervised LDA and MedLDA^c which uses unsupervised LDA as the underlying topic model. For the sLDA regression model, it takes even more training time because of the mismatch between its normal assumption and the non-Gaussian binary

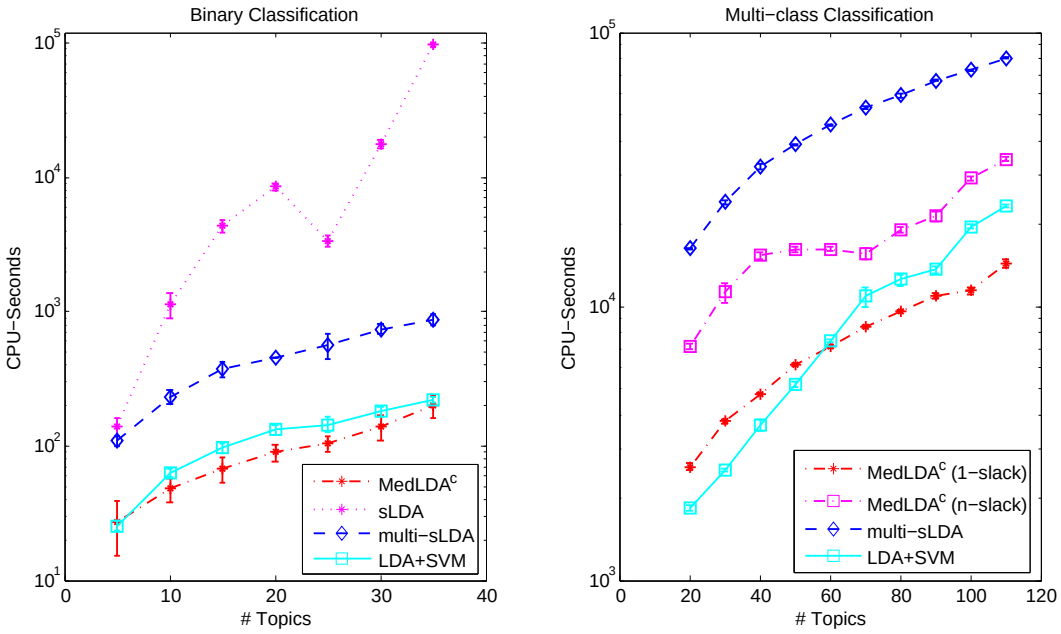
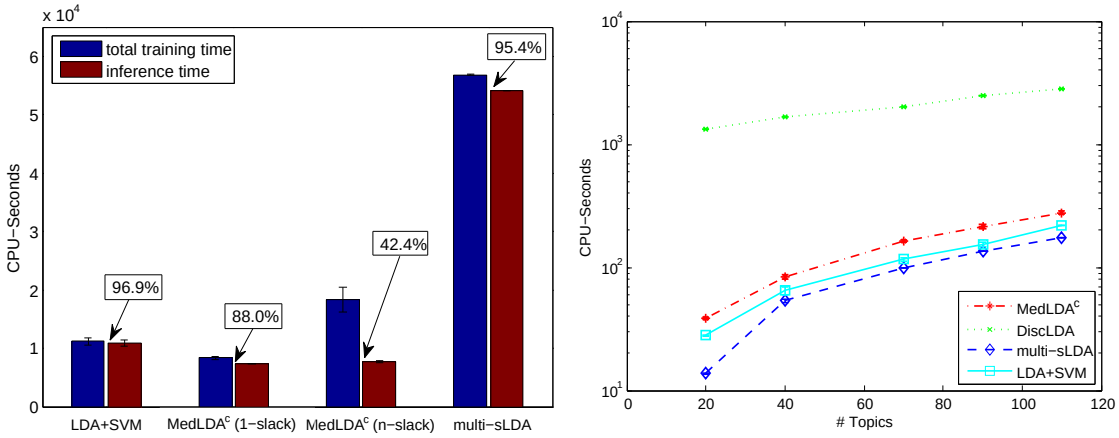


Figure 10: Training time (CPU seconds in log-scale) of different models with respect to the number of topics for both (Left) binary and (Right) multi-class classification.

response variables, which prolongs the E-step. In contrast, MedLDA^c does not have such a normal assumption.

For multi-class classification, the training time of MedLDA^c is mainly dependent on solving a multi-class SVM problem. Here, we implemented both 1-slack and n-slack versions of multi-class SVM (Joachims et al., 2009) for solving the sub-problem of estimating $q(\boldsymbol{\eta})$ and Lagrangian multipliers in MedLDA^c. As we can see from Figure 10, the MedLDA^c with 1-slack SVM as the sub-solver can be very efficient, comparable to unsupervised LDA+SVM. The MedLDA^c with n-slack SVM solvers is about 3 times slower. Similar to the binary case, for the multi-class supervised sLDA (Wang et al., 2009), because of the normalization factor in the category probability model (i.e., a softmax function), the posterior inference on different topic assignment variables (in the same document) are strongly correlated. Therefore, the inference is (about 10 times) slower than that on unsupervised LDA and MedLDA^c which takes an unsupervised LDA as the underlying topic model. For regression, the training time of MedLDA^r is comparable to that of sLDA, while MedLDA^r_p is more efficient.

We also show the time spent on inference (i.e., E-step) and the ratio it takes over the total training time for different models in Figure 11(a). We can clearly see that the difference between 1-slack MedLDA^c and n-slack MedLDA^c is on the learning of SVMs (i.e., M-step). Both methods have similar inference time. We can also see that for LDA+SVM and multi-sLDA, more than 95% of the training time is spent on inference, which is very expensive for multi-sLDA. Note that LDA+SVM takes a longer inference time than MedLDA^c. This is because we use more data (both training and testing) to learn unsupervised LDA. The SVM classifiers built on raw input word count features are generally much more faster than



(a) (b)

Figure 11: (a) The inference time (CPU seconds in linear scale) and total training time for learning different models, as well as the ratio of inference time over total training time. For MedLDA^c, we consider both the 1-slack and n-slack formulations; for LDA+SVM, the SVM classifier is by default the 1-slack formulation; and (b) Testing time (CPU seconds in log-scale) of different models with respect to the number of topics for multi-class classification.

all the topic models. For instance, it takes about 230 seconds to train a 1-slack multi-class SVM on the 20 Newsgroups training data, or about 1000 seconds to train a n-slack multi-class SVM on the same training set; both are faster than the fastest topic model 1-slack MedLDA^c. This is reasonable because SVM classifiers do not spend time on inferring the latent topic representations.

5.3.2 Testing Time

Figure 11(b) shows the average testing time with standard deviation on 20 Newsgroup testing data with 5 randomly initialized runs. We can see that MedLDA^c, multi-class sLDA and unsupervised LDA are comparable in testing time, faster than that of DiscLDA. This is because all the three models of MedLDA^c, multi-class sLDA and LDA are *downstream* models (See the Introduction for definition). In testing, they do exactly the same tasks, that is, to infer the overall latent topical representation and do prediction with a linear model. Therefore, they have comparable testing time. However, DiscLDA is an *upstream* model, for which the prediction task is done with multiple times of doing inference to find the category-dependent latent topical representations. Therefore, in principle, the testing time of an upstream topic model is about $|\mathcal{C}|$ times slower than that of its downstream counterpart model, where $\mathcal{C}$ is the finite set of categories. The results in Figure 11(b) show that DiscLDA is roughly about 20 times slower than other downstream models. Of course, the different inference algorithms can also make the testing time different.

6. Conclusions and Discussions

We have presented maximum entropy discrimination LDA (MedLDA), a supervised topic model that uses the discriminative max-margin principle to estimate model parameters such as topic distributions underlying a corpus, and infer latent topical vectors of documents. MedLDA integrates the max-margin principle into the process of topic learning and inference via optimizing one single objective function with a set of *expected* margin constraints. The objective function is a tradeoff between the goodness of fit of an underlying topic model and the prediction accuracy of the resultant topic vectors on a max-margin classifier. We provide empirical evidence as well as theoretical insights, which appear to demonstrate that this integration could yield predictive topical representations that are suitable for prediction tasks, such as regression and classification. We also present a general formulation of learning maximum entropy discrimination topic models, which allows any form of likelihood based topic models to be discriminatively trained. Although the general max-margin framework can be approximately solved with different methods, we concentrate on developing efficient variational methods for MedLDA in this paper. Our empirical results on movie review, hotel review and 20 Newsgroups data sets demonstrate that MedLDA is an attractive supervised topic model, which can achieve state of the art performance for topic discovery and prediction accuracy while needs fewer support vectors than competing max-margin methods that are built on raw text or the topical representations discovered by unsupervised LDA.

MedLDA represents the first step towards integrating the max-margin principle into supervised topic models, and under the general MedTM framework presented in Section 4, several improvements and extensions are in the horizon. Specifically, due to the nature of MedTM's joint optimization formulation, advances in either max-margin training or better variational bounds for inference can be easily incorporated. For instance, the mean field variational upper bound in MedLDA can be improved by using the tighter collapsed variational bound (Teh et al., 2006) that achieves results comparable to collapsed Gibbs sampling (Griffiths and Steyvers, 2004). Moreover, as the experimental results suggest, incorporation of a more expressive underlying topic model enhances the overall performance. Therefore, we plan to integrate and utilize other underlying topic models like the fully generative sLDA model in the classification case. However, as we have stated, the challenge in developing fully supervised MedLDA classification model lies in the hard posterior inference caused by the normalization factor in the category distribution model. Finally, advance in max-margin training would also results in more efficient training.

Acknowledgements

We thank David Blei for answering questions about implementing sLDA, Chong Wang for sharing his implementation of multi-class sLDA, Simon Lacoste-Julien for discussions on DiscLDA and feedbacks on our implementation of MedLDA, and the anonymous reviewers for valuable comments. This work was done while J.Z. was visiting CMU under a support from NSF DBI-0546594 and DBI-0640543 awarded to E.X.; J.Z. is supported by National Key Foundation R&D Projects 2012CB316301, Basic Research Foundation of Tsinghua National Laboratory for Information Science and Technology (TNList), a Starting Research

Fund from Tsinghua University, No. 553420003, and the 221 Basic Research Plan for Young Faculties at Tsinghua University.

Appendix A. Proof of Corollary 4

In this section, we prove the corollary 4.

Proof Since the variational parameters $(\boldsymbol{\gamma}, \boldsymbol{\phi})$ are fixed when solving for $q(\boldsymbol{\eta})$, we can ignore the terms in $\mathcal{L}^{bs}$ that do not depend on $q(\boldsymbol{\eta})$ and get the function

$$\begin{aligned}\mathcal{L}^{bs}_{[q(\boldsymbol{\eta})]} &\triangleq KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) - \sum_d \mathbb{E}_q[\log p(y_d|\bar{Z}_d, \boldsymbol{\eta}, \delta^2)] \\ &= KL(q(\boldsymbol{\eta})\|p_0(\boldsymbol{\eta})) + \frac{1}{2\delta^2}\Big(\mathbb{E}_{q(\boldsymbol{\eta})}[\boldsymbol{\eta}^\top \mathbb{E}[AA^\top]\eta - 2\boldsymbol{\eta}^\top \sum_{d=1}^{D} y_d \mathbb{E}[\bar{Z}_d]]\Big) + c,\end{aligned}$$

where c is a constant that does not depend on $q(\boldsymbol{\eta})$.

Let $U(\boldsymbol{\xi}, \boldsymbol{\xi}^*) = C\sum_{d=1}^{D}(\xi_d + \xi_d^*)$. Suppose $(q_0(\boldsymbol{\eta}), \boldsymbol{\xi}_0, \boldsymbol{\xi}_0^*)$ is the optimal solution of P1, then we have: for any feasible $(q(\boldsymbol{\eta}), \boldsymbol{\xi}, \boldsymbol{\xi}^*)$,

$$\mathcal{L}^{bs}_{[q_0(\boldsymbol{\eta})]} + U(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0^*) \leq \mathcal{L}^{bs}_{[q(\boldsymbol{\eta})]} + U(\boldsymbol{\xi}, \boldsymbol{\xi}^*).$$

From Corollary 3, we conclude that the optimum predictive parameter distribution is $q_0(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\lambda}_0, \Sigma)$, where $\Sigma = (I + 1/\delta^2 \mathbb{E}[A^\top A])^{-1}$ does not depend on $q(\boldsymbol{\eta})$. Since $q_0(\boldsymbol{\eta})$ is also normal, for any distribution[24] $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\lambda}, \Sigma)$, with several steps of algebra it is easy to show that

$$\mathcal{L}^{bs}_{[q(\boldsymbol{\eta})]} = \frac{1}{2}\boldsymbol{\lambda}^\top(I + \frac{1}{\delta^2}\mathbb{E}[A^\top A])\boldsymbol{\lambda} - \boldsymbol{\lambda}^\top(\sum_{d=1}^{D}\frac{y_d}{\delta^2}\mathbb{E}[\bar{Z}_d]) + c' = \frac{1}{2}\boldsymbol{\lambda}^\top\Sigma^{-1}\boldsymbol{\lambda} - \boldsymbol{\lambda}^\top(\sum_{d=1}^{D}\frac{y_d}{\delta^2}\mathbb{E}[\bar{Z}_d]) + c',$$

where c' is another constant that does not depend on $\boldsymbol{\lambda}$.

Thus, we can get: for any $(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\xi}^*)$, where

$$(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\xi}^*) \in \{(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\xi}^*) : \ y_d - \boldsymbol{\lambda}^\top\mathbb{E}[\bar{Z}_d] \leq \epsilon + \xi_d; \ -y_d + \boldsymbol{\lambda}^\top\mathbb{E}[\bar{Z}_d] \leq \epsilon + \xi_d^*; \text{ and } \boldsymbol{\xi}, \boldsymbol{\xi}^* \geq 0 \ \forall d\},$$

we have

$$\frac{1}{2}\boldsymbol{\lambda}_0^\top\Sigma^{-1}\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}_0^\top(\sum_{d=1}^{D}\frac{y_d}{\delta^2}\mathbb{E}[\bar{Z}_d]) + U(\boldsymbol{\xi}_0, \boldsymbol{\xi}_0^*) \leq \frac{1}{2}\boldsymbol{\lambda}^\top\Sigma^{-1}\boldsymbol{\lambda} - \boldsymbol{\lambda}^\top(\sum_{d=1}^{D}\frac{y_d}{\delta^2}\mathbb{E}[\bar{Z}_d]) + U(\boldsymbol{\xi}, \boldsymbol{\xi}^*),$$

which means the mean of the optimum posterior distribution under a Gaussian MedLDA is achieved by solving a primal problem as stated in the Corollary.

■

24. Although the feasible set of $q(\boldsymbol{\eta})$ in P1 is much richer than the set of normal distributions with the covariance matrix Σ, Corollary 3 shows that the solution is a restricted normal distribution. Thus, it suffices to consider only these normal distributions in order to learn the mean of the optimum distribution.

Appendix B. Max-Margin Learning of the Vanilla LDA for Regression

In Section 3.1, we have presented the MedLDA regression model that uses supervised sLDA (Blei and McAuliffe, 2007) to discover latent topic assignments Z and document-level topical representations $\boldsymbol{\theta}$. The same principle can be applied to perform joint maximum likelihood estimation and max-margin training for unsupervised LDA (Blei et al., 2003), which does not directly model side information such as user ratings y. In this section, we present this MedLDA model, which will be referred to as *MedLDA*^r_p. As in MedLDA^c, we assume that the supervised side information y is given, even though not included in the joint likelihood function defined in LDA[25].

A naïve approach to using unsupervised LDA for supervised prediction tasks (e.g., regression) is a two-stage procedure: 1) using unsupervised LDA to discover the latent topical representations of documents; and 2) feeding the low-dimensional topical representations into a regression model (e.g., SVR) for training and testing. This de-coupled approach can be rather sub-optimal because the side information of documents (e.g., rating scores of movie reviews) is not used in discovering the low-dimensional representations and thus can result in a sub-optimal representation for prediction tasks. Below, we present MedLDA^r_p, which integrates an unsupervised LDA for discovering topics with the SVR for regression. The inter-play between topic discovery and supervised prediction will result in more discriminative latent topical representations, similar as in MedLDA^r.

When the underlying topic model is unsupervised LDA, the likelihood is $p(\mathbf{W}|\boldsymbol{\alpha}, \boldsymbol{\beta})$, the same as in MedLDA^c. For regression, we apply the ϵ-insensitive support vector regression (SVR) (Smola and Schölkopf, 2003) approach as before. Again, we learn a distribution $q(\boldsymbol{\eta})$. The prediction rule is the same as in Eq. (8). The integrated learning problem is

$$\text{P6(MedLDA}^r_p): \quad \min_{q, q(\boldsymbol{\eta}), \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \quad \mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta}) + KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta})) + C\sum_{d=1}^{D}(\xi_d + \xi^*_d) \quad (43)$$

$$\forall d, \text{ s.t. }: \begin{cases} y_d - \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \leq \epsilon + \xi_d \\ -y_d + \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d] \leq \epsilon + \xi^*_d \\ \xi_d, \xi^*_d \geq 0 \end{cases},$$

where the KL-divergence is a regularizer that biases the estimate of $q(\boldsymbol{\eta})$ towards the prior. In MedLDA^r, this KL-regularizer is implicitly contained in the variational bound $\mathcal{L}^{bs}$ as shown in Eq. (9). The constrained problem is equivalent to the "unconstrained" problem by removing slack variables:

$$\min_{q, q(\boldsymbol{\eta}), \boldsymbol{\alpha}, \boldsymbol{\beta}} \mathcal{L}^u(q; \boldsymbol{\alpha}, \boldsymbol{\beta}) + KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta})) + C\sum_{d=1}^{D}\max(0, |y_d - \mathbb{E}[\boldsymbol{\eta}^\top \bar{Z}_d]| - \epsilon) \quad (44)$$

Variational Algorithm: For MedLDA^r_p, the unconstrained optimization problem (44) can be similarly solved with a coordinate-descent algorithm as in the case of MedLDA^r.

25. One could argue that this design is unreasonable because with y one should only consider sLDA. But we study fitting the vanilla LDA using y in an indirect way described below because of the popularity and historical importance of this scheme in many applied domains

Specifically, we assume that $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\}) = \prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d) \prod_{n=1}^{N} q(z_{dn}|\boldsymbol{\phi}_{dn})$, where the variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ have the same meanings as in MedLDA^r. Then, we alternately solve for each variable and get a variational algorithm which is similar to that of MedLDA^r.

Solve for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $q(\boldsymbol{\eta})$: the update rules of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the same as in the MedLDA^r. The parameter δ^2 is not used here. By using Lagrangian methods, we get that

$$q(\boldsymbol{\eta}) = \frac{p_0(\boldsymbol{\eta})}{Z} \exp\Big(\boldsymbol{\eta}^\top \sum_{d=1}^{D} (\hat{\mu}_d - \hat{\mu}_d^*)\mathbb{E}[\bar{Z}_d]\Big) \tag{45}$$

and the dual problem is the same as D2. Again, we can choose different priors to introduce some regularization effects. For the standard normal prior: $p_0(\boldsymbol{\eta}) = \mathcal{N}(0, I)$, the posterior is also a normal: $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\lambda}, I)$, where $\boldsymbol{\lambda} = \sum_{d=1}^{D}(\hat{\mu}_d - \hat{\mu}_d^*)\mathbb{E}[\bar{Z}_d]$ is the mean. This identity covariance matrix is much simpler than the covariance matrix Σ as in *MedLDA^r*, which depends on the latent topical representation Z. Since I is independent of Z, the prediction model in *MedLDA$_p^r$* is less affected by the latent topical representations. Together with the simpler update rule (48), we can conclude that the coupling between the max-margin estimation and the discovery of latent topical representations in *MedLDA$_p^r$* is looser than that of *MedLDA^r*. The looser coupling will lead to inferior empirical performance as we show in Section 5.2.

For the standard normal prior, the dual problem is a QP problem:

$$\begin{aligned} \max_{\boldsymbol{\mu}, \boldsymbol{\mu}^*} \quad & -\frac{1}{2}\|\boldsymbol{\lambda}\|_2^2 - \epsilon \sum_{d=1}^{D}(\mu_d + \mu_d^*) + \sum_{d=1}^{D} y_d(\mu_d - \mu_d^*) \\ \forall d, \text{ s.t. :} \quad & \mu_d, \mu_d^* \in [0, C], \end{aligned} \tag{46}$$

Similarly, we can derive its primal form, which is as a standard SVR problem:

$$\begin{aligned} \min_{\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \quad & \frac{1}{2}\|\boldsymbol{\lambda}\|_2^2 + C\sum_{d=1}^{D}(\xi_d + \xi_d^*) \\ \text{s.t. } \forall d: \quad & \begin{cases} y_d - \boldsymbol{\lambda}^\top \mathbb{E}[\bar{Z}_d] \le \epsilon + \xi_d \\ -y_d + \boldsymbol{\lambda}^\top \mathbb{E}[\bar{Z}_d] \le \epsilon + \xi_d^* \\ \xi_d, \xi_d^* \ge 0. \end{cases} \end{aligned} \tag{47}$$

Now, we can leverage recent developments in support vector regression (e.g., the public SVM-light package) to solve either the dual problem or the primal problem.

Solve for $q(\{\boldsymbol{\theta}_d, \mathbf{z}_d\})$: We have the same update rule for $\boldsymbol{\gamma}$ as in MedLDA^r. By using the similar one-step approximation strategy, we have:

$$\boldsymbol{\phi}_{dn} \propto \exp\Big(\mathbb{E}[\log \boldsymbol{\theta}_d|\boldsymbol{\gamma}_d] + \log p(w_{dn}|\boldsymbol{\beta}) + \frac{\mathbb{E}[\boldsymbol{\eta}]}{N}(\hat{\mu}_d - \hat{\mu}_d^*)\Big), \tag{48}$$

Again, we can see that how the max-margin constraints in P6 regularize the procedure of discovering latent topical representations through the last term in Eq. (48). Specifically, for a document d, which lies around the decision boundary, i.e., a support vector, either $\hat{\mu}_d$ or $\hat{\mu}_d^*$ is non-zero, and the last term biases $\boldsymbol{\phi}_{dn}$ towards a distribution that favors a more accurate prediction on the document. However, compared to Eq. (19), we can see that Eq.

(48) is simpler and does not have the complex third and fourth terms of Eq. (19). This simplicity suggests that the latent topical representation is less affected by the max-margin estimation (i.e., the prediction model's parameters).

References

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

David Blei and John Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

David Blei and Jon D. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

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

Gal Chechik and Naftali Tishby. Extracting relevant structures with side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2002.

Ning Chen, Jun Zhu, and Eric P. Xing. Predictive subspace learning for multi-view data: a large margin approach. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

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

Paul Damien and Stephen G. Walker. Sampling truncated Normal, Beta, and Gamma densities. *Journal of Computational and Graphical Statistics*, 10(2):206–215, 2001.

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

Pedro Felzenszwalb, Ross Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627 – 1645, 2010.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, (101):5228–5235, 2004.

William E. Griffiths. A Gibbs sampler for the parameters of a truncated multivariate normal distribution. *No 856, Department of Economics, University of Melbourne*, 2002.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. Hidden topic Markov models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

Xuming He and Richard S. Zemel. Learning hybrid models for image annotation with partially labeled data. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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.

Thorsten Joachims. Making large-scale SVM learning practical. *Advances in kernel methods–support vector learning, MIT-Press*, 1999.

Thorsten Joachims, Thomas Finley, and Chun-Nam Yu. Cutting-plane training of structural SVMs. *Machine Learning Journal*, 77(1), 2009.

Michael I. Jordan, Zoubin Ghahramani, Tommis Jaakkola, and Lawrence K. Saul. *An introduction to variational methods for graphical models.* M. I. Jordan (Ed.), Learning in Graphical Models, Cambridge: MIT Press, Cambridge, MA, 1999.

Simon Lacoste-Julien. *Discriminative Machine Learning with Structure.* PhD thesis, EECS Department, University of California, Berkeley, Jan 2009.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Dingcheng Li, Swapna Somasundaran, and Amit Chakraborty. A combination of topic models with max-margin learning for relation detection. In *ACL TextGraphs-6 Workshop*, 2011.

Li-Jia Li, Richard Socher, and Fei-Fei Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceddings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

Dmitry Pavlov, Alexandrin Popescul, David M. Pennock, and Lyle H. Ungar. Mixtures of conditional maximum entropy models. In *International Conference on Machine Learning (ICML)*, 2003.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceddings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009.

Gabriel Rodriguez-Yam, Richard Davis, and Louis Scharf. Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. *Technical Report, Department of Statistics, Columbia University*, 2004.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

Ruslan Salakhutdinov and Geoffrey Hinton. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Edward Schofield. *Fitting Maximum-Entropy Models on Large Sample Spaces*. PhD thesis, Department of Computing, Imperial College London, Jan 2007.

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

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceddings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

Erik Sudderth, Antonio Torralba, William Freeman, and Alan Willsky. Learning hierarchical models of scenes, objects, and parts. In *IEEE International Conference on Computer Vision (ICCV)*, 2005.

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

Yee Whye Teh, David Newman, and Max Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceddings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008.

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

Vladimir Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

Chong Wang, David Blei, and Li Fei-Fei. Simultaneous image classification and annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Yang Wang and G. Mori. Max-margin latent Dirichlet allocation for image classification and annotation. In *British Machine Vision Conference (BMVC)*, 2011.

Max Welling, Michal Rosen-Zvi, and Geoffrey Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Eric P. Xing, Rong Yan, and Alexander G. Hauptmann. Mining associated text and images with dual-wing Harmoniums. In *International Conference on Uncertainty in Artifical Intelligence (UAI)*, 2005.

Shuanghong Yang, Jiang Bian, and Hongyuan Zha. Hybrid generative/discriminative learning for automatic image annotation. In *International Conference on Uncertainty in Artifical Intelligence (UAI)*, 2010.

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

Bing Zhao and Eric P. Xing. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Jun Zhu, Amr Ahmed, and Eric P. Xing. MedLDA: Maximum margin supervised topic models for regression and classification. In *International Conference on Machine Learning (ICML)*, 2009.

Jun Zhu, Ning Chen, and Eric P. Xing. Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011a.

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

Jun Zhu, Li-Jia Li, Li Fei-Fei, and Eric P. Xing. Large margin training of upstream scene understanding models. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Jun Zhu and Eric P. Xing. Maximum entropy discrimination Markov networks. *Journal of Machine Learning Research*, (10):2531–2569, 2009.

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

Jun Zhu, Eric P. Xing, and Bo Zhang. Partially observed maximum entropy discrimination Markov networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.