# Supervised topic models

**David M. Blei**
Department of Computer Science
Princeton University
Princeton, NJ
blei@cs.princeton.edu

**Jon D. McAuliffe**
Department of Statistics
University of Pennsylvania,
Wharton School
Philadelphia, PA
mcjon@wharton.upenn.edu

## Abstract

We introduce supervised latent Dirichlet allocation (sLDA), a statistical model of labelled documents. The model accommodates a variety of response types. We derive a maximum-likelihood procedure for parameter estimation, which relies on variational approximations to handle intractable posterior expectations. Prediction problems motivate this research: we use the fitted model to predict response values for new documents. We test sLDA on two real-world problems: movie ratings predicted from reviews, and web page popularity predicted from text descriptions. We illustrate the benefits of sLDA versus modern regularized regression, as well as versus an unsupervised LDA analysis followed by a separate regression.

## 1 Introduction

There is a growing need to analyze large collections of electronic text. The complexity of document corpora has led to considerable interest in applying hierarchical statistical models based on what are called *topics*. Formally, a topic is a probability distribution over terms in a vocabulary. Informally, a topic represents an underlying semantic theme; a document consisting of a large number of words might be concisely modelled as deriving from a smaller number of topics. Such *topic models* provide useful descriptive statistics for a collection, which facilitates tasks like browsing, searching, and assessing document similarity.

Most topic models, such as latent Dirichlet allocation (LDA) [4], are unsupervised: only the words in the documents are modelled. The goal is to infer topics that maximize the likelihood (or the posterior probability) of the collection. In this work, we develop supervised topic models, where each document is paired with a response. The goal is to infer latent topics predictive of the response. Given an unlabeled document, we infer its topic structure using a fitted model, then form its prediction. Note that the response is not limited to text categories. Other kinds of document-response corpora include essays with their grades, movie reviews with their numerical ratings, and web pages with counts of how many online community members liked them.

Unsupervised LDA has previously been used to construct features for classification. The hope was that LDA topics would turn out to be useful for categorization, since they act to reduce data dimension [4]. However, when the goal is prediction, fitting unsupervised topics may not be a good choice. Consider predicting a movie rating from the words in its review. Intuitively, good predictive topics will differentiate words like "excellent", "terrible", and "average," without regard to genre. But topics estimated from an unsupervised model may correspond to genres, if that is the dominant structure in the corpus.

The distinction between unsupervised and supervised topic models is mirrored in existing dimension-reduction techniques. For example, consider regression on unsupervised principal components versus partial least squares and projection pursuit [7], which both search for covariate linear combinations most predictive of a response variable. These linear supervised methods have non-

parametric analogs, such as an approach based on kernel ICA [6]. In text analysis, McCallum et al. developed a joint topic model for words and categories [8], and Blei and Jordan developed an LDA model to predict caption words from images [2]. In chemogenomic profiling, Flaherty et al. [5] proposed "labelled LDA," which is also a joint topic model, but for genes and protein function categories. It differs fundamentally from the model proposed here.

This paper is organized as follows. We first develop the supervised latent Dirichlet allocation model (sLDA) for document-response pairs. We derive parameter estimation and prediction algorithms for the real-valued response case. Then we extend these techniques to handle diverse response types, using generalized linear models. We demonstrate our approach on two real-world problems. First, we use sLDA to predict movie ratings based on the text of the reviews. Second, we use sLDA to predict the number of "diggs" that a web page will receive in the `www.digg.com` community, a forum for sharing web content of mutual interest. The digg count prediction for a page is based on the page's description in the forum. In both settings, we find that sLDA provides much more predictive power than regression on unsupervised LDA features. The sLDA approach also improves on the lasso, a modern regularized regression technique.

## 2 Supervised latent Dirichlet allocation

In topic models, we treat the words of a document as arising from a set of latent topics, that is, a set of unknown distributions over the vocabulary. Documents in a corpus share the same set of $K$ topics, but each document uses a mix of topics unique to itself. Thus, topic models are a relaxation of classical document mixture models, which associate each document with a single unknown topic.

Here we build on latent Dirichlet allocation (LDA) [4], a topic model that serves as the basis for many others. In LDA, we treat the topic proportions for a document as a draw from a Dirichlet distribution. We obtain the words in the document by repeatedly choosing a topic assignment from those proportions, then drawing a word from the corresponding topic.

In *supervised latent Dirichlet allocation* (sLDA), we add to LDA a response variable associated with each document. As mentioned, this variable might be the number of stars given to a movie, a count of the users in an on-line community who marked an article interesting, or the category of a document. We jointly model the documents and the responses, in order to find latent topics that will best predict the response variables for future unlabeled documents.

We emphasize that sLDA accommodates various types of response: unconstrained real values, real values constrained to be positive (e.g., failure times), ordered or unordered class labels, nonnegative integers (e.g., count data), and other types. However, the machinery used to achieve this generality complicates the presentation. So we first give a complete derivation of sLDA for the special case of an unconstrained real-valued response. Then, in Section 2.3, we present the general version of sLDA, and explain how it handles diverse response types.

Focus now on the case $y \in \mathbb{R}$. Fix for a moment the model parameters: the $K$ topics $\beta_{1:K}$ (each $\beta_k$ a vector of term probabilities), the Dirichlet parameter $\alpha$, and the response parameters $\eta$ and $\sigma^2$. Under the sLDA model, each document and response arises from the following generative process:

1. Draw topic proportions $\theta \mid \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   (a) Draw topic assignment $z_n \mid \theta \sim \text{Mult}(\theta)$.
   (b) Draw word $w_n \mid z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \text{N}\left(\eta^\top \bar{z}, \sigma^2\right)$.

Here we define $\bar{z} := (1/N) \sum_{n=1}^{N} z_n$. The family of probability distributions corresponding to this generative process is depicted as a graphical model in Figure 1.

Notice the response comes from a normal linear model. The covariates in this model are the (unobserved) empirical frequencies of the topics in the document. The regression coefficients on those frequencies constitute $\eta$. Note that a linear model usually includes an intercept term, which amounts to adding a covariate that always equals one. Here, such a term is redundant, because the components of $\bar{z}$ always sum to one.

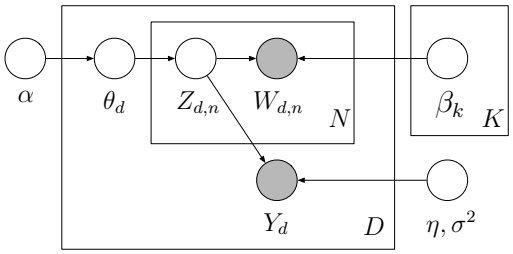
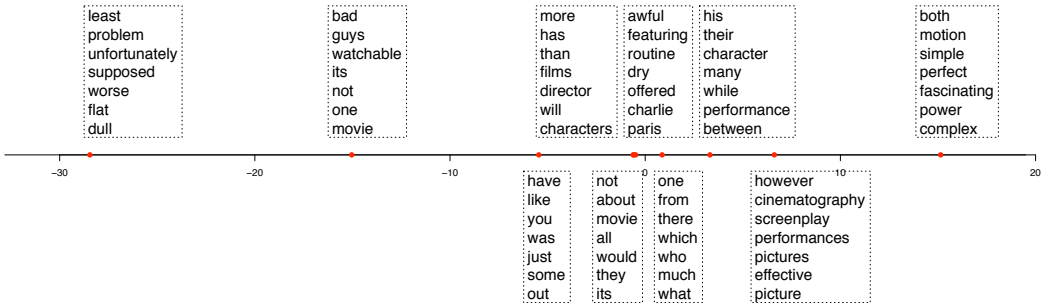



Figure 1: (Left) A graphical model representation of Supervised Latent Dirichlet allocation. (Bottom) The topics of a 10-topic sLDA model fit to the movie review data of Section 3.




By regressing the response on the empirical topic frequencies, we treat the response as non-exchangeable with the words. The document (i.e., words and their topic assignments) is generated first, under full word exchangeability; then, based on the document, the response variable is generated. In contrast, one could formulate a model in which $y$ is regressed on the topic proportions $\theta$. This treats the response and all the words as jointly exchangeable. But as a practical matter, our chosen formulation seems more sensible: the response depends on the topic frequencies which actually occurred in the document, rather than on the mean of the distribution generating the topics. Moreover, estimating a fully exchangeable model with enough topics allows some topics to be used entirely to explain the response variables, and others to be used to explain the word occurrences. This degrades predictive performance, as demonstrated in [2].

We treat $\alpha$, $\beta_{1:K}$, $\eta$, and $\sigma^2$ as unknown constants to be estimated, rather than random variables. We carry out approximate maximum-likelihood estimation using a variational expectation-maximization (EM) procedure, which is the approach taken in unsupervised LDA as well [4].

### 2.1 Variational E-step

Given a document and response, the posterior distribution of the latent variables is

$$p(\theta, z_{1:N} \mid w_{1:N}, y, \alpha, \beta_{1:K}, \eta, \sigma^2) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^N p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})\right) p(y \mid z_{1:N}, \eta, \sigma^2)}{\int d\theta\, p(\theta \mid \alpha) \sum_{z_{1:N}} \left(\prod_{n=1}^N p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})\right) p(y \mid z_{1:N}, \eta, \sigma^2)}. \quad (1)$$

The normalizing value is the marginal probability of the observed data, i.e., the document $w_{1:N}$ and response $y$. This normalizer is also known as the *likelihood*, or the *evidence*. As with LDA, it is not efficiently computable. Thus, we appeal to variational methods to approximate the posterior.

**Variational objective function.** We maximize the evidence lower bound (ELBO) $\mathcal{L}(\cdot)$, which for a single document has the form

$$\log p\big(w_{1:N}, y \mid \alpha, \beta_{1:K}, \eta, \sigma^2\big) \geq \mathcal{L}(\gamma, \phi_{1:N}; \alpha, \beta_{1:K}, \eta, \sigma^2) = \mathrm{E}[\log p(\theta \mid \alpha)] + \sum_{n=1}^N \mathrm{E}[\log p(Z_n \mid \theta)] + \sum_{n=1}^N \mathrm{E}[\log p(w_n \mid Z_n, \beta_{1:K})] + \mathrm{E}[\log p(y \mid Z_{1:N}, \eta, \sigma^2)] + \mathrm{H}(q). \quad (2)$$

Here the expectation is taken with respect to a variational distribution $q$. We choose the fully factorized distribution,

$$q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma) \prod_{n=1}^N q(z_n \mid \phi_n), \quad (3)$$

where $\gamma$ is a K-dimensional Dirichlet parameter vector and each $\phi_n$ parametrizes a categorical distribution over $K$ elements. Notice $\mathrm{E}[Z_n] = \phi_n$.

The first three terms and the entropy of the variational distribution are identical to the corresponding terms in the ELBO for unsupervised LDA [4]. The fourth term is the expected log probability of the response variable given the latent topic assignments,

$$\mathrm{E}[\log p(y \mid Z_{1:N}, \eta, \sigma^2)] = \quad = -\frac{1}{2}\log(2\pi\sigma^2) - \left(y^2 - 2y\eta^\top \mathrm{E}\big[\bar{Z}\big] + \eta^\top \mathrm{E}\big[\bar{Z}\bar{Z}^\top\big]\eta\right) \Big/ 2\sigma^2 \ . \tag{4}$$

The first expectation is $\mathrm{E}\big[\bar{Z}\big] = \bar{\phi} := (1/N)\sum_{n=1}^N \phi_n$, and the second expectation is

$$\mathrm{E}\big[\bar{Z}\bar{Z}^\top\big] \quad = \quad (1/N^2)\left(\sum_{n=1}^N \sum_{m\neq n} \phi_n\phi_m^\top + \sum_{n=1}^N \mathrm{diag}\{\phi_n\}\right). \tag{5}$$

To see (5), notice that for $m \neq n$, $\mathrm{E}[Z_n Z_m^\top] = \mathrm{E}[Z_n]\mathrm{E}[Z_m]^\top = \phi_n\phi_m^\top$ because the variational distribution is fully factorized. On the other hand, $\mathrm{E}[Z_n Z_n^\top] = \mathrm{diag}(\mathrm{E}[Z_n]) = \mathrm{diag}(\phi_n)$ because $Z_n$ is an indicator vector.

For a single document-response pair, we maximize (2) with respect to $\phi_{1:N}$ and $\gamma$ to obtain an estimate of the posterior. We use block coordinate-ascent variational inference, maximizing with respect to each variational parameter vector in turn.

**Optimization with respect to $\gamma$.** The terms that involve the variational Dirichlet $\gamma$ are identical to those in unsupervised LDA, i.e., they do not involve the response variable $y$. Thus, the coordinate ascent update is as in [4],

$$\gamma^{\,\mathrm{new}} \leftarrow \alpha + \textstyle\sum_{n=1}^N \phi_n. \tag{6}$$

**Optimization with respect to $\phi_j$.** Define $\phi_{-j} := \sum_{n\neq j}\phi_n$. Given $j \in \{1,\ldots,N\}$. In [3], we maximize the Lagrangian of the ELBO, which incorporates the constraint that the components of $\phi_j$ sum to one, and obtain the coordinate update

$$\phi_j^{\mathrm{new}} \propto \exp\left\{\mathrm{E}[\log\theta \mid \gamma\,] + \mathrm{E}[\log p(w_j \mid \beta_{1:K})] \ + \left(\frac{y}{N\sigma^2}\right)\eta - \frac{\big[2\big(\eta^\top\phi_{-j}\big)\eta + (\eta\circ\eta)\big]}{2N^2\sigma^2}\right\}. \tag{7}$$

Exponentiating a vector means forming the vector of exponentials. The proportionality symbol means the components of $\phi_j^{\mathrm{new}}$ are computed according to (7), then normalized to sum to one. Note that $\mathrm{E}[\log\theta_i \mid \gamma\,] = \Psi(\gamma_i) - \Psi(\sum\gamma_j)$, where $\Psi(\cdot)$ is the digamma function.

The central difference between LDA and sLDA lies in this update. As in LDA, the $j$th word's variational distribution over topics depends on the word's topic probabilities under the actual model (determined by $\beta_{1:K}$). But $w_j$'s variational distribution, and those of all other words, affect the probability of the response, through the expected residual sum of squares (RSS), which is the second term in (4). The end result is that the update (7) also encourages $\phi_j$ to decrease this expected RSS.

The update (7) depends on the variational parameters $\phi_{-j}$ of all other words. Thus, unlike LDA, the $\phi_j$ cannot be updated in parallel. Distinct occurrences of the same term are treated separately.

### 2.2 M-step and prediction

The corpus-level ELBO lower bounds the joint log likelihood across documents, which is the sum of the per-document log-likelihoods. In the E-step, we estimate the approximate posterior distribution for each document-response pair using the variational inference algorithm described above. In the M-step, we maximize the corpus-level ELBO with respect to the model parameters $\beta_{1:K}$, $\eta$, and $\sigma^2$. For our purposes, it suffices simply to fix $\alpha$ to $1/K$ times the ones vector. In this section, we add document indexes to the previous section's quantities, so $y$ becomes $y_d$ and $\bar{Z}$ becomes $\bar{Z}_d$.

**Estimating the topics.** The M-step updates of the topics $\beta_{1:K}$ are the same as for unsupervised LDA, where the probability of a word under a topic is proportional to the expected number of times that it was assigned to that topic [4],

$$\hat{\beta}_{k,w}^{\mathrm{new}} \propto \sum_{d=1}^D \sum_{n=1}^N 1(w_{d,n} = w)\phi_{d,n}^k. \tag{8}$$

Here again, proportionality means that each $\hat{\beta}_k^{\text{new}}$ is normalized to sum to one.

**Estimating the regression parameters.** The only terms of the corpus-level ELBO involving $\eta$ and $\sigma^2$ come from the corpus-level analog of (4).

Define $y = y_{1:D}$ as the vector of response values across documents. Let $A$ be the $D \times (K+1)$ matrix whose rows are the vectors $\bar{Z}_d^\top$. Then the corpus-level version of (4) is

$$\mathrm{E}[\log p(y \mid A, \eta, \sigma^2)] = -\frac{D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathrm{E}\left[(y - A\eta)^\top (y - A\eta)\right] . \tag{9}$$

Here the expectation is over the matrix $A$, using the variational distribution parameters chosen in the previous E-step. Expanding the inner product, using linearity of expectation, and applying the first-order condition for $\eta$, we arrive at an expected-value version of the normal equations:

$$\mathrm{E}\left[A^\top A\right]\eta = \mathrm{E}[A]^\top y \qquad \Rightarrow \qquad \hat{\eta}_{\text{new}} \leftarrow \left(\mathrm{E}\left[A^\top A\right]\right)^{-1}\mathrm{E}[A]^\top y . \tag{10}$$

Note that the $d$th row of $\mathrm{E}[A]$ is just $\bar{\phi}_d$, and all these average vectors were fixed in the previous E-step. Also, $\mathrm{E}\left[A^\top A\right] = \sum_d \mathrm{E}\left[\bar{Z}_d \bar{Z}_d^\top\right]$, with each term having a fixed value from the previous E-step as well, given by (5). We caution again: formulas in the previous section, such as (5), suppress the document indexes which appear here.

We now apply the first-order condition for $\sigma^2$ to (9) and evaluate the solution at $\hat{\eta}_{\text{new}}$, obtaining:

$$\hat{\sigma}^2_{\text{new}} \leftarrow (1/D)\{y^\top y - y^\top \mathrm{E}[A]\left(\mathrm{E}\left[A^\top A\right]\right)^{-1}\mathrm{E}[A]^\top y\} . \tag{11}$$

**Prediction.** Our focus in applying sLDA is prediction. Specifically, we wish to compute the expected response value, given a new document $w_{1:N}$ and a fitted model $\{\alpha, \beta_{1:K}, \eta, \sigma^2\}$:

$$\mathrm{E}[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] = \eta^\top \mathrm{E}[\bar{Z} \mid w_{1:N}, \alpha, \beta_{1:K}]. \tag{12}$$

The identity follows easily from iterated expectation. We approximate the posterior mean of $\bar{Z}$ using the variational inference procedure of the previous section. But here, the terms depending on $y$ are removed from the $\phi_j$ update in (7). Notice this is the same as variational inference for unsupervised LDA: since we averaged the response variable out of the right-hand side in (12), what remains is the standard unsupervised LDA model for $Z_{1:N}$ and $\theta$.

Thus, given a new document, we first compute $\mathrm{E}_q[Z_{1:N}]$, the variational posterior distribution of the latent variables $Z_n$. Then, we estimate the response with

$$\mathrm{E}[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] \approx \eta^\top \mathrm{E}_q[\bar{Z}] = \eta^\top \bar{\phi}. \tag{13}$$

### 2.3 Diverse response types via generalized linear models

Up to this point, we have confined our attention to an unconstrained real-valued response variable. In many applications, however, we need to predict a categorical label, or a non-negative integral count, or a response with other kinds of constraints. Sometimes it is reasonable to apply a normal linear model to a suitably transformed version of such a response. When no transformation results in approximate normality, statisticians often make use of a *generalized linear model*, or GLM [9].

In this section, we describe sLDA in full generality, replacing the normal linear model of the earlier exposition with a GLM formulation. As we shall see, the result is a generic framework which can be specialized in a straightforward way to supervised topic models having a variety of response types.

There are two main ingredients in a GLM: the "random component" and the "systematic component." For the random component, one takes the distribution of the response to be an *exponential dispersion family* with natural parameter $\zeta$ and dispersion parameter $\delta$:

$$p(y \mid \zeta, \delta) = h(y, \delta)\exp\left\{\frac{\zeta y - A(\zeta)}{\delta}\right\} . \tag{14}$$

For each fixed $\delta$, (14) is an exponential family, with base measure $h(y, \delta)$, sufficient statistic $y$, and log-normalizer $A(\zeta)$. The dispersion parameter provides additional flexibility in modeling the variance of $y$. Note that (14) need not be an exponential family jointly in $(\zeta, \delta)$.

In the systematic component of the GLM, we relate the exponential-family parameter $\zeta$ of the random component to a linear combination of covariates – the so-called *linear predictor*. For sLDA, the linear predictor is $\eta^\top \bar{z}$. In fact, we simply set $\zeta = \eta^\top \bar{z}$. Thus, in the general version of sLDA, the previous specification in step 3 of the generative process is replaced with

$$y \mid z_{1:N}, \eta, \delta \sim \text{GLM}(\bar{z}, \eta, \delta) \;, \tag{15}$$

so that

$$p(y \mid z_{1:N}, \eta, \delta) = h(y, \delta) \exp\left\{ \frac{\eta^\top (\bar{z} y) - A(\eta^\top \bar{z})}{\delta} \right\} \;. \tag{16}$$

The reader familiar with GLMs will recognize that our choice of systematic component means sLDA uses only canonical link functions. In future work, we will relax this constraint.

We now have the flexibility to model any type of response variable whose distribution can be written in exponential dispersion form (14). As is well known, this includes many commonly used distributions: the normal; the binomial (for binary response); the Poisson and negative binomial (for count data); the gamma, Weibull, and inverse Gaussian (for failure time data); and others. Each of these distributions corresponds to a particular choice of $h(y, \delta)$ and $A(\zeta)$. For example, it is easy to show that the normal distribution corresponds to $h(y, \delta) = (1/\sqrt{2\pi\delta}) \exp\{-y^2/(2\delta)\}$ and $A(\zeta) = \zeta^2/2$. In this case, the usual parameters $\mu$ and $\sigma^2$ just equal $\zeta$ and $\delta$, respectively.

**Variational E-step.** The distribution of $y$ appears only in the cross-entropy term (4). Its form under the GLM is

$$\text{E}[\log p(y \mid Z_{1:N}, \eta, \delta)] = \log h(y, \delta) + \frac{1}{\delta} \left[ \eta^\top \left( \text{E}\left[\bar{Z}\right] y \right) - \text{E}\left[A(\eta^\top \bar{Z})\right] \right] \;. \tag{17}$$

This changes the coordinate ascent step for each $\phi_j$, but the variational optimization is otherwise unaffected. In particular, the gradient of the ELBO with respect to $\phi_j$ becomes

$$\frac{\partial \mathcal{L}}{\partial \phi_j} = \text{E}[\log \theta \mid \gamma\,] + \text{E}[\log p(w_j \mid \beta_{1:K})] - \log \phi_j + 1 + \left(\frac{y}{N\delta}\right) \eta - \left(\frac{1}{\delta}\right) \frac{\partial}{\partial \phi_j} \left\{ \text{E}\left[A(\eta^\top \bar{Z})\right] \right\} . \tag{18}$$

Thus, the key to variational inference in sLDA is obtaining the gradient of the expected GLM log-normalizer. Sometimes there is an exact expression, such as the normal case of Section 2. As another example, the Poisson GLM leads to an exact gradient, which we omit for brevity.

Other times, no exact gradient is available. In a longer paper [3], we study two methods for this situation. First, we can replace $-\text{E}[A(\eta^\top \bar{Z})]$ with an adjustable lower bound whose gradient is known exactly; then we maximize over the original variational parameters plus the parameter controlling the bound. Alternatively, an application of the multivariate delta method for moments [1], plus standard exponential family theory, shows

$$\text{E}\left[A(\eta^\top \bar{Z})\right] \approx A(\eta^\top \bar{\phi}) + \text{Var}_{\text{GLM}}(Y \mid \zeta = \eta^\top \bar{\phi}) \cdot \eta^\top \text{Var}_q(\bar{Z}) \eta \;. \tag{19}$$

Here, $\text{Var}_{\text{GLM}}$ denotes the response variance under the GLM, given a specified value of the natural parameter—in all standard cases, this variance is a closed-form function of $\phi_j$. The variance-covariance matrix of $\bar{Z}$ under $q$ is already known in closed from from $\text{E}[\bar{Z}]$ and (5). Thus, computing $\partial/\partial\phi_j$ of (19) exactly is mechanical. However, using this approximation gives up the usual guarantee that the ELBO lower bounds the marginal likelihood. We forgo details and further examples due to space constraints.

The GLM contribution to the gradient determines whether the $\phi_j$ coordinate update itself has a closed form, as it does in the normal case (7) and the Poisson case (omitted). If the update is not closed-form, we use numerical optimization, supplying a gradient obtained from one of the methods described in the previous paragraph.

**Parameter estimation (M-step).** The topic parameter estimates are given by (8), as before. For the corpus-level ELBO, the gradient with respect to $\eta$ becomes

$$\frac{\partial}{\partial \eta} \left(\frac{1}{\delta}\right) \sum_{d=1}^{D} \left\{ \eta^\top \bar{\phi}_d y_d - \text{E}\left[A(\eta^\top \bar{Z}_d)\right] \right\} = \left(\frac{1}{\delta}\right) \left\{ \sum_{d=1}^{D} \bar{\phi}_d y_d - \sum_{d=1}^{D} \text{E}_q\left[\mu(\eta^\top \bar{Z}_d) \bar{Z}_d\right] \right\} . \tag{20}$$

The appearance of $\mu(\cdot) = \text{E}_{\text{GLM}}[Y \mid \zeta = \cdot\,]$ follows from exponential family properties. This GLM mean response is a known function of $\eta^\top \bar{Z}_d$ in all standard cases. However, $\text{E}_q[\mu(\eta^\top \bar{Z}_d) \bar{Z}_d]$ has

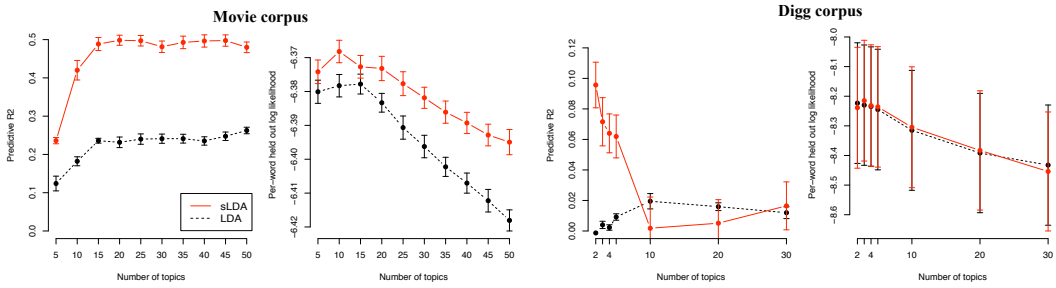



Figure 2: Predictive $R^2$ and per-word likelihood for the movie and Digg data (see Section 3).

an exact solution only in some cases (e.g. normal, Poisson). In other cases, we approximate the expectation with methods similar to those applied for the $\phi_j$ coordinate update. Reference [3] has details, including estimation of $\delta$ and prediction, where we encounter the same issues.

The derivative with respect to $\delta$, evaluated at $\hat{\eta}_{\text{new}}$, is

$$\left\{\sum_{d=1}^{D} \frac{\partial h(y_d, \delta)/\partial\delta}{h(y_d, \delta)}\right\} - \left(\frac{1}{\delta^2}\right)\left\{\sum_{d=1}^{D} \bar{\phi}_d y_d - \sum_{d=1}^{D} \mathrm{E}_q\left[\mu(\hat{\eta}_{\text{new}}^\top \bar{Z}_d)\bar{Z}_d\right]\right\}. \tag{21}$$

Given that the rightmost summation has been evaluated, exactly or approximately, during the $\eta$ optimization, (21) has a closed form. Depending on $h(y, \delta)$ and its partial with respect to $\delta$, we obtain $\hat{\delta}_{\text{new}}$ either in closed form or via one-dimensional numerical optimization.

**Prediction.** We form predictions just as in Section 2.2. The difference is that we now approximate the expected response value of a test document as

$$\mathrm{E}[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \delta] \approx \mathrm{E}_q[\mu(\eta^\top \bar{Z})]. \tag{22}$$

Again, this follows from iterated expectation plus the variational approximation. When the variational expectation cannot be computed exactly, we apply the approximation methods we relied on for the GLM E-step and M-step. We defer specifics to [3].

## 3 Empirical results

We evaluated sLDA on two prediction problems. First, we consider "sentiment analysis" of newspaper movie reviews. We use the publicly available data introduced in [10], which contains movie reviews paired with the number of stars given. While Pang and Lee treat this as a classification problem, we treat it as a regression problem. With a 5000-term vocabulary chosen by tf-idf, the corpus contains 5006 documents and comprises 1.6M words.

Second, we introduce the problem of predicting web page popularity on Digg.com. Digg is a community of users who share links to pages by submitting them to the Digg homepage, with a short description. Once submitted, other users "digg" the links they like. Links are sorted on the Digg homepage by the number of diggs they have received. Our Digg data set contains a year of link descriptions, paired with the number of diggs each received during its first week on the homepage. (This corpus will be made publicly available at publication.) We restrict our attention to links in the technology category. After trimming the top ten outliers, and using a 4145-term vocabulary chosen by tf-idf, the Digg corpus contains 4078 documents and comprises 94K words.

For both sets of response variables, we transformed to approximate normality by taking logs. This makes the data amenable to the continuous-response model of Section 2; for these two problems, generalized linear modeling turned out to be unnecessary. We initialized $\beta_{1:K}$ to uniform topics, $\sigma^2$ to the sample variance of the response, and $\eta$ to a grid on $[-1, 1]$ in increments of $2/K$. We ran EM until the relative change in the corpus-level likelihood bound was less than 0.01%. In the E-step, we ran coordinate-ascent variational inference for each document until the relative change in the

per-document ELBO was less than 0.01%. For the movie review data set, we illustrate in Figure 1 a matching of the top words from each topic to the corresponding coefficient $\eta_k$.

We assessed the quality of the predictions with "predictive $R^2$." In our 5-fold cross-validation (CV), we defined this quantity as the fraction of variability in the out-of-fold response values which is captured by the out-of-fold predictions: $pR^2 := 1 - (\sum(y - \hat{y})^2)/(\sum(y - \bar{y})^2)$.

We compared sLDA to linear regression on the $\bar{\phi}_d$ from unsupervised LDA. This is the regression equivalent of using LDA topics as classification features [4].Figure 2 (L) illustrates that sLDA provides improved predictions on both data sets. Moreover, this improvement does not come at the cost of document model quality. The per-word hold-out likelihood comparison in Figure 2 (R) shows that sLDA fits the document data as well or better than LDA. Note that Digg prediction is significantly harder than the movie review sentiment prediction, and that the homogeneity of Digg technology content leads the model to favor a small number of topics.

Finally, we compared sLDA to the lasso, which is $L_1$-regularized least-squares regression. The lasso is a widely used prediction method for high-dimensional problems. We used each document's empirical distribution over words as its lasso covariates, setting the lasso complexity parameter with 5-fold CV. On Digg data, the lasso's optimal model complexity yielded a CV $pR^2$ of 0.088. The best sLDA $pR^2$ was 0.095, an 8.0% relative improvement. On movie data, the best Lasso $pR^2$ was 0.457 versus 0.500 for sLDA, a 9.4% relative improvement. Note moreover that the Lasso provides only a prediction rule, whereas sLDA models latent structure useful for other purposes.

## 4 Discussion

We have developed sLDA, a statistical model of labelled documents. The model accommodates the different types of response variable commonly encountered in practice. We presented a variational procedure for approximate posterior inference, which we then incorporated in an EM algorithm for maximum-likelihood parameter estimation. We studied the model's predictive performance on two real-world problems. In both cases, we found that sLDA moderately improved on the lasso, a state-of-the-art regularized regression method. Moreover, the topic structure recovered by sLDA had higher hold-out likelihood than LDA on one problem, and equivalent hold-out likelihood on the other. These results illustrate the benefits of supervised dimension reduction when prediction is the ultimate goal.

## Acknowledgments

David M. Blei is supported by grants from Google and the Microsoft Corporation.

## References

[1] P. Bickel and K. Doksum. *Mathematical Statistics*. Prentice Hall, 2000.

[2] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR*, pages 127–134. ACM Press, 2003.

[3] D. Blei and J. McAuliffe. Supervised topic models. In preparation, 2007.

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

[5] P. Flaherty, G. Giaever, J. Kumm, M. Jordan, and A. Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15):3286–3293, 2005.

[6] K. Fukumizu, F. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. 2001.

[8] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *AAAI*, 2006.

[9] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

[10] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.