Journal of Machine Learning Research 14 (2013) 1005-1031 Submitted 9/12; Revised 1/13; Published 4/13

Variational Inference in Nonconjugate Models

Chong Wang CHONGW@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
Princeton, NJ, 08540, USA

Editor: Neil Lawrence

Abstract

Mean-field variational methods are widely used for approximate posterior inference in many probabilistic models. In a typical application, mean-field methods approximately compute the posterior with a coordinate-ascent optimization algorithm. When the model is conditionally conjugate, the coordinate updates are easily derived and in closed form. However, many models of interest—like the correlated topic model and Bayesian logistic regression—are nonconjugate. In these models, mean-field methods cannot be directly applied and practitioners have had to develop variational algorithms on a case-by-case basis. In this paper, we develop two generic methods for nonconjugate models, Laplace variational inference and delta method variational inference. Our methods have several advantages: they allow for easily derived variational algorithms with a wide class of nonconjugate models; they extend and unify some of the existing algorithms that have been derived for specific models; and they work well on real-world data sets. We studied our methods on the correlated topic model, Bayesian logistic regression, and hierarchical Bayesian logistic regression.

Keywords: variational inference, nonconjugate models, Laplace approximations, the multivariate delta method

1 Introduction

Mean-field variational inference lets us efficiently approximate posterior distributions in complex probabilistic models (Jordan et al., 1999; Wainwright and Jordan, 2008). Applications of variational inference are widespread. As examples, it has been applied to Bayesian mixtures (Attias, 2000; Corduneanu and Bishop, 2001), factorial models (Ghahramani and Jordan, 1997), and probabilistic topic models (Blei et al., 2003).

The basic idea behind mean-field inference is the following. First define a family of distributions over the hidden variables where each variable is assumed independent and governed by its own parameter. Then fit those parameters so that the resulting distribution is close to the conditional distribution of the hidden variables given the observations. Closeness is measured with the Kullback-Leibler divergence. Inference becomes optimization.

In many settings this approach can be used as a “black box” technique. In particular, this is possible when we can easily compute the conditional distribution of each hidden variable given all of the other variables, both hidden and observed. (This class contains the models mentioned above.) For such models, which are called *conditionally conjugate* models, it is easy to derive a coordinate ascent algorithm that optimizes the parameters of the variational distribution (Beal, 2003; Bishop, 2006). This is the principle behind software tools like VIBES (Bishop et al., 2003) and Infer.NET (Minka et al., 2010), which allow practitioners to define models of their data and immediately approximate the corresponding posterior with variational inference.

Many models of interest, however, do not enjoy the properties required to take advantage of this easily derived algorithm. Such *nonconjugate* models[1] include Bayesian logistic regression (Jaakkola and Jordan, 1997), Bayesian generalized linear models (Wells, 2001), discrete choice models (Braun and McAuliffe, 2010), Bayesian item response models (Clinton et al., 2004; Fox, 2010), and nonconjugate topic models (Blei and Lafferty, 2006, 2007). Using variational inference in these settings requires algorithms tailored to the specific model at hand. Researchers have developed a variety of strategies for a variety of models, including approximations (Braun and McAuliffe, 2010; Ahmed and Xing, 2007), alternative bounds (Jaakkola and Jordan, 1997; Blei and Lafferty, 2006, 2007; Khan et al., 2010), and numerical quadrature (Honkela and Valpola, 2004).

In this paper we develop two approaches to mean-field variational inference for a large class of nonconjugate models. First we develop *Laplace variational inference*. This approach embeds Laplace approximations—an approximation technique for continuous distributions (Tierney et al., 1989; MacKay, 1992)—within a variational optimization algorithm. We then develop *delta method variational inference*. This approach optimizes a Taylor approximation of the variational objective. The details of the algorithm depend on how the approximation is formed. Formed one way, it gives an alternative interpretation of Laplace variational inference. Formed another way, it is equivalent to using a multivariate delta approximation (Bickel and Doksum, 2007) of the variational objective.

Our methods are generic. Given a model, they can be derived nearly as easily as traditional coordinate-ascent inference. Unlike traditional inference, however, they place fewer conditions on the model, conditions that are less restrictive than conditional conjugacy. Our methods significantly expand the class of models for which mean-field variational inference can be easily applied.

We studied our algorithms with three nonconjugate models: Bayesian logistic regression (Jaakkola and Jordan, 1997), hierarchical logistic regression (Gelman and Hill, 2007), and the correlated topic model (Blei and Lafferty, 2007). We found that our methods give better results than those obtained through special-purpose techniques. Further, we found that Laplace variational inference usually outperforms delta method variational inference, both in terms of computation time and the fidelity of the approximate posterior.

Related work. We have described the various approaches that researchers have developed for specific models. There have been other efforts to examine generic variational inference in nonconjugate models. Paisley et al. (2012a) proposed a variational inference approach using stochastic search for nonconjugate models, approximating the intractable integrals with Monte Carlo methods. Gershman et al. (2012) proposed a nonparametric variational inference algorithm, which can be applied to nonconjugate models. Knowles and Minka (2011) presented a message passing algorithm for nonconjugate models, which has been implemented in Infer.NET (Minka et al., 2010); their technique applies to a subset of models described in this paper.[2]

[1] Carlin and Polson (1991) coined the term "nonconjugate model" to describe a model that does not enjoy full conditional conjugacy.

[2] It may be generalizable to the full set. However, one must determine how to compute the required expectations.

Laplace approximations have been used in approximate inference in more complex models, though not in the context of mean-field variational inference. Smola et al. (2003) used them to approximate the difficult-to-compute moments in expectation propagation (Minka, 2001). Rue et al. (2009) used them for inference in latent Gaussian models. Here we want to use them for variational inference, in a method that can be applied to a wider range of nonconjugate models.

Finally, we note that the delta method was first used in variational inference by Braun and McAuliffe (2010) in the context of the discrete choice model. Our method generalizes their approach.

Organization of this paper. In Section 2 we review mean-field variational inference and define the class of nonconjugate models to which our algorithms apply. In Section 3, we derive Laplace and delta-method variational inference and present our full algorithm for nonconjugate inference. In Section 4, we show how to use our generic method on several example models and in Section 5 we study its performance on these models. In Section 6, we summarize and discuss this work.

2 Variational Inference and a Class of Nonconjugate Models

We consider a generic model with observations x and hidden variables θ and z,

$$p(\theta,z,x) = p(x|z)p(z|\theta)p(\theta). \tag{1}$$

The distinction between the two hidden variables will be made clear below.

The inference problem is to compute the posterior,

$$p(\theta,z|x) = \frac{p(\theta,z,x)}{\int p(\theta,z,x)dzd\theta}.$$

This is intractable for many models because the denominator is difficult to compute; we must approximate the distribution. In variational inference, we approximate the posterior by positing a simple family of distributions over the latent variables $q(\theta,z)$ and then finding the member of that family which minimizes the Kullback-Leibler (KL) divergence to the true posterior (Jordan et al., 1999; Wainwright and Jordan, 2008).[3]

In this section we review variational inference and discuss mean-field variational inference for the class of conditionally conjugate models. We then define a wider class of nonconjugate models for which mean-field variational inference is not as easily applied. In the next section, we derive algorithms for performing mean-field variational inference in this larger class of models.

2.1 Mean-field Variational Inference

Mean-field variational inference is simplest and most widely used variational inference method. In mean-field variational inference we posit a fully factorized variational family,

$$q(\theta,z) = q(\theta)q(z). \tag{2}$$

[3] In this paper, we focus on mean-field variational inference where we minimize the KL divergence to the posterior. We note that there are other kinds of variational inference, with more structured variational distributions or with alternative objective functions (Wainwright and Jordan, 2008; Barber, 2012). In this paper, we use "variational inference" to indicate mean-field variational inference that minimizes the KL divergence.

In this family of distributions the variables are independent and each is governed by its own distribution. This family usually does not contain the posterior, where θ and z are dependent. However, it is very flexible—it can capture any set of marginals of the hidden variables.

Under the standard variational theory, minimizing the KL divergence between $q(\theta,z)$ and the posterior $p(\theta,z|x)$ is equivalent to maximizing a lower bound of the log marginal likelihood of the observed data x. We obtain this bound with Jensen's inequality,

$$\begin{aligned}\log p(x) &= \log \textstyle\int p(\theta,z,x)\mathrm{d}z\mathrm{d}\theta \\ &\geq \mathbb{E}_q\left[\log p(\theta,z,x)\right] - \mathbb{E}_q\left[\log q(\theta,z)\right] \\ &\triangleq \mathcal{L}(q),\end{aligned} \tag{3}$$

where $\mathbb{E}_q\left[\cdot\right]$ is the expectation taken with respect to q and note the second term is the entropy of q. We call $\mathcal{L}(q)$ the variational objective.

Setting $\partial\mathcal{L}(q)/\partial q = 0$ shows that the optimal solution satisfies the following,

$$q^*(\theta) \propto \exp\left\{\mathbb{E}_{q(z)}\left[\log p(z|\theta)p(\theta)\right]\right\}, \tag{4}$$

$$q^*(z) \propto \exp\left\{\mathbb{E}_{q(\theta)}\left[\log p(x|z)p(z|\theta)\right]\right\}. \tag{5}$$

Here we have combined the optimal conditions from Bishop (2006) with the particular factorization of Equation 1. Note that the variational objective usually contains many local optima.

These conditions lead to the traditional coordinate ascent algorithm for variational inference. It iterates between holding $q(z)$ fixed to update $q(\theta)$ from Equation 4 and holding $q(\theta)$ fixed to update $q(z)$ from Equation 5. This converges to a local optimum of the variational objective (Bishop, 2006).

When all the nodes in a model are *conditionally conjugate*, the coordinate updates of Equation 4 and Equation 5 are available in closed form. A node is conditionally conjugate when its conditional distribution given its Markov blanket (i.e., the set of random variables that it is dependent on in the posterior) is in the same family as its conditional distribution given its parents (i.e., its factor in the joint distribution). For example, in Equation 1 suppose the factor $p(\theta)$ is a Dirichlet and both factors $p(z\,|\,\theta)$ and $p(x\,|\,z)$ are multinomials. This means that the conditional $p(\theta|z)$ is also a Dirichlet and the conditional $p(z\,|\,x,\theta)$ is also a multinomial. This model, which is latent Dirichlet allocation (Blei et al., 2003), is conditionally conjugate. Many applications of variational inference have been developed for this type of model (Bishop, 1999; Attias, 2000; Beal, 2003).

However, if there exists any node in the model that is not conditionally conjugate then this coordinate ascent algorithm is not available. That setting arises in many practical models and does not permit closed-form updates or easy calculation of the variational objective. We will develop generic variational inference algorithms for a wide class of nonconjugate models. First, we define that class.

2.2 A Class of Nonconjugate Models

We present a wide class of nonconjugate models, still assuming the factorization of Equation 1.

1. We assume that θ is real-valued and the distribution $p(\theta)$ is twice differentiable with respect to θ. If we require $\theta > \theta_0$ (θ_0 is a constant), we may define a distribution over $\log(\theta - \theta_0)$. These assumptions cover exponential families, such as the Gaussian, Poisson and gamma, as well as more complex distributions, such as a student-t.

2. We assume the distribution $p(z \mid \theta)$ is in the exponential family (Brown, 1986),

$$p(z \mid \theta) = h(z) \exp\left\{\eta(\theta)^\top t(z) - a(\eta(\theta))\right\}, \tag{6}$$

   where $h(z)$ is a function of z; $t(z)$ is the sufficient statistic; $\eta(\theta)$ is the natural parameter, which is a function of the conditioning variables; and $a(\eta(\theta))$ is the log partition function. We also assume that $\eta(\theta)$ is twice differentiable; since θ is real-valued, this is satisfied in most statistical models. Unlike in conjugate models, these assumptions do not restrict $p(\theta)$ and $p(z \mid \theta)$ to be a conjugate pair; the conditional distribution $p(\theta \mid z)$ is not necessarily in the same family as the prior $p(\theta)$.

3. The distribution $p(x \mid z)$ is in the exponential family,

$$p(x \mid z) = h(x) \exp\left\{t(z)^\top \langle t(x), 1\rangle\right\}. \tag{7}$$

   We set up this exponential family so that the natural parameter for x is all but the last component of $t(z)$ and the last component is the negative log normalizer $-a(\cdot)$. Thus, the distribution of z is conjugate to the conditional distribution of x; the conditional $p(z \mid \theta, x)$ is in the same family as $p(z \mid \theta)$ (Bernardo and Smith, 1994).

Our terminology follows these assumptions: θ is the *nonconjugate variable*, z is the *conjugate variable*, and x is the *observation*.

This class of models is larger than the class of conditionally conjugate models. Our expanded class also includes nonconjugate models like the correlated topic model (Blei and Lafferty, 2007), dynamic topic model (Blei and Lafferty, 2006), Bayesian logistic regression (Jaakkola and Jordan, 1997; Gelman and Hill, 2007), discrete choice models (Braun and McAuliffe, 2010), Bayesian ideal point models (Clinton et al., 2004), and many others. Further, the methods we develop below are easily adapted more complicated graphical models, those that contain conjugate and nonconjugate variables whose dependencies are encoded in a directed acyclic graph. Appendix A outlines how to adapt our algorithms to this more general case.

Example: Hierarchical language modeling. We introduce the hierarchical language model, a simple example of a nonconjugate model to help ground our derivation of the general algorithms. Consider the problem of unigram language modeling. We are given a collection of documents $\mathcal{D} = x_{1:D}$ where each document x_d is a vector of word counts, observations from a discrete vocabulary of length V. We model each document with its own distribution over words and place a Dirichlet prior on that distribution. This model is used, for example, in the language modeling approach to information retrieval (Croft and Lafferty, 2003).

We want to place a prior on the Dirichlet parameters, a positive V-vector, that govern each document's distribution over terms. In theory, every exponential family distribution has a conjugate prior (Bernardo and Smith, 1994) and the prior to the Dirichlet is the multi-gamma distribution (Kotz et al., 2000). However, the multi-gamma is difficult to work with because its log normalizer is not easy to compute. As an alternative, we place a log normal distribution on the Dirichlet parameters. This is not the conjugate prior.

The full generative process is as follows:

1. Draw log Dirichlet parameters $\theta \sim \mathcal{N}(0, I)$.
2. For each document d, $1 \leq d \leq D$:

(a) Draw multinomial parameter $z_d \,|\, \theta \sim \text{Dirichlet}(\exp\{\theta\})$.
(b) Draw word counts $x_d \sim \text{Multinomial}(N, z_d)$.

Given a collection of documents, our goal is to compute the posterior distribution $p(\theta, z_{1:D} \,|\, x_{1:D})$. Traditional variational or Gibbs sampling methods cannot be easily used because the normal prior on the parameters θ is not conjugate to the Dirichlet($\exp\{\theta\}$) likelihood.

This language model fits into our model class. In the notation of the joint distribution of Equation 1, $\theta = \theta$, $z = z_{1:D}$, and $x = x_{1:D}$. The per-document multinomial parameters z and word counts x are conditionally independent given the Dirichlet parameters θ,

$$\begin{aligned} p(z \,|\, \theta) &= \textstyle\prod_d p(z_d \,|\, \theta), \\ p(x \,|\, z) &= \textstyle\prod_d \prod_n p(x_{dn} \,|\, z_d). \end{aligned}$$

In this case, the natural parameter $\eta(\theta) = \exp\{\theta\}$. This model satisfies the assumptions: the log normal $p(\exp\{\theta\})$ is not conjugate to the Dirichlet $p(z_d \,|\, \exp\{\theta\})$ but is twice differentiable; the Dirichlet is conjugate to the multinomial $p(x_d \,|\, z_d)$ and the multinomial is in the exponential family.

Below we will use various components of the exponential family form of the Dirichlet:

$$h(z_d) = \textstyle\prod_i z_{di}^{-1}; \;\; t(z_d) = \log z_d; \;\; a_d(\eta(\theta)) = \sum_i \log\Gamma(\exp\{\theta_i\} - \log\Gamma\left(\sum_i \exp\{\theta_i\}\right). \qquad (8)$$

We will return to this model as a simple running example.

3 Laplace and Delta Method Variational Inference

We have defined a class of nonconjugate models. Variational inference is difficult to derive for these models because $p(\theta)$ is not conjugate to $p(z \,|\, \theta)$. Specifically, the update in Equation 4 does not necessarily have the form of an exponential family we can work with and it is difficult to use $\mathbb{E}_{q(\theta)}[\log p(z \,|\, \theta)]$ in the update of Equation 5.

We will develop two variational inference algorithms for this class: *Laplace variational inference* and *delta method variational inference*. Both use coordinate ascent to optimize the variational parameters, iterating between updating $q(\theta)$ and $q(z)$. They differ in how they update the variational distribution of the nonconjugate variable $q(\theta)$. In Laplace variational inference, we use Laplace approximations (MacKay, 1992; Tierney et al., 1989) within the coordinate ascent updates of Equation 4 and Equation 5. In delta method variational inference, we apply Taylor approximations to approximate the variational objective in Equation 3 and then derive the corresponding updates. Different ways of taking the Taylor approximation lead to different algorithms. Formed one way, this recovers the Laplace approximation. Formed another way, it is equivalent to using a multivariate delta approximation (Bickel and Doksum, 2007) of the variational objective function.

In both variants, the variational distribution is the mean-field family in Equation 2. The variational distribution of the nonconjugate variable $q(\theta)$ is a Gaussian; the variational distribution of the conjugate variable $q(z)$ is in the same family as $p(z \,|\, \eta(\theta))$. In Laplace inference, these forms emerge from the derivation. In delta method inference, they are assumed. The complete variational family is,

$$q(\theta, z) = q(\theta \,|\, \mu, \Sigma) q(z \,|\, \phi).$$

where (μ,Σ) are the parameters for a Gaussian distribution and ϕ is a natural parameter for z. For example, in the hierarchical language model of Section 2.2, ϕ is a collection of D Dirichlet parameters. We will sometimes suppress the parameters, writing $q(\theta)$ for $q(\theta\,|\,\mu,\Sigma)$.

Our algorithms are coordinate ascent algorithms, where we iterate between updating the nonconjugate variational distribution $q(\theta)$ and updating the conjugate variational distribution $q(z)$. In the subsections below, we derive the update for $q(\theta)$ in each algorithm. Then, for both algorithms, we derive the update for $q(z)$. The full procedure is described in Section 3.4 and Figure 1.

3.1 Laplace Variational Inference

We first review the Laplace approximation. Then we show how to use it in variational inference.

3.1.1 The Laplace Approximation

Laplace approximations use a Gaussian to approximate an intractable density. Consider approximating an intractable posterior $p(\theta\,|\,x)$. (There is no hidden variable z in this set up.) Assume the joint distribution $p(x,\theta) = p(x\,|\,\theta)p(\theta)$ is easy to compute. Laplace approximations use a Taylor approximation around the maximum a posterior (MAP) point to construct a Gaussian proxy for the posterior. They are used for continuous distributions.

First, notice the posterior is proportional to the exponentiated log joint

$$p(\theta\,|\,x) = \exp\{\log p(\theta\,|\,x)\} \propto \exp\{\log p(\theta, x)\}.$$

Let $\hat{\theta}$ be the MAP of $p(\theta\,|\,x)$, found by maximizing $\log p(\theta,x)$. A Taylor expansion around $\hat{\theta}$ gives

$$\log p(\theta\,|\,x) \approx \log p(\hat{\theta}\,|\,x) + \tfrac{1}{2}(\theta-\hat{\theta})^\top H(\hat{\theta})(\theta-\hat{\theta}). \tag{9}$$

The term $H(\hat{\theta})$ is the Hessian of $\log p(\theta\,|\,x)$ evaluated at $\hat{\theta}$, $H(\hat{\theta}) \triangleq \nabla^2 \log p(\theta\,|\,x)|_{\theta=\hat{\theta}}$.

In the Taylor expansion of Equation 9, the first-order term $(\theta-\hat{\theta})^\top \nabla \log p(\theta\,|\,x)|_{\theta=\hat{\theta}}$ equals zero. The reason is that $\hat{\theta}$ is the maximum of $\log p(\theta\,|\,x)$ and so its gradient $\nabla \log p(\theta\,|\,x)|_{\theta=\hat{\theta}}$ is zero. Exponentiating Equation 9 gives the approximate Gaussian posterior

$$p(\theta\,|\,x) \approx \tfrac{1}{C}\exp\left\{-\tfrac{1}{2}(\theta-\hat{\theta})^\top \left(-H(\hat{\theta})\right)(\theta-\hat{\theta})\right\},$$

where C is a normalizing constant. In other words, $p(\theta\,|\,x)$ can be approximated by

$$p(\theta\,|\,x) \approx \mathcal{N}(\hat{\theta}, -H(\hat{\theta})^{-1}).$$

This is the Laplace approximation. While powerful, it is difficult to use in multivariate settings, for example, when there are discrete hidden variables. Now we describe how we use Laplace approximations as part of a variational inference algorithm for more complex models.

3.1.2 Laplace Updates in Variational Inference

We adapt the idea behind Laplace approximations to update the variational distribution $q(\theta)$. First, we combine the coordinate update in Equation 4 with the exponential family assumption in Equation 6,

$$q(\theta) \propto \exp\left\{\eta(\theta)^\top \mathbb{E}_{q(z)}\left[t(z)\right] - a(\eta(\theta)) + \log p(\theta)\right\}. \tag{10}$$

Define the function $f(\theta)$ to contain the terms inside the exponent of the update,

$$f(\theta) \triangleq \eta(\theta)^{\top} \mathbb{E}_{q(z)}[t(z)] - a(\eta(\theta)) + \log p(\theta). \tag{11}$$

The terms of $f(\theta)$ come from the model and involve $q(z)$ or θ. Recall that $q(z)$ is in the same exponential family as $p(z \mid \theta)$, $t(z)$ are its sufficient statistics, and ϕ is the variational parameter. We can compute $\mathbb{E}_{q(z)}[t(z)]$ from a basic property of the exponential family (Brown, 1986),

$$\mathbb{E}_{q(z)}[t(z)] = \nabla a(\phi).$$

Seen another way, $f(\theta) = \mathbb{E}_{q(z)}[\log p(\theta, z)]$. This function will be important in both Laplace and delta method inference.

The problem with nonconjugate models is that we cannot update $q(\theta)$ exactly using Equation 10 because $q(\theta) \propto \exp\{f(\theta)\}$ cannot be normalized in closed form. We approximate the update by taking a second-order Taylor approximation of $f(\theta)$ around its maximum, following the same logic as from the original Laplace approximation in Equation 9. The Taylor approximation for $f(\theta)$ around $\hat{\theta}$ is

$$f(\theta) \approx f(\hat{\theta}) + \triangledown f(\hat{\theta})(\theta - \hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^{\top} \triangledown^2 f(\hat{\theta})(\theta - \hat{\theta}), \tag{12}$$

where $\triangledown^2 f(\hat{\theta})$ is the Hessian matrix evaluated at $\hat{\theta}$. Now let $\hat{\theta}$ be the value that maximizes $f(\theta)$. This implies that $\triangledown f(\hat{\theta}) = 0$ and Equation 12 simplifies to

$$q(\theta) \propto \exp\{f(\theta)\} \approx \exp\left\{f(\hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^{\top} \triangledown^2 f(\hat{\theta})(\theta - \hat{\theta})\right\}.$$

Thus the approximate update for $q(\theta)$ is to set it to

$$q(\theta) \approx \mathcal{N}\left(\hat{\theta}, -\triangledown^2 f(\hat{\theta})^{-1}\right). \tag{13}$$

Note we did not assume $q(\theta)$ is Gaussian. Its Gaussian form stems from the Taylor approximation.

The update in Equation 13 can be used in a coordinate ascent algorithm for a nonconjugate model. We iterate between holding $q(z)$ fixed while updating $q(\theta)$ from Equation 13, and holding $q(\theta)$ fixed while updating $q(z)$. (We derive the second update in Section 3.3.) Each time we update $q(\theta)$ we must use numerical optimization to obtain $\hat{\theta}$, the optimal value of $f(\theta)$.

We return to the hierarchical language model of Section 2.2, where θ are the log of the parameters to the Dirichlet distribution. Implementing the algorithm to update $q(\theta)$ involves forming $f(\theta)$ for the model at hand and deriving an algorithm to optimize it.

With the model equations in Equation 8, we have

$$f(\theta) = \exp(\theta)^{\top} \mathbb{E}_{q(z)}[t(z)] - D\left(\textstyle\sum_i \log \Gamma(\exp(\theta_i) - \log \Gamma\left(\textstyle\sum_i \exp(\theta_i)\right)\right) - (1/2)\theta^{\top}\theta. \tag{14}$$

The expected sufficient statistics of the conjugate variable are

$$\mathbb{E}_{q(z)}[t(z)] = \textstyle\sum_d \mathbb{E}_{q(z_d)}[t(z_d)] = \textstyle\sum_d \Psi(\phi_d) - \Psi(\textstyle\sum_i \phi_{di}),$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of $\log \Gamma(\cdot)$. (This function will also arise in the gradient.) It is straightforward to use numerical methods, such as conjugate gradient (Bertsekas, 1999), to optimize Equation 14. We can then use Equation 13 to update the nonconjugate variable.

3.2 Delta Method Variational Inference

In Laplace variational inference, the variational distribution $q(\theta)$ Equation 13 is solely a function of $\hat{\theta}$, the maximum of $f(\theta)$ in Equation 11. A natural question is, would other values of θ be suitable as well? To consider such alternatives, we describe a different technique for variational inference. We approximate the variational objective $\mathcal{L}$ in Equation 3 and then optimize that approximation.

Again we focus on updating $q(\theta)$ in a coordinate ascent algorithm and postpone the discussion of updating $q(z)$. We set the variational distribution $q(\theta)$ to be a Gaussian $\mathcal{N}(\mu,\Sigma)$, where the parameters are free variational parameters fit to optimize the variational objective. (Note that in Laplace inference, this Gaussian family came out of the derivation.) We isolate the terms of the objective in Equation 3 related to $q(\theta)$, and we substitute the exponential family form of $p(z\,|\,\theta)$ in Equation 6,

$$\mathcal{L}(q(\theta)) = \mathbb{E}_{q(\theta)}\left[\eta(\theta)^\top \mathbb{E}_{q(z)}\left[t(z)\right] - a(\eta(\theta)) + \log p(\theta)\right] + \tfrac{1}{2}\log|\Sigma|.$$

The second term comes from the entropy of the Gaussian,

$$-\mathbb{E}_{q(\theta)}\left[\log q(\theta)\right] = \frac{1}{2}\log|\Sigma| + C,$$

where C is a constant and is excluded from the objective. The first term is $\mathbb{E}_{q(\theta)}\left[f(\theta)\right]$, where $f(\cdot)$ is the same as defined for Laplace inference in Equation 11. Thus,

$$\mathcal{L}(q(\theta)) = \mathbb{E}_{q(\theta)}\left[f(\theta)\right] + \tfrac{1}{2}\log|\Sigma|.$$

We cannot easily compute the expectation in the first term. So we use a Taylor approximation of $f(\theta)$ around a chosen value $\hat{\theta}$ (Equation 12) and then take the expectation,

$$\mathcal{L}(q(\theta)) \approx f(\hat{\theta}) + \nabla f(\hat{\theta})^\top(\mu-\hat{\theta}) + \tfrac{1}{2}(\mu-\hat{\theta})^\top \nabla^2 f(\hat{\theta})(\mu-\hat{\theta})]] + \tfrac{1}{2}\left(\mathrm{Tr}\left\{\nabla^2 f(\hat{\theta})\Sigma\right\} + \log|\Sigma|\right), \quad (15)$$

where $\mathrm{Tr}(\cdot)$ is the Trace operator. In the coordinate update of $q(\theta)$, this is the function we optimize with respect to its variational parameters $\{\mu,\Sigma\}$.

To fully specify the algorithm we must choose $\hat{\theta}$, the point around which to approximate $f(\theta)$. We will discuss three choices. The first is to set $\hat{\theta}$ to be the maximum of $f(\theta)$. With this choice, maximizing the approximation in Equation 15 gives $\mu = \hat{\theta}$ and $\Sigma = -\nabla^2 f(\hat{\theta})^{-1}$. Notice this is the update derived in Section 3.1. We have given a different derivation of Laplace variational inference.

The second choice is to set $\hat{\theta}$ as the mean of the variational distribution from the previous iteration of coordinate ascent. If the prior $p(\theta)$ is Gaussian, this recovers the updates derived in Ahmed and Xing (2007) for the correlated topic model.[4] In our study, we found this algorithm did not work well. It did not always converge, possibly due to the difficulty of choosing an appropriate initial $\hat{\theta}$.

The third choice is to set $\hat{\theta} = \mu$, that is, the mean of the variational distribution $q(\theta)$. With this choice, the variable around which we center the Taylor approximation becomes part of the optimization problem. The objective is

$$\mathcal{L}(q(\theta)) \approx f(\mu) + \tfrac{1}{2}\mathrm{Tr}\left\{\nabla^2 f(\mu)\Sigma\right\} + \tfrac{1}{2}\log|\Sigma|. \quad (16)$$

[4] This is an alternative derivation of their algorithm. They derived these updates from the perspective of generalized mean-field theory (Xing et al., 2003).

This is the multivariate delta method for evaluating $\mathbb{E}_{q(\theta)}[f(\theta)]$ (Bickel and Doksum, 2007). *Delta method variational inference* optimizes this objective in the coordinate update of $q(\theta)$.

In more detail, we first optimize μ with gradient methods and then optimize Σ in closed form $\Sigma = -\nabla^2 f(\mu)^{-1}$. Note this is more expensive than Laplace variational inference because optimizing Equation 16 requires the third derivative $\nabla^3 f(\theta)$. Braun and McAuliffe (2010) were the first to use the delta method in a variational inference algorithm, developing this technique for the discrete choice model. If we assume the prior $p(\theta)$ is Gaussian then we recover their algorithm. With the ideas presented here, we can now use this strategy in many models.

We return briefly to the unigram language model. The delta method update for $q(\theta)$ optimizes Equation 16, using the specific $f(\cdot)$ found in Equation 14. While Laplace inference required the digamma function and $\log\Gamma$ function, delta method inference will further require the trigamma function.

3.3 Updating the Conjugate Variable

We derived variational updates for $q(\theta)$ using two methods. We now turn to the update for the variational distribution of the conjugate variable $q(z)$. We show that both Laplace inference (Section 3.1) and delta method inference (Section 3.2) lead to the same update. Further, we have implicitly assumed that $\mathbb{E}_{q(z)}[t(z)]$ in Equation 11 is easy to compute. We will confirm this as well.

We first derive the update for $q(z)$ when using Laplace inference. We apply the exponential family form in Equation 6 to the exact update of Equation 5,

$$\log q(z) = \log p(x\,|\,z) + \log h(z) + \mathbb{E}_{q(\theta)}[\eta(\theta)]^\top t(z) + C,$$

where C is a constant not depending on z. Now we use $p(x\,|\,z)$ from Equation 7 to obtain

$$q(z) \propto h(z)\exp\left\{\left(\mathbb{E}_{q(\theta)}[\eta(\theta)] + t(x)\right)^\top t(z)\right\}, \tag{17}$$

which is in the same family as $p(z\,|\,\theta)$ in Equation 6. This is the update for $q(z)$.

Recall that $\eta(\theta)$ maps the nonconjugate variable θ to the natural parameter of the conjugate variable z. The update for $q(z)$ requires computing $\mathbb{E}_{q(\theta)}[\eta(\theta)]$. For some models, this expectation is computable. If not, we can take a Taylor approximation of $\eta(\theta)$ around the variational parameter μ,

$$\eta_i(\theta) \approx \eta_i(\mu) + \nabla\eta(\mu)_i^\top(\theta-\mu) + \tfrac{1}{2}(\theta-\mu)^\top\nabla^2\eta_i(\mu)(\theta-\mu),$$

where $\eta(\theta)$ is a vector and i indexes the ith component. This requires $\eta(\theta)$ is twice differentiable, which is satisfied in most models. Since $q(\theta) = \mathcal{N}(\mu,\Sigma)$, this means that

$$\mathbb{E}_{q(\theta)}[\eta_i(\theta)] \approx \eta_i(\mu) + \tfrac{1}{2}\mathrm{Tr}\left\{\nabla^2\eta_i(\mu)\Sigma\right\}. \tag{18}$$

(Note that the linear term $\mathbb{E}_{q(\theta)}\left[\nabla\eta_i(\mu)^T(\theta-\mu)\right] = 0$.)

Using delta method variational inference to update $q(\theta)$, the update for $q(z)$ is identical to that in Laplace variational inference. We isolate the relevant terms in Equation 3,

$$\mathcal{L}(q(z)) = \mathbb{E}_{q(z)}\left[\log p(x\,|\,z) + \log h(z) + \mathbb{E}_{q(\theta)}[\eta(\theta)]^\top t(z)\right] - \mathbb{E}_{q(z)}[\log q(z)].$$

```
Initialize variational distributions $q(\theta \mid \mu, \Sigma)$ and $q(z \mid \phi)$.
repeat
    Compute the statistics $\mathbb{E}_{q(z)}[t(z)]$.
    Update $q(\theta)$ using one of the methods:
            a) For Laplace inference, compute Equation 13.
            b) For delta method inference, optimize Equation 16.
    Compute $\mathbb{E}_{q(\theta)}[\eta(\theta)]$ or approximate it as in Equation 18.
    Update $q(z)$ from Equation 17.
until convergence. (See Section 3.4 for the criterion.)
return $q(\theta)$ and $q(z)$.
```

Figure 1: Nonconjugate variational inference

Setting the partial gradient $\partial \mathcal{L}(q(z))/\partial q(z) = 0$ gives the same optimal $q(z)$ of Equation 5. Computing this update reduces to the approach for Laplace variational inference in Equation 17.

We return again to the unigram language model with log normal priors on the Dirichlet parameters. In this model, we can compute $\mathbb{E}_{q(\theta)}[\eta(\theta)]$ exactly by using properties of the log normal,

$$\mathbb{E}_{q(\theta)}[\eta(\theta)] = \mathbb{E}_{q(\theta)}[\exp\{\theta\}] = \exp\{\mu + \mathrm{diag}(\Sigma)/2\}.$$

Recall that x_d are the word counts for document d and note that it is its own sufficient statistic in a multinomial count model. Given the calculation of $\mathbb{E}_{q(\theta)}[\eta(\theta)]$ and the model-specific calculations in Equation 8, the update for $q(z_d)$ is

$$q(z_d) = \mathrm{Dirichlet}\left(\exp(\mu + \mathrm{diag}(\Sigma)/2) + x_d\right).$$

This completes our derivation in the example model. To implement nonconjugate inference we need this update for $q(z)$ and the definition of $f(\cdot)$ in Equation 14.

3.4 Nonconjugate Variational Inference

We now present the full algorithm for nonconjugate variational inference. In this section, we will be explicit about the variational parameters. Recall that the variational distribution of the nonconjugate variable is a Gaussian $q(\theta \mid \mu, \Sigma)$; the variational distribution of the conjugate variable is $q(z \mid \phi)$, where ϕ is a natural parameter in the same family as $p(z \mid \eta(\theta))$.

The algorithm is as follows. Begin by initializing the variational parameters. Iterate between updating $q(\theta)$ and updating $q(z)$ until convergence. Update $q(\theta)$ by either Equation 13 (Laplace inference) or optimizing Equation 16 (Delta method inference). Update $q(z)$ from Equation 17. Assess convergence by measuring the L_2 norm of the mean of the nonconjugate variable, $\mathbb{E}_q[\theta]$.

This algorithm is summarized in Figure 1. In either Laplace or delta method inference, we have reduced deriving variational updates for complicated nonconjugate models to mechanical work—calculating derivatives and calling a numerical optimization library. We note that Laplace inference is simpler to derive because it only requires second derivatives of the function in Equation 11;

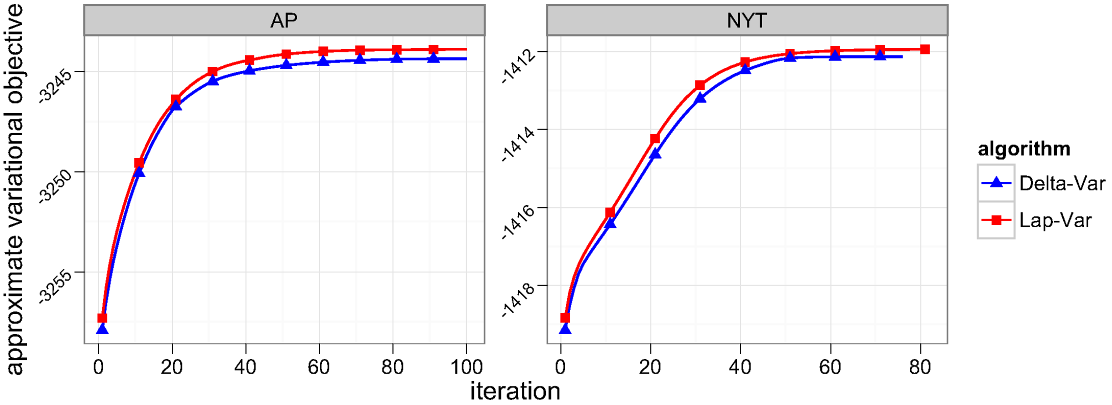

Figure 2: The approximate variational objective from Equation 19 goes up as a function of the iteration. This is for document-level inference in the correlated topic model. The left plot is for a collection from the *Associated Press*; the right plot is for a collection from the *New York Times*. (See Section 4.1 and Section 5.1 for details about the model and data.)

delta method inference requires third derivatives. We study the empirical difference between these methods in Section 5.

Our algorithm (in either setting) is based on approximately optimal coordinate updates for the variational objective, but we cannot compute that objective. However, we can compute an approximate objective at each iteration with the same Taylor approximation used in the coordinate steps, and this can be monitored as a proxy. The approximate objective is

$$\mathcal{L} \approx f(\hat\theta) + \nabla f(\hat\theta)^\top(\mu - \hat\theta) + \tfrac{1}{2}(\mu - \hat\theta)^\top \nabla^2 f(\hat\theta)(\mu - \hat\theta) \\ + \tfrac{1}{2}\left(\mathrm{Tr}\left\{\nabla^2 f(\hat\theta)\Sigma\right\} + \log|\Sigma|\right) - \mathbb{E}_{q(z)}\left[\log q(z)\right] \tag{19}$$

where $f(\theta)$ is defined in Equation 11 and $\hat\theta$ is defined as for Laplace or delta method inference.[5]

Figure 2 shows this score at each iteration for two runs of inference in the correlated topic model. (See Section 4.1 for details about the model.) The approximate objective increases as the algorithm proceeds, and these plots were typical. In practice, as did Braun and McAuliffe (2010) in their setting, we found that this is a good score to monitor.

4 Example Models

We have described a generic algorithm for approximate posterior inference in nonconjugate models. In this section we derive this algorithm for several nonconjugate models from the research literature: the correlated topic model (Blei and Lafferty, 2007), Bayesian logistic regression (Jaakkola and Jordan, 1997), and hierarchical Bayesian logistic regression (Gelman and Hill, 2007). For each

[5] We note again that Equation 19 is not the function we are optimizing. Even the simpler Laplace approximation is not clearly minimizing a well-defined distance function between the approximate Gaussian and true posterior (MacKay, 1992). Thus, while this approach is an approximate coordinate ascent algorithm, clearly characterizing the corresponding objective function is an open problem.

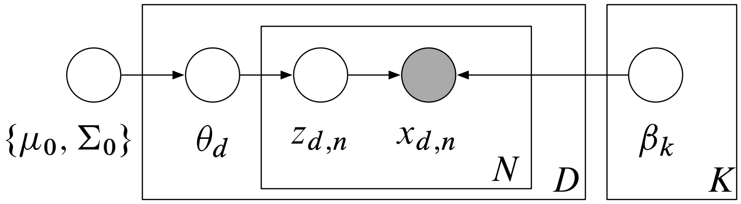

Figure 3: The graphical representation of the correlated topic model (CTM). The nonconjugate variable is θ; the conjugate variable is the collection $z = z_{1:N}$; the observation is the collection of words $x = x_{1:N}$.

model, we identify the variables—the nonconjugate variable θ, conjugate variable z, and observations x—and we calculate $f(\theta)$ from Equation 11. (The calculations of $f(\theta)$ are in the appendices.) In the next section, we study how our algorithms perform when analyzing data under these models.[6]

4.1 The Correlated Topic Model

Probabilistic topic models are models of document collections. Each document is treated as a group of observed words that are drawn from a mixture model. The mixture components, called "topics," are distributions over terms that are shared for the whole collection; each document exhibits them with individualized proportions.

Conditioned on a corpus of documents, the posterior topics place high probabilities on words that are associated under a single theme; for example, one topic may contain words like "bat," "ball," and "pitcher." The posterior topic proportions reflect how each document exhibits those themes; for example, a document may combine the topics of *sports* and *health*. This posterior decomposition of a collection can be used for summarization, visualization, or forming predictions about a document. See Blei (2012) for a review of topic modeling.

The per-document topic proportions are a latent variable. In latent Dirichlet allocation (LDA) (Blei et al., 2003)—which is the simplest topic model—these are given a Dirichlet prior, which makes the model conditionally conjugate. Here we will study the correlated topic model (CTM) (Blei and Lafferty, 2007). The CTM extends LDA by replacing the Dirichlet prior on the topic proportions with a logistic normal prior (Aitchison, 1982). This is a richer prior that can capture correlations between occurrences of the components. For example, a document about *sports* is more likely to also be about *health*. The CTM is not conditionally conjugate. But it is a more expressive model: it gives a better fit to texts and provides new kinds of exploratory structure.

Suppose there are K topic parameters $\beta_{1:K}$, each of which is a distribution over V terms. Let $\pi(\theta)$ denote the multinomial logistic function, which maps a real-valued vector to a point on the simplex with the same dimension, $\pi(\theta) \propto \exp\{\theta\}$. The CTM assumes a document is drawn as follows:

1. Draw log topic proportions $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$.
2. For each word n:

    (a) Draw topic assignment $z_n \,|\, \theta \sim \text{Mult}(\pi(\theta))$.

[6] Python implementations of our algorithms are available at `http://www.cs.cmu.edu/~chongw/software/nonconjugate_inference.tar.gz`.

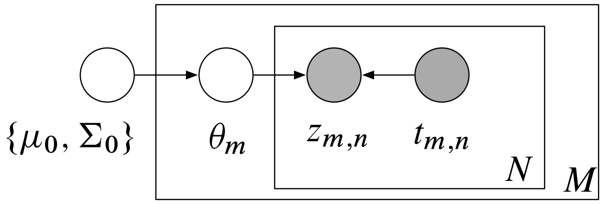

Figure 4: The graphical representation of hierarchical logistic regression. (When $M = 1$, this is standard Bayesian logistic regression.)The nonconjugate variable is the vector of coefficients θ_m, the conjugate variable is the collection of observed classes for each data point, $z_m = z_{m,1:N}$. (In this case there is no additional observation x downstream.)

(b) Draw word $x_n \,|\, z_n, \beta \sim \text{Mult}(\beta_{z_n})$.

Figure 3 shows the graphical model. The topic proportions $\pi(\theta)$ are drawn from a logistic normal distribution; their correlation structure is captured in its covariance matrix Σ_0. The topic assignment variable z_n indicates from which topic the nth word is drawn.

Holding the topics $\beta_{1:K}$ fixed, the main inference problem in the CTM is to infer the conditional distribution of the document-level hidden variables $p(\theta, z_{1:N} \,|\, x_{1:N}, \beta_{1:K})$. This calculation is important in two contexts: it is used when forming predictions about new data; and it is used as a subroutine in the variational expectation maximization algorithm for fitting the topics and logistic normal parameters (mean μ_0 and covariance Σ_0) with maximum likelihood. The corresponding per-document inference problem is straightforward to solve in LDA, thanks to conditional conjugacy. In the CTM, however, it is difficult because the logistic normal on θ is not conjugate to the multinomial on z. Blei and Lafferty (2007) used a Taylor approximation designed specifically for this model. Here we apply the generic algorithm from Section 3.

In terms of the earlier notation, the nonconjugate variable is the topic proportions θ, the conjugate variable is the collection of topic assignments $z = z_{1:N}$, and the observation is the collection of words $x = x_{1:N}$. The variational distribution for the topic proportions θ is Gaussian, $q(\theta) = \mathcal{N}(\mu, \Sigma)$; the variational distribution for the topic assignments is discrete, $q(z) = \prod_n q(z_n \,|\, \phi_n)$ where each ϕ_n is a distribution over K elements. In delta method inference, as in Braun and McAuliffe (2010), we restrict the variational covariance Σ to be diagonal to simplify the derivative of Equation 16. Laplace variational inference does not require this simplification. Appendix B gives the detailed derivations of the algorithm.

Besides the CTM, this approach can be adapted to a variety of nonconjugate topic models, including the topic evolution model (Xing, 2005), Dirichlet-multinomial regression (Mimno and McCallum, 2008), dynamic topic models (Blei and Lafferty, 2006; Wang et al., 2008), and the discrete infinite logistic normal distribution (Paisley et al., 2012b).

4.2 Bayesian Logistic Regression

Bayesian logistic regression is a well-studied model for binary classification (Jaakkola and Jordan, 1997). It places a Gaussian prior on a set of coefficients and draws class labels, conditioned on covariates, from the corresponding logistic. Let t_n is be a p-dimensional observed covariate vector for the nth sample and z_n be its class label (an indicator vector of length two). Let θ be the real-valued coefficients in $\mathbb{R}^p$; there is a coefficient for each feature. Bayesian logistic regression assumes the following conditional process:

1. Draw coefficients $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$.
2. For each data point n and its covariates t_n, draw its class label from

$$z_n \,|\, \theta, t_n \sim \text{Bernoulli}\left(\sigma(\theta^\top t_n)^{z_{n,1}} \sigma(-\theta^\top t_n)^{z_{n,2}}\right),$$

where $\sigma(y) \triangleq 1/(1+\exp(-y))$ is the logistic function.

Figure 4 shows the graphical model. Given a data set of labeled feature vectors, the posterior inference problem is to compute the conditional distribution of the coefficients $p(\theta \,|\, z_{1:N}, t_{1:N})$. The issue is that the Gaussian prior on the coefficients is not conjugate to the conditional likelihood of the label.

This is a subset of the model class in Section 2.2. The nonconjugate variable θ is identical and the variable z is the collection of observed classes of each data point, $z_{1:N}$. Note there is no additional observed variable x downstream. The variational distribution need only be defined for the coefficients, $q(\theta) = \mathcal{N}(\mu, \Sigma)$. Using Laplace variational inference, our approach recovers the standard Laplace approximation for Bayesian logistic regression (Bishop, 2006). This gives a connection between standard Laplace approximation and variational inference. Delta method variational inference provides an alternative. Appendix C gives the detailed derivations.

An important extension of Bayesian logistic regression is hierarchical Bayesian logistic regression (Gelman and Hill, 2007). It simultaneously models related logistic regression problems, and estimates the hyperparameters of the shared prior on the coefficients. With M related problems, we construct the following hierarchical model:

1. Draw the global hyperparameters,

$$\Sigma_0^{-1} \sim \text{Wishart}(\nu, \Phi_0), \tag{20}$$

$$\mu_0 \sim \mathcal{N}(0, \Phi_1). \tag{21}$$

2. For each problem m:
   (a) Draw coefficients $\theta_m \sim \mathcal{N}(\mu_0, \Sigma_0)$.
   (b) For each data point n and its covariates t_{mn}, draw its class label,

$$z_{mn} \,|\, \theta_m, t_{mn} \sim \text{Bernoulli}(\sigma(\theta_m^\top t_{mn})^{z_{mn,1}} \sigma(-\theta_m^\top t_{mn})^{z_{mn,2}}).$$

As for the CTM, we use nonconjugate inference as a subroutine in a variational EM algorithm (where the M step is regularized). We construct $f(\theta_m)$ in Equation 11 separately for each problem m, and fit the hyperparameters μ_0 and Σ_0 from their approximate expected sufficient statistics (Bishop, 2006). This amounts to MAP estimation with priors as specified above. See Appendix C for the complete derivation.

Finally, we note that logistic regression is a generalized linear model with a binary response and canonical link function (McCullagh and Nelder, 1989). It is straightforward to use our algorithms with other Bayesian generalized linear models (and their hierarchical forms).

5 Empirical Study

We studied nonconjugate variational inference with correlated topic models and Bayesian logistic regression. We found that nonconjugate inference is more accurate than the existing methods tailored to specific models. Between the two nonconjugate inference algorithms, we found that Laplace inference is faster and more accurate than delta method inference.

5.1 The Correlated Topic Model

We studied Laplace inference and delta method inference in the CTM. We compared it to the original inference algorithm of Blei and Lafferty (2007).

We analyzed two collections of documents. The *Associated Press* (AP) collection contains 2,246 documents from the *Associated Press*. We used a vocabulary of 10,473 terms, which gave a total of 436K observed words. The *New York Times* (NYT) collection contains 9,238 documents from the *New York Times*. We used a vocabulary of 10,760 terms, which gave a total of 2.3 million observed words. For each corpus we used 80% of the documents to fit models and reserved 20% to test them.

We fitted the models with variational EM. At each iteration, the algorithm has a set of topics $\beta_{1:K}$ and parameters to the logistic normal $\{\mu_0, \Sigma_0\}$. In the E-step we perform approximate posterior inference with each document, estimating its topic proportions and topic assignments. In the M-step, we re-estimate the topics and logistic normal parameters. We fit models with different kinds of E-steps, using both of the nonconjugate inference methods from Section 3 and the original approach of Blei and Lafferty (2007). To initialize nonconjugate inference we set the variational mean parameter $\mu = 0$ for log topic proportions θ and computed the corresponding updates for the topic assignments z. We initialize the topics in variational EM to random draws from a uniform Dirichlet.

With nonconjugate inference in the E-step, variational EM approximately optimizes a bound on the marginal probability of the observed data. We can calculate an approximation of this bound with Equation 19 summed over all documents. We monitor this quantity as we run variational EM.

To test our fitted models, we measured predictive performance on held-out data with predictive distributions derived from the posterior approximations. We follow the testing framework of Asuncion et al. (2009) and Blei and Lafferty (2007). We fix fitted topics and logistic normal parameters $M = \{\beta_{1:K}, \mu_0, \Sigma_0\}$. We split each held-out document in to two halves $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ and form the approximate posterior log topic proportions $q_{\boldsymbol{w}_1}(\theta)$ using one of the approximate inference algorithms and the first half of the document $\boldsymbol{w}_1$. We use this to form an approximate predictive distribution,

$$p(w \,|\, \boldsymbol{w}_1, M) \approx \int_\theta \sum_z p(w \,|\, z, \beta_{1:K}) q_{\boldsymbol{w}_1}(\theta) d\theta \approx \sum_{k=1}^K \beta_{kw} \pi_k,$$

where $\pi_k \propto \exp\{\mathbb{E}_q[\theta_k]\}$. Finally, we evaluate the log probability of the second half of the document using that predictive distribution; this is the *held out log likelihood*. A better model and inference method will give higher predictive probabilities of the unseen words. Note that this testing framework puts the approximate posterior distributions on the same playing field. The quantities are comparable regardless of how the approximate posterior is formed.

Figure 5 shows the per-word approximate bound and the per-word held out likelihood as functions of the number of topics. Figure 5 (a) indicates that the approximate bounds from nonconjugate inference generally go up as the number of topics increases. This is a property of a good approximation because the marginal certainly goes up as the number of parameters increases. In contrast, Blei and Lafferty's (2007) objective (which is a true bound on the marginal of the data) behaves erratically. This is illustrated for the *New York Times* corpus; on the *Associated Press* corpus, it does not come close to the approximate bound and is not plotted.

Figure 5 (b) shows that on held out data, Blei and Lafferty's approach, tailored for this model, performed worse than both of our algorithms. Our conjecture is that while this method gives a strict lower bound on the marginal, it might be a loose bound and give poor predictive distributions. Our methods use an approximation which, while not a bound, might be closer to the objective and give better predictive distributions. The held out likelihood plots also show that when the number of topics increases the algorithms eventually overfit the data. Finally, note that Laplace variational inference was always better than both other algorithms.

Finally, Figure 6 shows the approximate bound and the held out log likelihood as functions of running time.[7] From Figure 6 (a), we see that even though variational EM is not formally optimizing this approximate objective (see Equation 19), the increase at each iteration suggests that the marginal probability is also increasing. The plot also shows that Laplace inference converges faster than delta method inference. Figure 6 (b) confirms that Laplace inference is both faster and gives better predictive performance.

5.2 Bayesian Logistic Regression

We studied our algorithms on Bayesian logistic regression in both standard and hierarchical settings. In the standard setting, we analyzed two data sets. With the *Yeast* data (Elisseeff and Weston, 2001), we form a predictor of gene functional classes from features composed of micro-array expression data and phylogenetic profiles. The data set has 1,500 genes in the training set and 917 genes in the test set. For each gene there are 103 covariates and up to 14 different gene functional classes (14 labels). This corresponds to 14 independent binary classification problems. With the *Scene* data (Boutell et al., 2004), we form a predictor of scene labels from image features. It contains 1,211 images in the training set and 1,196 images in the test set. There are 294 images features and up to 6 scene labels per image. This corresponds to 6 independent binary classification problems.[8]

We used two performance measures. First we measured accuracy, which is the proportion of test-case examples correctly labeled. Second, we measured average log predictive likelihood. Given a test-case input t with label z, we compute the log predictive likelihood,

$$\log p(z \mid \mu, t) = z_1 \log \sigma(\mu^\top t) + z_2 \log \sigma(-\mu^\top t),$$

where μ is the mean of variational distribution $q(\theta) = \mathcal{N}(\mu, \Sigma)$. Higher likelihoods indicate a better fit. For both accuracy and predictive likelihood, we used cross validation to estimate the generalization performance of each inference algorithm. We set the priors $\mu_0 = 0$ and $\Sigma_0 = I$.

We compared Laplace inference (Section 3.1), delta method inference (Section 3.2), and the method of Jaakkola and Jordan (1997). Jaakkola and Jordan's (1996) method preserves a lower bound on the marginal likelihood with a first-order Taylor approximation and was developed specifically for Bayesian logistic regression. (We note that Blei and Lafferty's bound-preserving method for the CTM was built on this technique.)

[7] We did not formally compare the running time of Blei and Lafferty's (2007) method because we used the authors' C implementation, while ours is in Python. We observed that their method took more than five times longer than ours.

[8] The *Yeast* and *Scene* data are at `http://mulan.sourceforge.net/datasets.html`.

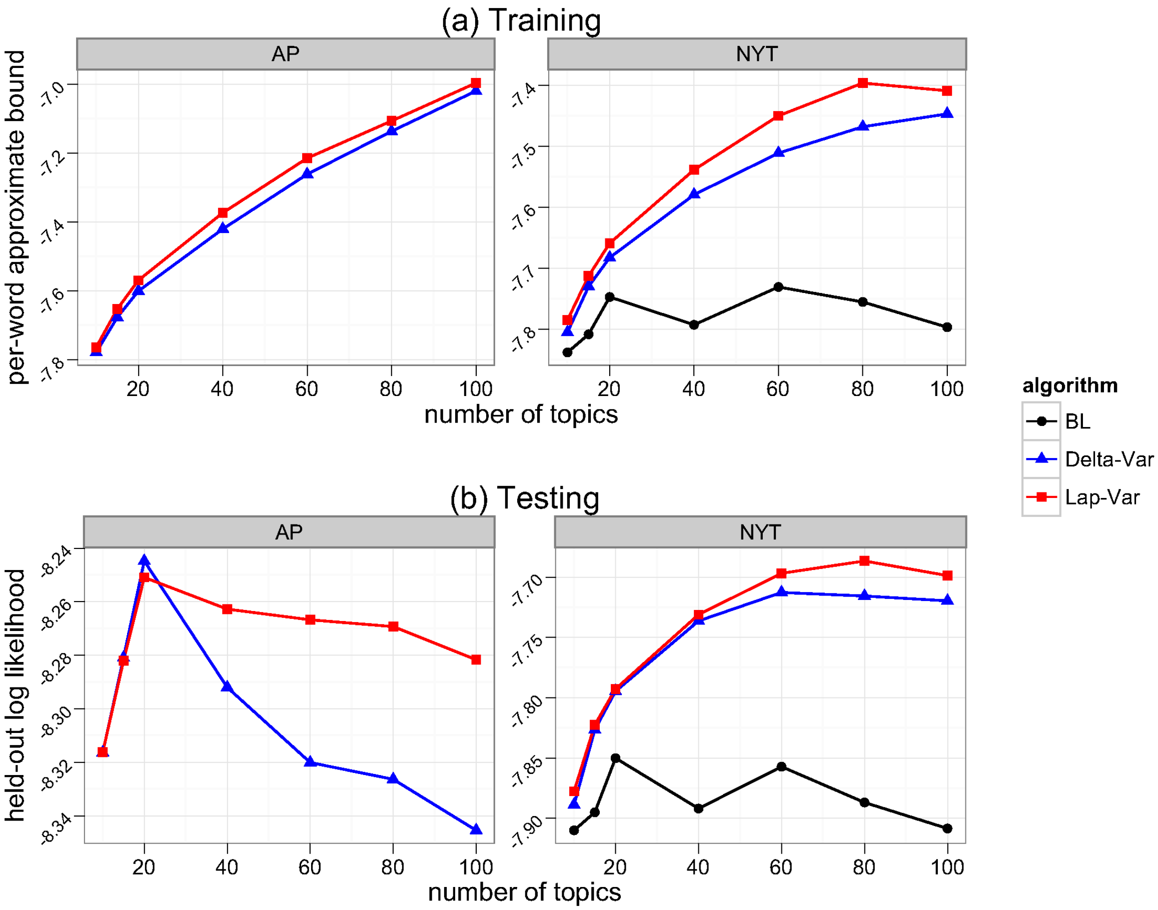

Figure 5: Laplace variational inference is "Lap-Var"; delta method variational inference is "Delta-Var"; Blei and Lafferty's method is "BL." (a) Approximate per-word lower bound against the number of topics. A good approximation will go up as the number of topics increases, but not necessarily indicate a better predictive performance on the held out data. (b) Per-word held-out log likelihood against the number of topics. Higher numbers are better. Both nonconjugate methods perform better than Blei and Lafferty's method. Laplace inference performs best. Blei and Lafferty's method was erratic in both collections. (It is not plotted for the AP collection.)

Table 1 gives the results. To compare methods we compute the difference in score (accuracy or log likelihood) on the independent binary classification problems, and then perform a standard t-test (at level 0.05) to test if the mean of the differences is larger than 0. Laplace inference and delta method inference gave slightly better accuracy than Jaakkola and Jordan's method, and much

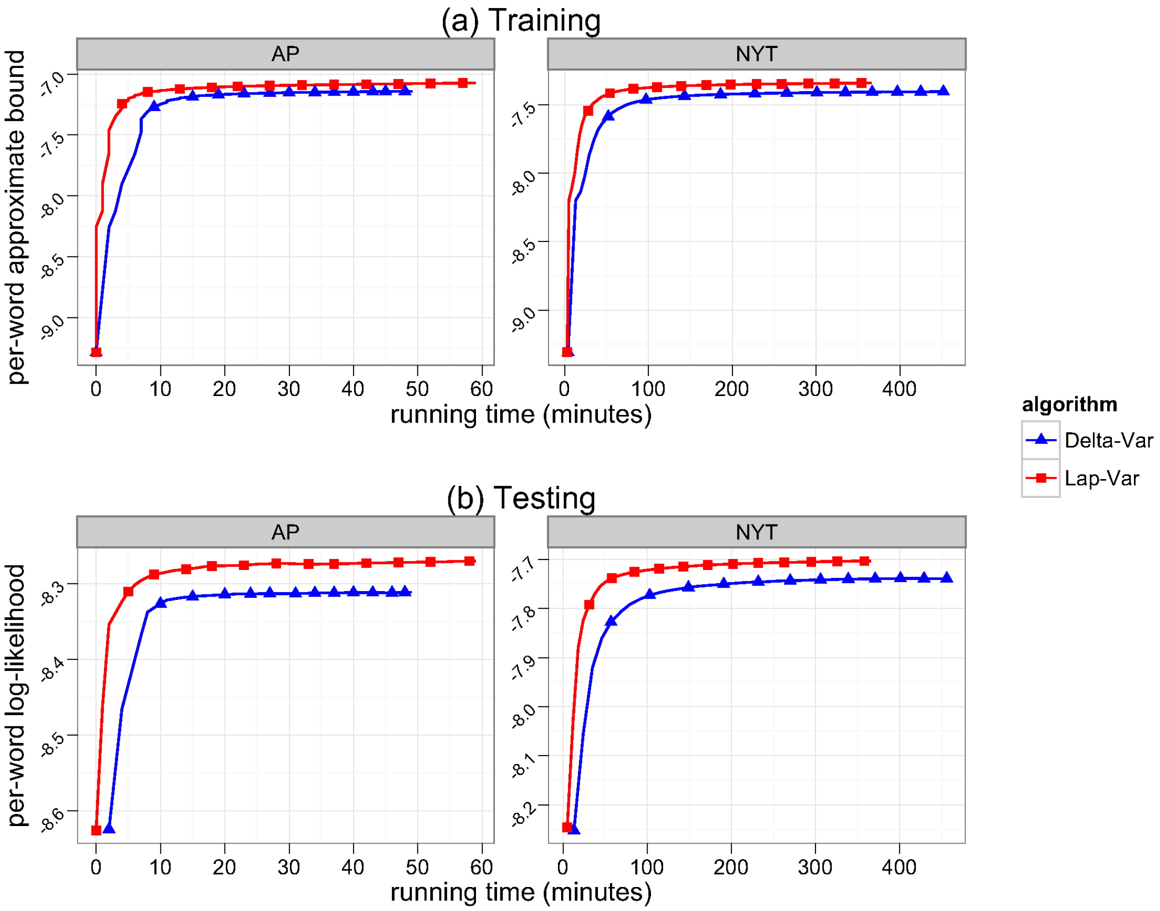

Figure 6: In this figure, we set the number of topics as $K = 60$. (Others are similar.) (a) The per-word approximate bound during model fitting with variational EM. Though it is an approximation of the variational EM objective, it converges in practice. (b) The per-word held out likelihood during the model fitting with variational EM. Laplace inference performs best in terms of speed and predictive performance.

better log predictive likelihood.[9] The t-test showed that both Laplace and delta method inference are better than Jaakkola and Jordan's method.

We next examined a data set of student performance in a collection of schools. With the *School* data, our goal is to use various features of a student to predict whether he or she will perform above or below the median on a standardized exam.[10] The data came from the Inner London Education Authority. It contains examination records from 139 secondary schools for the years 1985, 1986 and 1987. It is a random 50% sample with 15,362 students. The students' features contain four student-dependent features and school-dependent features. The student dependent features are

[9] Previous literature, for example, Xue et al. (2007) and Archambeau et al. (2011) treat *Yeast* and *Scene* as multi-task problems. In our study, we found that our standard Bayesian logistic regression algorithms performed the same as the algorithms developed in these papers.

[10] The data is available at `http://multilevel.ioe.ac.uk/intro/datasets.html`.

	Yeast		*Scene*	
	Accuracy	Log Likelihood	Accuracy	Log Likelihood
Jaakkola and Jordan (1996)	79.7%	-0.678	87.4%	-0.670
Laplace inference	**80.1%**	**-0.449**	**89.4%**	**-0.259**
Delta method inference	**80.2%**	**-0.450**	**89.5%**	-0.265

Table 1: Comparison of the different methods for Bayesian logistic regression using accuracy and averaged log predictive likelihood. Higher numbers are better. These results are averaged from five random starts. (The variance is too small to report.) Bold results indicate significantly better performance using a standard t-test. Laplace and delta method inference perform best.

the year of the exam, gender, VR band (individual prior attainment data), and ethnic group; the school-dependent features are the percentage of students eligible for free school meals, percentage of students in VR band 1, school gender, and school denomination. We coded the binary indicator of whether each was below the median ("bad") or above ("good"). We use the same 10 random splits of the data as Argyriou et al. (2008).

In this data, we can either treat each school as a separate classification problem, pool all the schools together as a single classification problem, or analyze them with hierarchical logistic regression (Section 4.2). The hierarchical model allows the predictors for each school to deviate from each other, but shares statistical strength across them. Let p be the number of covariates. We set the prior on the hyperparameters to the coefficients to $\nu = p + 100$, $\Phi_0 = 0.01I$, and $\Phi_1 = 0.01I$ (see Equation 20 and Equation 21) to favor sparsity. We initialized the variational distributions to $q(\theta) = \mathcal{N}(0, I)$.

Table 2 gives the results. A standard t-test (at level 0.05) showed that the hierarchical models are better than the non-hierarchical models both in terms of accuracy and predictive likelihood. With predictive likelihood, Laplace variational inference in the hierarchical model is significantly better than all other approaches.

6 Discussion

We developed Laplace and delta method variational inference, two strategies for variational inference in a large class of nonconjugate models. These methods approximate the variational objective function with a Taylor approximation, each in a different way. We studied them in two nonconjugate models and showed that they work well in practice, forming approximate posteriors that lead to good predictions. In the examples we analyzed, our methods worked better than methods tailored for the specific models at hand. Between the two, Laplace inference was better and faster than delta method inference. These methods expand the scope of variational inference.

		Accuracy	Log Likelihood
Separate			
	Jaakkola and Jordan (1996)	70.5%	-0.684
	Laplace inference	70.8%	-0.569
	Delta inference	70.8%	-0.571
Pooled			
	Jaakkola and Jordan (1996)	71.2%	-0.685
	Laplace inference	71.3%	-0.557
	Delta inference	71.3%	-0.557
Hierarchical			
	Jaakkola and Jordan (1996)	71.3%	-0.685
	Laplace inference	**71.9%**	**-0.549**
	Delta inference	**71.9%**	-0.559

Table 2: Comparison of the different methods on the *School* data using accuracy and averaged log predictive likelihood. Results are averaged from 10 random splits. (The variance is too small to report.) We compared Laplace inference, delta inference and Jaakkola and Jordan's (1996) method in three settings: separate logistic regression models for each school, a pooled logistic regression model for all schools, and the hierarchical logistic regression model in Section 4.2. Bold indicates significantly better performance by a standard t-test (at level 0.05). The hierarchical model performs best.

Acknowledgments

We thank Jon McAuliffe and the anonymous reviewers for their valuable comments. Chong Wang was supported by Google Ph.D. and Siebel Scholar Fellowships. David M. Blei is supported by NSF IIS-0745520, NSF IIS-1247664, NSF IIS-1009542, ONR N00014-11-1-0651, and the Alfred P. Sloan foundation.

Appendix A. Generalization to Complex Models

We describe how we can generalize our approaches to more complex models. Suppose we have a directed probabilistic model with latent variables $\theta = \theta_{1:m}$ and observations x. (We will not differentiate notation between conjugate and nonconjugate variables.) The log joint likelihood of all latent and observed variables is

$$\log p(\theta, x) = \textstyle\sum_{i=1}^{m} \log p(\theta_i \,|\, \theta_{\pi_i}) + \log p(x \,|\, \theta),$$

where π_i are the indices of the parents of θ_i, the variables it depends on.

Our goal is to approximate the posterior distribution $p(\theta \,|\, x)$. Similar to the main paper, we use mean-field variational inference (Jordan et al., 1999). We posit a fully-factorized variational family

$$q(\theta) = \textstyle\prod_{i=1}^{m} q(\theta_i),$$

and optimize ach factor $q(\theta_i)$ to find the member closest in KL-divergence to the posterior.

As in the main paper, we solve this optimization problem with coordinate ascent, iteratively optimizing each variational factor while holding the others fixed. Recall that Bishop (2006) shows that this leads to the following update

$$q(\theta_i) \propto \exp\left\{\mathrm{E}_{-i}\left[\log p(\theta, x)\right]\right\}, \tag{22}$$

where $\mathrm{E}_{-i}[\cdot]$ denotes the expectation with respect to $\prod_{j, j \neq i} q(\theta_j)$.

Many of the terms of the log joint will be constant with respect to θ_i and absorbed into the constant of proportionality. This allows us to simplify the update in Equation 22 to be $q(\theta_i) \propto \exp\{f(\theta_i)\}$ where

$$f(\theta_i) = \mathrm{E}_{-i}\left[\log p(\theta_i \,|\, \theta_{\pi_i})\right] + \textstyle\sum_{\{j : i \in \pi_j\}} \mathrm{E}_{-i}\left[\log p(\theta_j \,|\, \theta_{\pi_j})\right] + \mathrm{E}_{-i}\left[\log p(x \,|\, \theta)\right]. \tag{23}$$

As in the main paper, this update is not tractable in general. We use Laplace variational inference (Section 3.1) to approximate it, although delta method variational inference (Section 3.2) is also applicable. In Laplace variational inference, we take a Taylor approximation of $f(\theta_i)$ around its maximum $\hat{\theta}_i$. This naturally leads to $q(\theta_i)$ as a Gaussian factor,

$$q^*(\theta_i) \approx \mathcal{N}(\hat{\theta}_i, -\nabla^2 f(\hat{\theta}_i)^{-1}).$$

The main paper considers the case where θ is a single random variable and updates its variational distribution. In the more general coordinate ascent setting considered here, we need to compute or approximate the expected log probabilities (and their derivatives) in Equation 23.

Now suppose each factor is in the exponential family. (This is weaker than the conjugacy assumption, and describes most graphical models from the literature.) The log joint likelihood becomes

$$\log p(\theta, x) = \textstyle\sum_{i=1}^{m} \left(\eta(\theta_{\pi_i})^\top t(\theta_i) - a(\eta(\theta_{\pi_i}))\right) + \log p(x \,|\, \theta),$$

where $\eta(\cdot)$ are natural parameters, $t(\cdot)$ are sufficient statistics, and $a(\eta(\cdot))$ are log normalizers. (All are overloaded.) Substituting the exponential family assumptions into $f(\theta_i)$ gives

$$\begin{aligned} f(\theta_i) = {} & \mathrm{E}_{-i}\left[\eta(\theta_{\pi_i})\right]^\top t(\theta_i) \\ & + \textstyle\sum_{\{j: i \in \pi_j\}} \left(\mathrm{E}_{-i}\left[\eta(\theta_{\pi_j})\right]^\top \mathrm{E}_{-i}\left[t(\theta_j)\right] - \mathrm{E}_{-i}\left[a(\eta(\theta_{\pi_j}))\right]\right) \\ & + \mathrm{E}_{-i}\left[t(\theta)\right]^\top t(x) - \mathrm{E}_{-i}\left[a(\eta(\theta))\right]. \end{aligned}$$

Here we can use further Taylor approximations of the natural parameters $\eta(\cdot)$, sufficient statistics $t(\cdot)$, and log normalizers $a(\cdot)$ in order to easily take their expectations.

Finally, for some variables we may be able to exactly compute $f(\theta_i)$ and form the $q^*(\theta_i)$ without further approximations. (These are conjugate variables for which the complete conditional $p(\theta_i \,|\, \theta_{-i}, x)$ is available in closed form.) These variables were separated out in the main paper; here we note that they can be updated exactly in the coordinate ascent algorithm.

Appendix B. The Correlated Topic Model

The correlated topic model is described in Section 4.1. We identify the quantities from Equation 6 and Equation 7 that we need to compute $f(\theta)$ in Equation 11,

$$\begin{aligned} h(z) = 1, \;\; t(z) &= \textstyle\sum_n z_n, \\ \eta(\theta) &= \theta - \log\left\{\textstyle\sum_k \exp\{\theta_k\}\right\}, \\ a(\eta(\theta)) &= 0. \end{aligned}$$

With this notation,

$$f(\theta) = \eta(\theta)^\top \mathbb{E}_{q(z)}\left[t(z)\right] - \tfrac{1}{2}(\theta - \mu_0)^\top \Sigma_0^{-1}(\theta - \mu_0),$$

where $\mathbb{E}_{q(z)}\left[t(z)\right]$ is the expected word counts of each topic under the variational distribution $q(z)$.

Let $\pi \propto \exp\{\eta(\theta)\}$ be the topic proportions. Using $\partial\pi_i / \partial\theta_j = \pi_i(1_{[i=j]} - \pi_j)$, we obtain the gradient and Hessian of the function $f(\theta)$ in the CTM,

$$\begin{aligned} \nabla f(\theta) &= \mathbb{E}_{q(z)}\left[t(z)\right] - \pi \textstyle\sum_{k=1}^{K} \left[\mathbb{E}_{q(z)}\left[t(z)\right]\right]_k - \Sigma_0^{-1}(\theta - \mu_0), \\ \nabla^2 f(\theta)_{ij} &= (-\pi_i 1_{[i=j]} + \pi_i \pi_j) \textstyle\sum_{k=1}^{K} \left[\mathbb{E}_{q(z)}\left[t(z)\right]\right]_k - (\Sigma_0^{-1})_{ij}. \end{aligned}$$

where $1_{[i=j]} = 1$ if $i = j$ and 0 otherwise. Note that $\nabla f(\theta)$ is all we need for Laplace inference.

In delta method variational inference, we also need to compute the gradient of

$$\mathrm{Trace}\left\{\nabla^2 f(\theta)\Sigma\right\} = \left(-\textstyle\sum_{k=1}^{K} \pi_k \Sigma_{kk} + \pi^T \Sigma \pi\right) \textstyle\sum_{k=1}^{K} \left[\mathbb{E}_{q(z)}\left[t(z)\right]\right]_k - \mathrm{Trace}(\Sigma_0^{-1}\Sigma).$$

Following Braun and McAuliffe (2010), we assume Σ is diagonal in the delta method. (In Laplace inference, we do not need this assumption.) This gives

$$\frac{\partial \mathrm{Trace}\left\{\nabla^2 f(\theta)\Sigma\right\}}{\partial\theta_i} = \pi_i(1 - 2\pi_i)(\textstyle\sum_k \pi_k \Sigma_{kk} - 1).$$

These quantities let us implement the algorithm in Figure 1 to infer the per-document posterior of the CTM hidden variables.

As we discussed Section 4.1, we use this algorithm in variational EM for finding maximum likelihood estimates of the model parameters. The E-step runs posterior inference on each document. Since the variational family is the same, the M-step is as described in Blei and Lafferty (2007).

Appendix C. Bayesian Logistic Regression

Bayesian logistic regression is described in Section 4.2.

The distribution of the observations $z_{1:N}$ fit into the exponential family as follows,

$$\begin{aligned} h(z) &= 1, \;\; t(z) = [z_1, \ldots, z_N], \\ \eta(\theta) &= [\log \sigma(\theta^\top t_n), \log \sigma(-\theta^\top t_n)]_{n=1}^N, \\ a(\eta(\theta)) &= 0. \end{aligned}$$

In this set up, $t(z)$ represents the whole set of labels. Since z is observed, its "expectation" is just itself. With this notation, $f(\theta)$ from Equation 11 is

$$f(\theta) = \eta(\theta)^\top t(z) - \tfrac{1}{2}(\theta - \mu_0)^\top \Sigma_0^{-1}(\theta - \mu_0).$$

The gradient and Hessian of $f(\theta)$ are

$$\begin{aligned} \triangledown f(\theta) &= \textstyle\sum_{n=1}^N t_n \left(z_{n,1} - \sigma(\theta^T t_n)\right) - \Sigma_0^{-1}(\theta - \mu_0), \\ \triangledown^2 f(\theta) &= -\textstyle\sum_{n=1}^N \sigma(\theta^T t_n)\sigma(-\theta^T t_n) t_n t_n^T - \Sigma_0^{-1}. \end{aligned} \tag{24}$$

This is the standard Laplace approximation to Bayesian logistic regression (Bishop, 2006).

For delta variational inference, we also need the gradient for Trace$\left\{\triangledown^2 f(\theta)\Sigma\right\}$. It is

$$\frac{\partial \text{Trace}\left\{\triangledown^2 f(\theta)\Sigma\right\}}{\partial \theta_i} = -\sum_{n=1}^N \sigma(\theta^T t_n)\sigma(-\theta^T t_n)(1 - 2\sigma(\theta^T t_n)) t_n t_n^T \Sigma t_n.$$

Here we do not need to assume Σ is diagonal, since the special structure of the Hessian in Equations 24 makes the computation of Trace$\left\{\triangledown^2 f(\theta)\Sigma\right\}$ fairly simple.

C.1 Hierarchical Logistic Regression

Here we describe how we update the global hyperparameters (μ_0, Σ_0) (Equations 20 and 21) in hierarchical logistic regression. At each iteration, we first compute the variational distribution of coefficients θ_m for each problem $m = 1, ..., M$,

$$q(\theta_m) = \mathcal{N}(\mu_m, \Sigma_m).$$

We then estimate the global hyperparameters (μ_0, Σ_0) using the MAP estimate. These come from the following update equations,

$$\begin{aligned} \mu_0 &= \left(\frac{\Sigma_0 \Phi_1^{-1}}{M} + I_p\right)^{-1} \frac{\sum_{m=1}^M \mu_m}{M}, \\ \Sigma_0 &= \frac{\Phi_0^{-1} + \sum_{m=1}^M (\mu_m - \mu_0)(\mu_m - \mu_0)^\top}{M + \nu - p - 1}, \end{aligned}$$

where p is the dimension of coefficients θ_m.

References

A. Ahmed and E. Xing. On tight approximate inference of the logistic normal topic admixture model. In *Workshop on Artificial Intelligence and Statistics*, 2007.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44(2):139–177, 1982.

C. Archambeau, S. Guo, and O. Zoeter. Sparse Bayesian multi-task learning. In *Advances in Neural Information Processing Systems*, 2011.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Maching Learning*, 73:243–272, December 2008.

A. Asuncion, M. Welling, P. Smyth, and Y. Teh. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence*, 2009.

H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems*, 2000.

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons Ltd., Chichester, 1994.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Pearson Prentice Hall, Upper Saddle River, NJ, 2nd edition, 2007.

C. Bishop. Variational principal components. In *International Conference on Artificial Neural Networks*, volume 1, pages 509–514. IET, 1999.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer New York., 2006.

C. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems*, 2003.

D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

D. Blei and J. Lafferty. Dynamic topic models. In *International Conference on Machine Learning*, 2006.

D. Blei and J. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1): 17–35, 2007.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 2010.

L. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, CA, 1986.

B. Carlin and N. Polson. Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics*, 19(4):399–405, 1991.

J. Clinton, S. Jackman, and D. Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370, 2004.

A. Corduneanu and C. Bishop. Variational Bayesian model selection for mixture distributions. In *International Conference on Artifical Intelligence and Statistics*, 2001.

W. Croft and J. Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems*, 2001.

J. Fox. *Bayesian Item Response Modeling: Theory and Applications*. Springer Verlag, 2010.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

S. Gershman, M. Hoffman, and D. Blei. Nonparametric variational inference. In *International Conference on Machine Learning*, 2012.

Z. Ghahramani and M. Jordan. Factorial hidden Markov models. *Machine Learning*, 31(1), 1997.

A. Honkela and H. Valpola. Unsupervised variational Bayesian learning of nonlinear models. In *Advances in Neural Information Processing Systems*, 2004.

T. Jaakkola and M. Jordan. Bayesian logistic regression: A variational approach. In *Artificial Intelligence and Statistics*, 1997.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

M. Khan, B. Marlin, G. Bouchard, and K. Murphy. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems*, 2010.

D. Knowles and T. Minka. Non-conjugate variational message passing for multinomial and binary regression. In *Neural Information Processing Systems*, 2011.

S. Kotz, N. Balakrishnan, and N. Johnson. *Continuous Multivariate Distributions, Models and Applications*, volume 334. Wiley-Interscience, 2000.

D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

P. McCullagh and J. Nelder. *Generalized Linear Models*. London: Chapman and Hall, 1989.

D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, 2008.

T. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, 2001.

T. Minka, J. Winn, J. Guiver, and D. Knowles. Infer.NET 2.4, 2010. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

J. Paisley, D. Blei, and M. Jordan. Stick-breaking beta processes and the Poisson process. In *Artificial Intelligence and Statistics*, 2012a.

J. Paisley, C. Wang, and D. Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(2):235–272, 2012b.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B (Methodological)*, 71(2):319–392, 2009.

A. Smola, V. Vishwanathan, and E. Eskin. Laplace propagation. In *Advances in Neural Information Processing Systems*, 2003.

L. Tierney, R. Kass, and J. Kadane. Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of American Statistical Association*, 84(407), 1989.

M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence*, 2008.

M. Wells. Generalized linear models: A Bayesian perspective. *Journal of American Statistical Association*, 96(453):339–355, 2001.

E. Xing. On topic evolution. *CMU-ML TR-05-115*, 2005.

E. Xing, M. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*, 2003.

Y. Xue, D. Dunson, and L. Carin. The matrix stick-breaking process for flexible multi-task learning. In *International Conference on Machine Learning*, 2007.