Journal of Machine Learning Research (2012) Submitted 00/00; Published 00/00

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:

Abstract

Mean-field variational inference is widely used for approximate posterior inference in many probabilistic models. When the model is conditionally conjugate, variational updates are in closed-form. However, many models of interest are nonconjuate and practitioners may face the challenges of deriving the corresponding variational updates. In this paper, we develop and study two generic variational strategies for nonconjugate models—Laplace variational inference and delta method variational inference—which place minimal conditions on the model. These strategies extend and unify existing methods that were derived for specific models. We illustrate our approach on the correlated topic models, Bayesian logistic regression, and hierarchical Bayesian logistic regression. Our experimental results show that our methods work well on real-world datasets.

Keywords: Variational inference, Nonconjugate models, Laplace approximations, The 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 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 the resulting distribution is close to the

conditional distribution of the hidden variables given the observations, i.e., the posterior. Closeness is measured with Kullback-Leibler divergence. Inference becomes optimization.

In many settings this approach can be used as a "black box" technique, specifically 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 include Bayesian logistic regression (Jaakkola and Jordan, 1996), Bayesian generalized linear models (Wells, 2001), discrete choice models (Braun and McAuliffe, 2007), Bayesian item response models (Clinton et al., 2004; Fox, 2010), and nonconjugate topic models (Blei and Lafferty, 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, 2007; Ahmed and Xing, 2007), alternative bounds (Blei and Lafferty, 2007; Jaakkola and Jordan, 1996; Khan et al., 2010), and numerical quadrature (Honkela and Valpola, 2004).

There have been some other recent efforts to examine 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. Finally, the recent work of Knowles and Minka (2011) presents a message passing algorithm for nonconjugate models, which has been implemented in Infer.NET (Minka et al., 2010). This technique applies to a subset of models described in this paper.[1]

In this paper we develop generic approaches to mean-field variational inference for a large class of nonconjugate models. We develop two related strategies, both based on Taylor approximations. We first develop *Laplace variational inference*. This approach embeds Laplace approximations—an approximation technique for univariate distributions (MacKay, 1992)—within a variational optimization algorithm. We then develop the *delta method variational inference*, which 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.[2]

1. It may be generalizable to the full set. However, how to find the required expectations would still have to be determined.
2. The delta method was first used in variational inference by Braun and McAuliffe (2007) in the context of the discrete choice model. Our method generalizes their approach.

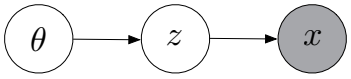

Figure 1: A general graphical model with hidden variables θ (real-valued) and z, and observed variable x. (This graphical model We assume that $p(z \mid \theta)$ is conjugate to $p(x \mid z)$, but we do not assume that $p(\theta)$ is conjugate to $p(z \mid \theta)$. We omit fixed hyper-parameters for simplicity. It is straightforward to embed this model into more complex models.

We studied our algorithms with two nonconjugate models: Bayesian logistic regression (Jaakkola and Jordan, 1996) (and its hierarchical extension) and correlated topic models (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 fidelity of the approximate posterior. These methods significantly expand the class of models for which mean-field variational inference can be easily applied.

2 Variational Inference in Nonconjugate models

Consider a generic model with the following joint distribution,

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

We assume this model has *hidden* variables θ and z, and *observed* variable x. (The distinction between the two hidden variables will be made clear below.) This is illustrated as a graphical model in Figure 1.

The inference problem is to compute the posterior of θ and z, $p(z, \theta|x)$. This is intractable for many models and we must resort to approximation. 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 $p(\theta, z|x)$ (Jordan et al., 1999).

Mean-field variational inference is simplest and most widely used variational inference method. It uses a fully factorized variational family,

$$q(z, \theta) = q(z)q(\theta).$$

Under the standard variational theory, minimizing the KL-divergence between $q(z, \theta)$ and $p(\theta, z|x)$ 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,

$$\log p(x) = \log \int p(x, z, \theta) \mathrm{d}z \mathrm{d}\theta \geq \mathbb{E}_q \left[\log p(x, Z, \theta)\right] + H[q] = \mathcal{L}(q), \tag{2}$$

where $\mathbb{E}_q[\cdot]$ is the expectation taken with respect to q and $H[q]$ the entropy. Setting $\partial\mathcal{L}(q)/\partial q = 0$ shows that the optimal solution satisfies the following,

$$q^*(\theta) \propto \exp\left\{\mathbb{E}_{q(z)}\left[\log p(Z|\theta)p(\theta)\right]\right\}, \tag{3}$$

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

Here we have combined the optimal conditions from Bishop (2006) with the particular factorization of Equation 1.

A model is *conditionally conjugate* when $p(\theta)$ is a conjugate prior to $p(z|\theta)$ and $p(z|\theta)$ is a conjugate prior to $p(x|z)$. By *conjugate*, we mean that the conditional distribution of a variable given its Markov blanket is in the same family as the distribution of the variable in the model (Bernardo and Smith, 1994). For example, the model is conditionally conjugate if $p(\theta)$ is a Dirichlet and both $p(z|\theta)$ and $p(x|z)$ are multinomials.

For such models, both $q^*(\theta)$ and $q^*(z)$ are available in closed-form (holding the other distribution fixed) and are in the same family as their corresponding distributions in the model, $p(\theta)$ and $p(z|\theta)$. This leads to the traditional coordinate ascent algorithm for mean-field variational inference, where we alternately update $q(\theta)$ and $q(z)$ using Equation 3 and Equation 4 (Bishop, 2006).

However, this simple coordinate ascent algorithm is not available when the variable θ is not part of a conjugate pair. In that setting, which arises in many practical models, we cannot compute closed-form updates for either distribution. Further, the nonconjugacy makes it difficult to evaluate the lower bound $\mathcal{L}(q)$ in Eq. 2. In the next sections, we will define a wider class of models beyond those that are conditionally conjugate, and we will develop generic variational inference for this more general class.

2.1 A Class of Nonconjugate Models

We present a wider class of models through a set of assumptions with respect to Eq. 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 $\log(\theta - \theta_0)$.
2. We assume the distribution $p(z|\theta)$ is in the exponential family (Brown, 1986),

$$p(z|\theta) = h(z)\exp\left\{\eta(\theta)^\top t(z) - a(\eta(\theta))\right\}, \tag{5}$$

   where $h(z)$ is a function of z; $t(z)$ is the sufficient statistic; $\eta(\theta)$ is the natural parameter; and $a(\eta(\theta))$ is log partition function. We emphasize that $p(\theta)$ is not necessarily the conjugate prior.
3. The distribution $p(x|z)$ is in the exponential family with z as the natural parameter,

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

and we require the distribution $p(z|\theta)$ is conjugate (Bernardo and Smith, 1994). Consequently, the term $t(z)$ in Eq. 5 is

$$t(z) = [z, -a(z)]. \tag{7}$$

Note that $t(x)$ and $a(z)$ are overloaded. They are different from $t(z)$ and $a(\eta(\theta))$.

These assumptions are weaker than those of conditional conjugacy. This family includes nonconjugate models like the correlated topic model (CTM) (Blei and Lafferty, 2007), Bayesian logistic regression (Jaakkola and Jordan, 1996), discrete choice models (Braun and McAuliffe, 2007), Bayesian ideal point models (Clinton et al., 2004), and many others. Variational inference is not as straightforward in these models because $p(\theta)$ is not conjugate to $p(z \mid \theta)$. Specifically, the update in Equation 3 does not necessarily have the form of an exponential family we can work with and it is difficult to use $\mathbb{E}_{q(\theta)}\left[p(z \mid \theta)\right]$ in Equation 4.

We will develop two variational inference algorithms for this class of models: *Laplace variational inference* and *delta method variational inference*. Both use coordinate ascent to optimize the variational parameters, iterating between updating $q(\theta)$ and $q(z)$. In Laplace variational inference, we use Laplace approximations (MacKay, 1992) within the coordinate ascent updates of Equation 3 and Equation 4. In delta method variational inference, we apply Taylor approximations to approximate the variational objective in Equation 2 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 algorithms, the variational distribution $q(\theta)$ is a Gaussian and the variational distribution $q(z)$ is in the same family as $p(z|\theta)$ in Eq. 5. In Laplace variational inference, these forms emerge from the derivation. In delta method variational inference, they are assumed.

As for traditional variational inference, our algorithms alternate between updating $q(\theta)$ and $q(z)$. In the following sections, we derive each algorithm for updating $q(\theta)$. We then show how to update $q(z)$.

2.2 Laplace Variational Inference

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

The Laplace Approximation. Laplace approximations use a Gaussian to approximate an intractable density (MacKay, 1992). 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) to construct a Gaussian proxy for the posterior. They are typically used in univariate settings.

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

$$\log p(\theta|x) \approx \log p(\hat{\theta}|x) + \tfrac{1}{2}(\theta - \hat{\theta})^\top H(\hat{\theta})(\theta - \hat{\theta}). \tag{8}$$

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 Eq. 8, 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 Eq. 8 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

$$p(\theta|x) \approx \mathcal{N}(\hat{\theta}, -H(\hat{\theta})^{-1}). \tag{9}$$

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 variational inference for a complex model.

Laplace updates in variational inference. Laplace approximations have been used in approximate inference in more complex models. 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 latent Gaussian models. Here we want to use them for variational inference that can be applied to a wider range of nonconjugate models, including those with discrete hidden variables.

We use Laplace approximations to update the variational distribution $q(\theta)$. First, we combine the coordinate update in Eq. 3 with the exponential family assumption in Eq. 5,

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

Now define

$$\bar{t}_z \triangleq \mathbb{E}_{q(z)}\left[t(Z)\right], \tag{11}$$

$$f(\theta) \triangleq \eta(\theta)^\top \bar{t}_z - a(\eta(\theta)) + \log p(\theta). \tag{12}$$

We approximate $q(\theta)$ by taking a second-order Taylor approximation of $f(\theta)$ around its maximum, following the same logic as from Equation 8. Let $\hat{\theta}$ be the value that maximizes

$f(\theta)$ and $\nabla^2 f(\hat{\theta})$ be the Hessian matrix evaluated at $\hat{\theta}$. Adapting Eq. 9 and Eq. 10 to this setting gives

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

The Gaussian form of $q(\theta)$ stems from using the Taylor approximation. This is an approximate update for $q(\theta)$ and can be embedded in a coordinate ascent algorithm for a nonconjugate model. Notice we need to use numerical optimization to obtain $\hat{\theta}$.

2.3 Delta Method Variational Inference

In Laplace variational inference, the variational distribution $q(\theta)$ Eq. 13 is solely a function of $\hat{\theta}$, the maximum of $f(\theta)$ in Eq. 12. A natural question is, would other values of θ be suitable as well? To consider such alternatives, we describe a different way of performing variational inference. We approximate the variational objective $\mathcal{L}$ in Equation 2 and then optimize that approximation.

Again we focus on $q(\theta)$ and postpone the discussion of $q(z)$. The variational distribution $q(\theta)$ is a Gaussian $q(\theta) = \mathcal{N}(\mu, \Sigma)$. We isolate the terms related to $q(\theta)$ in the objective, then substitute the exponential family assumption about $p(z|\theta)$ in Equation 5 into $\mathcal{L}(q)$ in Equation 2,

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

The entropy of the Gaussian is $H[q(\theta)] = \frac{1}{2}\log|\Sigma| + C$, where C is a constant. Notice the first three terms are the same function $f(\theta)$ defined in Eq. 12. We simplify the lower bound $\mathcal{L}(q(\theta))$,

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

We cannot easily compute the expectation in the first term. We use a Taylor expansion of $f(\theta)$ around a chosen value $\hat{\theta}$,

$$f(\theta) \approx f(\hat{\theta}) + \nabla f(\hat{\theta})(\theta - \hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^\top \nabla^2 f(\hat{\theta})(\theta - \hat{\theta}).$$

With this Taylor approximation, $\mathcal{L}(q(\theta))$ can be approximated with

$$\begin{aligned}\mathcal{L}(q(\theta)) \approx\; & f(\hat{\theta}) + \nabla f(\hat{\theta})^\top(\mu - \hat{\theta}) + \tfrac{1}{2}(\mu - \hat{\theta})^\top \nabla^2 f(\hat{\theta})(\mu - \hat{\theta}) \\ & + \tfrac{1}{2}\left(\mathrm{Tr}\left\{\nabla^2 f(\hat{\theta})\Sigma\right\} + \log|\Sigma|\right),\end{aligned} \tag{14}$$

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

The form of this optimization, however, depends on how we choose $\hat{\theta}$, the point around which to approximate $f(\theta)$. Unlike in Laplace variational inference, $\hat{\theta}$ can be any value. We will describe three options.

One choice is to set $\hat{\theta}$ to be the maximum of $f(\theta)$. Then maximizing the approximation in Equation 14 gives $\mu = \hat{\theta}$ and $\Sigma = -\nabla^2 f(\hat{\theta})^{-1}$. Notice this is the update derived in Section 2.2. It gives a different derivation of Laplace variational inference.

A second choice is to set $\hat{\theta} = \mu$, i.e., 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 in Eq. 14 is

$$\mathcal{L}(q(\theta)) \approx f(\mu) + \tfrac{1}{2}\mathrm{Tr}\left\{\nabla^2 f(\mu)\Sigma\right\} + \tfrac{1}{2}\log|\Sigma|. \tag{15}$$

This is the multivariate delta method for evaluating $\mathbb{E}_{q(\theta)}[f(\theta)]$ (Bickel and Doksum, 2007) and so we call this choice *delta method variational inference*.

In delta method variational inference, we optimize $\mathcal{L}(q(\theta))$ with coordinate ascent on μ and Σ. We use gradient methods to find μ. (Note this is more expensive than Laplace variational inference because it requires the third derivative $\nabla^3 f(\theta)$.) Then, given a value of μ, we optimize Σ in closed form $\Sigma = -\nabla^2 f(\mu)^{-1}$. Braun and McAuliffe (2007) were the first to use the delta method in a variational inference algorithm, developing this technique for the discrete choice model. If we assume $p(\theta)$ is Gaussian then we recover their algorithm. With the ideas presented here, we can now use this strategy in many models.

A final choice is to guess $\hat{\theta}$, for example, as the mean of the variational distribution from the previous iteration of coordinate ascent. If $p(\theta)$ is a Gaussian distribution, this recovers the updates derived in Ahmed and Xing (2007) for the correlated topic model.[3] We found this simple guess did not work very well on our experiments. (It did not always converge, probably due to the difficulty of choosing an appropriate initial $\hat{\theta}$.) We thus focus on Laplace and delta method variational inference.

2.4 Updating $q(z)$

We have derived variational updates for $q(\theta)$ using two methods. We now turn to the update for $q(z)$. We will show that Laplace (Section 2.2) and delta method variational inference (Section 2.3) lead to the same update form for $q(z)$. Further, we have implicitly assumed that $\mathbb{E}_{q(z)}[t(z)]$ in Eq. 11 and Eq. 12 is easy to compute. We will confirm this as well.

Laplace variational inference. First, we apply the exponential family form in Equation 5 to the exact update of Eq. 4,

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

where C is a constant not depending on z. Now we use $p(x|z)$ from Eq. 6 and $t(z)$ from Eq. 7 to obtain

$$q(z) \propto h(z)\exp\left\{\left(\mathbb{E}_{q(\theta)}[\eta(\theta)] + [t(x), 1]\right)^\top t(z)\right\}, \tag{16}$$

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

1: Initialize variational distribution $q(\theta)$ and $q(z)$.
2: **repeat**
3: Compute the statistics $\mathbb{E}_{q(z)}[t(z)]$ in Eq. 11.
4: To obtain $q(\theta)$,

Option a): in Laplace variational inference, compute Eq. 13.

Option b): in delta method variational inference, optimize Eq. 15.

5: Approximate the statistics $\mathbb{E}_{q(\theta)}[\eta(\theta)]$ in Eq. 17.
6: Update $q(z)$ in Eq. 16.
7: **until** Convergence (See Section 2.5 for the criteria).
8: **return** $q(\theta)$ and $q(z)$.

Figure 2: The approximation framework in variational inference.

which is in the same family as $p(z|\theta)$ in Eq. 5. This is the update for $q(z)$. Further, because it is in a simple exponential family form, we can compute $\mathbb{E}_{q(z)}[t(z)]$ from Eq. 11.

We have additionally assumed the tractability of $\mathbb{E}_{q(\theta)}[\eta(\theta)]$. To approximate this, we take a Taylor approximation of $\eta(\theta)$ around the variational parameter μ,

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

Since $q(\theta) = \mathcal{N}(\mu, \Sigma)$, this means that

$$\mathbb{E}_{q(\theta)}[\eta(\theta)] \approx \eta(\mu) + \tfrac{1}{2}\mathrm{Tr}\left\{\nabla^2\eta(\mu)\Sigma\right\}. \tag{17}$$

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

Delta method variational inference. 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 Eq. 2,

$$\mathcal{L}(q(z)) = \mathbb{E}_{q(z)}\left[\log p(x|z) + \log h(z) + \mathbb{E}_{q(\theta)}[\eta(\theta)]^\top t(z)\right] + H[q(z)]. \tag{18}$$

Setting the partial gradient $\partial\mathcal{L}(q(z))/\partial q(z) = 0$ gives the same optimal $q(z)$ of Eq. 4. Computing this update reduces to the approach for Laplace variational inference in Equation 16.

2.5 The Algorithm

We have described updates for $q(\theta)$ and $q(z)$, which can be embedded in an iterative algorithm. See Figure 2 for an outline of this algorithm. We have reduced deriving variational updates for complicated nonconjugate models to somewhat mechanical work—calculating derivatives and calling a numerical optimization library.[4] Regarding the variants of our algorithm,

4. Python implementations of our algorithm for the example models described in Section 3 are available at `http://www.cs.cmu.edu/~chongw/software/nonconjugate-inference.tar.gz`.

Laplace variational inference is simpler to derive because it only requires first derivatives of the function in Eq. 12, while delta variational inference requires third derivatives. We empirically study the differences between these methods in Section 4.

We monitor the convergence of our algorithm in Figure 2 using the following criteria,

- The change of the mean of $|\mathbb{E}_q[\theta]|$ is less than 0.00001.
- The relative change of an approximate variational objective (explained below) is less than 0.00001.

Our approximate variational objective is,

$$\begin{aligned}\mathcal{L}(q(z,\theta)) \approx & f(\hat{\theta}) + \nabla f(\hat{\theta})^\top(\mu - \hat{\theta}) + \tfrac{1}{2}(\mu-\hat{\theta})^\top \nabla^2 f(\hat{\theta})(\mu - \hat{\theta}) \\ & + \tfrac{1}{2}\left(\mathrm{Tr}\left\{\nabla^2 f(\hat{\theta})\Sigma\right\} + \log|\Sigma|\right) + H(q(z)),\end{aligned} \tag{19}$$

where $f(\theta)$ is defined in Eq. 12, $q(\theta) = \mathcal{N}(\mu, \Sigma)$ and $H(q(z))$ is the entropy of the variational distribution $q(z)$. We obtain this approximation by combining the analysis for $q(\theta)$ in Eq. 14 and for $q(z)$ in Eq. 18. As outlined in Section 2.3, the choice of $\hat{\theta}$ determines whether we use Laplace or delta method variational inference. In both cases, the approximate lower bound is computed the same way. Finally we limit the maximum number of iterations to 100.

We note that our algorithms are not associated with a strict lower bound of the marginal probability of the observations. However, as also observed in Braun and McAuliffe (2007), we found that they converge in practice by monitoring Eq. 19. Figure 3 shows the plots of this quatity for *one document* during the training for the correlated topic model (CTM). (For CTM, the algorithm in Figure 2 applies to a single document. See Section 3.1 for more details.) Further, Eq. 19 is not the function we are optimizing—in fact the true global objective function is not clear. 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.

In the following extreme case, Eq. 19 becomes the true global objective; if we assume the variational distribution $q(\theta) = \delta(\theta - \mu)$, i.e., a delta distribution,[5] both Laplace and delta method variational inference reduce to a maximum a posterior (MAP) estimate of θ.

3 Example Models

We have described a generic algorithm for approximate posterior inference in nonconjugate models. We now discuss several nonconjugate models from the research literature for which we can apply our method. For each model, we identify the hidden variables z and θ from Eq. 5, the observation x from Eq. 6, and the form of $f(\theta)$ from Eq. 12. (We give its derivatives in the appendix.) In the next section, we study how our algorithm perform when analyzing data under these models.

5. In this case, the entropy of $q(\theta)$ is 0. See Welling et al. (2008) for more discussions on using delta distributions in variational inference.

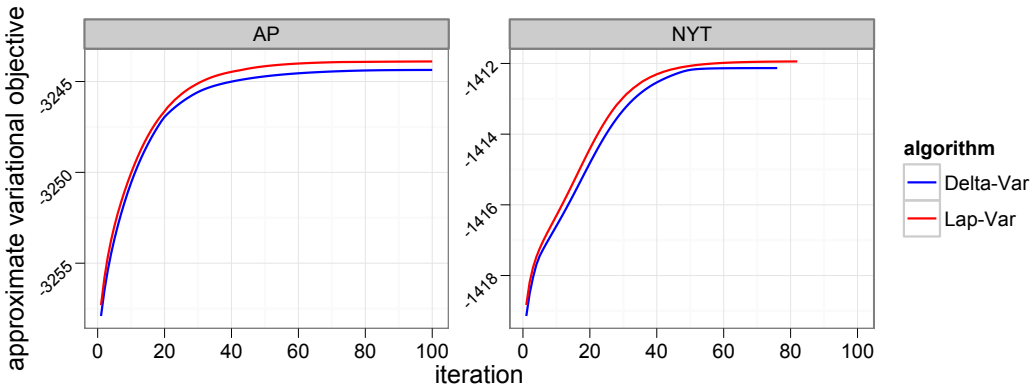

Figure 3: The approximate variational objective from Eq. 19 as a function of the iteration. This is for the correlated topic model on a single document (see Section 3.1).

3.1 Correlated Topic Models

The correlated topic model (CTM) (Blei and Lafferty, 2007) is an extension of latent Dirichlet allocation (LDA) (Blei et al., 2003). A topic model is a probabilistic model of a document collection. Each document is treated as a collection 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; however, each document exhibits them with individualized proportions.

Conditioned on a corpus of documents, the posterior topics place high probability on words that are associated under a single theme (e.g., *sports* or *health* or *business*); the posterior topic proportions reflect how each document exhibits those themes. 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 LDA these are given a Dirichlet prior, which makes the model conditionally conjugate. 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*.) While the CTM is more expressive, it is no longer conditionally conjugate.

We now cast the CTM as a type of generic model from Section 2. Suppose there are K topic parameters $\beta_{1:K}$ (fixed for now), 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 of the simplex with the same dimension, $\pi(\theta) \propto \exp\{\theta\}$. The CTM assumes the following generative process of a document:

1. Draw log topic proportions $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$.

2. For each word n:
   (a) Draw topic assignment $z_n \mid \theta \sim \text{Mult}(\pi(\theta))$.
   (b) Draw word $x_n \mid z_n, \beta \sim \text{Mult}(\beta_{z_n})$.

The topic proportions $\pi(\theta)$ are drawn from a logistic normal distribution; their correlation structure is captured in Σ_0. The topic assignment variable z_n indicates which topic the nth word is drawn from. In terms of our earlier notation, θ is identical; the earlier z is the collection of topic assignments $z_{1:N}$; the observation x is the collection of words $x_{1:N}$. This model satisfies the conditions in Section 2.1.

The variational distribution for parameter θ is $q(\theta) = \mathcal{N}(\mu, \Sigma)$; and the variational distribution for the z variables is discrete. In delta method variational inference, as in Braun and McAuliffe (2007), we restrict the covariance matrix Σ to be diagonal. This is because a full covariance matrix makes the derivative of Eq. 15 (with respect to μ) expensive to compute. Laplace variational inference does not require this simplification. The detailed derivations of the algorithm are in Appendix A. Note that, besides CTM, this approach can be used in a variety of nonconjugate topic models, including Dirichlet-multinomial regression (DMR) (Mimno and McCallum, 2008), Dynamic topic models (DTM) (Blei and Lafferty, 2006; Wang et al., 2008) and the discrete infinite logistic normal distribution (DILN) (Paisley et al., 2012b)

Note that for CTM, this algorithm is only for inference at the document level. As in Blei and Lafferty (2007), we estimate the corpus-level topic parameters $\beta_{1:K}$ and logistic normal parameters (μ_0, Σ_0) with variational EM.

3.2 Bayesian Logistic regression

Bayesian logistic regression is a well-studied model for binary classification (Jaakkola and Jordan, 1996). Let t_n is be a p-dimensional observed feature vector for the nth sample and z_n be its class (an indicator vector of length two). Let θ be the real-valued parameter vector in $\mathbb{R}^p$; there is a component for each feature. The usual Bayesian logistic regression is the following:

$$\begin{aligned} p(\theta) &= \mathcal{N}(\mu_0, \Sigma_0), \\ p(z_n \mid \theta, t_n) &= \sigma(\theta^\top t_n)^{z_{n,1}} \sigma(-\theta^\top t_n)^{z_{n,2}}, \ \forall n \end{aligned} \tag{20}$$

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

This is a subset of the model class in Section 2. In our earlier notation, θ is identical and z are the observed classes of each data point. There is no additional no additional observed variable x downstream. The variational distribution need only be defined for θ, $q(\theta) = \mathcal{N}(\mu, \Sigma)$. Using Laplace variational inference, our approach recovers the standard Laplace approximation for Bayesian logistic regression (Bishop, 2006). (This gives a connection between standard Laplace variational inference and variational inference.) Delta method variational inference provides an alternative. The detailed derivations are in Appendix B.

An important extension of Bayesian logistic regression (and Bayesian generalized models) is *hierarchical Bayesian logistic regression*. The hierarchical variant considers multiple related logistic regression problems and solves them simultaneously (Gelman and Hill, 2007). We will take an empirical Bayes approach. We treat the parameters for each problem from a shared prior, and then estimate the hyperparameters of that prior with maximum likelihood or MAP. With M related problems, we construct the following hierarchical model:

1. Draw the global hyperparameters:

$$\Sigma_0^{-1} \sim \text{Wishart}(\hat{\nu}, \widehat{\Phi}_0)$$
$$\mu_0 \sim \mathcal{N}(0, \widehat{\Sigma}_0)$$

2. For each problem m:
   (a) Draw coefficients $\theta_m \sim \mathcal{N}(\mu_0, \Sigma_0)$
   (b) Draw class labels $z_{1:N}$, conditioned on covariates $t_{1:N}$:

$$p(z_{mn} \mid \theta_m, t_{mn}) = \sigma(\theta_m^\top t_{mn})^{z_{mn,1}} \sigma(-\theta_m^\top t_{mn})^{z_{mn,2}}.$$

We construct $f(\theta_m)$ in Eq. 12 separately for each m, and fit the hyperparameters μ_0 and Σ_0 using regularized variational EM (Bishop, 2006). This amounts to MAP estimation, with priors as specified above. 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 variants).

4 Empirical Study

We studied the Laplace and delta method variational inference for the CTM and Bayesian logistic regression model on several data sets.

Correlated topic models (CTM). For the CTM, we analyzed two document collections. The *Associated Press* data contains 2,246 documents from the *Associated Press*, with 436K observed words and a vocabulary size of 10,473 terms. The *New York Times* data contains 9,238 documents from the *New York Times*, with 2.3 million observed words and a vocabulary size of 10,760 terms.

We measured per-word held-out log likelihood to evaluate the models' fit to the data. For each collection, we held out 20% of the documents as a test set. We split each document $\boldsymbol{w}_d$ in $\mathcal{D}_{\text{test}}$ into halves, $\boldsymbol{w}_i = (\boldsymbol{w}_{d1}, \boldsymbol{w}_{d2})$, and computed the log likelihood of the words in $\boldsymbol{w}_{d2}$ (independently) conditioned on $\boldsymbol{w}_{d1}$ and $\mathcal{D}_{\text{train}}$, similar to Asuncion et al. (2009) and Blei and Lafferty (2007). A better predictive distribution gives higher likelihood to the second half.

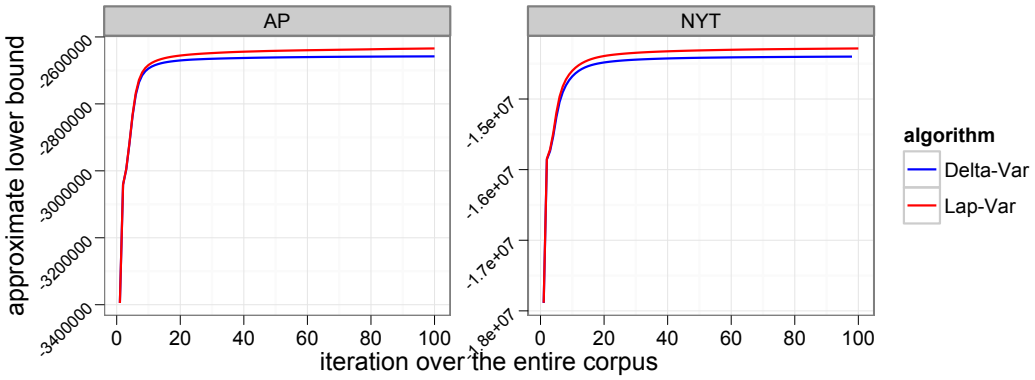

Figure 4: The approximate lower bound (y axis) during training for CTM when the number of topics $K = 60$ (Others are similar); this approximate lower bound is the sum of the per-document approximate lower bound for each document in Eq. 19. The iteration (x axis) indicates the overall number of passes of the entire corpus. It converged in practice.

Let $\bar{\pi}_d$ be the variational expectation of the topic proportions given $\boldsymbol{w}_{d1}$. The predictive probability of $\boldsymbol{w}_{d2}$ given $\boldsymbol{w}_{d1}$ is approximated by $p(\boldsymbol{w}_{d2}|\boldsymbol{w}_{d1}, \mathcal{D}_{\text{train}}) \approx \prod_{w \in \boldsymbol{w}_{d2}} \sum_k \bar{\pi}_{dk} \beta_{kw}$. Then the per-word held-out log likelihood is

$$\text{likelihood}_{\text{pw}} \triangleq \sum_{d \in \mathcal{D}_{\text{test}}} \log p(\boldsymbol{w}_{d2}|\boldsymbol{w}_{d1}, \mathcal{D}_{\text{train}}) / \sum_{\boldsymbol{d} \in \mathcal{D}_{\text{test}}} |\boldsymbol{w}_{d2}|.$$

This evaluation lets us compare methods regardless of whether they provide a bound. Here $|\boldsymbol{w}_{d2}|$ is the number of tokens in $\boldsymbol{w}_{d2}$. It is a measure of the quality of the estimated predictive distribution. Exact computation is intractable, and so we use the following approximation. Let $\bar{\pi}_d$ be the variational expectation of the topic proportions given $\boldsymbol{w}_{d2}$. The predictive probability of $\boldsymbol{w}_{d2}$ given $\boldsymbol{w}_{d1}$ is approximated by

$$p(\boldsymbol{w}_{d2}|\boldsymbol{w}_{d1}, \mathcal{D}_{\text{train}}) \approx \prod_{w \in \boldsymbol{w}_{d2}} \sum_k \bar{\pi}_{dk} \beta_{kw}.$$

We compared the original method of Blei and Lafferty (2007) to Laplace and delta method variational inference. We varied the number of topics from 20 to 80. We stopped fitting when the relative change of the approximate lower bound in the corpus-level EM algorithm was smaller than 10e-5. In CTM, this approximate lower bound amounts to be the sum of the per-document approximate lower bound for each document in Eq. 19. Figure 4 shows some example approximate lower bounds during training.

Figure 5(a) shows per-word held-out log likelihood as a function of the number of topics. On both data sets, the original algorithm of Blei and Lafferty (2007), tailored for this model, usually performed worse than both generic algorithms. Our conjecture is that Blei and Lafferty (2007) gives a strict lower bound, which might be loose; our method uses an approximation which, while not a bound, might be closer to the objective and give better

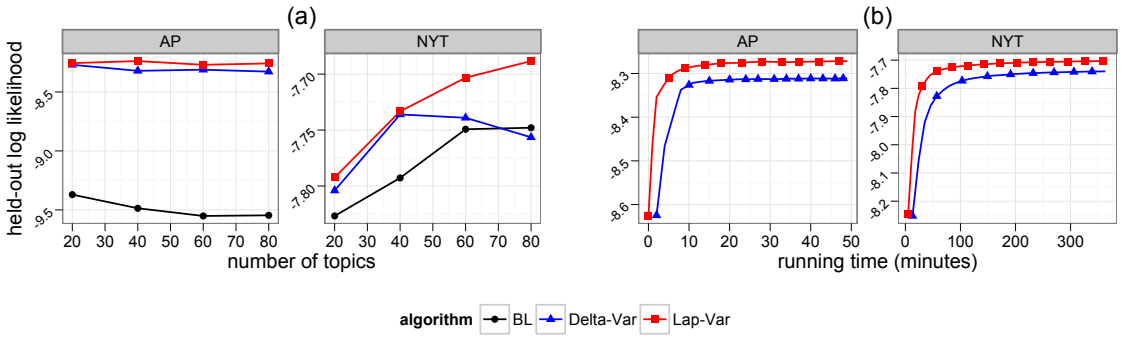

Figure 5: Comparison on per-word held-out log likelihood. Higher numbers are better. Laplace variational inference is "Lap-Var"; delta method variational inference is "Delta-Var"; the method by Blei and Lafferty in Blei and Lafferty (2007) is BL. (**a**) Held-out likelihood against number of topics. Delta-Var performs better BL. Lap-Var performs best. (**b**) Held-out likelihood against running time when $K = 60$. Lap-Var is faster and gives performs better than Delta-Var.

predictive distributions. Laplace variational inference was always better than both other algorithms.

We also compared the running times of Laplace variational inference and delta method inference. For a model with 60 topics, Figure 5(b) shows the per-word held-out log likelihood as a function of time.[6] (Other topic settings looked similar.) In addition to finding a better fit to the data, Laplace variational inference is faster.

Bayesian logistic regression We studied our algorithms on Bayesian logistic regression in both standard and hierarchical settings. In the standard setting, we analyzed two datasets. The *Yeast* data (Elisseeff and Weston, 2001) is formed by micro-array expression data and phylogenetic profiles. There are 1,500 genes in the training set and 917 genes in the test set. The input dimension is 103. One gene is associated with up to 14 different edges (14 labels), corresponding to 14 independent binary classification problems. The *Scene* data (Boutell et al., 2004) contains 1,211 training and 1,196 test images, with 294 images features and up to 6 scene labels per image, corresponding to 6 independent binary classification problems.[7]

We used two metrics. 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 of the predictive likelihood,

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

6. We did not formally compare the running time of Blei and Lafferty (2007)'s method because we used the authors' C implementation, while ours is in Python. We observed, however, that their method took more than five times longer than ours.
7. The *Yeast* and *Scene* data are at http://mulan.sourceforge.net/datasets.html.

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 generic Laplace variational inference (Section 2.2), delta method variational inference (Section 2.3), and the method of Jaakkola and Jordan (1996). This last method preserves a lower bound on the marginal likelihood with a first-order Taylor approximation, and was developed specifically for Bayesian logistic regression. Table 1 gives the results. Laplace variational inference and delta method variational inference gave slightly better accuracy than Jaakkola and Jordan's method, and much better log predictive likelihood.[8]

To study hierarchical Bayesian logistic regression, we analyzed the *School* data from the Inner London Education Authority.[9] This data 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 (year of the exam, gender, VR band (individual prior attainment data), and ethnic group) and four school-dependent features (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 big school), or analyze them with the hierarchical logistic regression structure described in Section 3.2. For the non-hierarchical settings, we studied Jaakkola and Jordan's method, Laplace variational inference, and delta method variational inference. When treating the problem hierarchically, we studied the corresponding Laplace and delta method variational inference algorithms. Table 2 gives the results. Hierarchical logistic regression performed best in terms of both accuracy and predictive log likelihood.

5 Conclusions

We have 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. Further, we showed that Laplace variational inference is better and faster than delta method variational inference.

8. Previous literature, e.g., (Xue et al., 2007; Archambeau et al., 2011) also treat *Yeast* and *Scene* as multi-task problems. This is slightly non-standard, since each sub-problem shares the same input features. In our experiments, we found that standard Bayesian logistic regression performed competitively with the multi-task algorithms reported in Archambeau et al. (2011).
9. The data is available at http://multilevel.ioe.ac.uk/intro/datasets.html.

Table 1: Comparison of different methods on standard Bayesian logistic regression using accuracy (Acc.) and averaged log predictive likelihood (Lik.)—the higher, the better. Results are averaged from five random starts. Variance is too small to show. The best results are highlighted. Lap-Var and Delta-Var gives slightly better accuracy but much better predictive log likelihood.

	Yeast		Scene	
Algorithm	Acc.	Lik.	Acc.	Lik.
Jaakkola	79.7%	-0.678	87.4%	-0.670
Lap-Var	80.1%	**-0.449**	89.4%	**-0.259**
Delta-Var	**80.2%**	-0.450	**89.5%**	-0.265

Table 2: Comparison of different methods on School data using accuracy (Acc.) and averaged log predictive likelihood (Lik.). Results are averaged from 10 random splits. Variance is too small to show. The best results are highlighted. Lap-Var-all indicates the case where we treats all schools together. The hierarchical approach performs the best.

	School	
Algorithm	Acc.	Lik.
Jaakkola	70.5%	-0.684
Lap-Var	70.8%	-0.569
Delta-Var	70.8%	-0.571
Jaakkola-all	71.2%	-0.685
Lap-Var-all	71.3%	-0.557
Delta-Var-all	71.3%	-0.557
Hier-Lap-Var	**71.9%**	**-0.549**
Hier-Delta-Var	**71.9%**	-0.559

This method expands the scope of variational inference. One area of future work is to further expand that scope, by relaxing the assumption of full conditional conjugacy in z. In such a setting, we may be able to use moment matching (Tierney et al., 1989) to develop efficient variational algorithms.

Appendix A: The correlated topic model

In correlated topic models (CTM) (Blei and Lafferty, 2007), there are K topic parameters $\beta_{1:K}$ (fixed for now), each of which is a distribution over V terms. Let π be the topic proportions for a document and n be the index of an observed word x_n. The CTM assumes

the following generative process of a document,

$$\begin{aligned}\theta &\sim \mathcal{N}(\mu_0, \Sigma_0),\\ \pi &\propto \exp(\theta)\\ z_n \,|\, \pi &\sim \text{Mult}(\pi),\\ x_n \,|\, z_n, \beta &\sim \text{Mult}(\beta_{z_n}).\end{aligned}$$

In this model, the topic proportions π are drawn from a logistic normal distribution. Their correlation structure is captured in Σ_0. The variable z_n indicates which topic the nth word is drawn from. We can now identify the quantities from Eq. 5, Eq. 6 and Eq. 7 that we need to compute $f(\theta)$ in Eq. 12

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

With this notation, function $f(\theta)$ defined in Eq. 12 is,

$$f(\theta) = \eta(\theta)^\top \bar{t}_z - \tfrac{1}{2}(\theta - \mu_0)^\top \Sigma_0^{-1}(\theta - \mu_0),$$

where $\bar{t}_z$ is the expected word counts of each topic under the variational distribution $q(z)$.

Using $\partial\pi_i/\partial\theta_j = \pi_i(1_{[i=j]} - \pi_j)$, we obtain the gradient and Hessian for function $f(\theta)$ in CTM,

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

where $1_{[i=j]} = 1$ if $i = j$ and 0 otherwise. Note the first $\nabla f(\theta)$ is enough for Laplace variational inference.

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

$$\text{Trace}\left\{\nabla^2 f(\theta)\Sigma\right\} = \left(-\textstyle\sum_{k=1}^K \pi_k\Sigma_{kk} + \pi^T\Sigma\pi\right) \textstyle\sum_{k=1}^K [\bar{t}_z]_k - \text{Trace}(\Sigma_0^{-1}\Sigma).$$

We assume Σ is diagonal for the delta method for simplicity (Braun and McAuliffe, 2007). (Note for Laplace method, we don't need this assumption.) This gives us

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

The algorithm in Fig Figure 2 for CTM only applies to a single document. As we discussed in the main text, to complete whole algorithm for CTM, we employ the same variational EM algorithm as in Blei and Lafferty (2007). The E-step corresponds to the algorithm in Figure 2. The M-step, i.e., the estimation of $\beta_{1:K}$, amounts to,

$$\hat{\beta}_{kw} \propto \sum_d \sum_{n=1}^{N_d} 1[x_{dn} = w] q(z_{dn} = k).$$

where d is the index of the documents in the collection, N_d is the number of words in document d and w is the word index in a predefined vocabulary.

Appendix B: Bayesian logistic regression

In Bayesian logistic regression, let t_n is be a p-dimensional observed feature vector for the nth sample and z_n be its class (represented as an indicator vector of length two). Let θ be the real-valued parameter vector in $\mathbb{R}^p$; there is a component for each feature. The usual Bayesian logistic regression model is the following:

$$\begin{aligned} p(\theta) &= \mathcal{N}(\mu_0, \Sigma_0), \\ p(z_n \,|\, \theta, t_n) &= \sigma(\theta^\top t_n)^{z_{n,1}} \sigma(-\theta^\top t_n)^{z_{n,2}}, \ \forall n \end{aligned}$$

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

In Bayesian logistic regression, we can fit the distribution of the observations $z_{1:N}$ into the exponential family with

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

In this set up, $\bar{t}_z$ represents the whole set of labels. With this notation, the $f(\theta)$ defined in Eq. 12 is

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

The gradient and Hessian for function $f(\theta)$ in Bayesian logistic regression are

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

Note the first $\nabla f(\theta)$ is enough for Laplace variational inference. This is the standard Laplace approximation to Bayesian logistic regression (Bishop, 2006).

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

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

Note that the "diagonal" assumption for Σ is not needed.

References

A. Ahmed and E. Xing. On tight approximate inference of the logistic normal topic admixture model. In *International Conference on Artifical 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 12*, 2000.

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.

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. *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 American Statistical Association*, 105(489), 2007.

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

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.

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 Univ. 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. A variational approach to Bayesian logistic regression models and their extensions. In *International Workshop on Artificial Intelligence and Statistics*, 1996.

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 *Advances in Neural Information Processing Systems*, 2011.

D. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4:448–472, May 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*, pages 362–369, 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. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 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 (Statistical Methodology)*, 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 (UAI)*, 2008.

M. Welling, C. Chemudugunta, and N. Sutter. Deterministic latent variable models and their pitfalls. In *SIAM International Conference on Data Mining*, 2008.

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

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.