On Tight Approximate Inference of the Logistic-Normal Topic Admixture Model

Amr Ahmed
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
amahmed@cs.cmu.edu

Eric P. Xing
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
epxing@cs.cmu.edu

Abstract

The Logistic-Normal Topic Admixture Model (LoNTAM), also known as *correlated topic model* (Blei and Lafferty, 2005), is a promising and expressive admixture-based text model. It can capture topic correlations via the use of a logistic-normal distribution to model non-trivial variabilities in the topic mixing vectors underlying documents. However, the non-conjugacy caused by the logistic-normal makes posterior inference and model learning significantly more challenging. In this paper, we present a new, tight approximate inference algorithm for LoNTAM based on a multivariate quadratic Taylor approximation scheme that facilitates elegant closed-form message passing. We present experimental results on simulated data as well as on the NIPS17 and PNAS document collections, and show that our approach is not only simple and easy to implement, but also it converges faster, and leads to more accurate recovery of the semantic truth underlying documents and estimates of the parameters comparing to previous methods.

1 Introduction

Statistical admixture models have recently gained much popularity in managing large collection of discrete objects. Via an admixture model, one can project such objects into a low dimensional space where their latent semantic (such as topical aspects) can be captured. This low dimensional representation can then be used for tasks like classifications and clustering or merely as a tool to structurally browse the otherwise unstructured collection.

An admixture model posits that each object is sampled from a mixture model according to the object's specific mixing vector over the mixture components. Special instances of this formalism have been used for population genetics (Pritchard *et al.*, 2000), vision (Sivic *et al.*, 2005), text modeling (Blei *et al.*, 2003) and machine translation (Zhao and Xing, 2006). When applied to text, the objects correspond to documents, and the mixture components are known as *topics* which are often represented as a multinomial distribution over a given vocabulary. Under this formalism, each document is succinctly represented as a mixing vector over the set of *topics*, and the *topic mixing vector* reflects the semantics of the document.

Much of the expressiveness of admixture-based text models lies in how they model the variabilities in the topic mixing vectors underlying documents. For instance, when the variability is modeled via a Dirichlet distribution, the model is known as *latent Dirichlet allocation* (Blei *et al.*, 2003). The Dirichlet indeed is an appealing choice because its conjugacy to the multinomial allows for computational advantages in inference and learning. However, the Dirichlet can only capture variations in each topic's intensity (almost) independently, and fails to model the fact that some topics are highly correlated and can arise synergistically. Failure to model such correlation limits the model's ability to discover subtle topical structures underlying the data.

Apart from the Dirichlet, another popular distribution over the *simplex* (the space of normalized weight vectors) is the logistic-normal (LN) distribution which has the sought-after property of being able to model correlations between the components of the vectors drawn from it (Aitchison and Shen, 1980). When the logistic-normal is used instead of the Dirichlet, we call the resulting model a *Logistic-Normal Topic Admixture Model* (LoNTAM) which is also known as the *correlated topic model* (Blei and Lafferty, 2005). Unfortunately, this added expressivity comes with a price, because the non-conjugacy of LN to the multinomial distribution makes posterior inference and parameter estimation significantly more difficult.

In the sequel, we present a new, approximate inference algorithm for LoNTAM which offers a simple and efficient way of capturing correlated topic posteriors. (A

longer version, which contains derivational details and additional extensions/generalizations can be found in a earlier Technical Report (Xing, 2005).) We begin by outlining LoNTAM; then we proceed to describe our approximate inference method that overcomes the non-conjugacy of LN via the use of a *multivariate* quadratic Taylor approximation to LN, which enables an elegant closed-form variational message passing algorithm. For completeness, we also describe an earlier inference algorithm for LoNTAM given by Blei and Lafferty (2005) and highlight key differences between the two approaches. Finally we present experimental results on simulated text datasets as well as real text collections from NIPS and PNAS. Our results show that the proposed algorithm is not only simpler and easier to implement, but also leads tighter approximation to the topic posterior, more accurate estimations of the parameters, and faster convergence, comparing to the previous method based on linear approximation and numerical procedure.

2 Logistic-Normal Topic Admixture

We begin with a brief recap of the general admixture formalism, and the LoNTAM model for text documents represented as bags of words.

2.1 Admixture Model

Statistically, an object x is said to be derived from an *admixture* if it consists of a bag of elements, say $\{x_1, \ldots, x_N\}$, each sampled independently or coupled in some way, from a mixture model, according to an *admixing coefficient vector* $\vec{\theta}$, which represents the (normalized) fraction of contribution from each of the mixture components to the object being modeled. In a typical text modeling setting, each document corresponds to an object, the words thereof correspond to the elements constituting the object, and the document-specific admixing coefficient vector is often known as a *topic mixing vector* (or simply, *topic vector*). Generatively, to sample a document according to an admixture model, we first sample a topic vector from some admixing prior, then a latent topic is sampled for each word based on the topic vector to induce topic-specific word instantiations. Since the *admixture formalism* enables an object to be synthesized from elements drawn from a mixture of multiple sources, it is also known as *mixed membership model* in the statistical community (Erosheva *et al.*, 2004).

2.2 Logistic-Normal Topic Admixture Model

Much of the expressiveness of admixture-based text models lies in the choice of the prior for the documents topic vectors. To capture non-trivial correlations among the weights of all possible topics underlying a document (i.e., the elements of the topic vector $\vec{\theta}$), instead of using a Dirichlet as in LDA, a LoNTAM model employs a logistic-normal distribution for the topic vectors of a study corpus. As discussed in Blei and Lafferty (2005), this prior captures a much richer abundance of correlation patterns in a topic simplex; but as a cost, the non-conjugacy between the logistic-normal prior (for topic vectors) and the multinomial likelihood (for topic instantiations) makes posterior inference and parameter estimation extremely hard. For example, variational inferences algorithms commonly used for generalized linear models (GLIMs) won't have close-form fixed point iterative formula in this case. Indeed, the approximate inference scheme adopted so far fail to capture the much-desired correlation structure in the posterior distribution of the topic vectors.

Before presenting our tight approximate inference algorithm, which offers a simple and efficient way for obtaining truly correlated topic posteriors under LoNTAM, in the following we outline the details of this model for later reference. As illustrated in Figure 1, in a LoNTAM, each topic, say topic k, is represented by an M-dimensional word frequency vector $\vec{\beta}_k$, which parameterizes a topic-specific multinomial distribution. A document is generated by sampling its topic vector from a logistic normal distribution with K-dimensional mean μ and covariance Σ, that is $\mathrm{LN}(\mu, \Sigma)$, and then words are sampled based on this topic vector. More formally, to generate a document $\mathbf{w}_d = \{w_{d,1}, w_{d,2}, \ldots, w_{d,N}\}$, we proceed as bellow:

1. Draw $\vec{\theta}_d \sim \mathrm{LN}(\mu, \Sigma)$
2. For each word $w_{d,n}$ in $\mathbf{w}_d$
   - Draw latent topic $z_{d,n} \sim \mathrm{Multinomial}(\vec{\theta}_d)$
   - Draw $w_{d,n}|z_{d,n} = k \sim \mathrm{Multinomial}(\vec{\beta}_k)$

The first step can be broken down into two sub-steps: first draw $\vec{\gamma}_d \sim \mathrm{Normal}(\mu, \Sigma)$; then map it to the simplex via the following *logistic* transformation:

$$\theta_{d,k} = \exp\{\gamma_{d,k} - C(\vec{\gamma}_d)\}, \quad \forall k = 1, \ldots, K \tag{1}$$

$$\text{where} \qquad C(\vec{\gamma}_d) = \log\Big(\sum_{k=1}^{K} \exp\{\gamma_{d,k}\}\Big). \tag{2}$$

Here $C(\vec{\gamma}_d)$ is a normalization constant (i.e., the log partition function). Furthermore, due to the normalizability constrain on the multinomial parameters, $\vec{\theta}_d$ only has $K-1$ degree of freedom. Thus we only need to draw the first $K-1$ components of $\vec{\gamma}_d$ from a $(K-1)$-dimensional multivariate Gaussian, and leave $\gamma_{d,K} = 0$. For simplicity, we omit this technicality in the forth coming general operation of our model. Putting everything together, the marginal probability of a document $\mathbf{w}$ can be written as follows (for simplicity, in the sequel we omit document index "d" when our statements and/or expressions apply to all documents):

$$p(\mathbf{w}) = \int_{\vec{\gamma}} \Big(\prod_{n=1}^{N} \sum_{z_n=1}^{K} p(w_n|z_n; \vec{\beta}_{1:K}) \times p(z_n|\mathrm{logistic}(\vec{\gamma})) \Big) \mathcal{N}(\vec{\gamma}|\mu, \Sigma) d\vec{\gamma}, \tag{3}$$

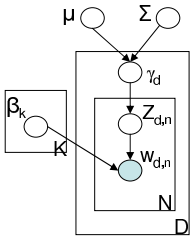

Figure 1: The Graphical Model.

where $p(w_n|z_n;\vec{\beta}_{1:K})$ and $p(z_n|\text{logistic}(\vec{\gamma}))$ are both multinomial distributions parameterized by $\vec{\beta}_{z_n}$ and $\vec{\theta} = \text{logistic}(\vec{\gamma})$, respectively.

3 Tight Approximate Inference and Learning on LoNTAM

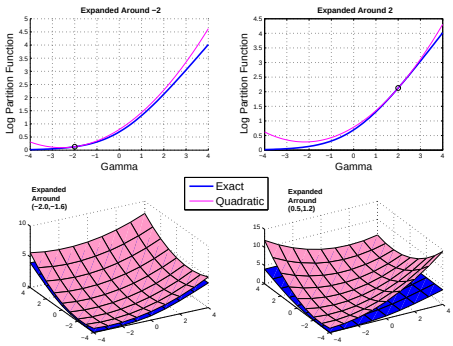

Figure 2: Approximating the log partition function using a truncated quadratic Taylor expansion. Top for K=2 and bottom for K=3.

3.1 Variational Inference Under Taylor-Approximated Conjugacy

Given a document $\mathbf{w}$, the inference task is to find the posterior distribution over the latent variables. Unfortunately, conditioned on $w_{1:N}$, the posterior over $\{\vec{\gamma}, z_{1:N}\}$ under LoNTAM is intractable. Therefore following a variational principle, such as in Xing *et al.* (2003), we approximate $p(\vec{\gamma}, z_{1:N}|\mathbf{w})$ with a product of simpler marginals, each on a cluster of latent variable subset, i.e., $\{\vec{\gamma}\}$ and $\{z_{1:N}\}$.

Based on the generalized mean field (GMF) theorem given in Xing *et al.* (2003), the optimal solution of each marginal, $q(\mathbf{X}_C|\Theta)$, over cluster of variables, $\mathbf{X}_C$, is isomorphic to the true conditional distribution of $\mathbf{X}_C$ given its *Markov blanket* (MB)– that is $p(\mathbf{X}_C|\mathbf{X}_{\text{MB}})$. This optimal variational marginal can be symbolically written down from the original joint model (such as our LoNTAM for $\{\vec{\gamma}, z_{1:N}, w_{1:N}\}$), except that in $q(\cdot|\cdot)$ we replace the $\mathbf{X}_{\text{MB}}$ by a set of "GMF messages" related to $\mathbf{X}_{\text{MB}}$. That is, $q^*(\mathbf{X}_C|\Theta) = p\big(\mathbf{X}_C|\text{GMF}(\mathbf{X}_{\text{MB}})\big)$. These GMF messages can be thought of as surrogates of the dependencies of $\mathbf{X}_C$ on $\mathbf{X}_{\text{MB}}$. Xing *et al.* (2003) showed that in case the joint model is a GLIM, the GMF messages correspond to an expectation of the *sufficient statistics* of the relevant Markov blanket variables under their own associated GMF cluster marginals. In the sequel, we use $\langle S_{\text{x}}\rangle_{q_{\text{x}}}$ to denote the GMF message due to latent variable x; thus the optimal GMF approximation to $p(\mathbf{X}_C)$ is:

$$q^*(\mathbf{X}_C) = p\big(\mathbf{X}_C|\langle S_{\text{y}}\rangle_{q_{\text{y}}} : \forall \text{y} \in \mathbf{X}_{\text{MB}}\big). \quad (4)$$

Inspecting our LoNTAM depicted in in Figure 1, we would approximate the posterior over $\{\vec{\gamma}, z_{1:N}\}$ using $q(\vec{\gamma}, z_{1:N}) = q_{\vec{\gamma}}(\vec{\gamma})q_{\text{z}}(z_{1:N})$. (The same approximating scheme was also adopted in Blei and Lafferty (2005), but as we explain soon, different techniques for seeking optimal $q_{\vec{\gamma}}(\cdot)$ and $q_{\text{z}}(\cdot)$ have led to remarkably different results.) Using Eq. (4), we can write down the the GMF approximation to the marginal posterior of the (inverse-logistic-transformed) topic vector $\vec{\gamma}$ as:

$$\begin{aligned} q_{\vec{\gamma}}(\vec{\gamma}) &= p(\vec{\gamma}|\mu, \Sigma, \text{z} \rightarrow \langle S_{\text{z}}\rangle_{q_z}) \\ &\propto p(\vec{\gamma}, \text{z} \rightarrow \langle S_{\text{z}}\rangle_{q_z}|\mu, \Sigma) \\ &\propto \mathcal{N}(\vec{\gamma}; \mu, \Sigma) \times p(\text{z} \rightarrow \langle S_{\text{z}}\rangle_{q_z}|\vec{\gamma}), \end{aligned} \quad (5)$$

where "$\rightarrow$" donates a symbolic replacement of the argument on the left with the one on the right. Note that Figure 1 makes explicit that the MB for $\vec{\gamma}$ is $\{z_{1:N}\}$, thus we replace $\{z_{1:N}\}$ with its GMF message, which corresponds to its expected sufficient statistics. It can be easily shown that the second term in Eq. 5 is given by:

$$p(\text{z} \rightarrow \langle S_{\text{z}}\rangle_{q_z}|\vec{\gamma}) = \exp\{\langle m\rangle_{q_z}\vec{\gamma} - N \times C(\vec{\gamma})\}, \quad (6)$$

where N is the number of words in $\mathbf{w}$, and $\langle m\rangle_{q_z}$ is a K-dimensional (row) vector that represents the expected histogram of topic occurrences in $\mathbf{w}$. More formally:

$$m_k = \sum_{i=1}^{N} I(z_i = k) \quad \text{and} \quad \langle m_k\rangle_{q_z} = \sum_{i=1}^{N} q_{\text{z}}(z_i = k). \quad (7)$$

Now we can have a glimpse of why LoNTAM is difficult to handle. First, the two factors in Eq. (5) are not conjugate, thus their product does not emerge as an easy close-form distribution such as a Gaussian; second, the second factor contains a nasty $C(\vec{\gamma})$, which is a complex function of the argument of our approximating distribution $q_{\vec{\gamma}}(\cdot)$. Thus $q_{\vec{\gamma}}(\vec{\gamma})$ as defined above is not integrable during inference (e.g. to calculate expectations of $\vec{\gamma}$ as output of our low-dimensional representation of the document, and as the GMF message sent to the $\{z_{1:N}\}$ cluster as needed bellow).

To circumvent the non-conjugacy and non-integrability of our variational cluster marginal caused by $C(\vec{\gamma})$, we introduce a truncated Taylor-appximotation to $C(\vec{\gamma})$ to make it algebraically manageable. The rational is that, in a multivariate Gaussian, we have only linear and quadratic terms of the argument. Inspecting the forms of the argument

(i.e, $\vec{\gamma}$) in the distribution defined by Eq. (5), all except $C(\vec{\gamma})$ are either linear or quadratic. If we can approximate $C(\vec{\gamma})$ up to a quadratic form, then the two factors in Eq. (5) will become "conjugate" and we can rearrange the resulting approximation to $q_{\vec{\gamma}}(\vec{\gamma})$ into a reparameterized multivariate Gaussian! Fortunately, this turns out to be feasible, and indeed it leads to a very general second-order approximate scheme superior to the tangent approximations underlying many extant variational inference algorithms.

Specifically, using a quadratic Taylor expansion of $C(\vec{\gamma})$ with respect to some $\hat{\gamma}$ [1],we have:

$$C(\vec{\gamma}) \approx C(\hat{\gamma}) + g'_{\vec{\gamma}}(\vec{\gamma} - \hat{\gamma}) + \frac{1}{2}(\vec{\gamma} - \hat{\gamma})' H_{\vec{\gamma}}(\vec{\gamma} - \hat{\gamma}), \quad (8)$$

where $\mathbf{g} = (g_1, \cdots, g_K)$ is the gradient and $\mathbf{H} = \{h_{ij}\}$ is the Hessian matrix of C w.r.t. $\vec{\gamma}$. Figure 2 demonstrates this expansion for K=2 and 3. As clear from the figure, this expansion provides a tight *local* approximation to $C(\vec{\gamma})$ for "practical" values of $\vec{\gamma}$. [2]

Combining Eqs. (5,6,8), it is easy to show that (see (Xing, 2005)) $q_{\vec{\gamma}}(\vec{\gamma})$ can now be expressed as a Gaussian $\mathcal{N}(\mu_{\hat{\gamma}}, \Sigma_{\hat{\gamma}})$ where:

$$\Sigma_{\hat{\gamma}} = \text{inv}\big(\Sigma^{-1} + NH(\hat{\gamma})\big), \quad (9)$$

$$\mu_{\hat{\gamma}} = \Sigma_{\hat{\gamma}}(\Sigma^{-1}\mu + NH(\hat{\gamma})\hat{\gamma} + \langle m \rangle_{q_z} - Ng(\hat{\gamma})). \quad (10)$$

Now we turn to $q_{\text{z}}(z_{1:N})$. Again from Figure 1, we see that the MB of $\{z_{1:N}\}$ is $\vec{\gamma} \cup \{w_{1:N}\}$, of which $\vec{\gamma}$ needs to be replaced by its GMF message. Thus we have:

$$q_{\text{z}}(z_{1:N}) = \prod_{n=1}^{N} q_{\text{z}}(z_n) = \prod_{i=1}^{N} p(z_n | \vec{\gamma} \to \langle S_{\vec{\gamma}} \rangle_{q_{\vec{\gamma}}}, w_n, \vec{\beta}_{1:K}). \quad (11)$$

For notational simplicity we drop the word-index "n" and give a generic formula for the variational approximation to a singleton marginal:

$$p(z^k | \langle S_{\vec{\gamma}} \rangle_{q_{\vec{\gamma}}}, w^j, \vec{\beta}_k) \propto p(z^k | \langle S_{\vec{\gamma}} \rangle_{q_{\vec{\gamma}}}) \times p(w^j | z^k, \vec{\beta}_k)$$

$$\propto \exp\{\langle \gamma_k \rangle_{q_{\vec{\gamma}}}\} \beta_{kj} = \exp\{\mu_{\hat{\gamma},k}\} \beta_{kj}, \quad (12)$$

where z^k and w^j are notational shorthands for $z = k$ (i.e., z picks the kth topic) and $w = j$ (i.e., w represents the jth word), respectively, β_{kj} is the probability of word j under topic k, and $\mu_{\hat{\gamma},k}$ is k^{th} component of the expectation of $\vec{\gamma}$ given in Eq. (10).

The above two GMF marginals given in Eqs. (9,10) and Eq. (12) are coupled and thus constitute a set of fixed-point equations. Thus, we can iteratively update each marginal until convergence. This approximation can be shown to minimize the KL divergence between the variational posterior and the true posterior of latent variables at convergence (Xing *et al.*, 2003). To diagnose convergence, one can either monitor the relative change in $\mu_{\hat{\gamma}}$ or the relative change of the log-likelihood of $\mathbf{w}$ (log of the integral in Eq. 3). Under our factorized approximation to the true distribution, the integral in Eq. (3) is computable (Xing, 2005).

3.2 Parameter Estimation via variational EM

Given a corpus of documents $\{\mathbf{w}_{1:D}\}$, the learning task is to find model parameters $\{\mu, \Sigma, \vec{\beta}_{1:K}\}$ that maximize the log likelihood of the data. We use a Variational Expectation-Maximization (VEM) algorithm to fit the model parameters. VEM alternates between two steps: in the E-Steps, the variational approximation in §3.1 is used to compute expectations over hidden variables $\{\vec{\gamma}_d, z_{d,1:N}\}_{1:D}$; then in the M-Step, model parameters are updated using their expected sufficient statistics from the E-step.

As it is always the case, details are important with VEM. As pointed out by Welling *et al.* (2004), strong conditional dependencies between hidden variables in directed models can seriously affect the model performance in VEM-based learning, because such dependendies lead to poor approximation to the posterior on the hidden variables, and can cause difficulties to escape local optima. To remedy this we used deterministic annealing (DA) (Ueda and Nakano, 1998).

DA-VEM seeks to maximize an exponentiated version of the likelihood, $E_{q(Y|X)}[p(X,Y)^T]$, where X and Y are observed and hidden variables respectively, and T is the temperature parameter. DA-VEM starts by maximizing a concave function (small values of T) and maintains local maxima while gradually morphing the function to the desired non-concave likelihood function when T=1. It is a continuation method in which the estimate at the end of each annealing step is used to initialize the search over the next step. In our experiments we used four annealing steps with temperatures $(0.1, 0.25, 0.5, 1)$. We also found that a fast annealing on $\vec{\beta}_{1:K}$ results in a faster convergence with no performance loss, thus we used the following temperature schedule for $\vec{\beta}_{1:K}$: $(0.25, 0.75, 1, 1)$.

3.3 Relation to Previous Work

Blei and Lafferty (2005) gave an alternative mean-field variational approximation to the same model. The factorized distribution in their work is:

$$q^{BL}(\vec{\gamma}, z_{1:N} | \lambda_{1:K}, \nu_{1:K}, \phi_{1:N}) = \prod_{k=1}^{K} q^{BL}(\gamma_k | \lambda_k, \nu_k) \prod_{n=1}^{N} q^{BL}(z_n | \phi_n), \quad (13)$$

[1] The $\hat{\gamma}$ is replaced with the mean of the variational distribution over $\vec{\gamma}$ from the pervious iteration.

[2] Empirically, values of $\vec{\gamma}$ outside of the "practical" range would result in a skewed $\vec{\theta}$ on the topic simplex. Our experimental results on real data sets confirm that the shown range is the operational one.

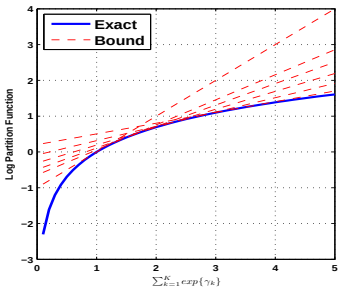

Figure 3: Upper bounds for the log partition function. The function is viewed as a logarithm function of a single parameter equals to $\sum_{k=1}^{K} \exp\{\gamma_k\}$ and each dotted line represents a linear upper bound.

where $\lambda_{1:K}, \nu_{1:K}$ and $\phi_{1:N}$ are the variational parameters. It is noteworthy that, compare to our approximate posterior of $\vec{\gamma}$ defined by Eq. (5), in this approximation the posterior of $\vec{\gamma}$ is a fully factored distribution over each element of $\vec{\gamma}$. Each $q^{BL}(\gamma_k|\lambda_k, \nu_k)$ is a univariate Gaussian. So in fact this approximate posterior does not capture correlations between elements in $\vec{\gamma}$, which might not result in a tight approximation during inference and learning.

With $q^{BL}(z_n|\phi_n)$ set to be a multinomial, the variational parameters $\{\lambda_{1:K}, \nu_{1:K}, \phi_{1:N}\}$ are fit by maximizing the following lower bound on the log-likelihood:

$$\log p(\mathbf{w}) \geq \mathrm{H}(q^{BL}) + \mathrm{E}_{q^{BL}}[\log p(\vec{\gamma}|\mu, \Sigma)] + \sum_{n=1}^{N} \Big(\mathrm{E}_{q^{BL}}[\log p(z_n|\vec{\gamma})] + \mathrm{E}_{q^{BL}}[\log p(w_n|z_n, \vec{\beta}_{1:K})] \Big), \quad (14)$$

where H is the entropy function. As discussed before, the expectation of the term $\log p(z_n|\vec{\gamma}) = z_n'\vec{\gamma} - C(\vec{\gamma})$ can not be computed due to the non-conjugacy of the logistic normal distribution. To deal with that, a first-order conjugate dual approximation (Jordan *et al.*, 1999) was used to upper bound $C(\vec{\gamma})$ as follows:

$$\log\Big(\sum_{k=1}^{K} \exp\{\gamma_k\}\Big) \leq \zeta^{-1}\Big(\sum_{k=1}^{K} \exp\{\gamma_k\}\Big) - 1 + \log(\zeta), \quad (15)$$

where ζ is a new variational parameter. Note that the bound in Eq. (15) views $C(\vec{\gamma})$ as a logarithm function of a unary parameter equals to $\sum_{k=1}^{K} \exp\{\gamma_k\}$. As shown in Figure (3), each value of ζ corresponds to an upper bound. The variational parameters are fit by maximizing the bound in Eq. (14), the maximizer $\phi^*_{1:N}$ and ζ^* has a closed form solution, whereas the maximizer $\lambda^*_{1:K}, \nu^*_{1:K}$ are found numerically using conjugate gradient and Newton's methods respectively.

There are two important differences between the aforementioned approach and our tight approximate inference presented here. First, q^{BL} posited that the posterior over $\vec{\gamma}$ has a *diagonal* covariance[3], which contradicts the original motivation of a "correlated topic model" as is LoNTAM, which is to capture correlations between topic weights. In particular, since model learning (e.g., for μ, Σ and $\vec{\beta}_{1:K}$) would iteratively uses the posterior estimation of $\vec{\gamma}$, a fully de-correlated estimator of $\vec{\gamma}$ as resulted from q^{BL} might possibly mislead the parameter estimation and eventually fail to *accurately* estimate the correlation over topics as introduced by the LoNTAM model. In our newly proposed method, we have no restriction on the covariance in our approximate posterior of $\vec{\gamma}$ (see Eq. 9). Second, the two approaches differ in the way they deal with $C(\vec{\gamma})$. In their work, $C(\vec{\gamma})$ was viewed as a unary function and was *upper bounded* using a tangent approximation. In contrast, we view $C(\vec{\gamma})$ as a multivariate function of $\vec{\gamma}$ and *approximate* it using a *multivariate* quadratic Taylor expansion. To make this difference clear, note that $C(\vec{\gamma})$ is used in two places: first to approximate the log-likelihood of $\mathbf{w}$ and second to fit the posterior distribution over $\vec{\gamma}$. For log-likelihood computation, viewing $C(\vec{\gamma})$ as a unary function is sufficient to get a close bound; however, for updating the posterior over γ during the variational fixed point iterations, it is important to keep the coupling between the components of $\vec{\gamma}$ as represented in $C(\gamma)$. In the early method, there is no clear way of how to deal with $C(\gamma)$ differently based on its (different) roles. However, in our work, we can deal with $C(\vec{\gamma})$ differently based on its role. As will be shown in the next section, modeling the interaction between the components of $\vec{\gamma}$ via the multivariate quadratic Taylor expansion leads to much tighter approximation and faster convergence.

4 Experimental Results

We validate our inference algorithms on both simulated text corpus (sampled according to hand-specified topics and admixing priors); and the NIPS dataset.

4.1 Experiments on Simulated Data

We first tested our approach (referred to as AX) over controlled settings (where the ground truth is known) and compared it to that of Blei and Lafferty (2005) (refereed to as BL). To test the accuracy of both inference algorithms, different settings were simulated by varying one of the three model aspects — 1) K: number of topics; 2) M: size of the vocabulary; and 3) N: number of words per document) — while fixing the other two. The model parameters $\{\mu, \Sigma, \vec{\beta}_{1:K})\}$ were drawn randomly in each case from some prespecified distributions. For each setting, 200 documents were sampled and we run both algorithms under each model setting until the relative change in the bound of the log-likelihood is less than 10^{-6}.

[3]We believe that this assumption was made to make possible that $\nu_{1:K}$ can be fit numerically.

Accuracy of posterior inference: As shown in the first row of Figure 4, our approach achieves higher accuracy in recovering the true $\vec{\theta}$ (the logistic transformation of $\vec{\gamma}$) simulated for each document across all settings, using the posterior mean of $\vec{\gamma}$ inferred from $\mathbf{w}$. As we noted in Section 3.2, this is due to the tightness of the *multivariate* quadratic approximation we use. As N increases (i.e. the longer the document), the task becomes easy and the difference in performance between the two approaches decreases. This is because in longer document $\langle m \rangle_{q_z}$ (the expected topic histogram) becomes the dominant factor in recovering $\vec{\gamma}$. The second row in Figure 4 shows the *absolute* difference in the error in recovering $\vec{\theta}$ between the two approaches, that is $\text{Error}(BL) - \text{Error}(AX)$ on a per document level. Our approach always results in an improvement in the order of 10% absolute difference.

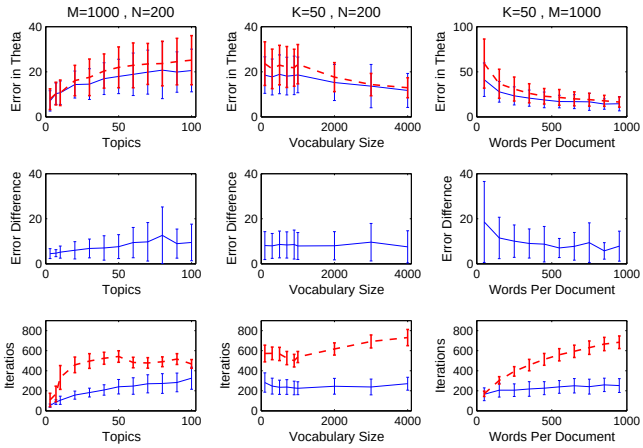

Figure 4: Inference On Simulated Data. Dotted and solid lines correspond to the BL and AX approaches respectively. Each column represents an experiment in which one dimension is varied.**Top row**:Average L2 error in recovering $\vec{\theta}$. **Middle row**:Error difference (L2(BL)-L2(AX)) in recovering $\vec{\theta}$ on a per document level.**Bottom row**: Number of iterations needed by each approach to converge.

Convergence rate: The third row in Figure 4 shows the number of iterations consumed by each algorithm until the bound converges. Again, our approach converges significantly faster in almost all model settings. It is interesting to note how convergence is affected when K is fixed and the other model aspects are varied. One might expect that as $\vec{\theta}$ scales only with K, the other aspects should have little effect on the convergence of the algorithm. Indeed this was roughly the case with our approach, yet for the BL approach, the number of iterations until convergence is highly affected by other aspects, especially N. To understand why this happens, we need to examine the messages communicated during the fixed point update equations. Changes in the posterior mean over $\vec{\gamma}$ are propagated exponentially to the posterior over z (Eq. (12)), then summed up over all $z_{1:N}$ and propagated back to $\vec{\gamma}$ via $\langle m \rangle_{q_z}$. Thus as N increases, this effect becomes more pronounced and prolongs the time needed until convergence. The reason why our approach does not suffer from this effect is because the use of the quadratic approximation damps the update in the posterior over $\vec{\gamma}$ quickly, and future iterations only fine tune it. Hence even when N increases, the compound effect described above does not result in a large difference between the $\langle m \rangle_{q_z}$ messages sent from the $z_{1:N}$ to γ across iterations.

One important point to be noted here is the cost of each iteration in the two approaches. Our approach scales as $O(K^3)$ per iteration due to matrix inversion in Eq. 9, while the BL approach scales as $O(LK^2)$ per iteration, where L is the number of derivative evaluations for the numerical optimization routines[4].

Parameter estimation: Does more accurate inference result in better parameter estimation? To answer this question, we started by a toy problem. Model dimensions were fixed as: $K = 3$, $N = 200$ and $M = 32$, other model parameters were generated randomly. The ground truth (topic distribution and the shape of the LN-density over the 3-topic simplex) is depicted in the top of Figure 5. We sampled 400 documents from this model and run both approaches on it until the relative change of the bound on the log-likelihood is less than 10^{-3}.[5] The estimated topics and LN-density shapes over the simplex are given in the Figure 5.

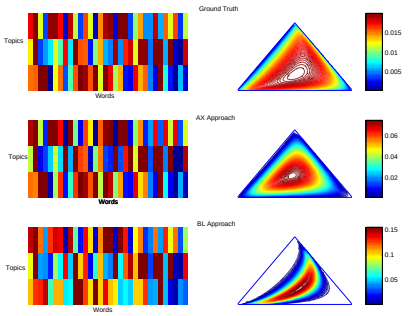

Figure 5: Parameter Estimation. Left panels represent topic distributions where each row is a topic, each column is a word, and colors correspond to probabilities. Right panels represent shapes of LN distribution over the simplex. Top row gives the ground truth model parameters, while middle and bottom rows give those estimated using the AX and BL approach respectively.

As shown in Figure 5, our approach results in a much more accurate density estimation over the simplex, and the topic-specific word frequencies $\vec{\beta}_{1:3}$. In fact, the BL approach puts no mass at areas where the ground truth puts high probability mass. The esti-

[4]We avoided comparing the two approaches using wall time because our code is written in matlab while the BL code we compare against is written in C++.

[5]Unless otherwise stated, this is the convergence criteria used for parameter estimation in the rest of this section.

mated topics by both models are acceptable, yet our approach was able to get more accurate results (note the difference in the bottom topic). To better understand this result, in Figure 6 we depict the ground truth $\vec{\theta}$ (represented by a vertical line which is partitioned into colored segments in proportion to the topic weights recorded by $\vec{\theta}$) and that recovered by each approach when the VEM algorithm converged over the 400 documents. Recall that the expected sufficient statistics for $\beta_{1:K}$ depend on the variational distribution over $z_{1:N}$. In turn the distribution over $z_{1:N}$ depends on $\vec{\theta}$. Thus the quality of the estimated topics depends on how well $\vec{\theta}$ is approximated. In fact the error in recovering θ was 13% for our approach and 19% for the BL one, which explains why both approaches got comparable topic estimates (the KL divergence between the estimated and true topic distribution is 0.02 for our approach and 0.09 for BL).

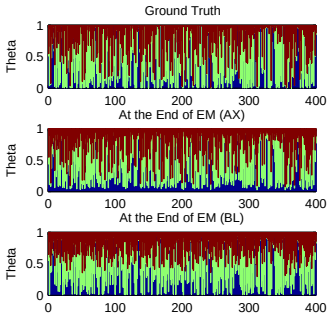

Figure 6: The Recovered Document Topic Mixing Vectors. Each vertical line represent a document mixing vector and each color corresponds to a topic.

In contrast, estimation of the LN parameters, and hence its density over the simplex, depends more directly on how well the components in the $\vec{\theta}$ vector are recovered, not just the overall error in recovering $\vec{\theta}$. As clear from Figure 6, our approach results in recovering finer details of the components in the $\vec{\theta}$ vector than the BL does — the difference can be easily seen by inspecting the red component (the top one).

4.2 Experiments on the NIPS Dataset

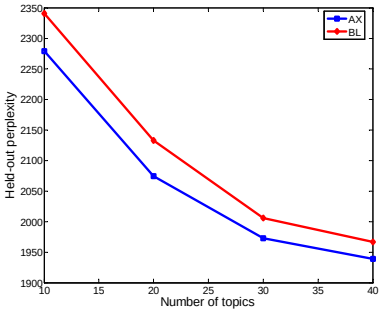

Figure 7: Test-set perplexities on the NIPS dataset.

In addition to simulation study, we conducted experiments on the NIPS17 dataset which contains the proceedings of the NIPS conference from 1988 to 2003.

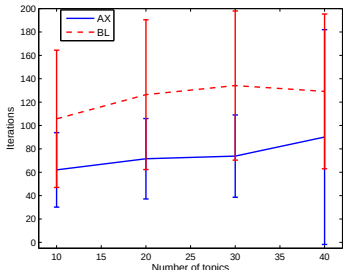

Figure 8: Number of Iterations to converge when performing inference on the held out documents in the NIPS17 collection.

This corpus has 2484 documents, vocabulary size (M) of 14036 words (after removing stop words), and and average of 1320 words per document. The dataset were divided into 2220 documents for training and 256 ones for testing. We fitted 4 models to this corpus with $K = 10, 20, 30, 40$ topics. We compared the two approached based on perplexity on the held out testset, where the perplexity of a test document $\mathbf{w}_{test}$ is defined as:

$$\text{Perplexity}(\mathbf{D}_{test}) = \exp\Big\{-\frac{\sum_{n=1}^{|\mathbf{D}_{test}|}\log p(w_n)}{\sum_{n=1}^{|\mathbf{D}_{test}|}|\mathbf{w}_n|}\Big\} \quad (16)$$

To avoid comparing bounds, the true marginal log-likelihood was estimated using importance sampling where each approach's posterior distribution over $\vec{\gamma}$ was used as the proposal. Figure 7 summarizes the results which show that we achieve better testset perplexities across all topics. The reason for this slight improvement over this dataset is due to the large number of words in each documents which reduces the effect of accurate prior estimation — similar results were explained in Section 4.1 with reference to Figure 4. In Figure 8 we depict the average number of iterations needed by each approach to converge when performing inference over the testset, which shows that our approach converges faster in terms of the number of iterations.

4.3 Experiments on the PNAS Dataset

We also compared the two approaches on a 4-way classification task over the abstracts of the Proceeding of the National Academy of Science (PNAS). These abstracts are labeled according to their scientific category. We selected 2500 abstracts from the period of 1997 to 2002 and we fitted 40 topics to the resulting corpus. We then used the resulting low dimensional topic mixing vectors induced by each approach as features for classification. Out of those 2500 abstracts, we only selected those having the required categories which results in 962 abstracts. We Then trained an SVM classifier over 85% of those selected abstracts and tested the accuracy over the remaining 15% ones. We present the classification accuracy of both approaches

in Table 1. As clear from this table, our approach results in a more accurate classifier. It should be noted that the improvement over the BL approach in this dataset is significant, as opposed to the slight improvement in perplexities over the NIPS dataset. This is due to the relatively small number of words per abstract in the PNAS datset — which was an average of 170 words per abstract. These results conform with those from the simulation study we conducted, and fully analyzed, in Section 4.1. Furthermore, inspecting the confusion matrix of both classifiers, we found that most of the errors in the BL approach were due to confusing the Biochemistry and Biophysics abstracts. In figure 9 we depict the low dimensional representation of the abstracts in both of these classes, as recovered by the two approaches. As it is clear from the figure, our approach results in a better separation of these two seemingly similar classes.

Table 1: Document classification accuracies

Category	Doc	BL	AX
Genetics	21	61.9	61.9
Biochemistry	86	65.1	77.9
Immunology	24	70.8	66.6
Biophysics	15	53.3	66.6
Total	146	64.3	72.6

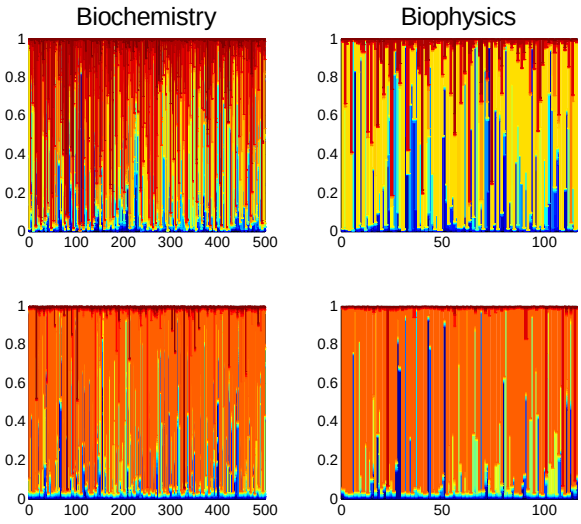

Figure 9: The reduced representation of abstracts in Biochemistry and Biophysics. Each line represents a document, and each color segment represents a topic contribution as recorded by $\vec{\theta}$. **Top row**: the AX approach; **Bottom row**: the BL approach.

5 Conclusions

In this paper we presented a novel approximate inference algorithm for the Logistic-Normal Topic Admixture Model. Our approach overcomes technical difficulties for inference/learning due to the non-conjugacy within the model via the use of a *multivariate* quadratic Taylor approximation to LN. Our method not only makes the variational fixed point equations for inference amenable to analytic closed-form solution, but also keeps the coupling between the components in the per document topic mixing vector. This results in a simple yet efficient tight approximate inference algorithm that enjoys nice representational and convergence properties.

We presented experimental results on simulated datasets as well as on the NIPS17 collection and the PNAS collection, and contrasted our approach with that given by Blei and Lafferty (2005). The results demonstrated that our approach results in tighter approximation in inference and learning especially when the number of words per document is relatively small.

Acknowledgements

We would like to thank David Blei for providing his code that we compared against in this paper, and Edo Airoldi for providing the PNAS data set. We would like also to thank the anonymous reviewers for their helpful comments and suggestions.

References

J. Aitchison and S. Shen. Logistic-normal distributions: Some properties and uses, Biometrika 67, 1980.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:9931022, January 2003.

D. Blei and J. Lafferty. Correlated topic models. In NIPS 18 , 2006

E. Erosheva, S. Fienberg, and J. Lafferty. Mixed membership models of scientific publications. PNAS, Vol. 101, Suppl. 1, April 6, 2004

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan (Ed.), Learning in Graphical Models, Cambridge: MIT Press, 1999.

J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. Genetics, 155:945959,June 2000.

J. Sivic, B. Rusell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. Technical report, CSAIL, MIT, 2005

N. Ueda and R. Nakano. Deterministic annealing EMalgorithm. Neural Networks, 11, 271282, 1998.

M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In NIPS 17, 2004.

E. Xing, M. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In Proc. of the 19th UAI, pages 583:591, 2003.

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

B. Zhao and E. Xing, BiTAM: Bilingual Topic AdMixture Models for Word Alignment, The joint conference of the Association for Computational Linguistics, (ACL 2006).