ÉCOLE POLYTECHNIQUE

INTERNSHIP REPORT

Correlated Topic Models with Conjugate-Computation Variational Inference

Author:
Salma EL ALAOUI TALIBI

Supervisor:
Dr. Emtiyaz KHAN

RIKEN Center for Advanced Intelligence Project, Tokyo
Approximate Bayesian Inference Team

June - September 2017

Abstract

Correlated Topic Models with Conjugate-Computation Variational Inference

The Correlated Topic Model is an extension of LDA, which models the topic correlation pattern through a non-conjugate logistic-normal prior. The CTM posterior is intractable, and the logistic-normal prior distribution induces a non-conjugate term in the model. Posterior inference is therefore challenging to perform using MCMC methods, and the variational inference algorithm developed by [Blei and Lafferty, 2007] presents some inefficiencies.

We derive a Conjugate-Computation variational inference algorithm for the CTM. CVI is an efficient and modular method for variational inference in non-conjugate models. It allows us to convert inference in the CTM into inference over two conjugate models: LDA and a linear model. This is advantageous as it enables us to use already-existing implementations of the two simpler models.

We test the algorithm on several corpora, and confirm that modeling the correlations between topics improves the predictive performance, and gives a higher accuracy in classification tasks. We also show that the CVI algorithm converges much faster to similar or better solutions compared with the traditional variational inference method for CTM. We further improve the convergence speed of the CVI algorithm by deriving a stochastic version that scales to large collections of documents.

Contents

List of Figures

Chapter 1

Introduction

The availability of large collections of documents provides an interesting opportunity for building tools to intelligently navigate these archives. Indeed, the scale and growth of these collections requires automating the tasks of annotating the documents, organizing them and retrieving content.

Topic models uncover the underlying topical structure of unstructured document collection by performing hierarchical Bayesian analysis of the original texts [Blei and Lafferty, 2009]. Examples of their application to various categories of documents include scientific abstracts [Griffiths and Steyvers, 2004], scientific journal archives [Blei and Lafferty, 2007] and newspaper corpora [Wei and Croft, 2006].

Latent Dirichlet Allocation (LDA) [Blei et al., 2003] is one of the most popular topic models to infer a semantic structure from text corpora. However, LDA uses a conjugate Dirichlet prior for the topic proportion, for computational convenience in deriving the posterior inference algorithm. The Dirichlet distribution implies that the occurrences of topics are independent, and therefore fails to capture the potentially rich topic correlations: an article about Neuroscience is more likely to also be about Medicine than about Geology. The Correlated Topic Model (CTM) [Blei and Lafferty, 2007] is a more flexible extension of LDA, which uses a non-conjugate logistic-normal prior, allowing to explicitly model correlation patterns with a Gaussian covariance matrix. As in LDA, the CTM posterior is intractable, and has to be approximated, but the enhanced expressiveness of CTM comes at the cost of a more complex procedure for approximate posterior inference. Indeed, the non-conjugacy renders usage of Gibbs sampling [Griffiths and Steyvers, 2004] impossible; while using a Metropolis-Hastings based simulation is challenging because of the large scale of the data.

We therefore resort to a variational inference approach to approximate the posterior distribution in the correlated topic model. The variational inference algorithm derived in [Blei and Lafferty, 2007], which performs coordinate ascent optimization, is inefficient because it requires performing a costly gradient-based optimization several times for each document in the collection. Furthermore, it analyses the whole dataset at each iteration, which is impractical for large datasets.
In this work, we present a more efficient algorithm, by deriving Conjugate-Computation variational inference (CVI) [Khan and Lin, 2017] for the correlated topic model. This methods combines (stochastic) gradients for the non-conjugate terms, while retaining the conjugate computations on the conjugate terms.

This report is organized as follows: in the first chapter, we introduce variational inference, and discuss its efficiency on models containing both conjugate and non-conjugate terms. We then present the Conjugate-computation variational inference and its extension to the mean-field approximation.

In the second Chapter, we provide the relevant background on topic models by presenting the traditional variational algorithm for LDA, and introducing the correlated topic model.
In the third chapter, we derive the CVI algorithm for the correlated topic model, and show that it reduces inference in the non-conjugate CTM model to inference over two conjugate, simpler models: LDA and a linear model. We then present a stochastic version of the CVI algorithm that scales to large collections of documents.
We conclude by comparing the performance of the algorithms on several datasets. We confirm that modeling correlation patterns improves document representation and gives more accurate predictions in classification tasks. We also show that the CVI algorithm converges much faster to solutions as good or better than those found by the coordinate ascent optimization algorithm of [Blei and Lafferty, 2007]. Finally, we show that the stochastic version of CVI significantly increases the convergence speed of batch CVI.

Chapter 2

Conjugate and Non-Conjugate Variational Inference

In this chapter [1], we first introduce Variational Inference (VI), a family of approximation techniques for evaluating the intractable posterior distribution in probabilistic models. We then define the notion of conjugacy and discuss the efficiency of the existing variational inference techniques on models containing conjugate and non-conjugate terms. Finally, we present Conjugate-Computation Variational inference, a new algorithm for Bayesian inference in non-conjugate models, which doesn't suffer from the issues encountered in many of the state-of-the-art methods.

2.1 Variational Inference

This section is based on [Bishop, 2006] and [Blei et al., 2016], and introduces *Variational Inference* (VI), a method widely used to approximate difficult-to-compute posterior probabilities in Bayesian models.

2.1.1 The approximate inference problem

We consider a Bayesian model, where $\mathbf{y}$ are the observations and $\mathbf{z}$ the latent variables. The joint density over $\mathbf{z}$ and $\mathbf{y}$ is given by Bayes' theorem:

$$p(\mathbf{z}, \mathbf{y}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \tag{2.1}$$

where $p(\mathbf{z})$ is the prior distribution on $\mathbf{z}$ and $p(\mathbf{y}|\mathbf{z})$ is the likelihood, which relates the observations to the hidden variables.

In Bayesian statistics, inference about unknown quantities amounts to computing the posterior distribution $p(\mathbf{z}|\mathbf{y})$, whose formula is also given by Bayes' theorem:

$$p(\mathbf{z}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{y})} \tag{2.2}$$

where the marginal distribution in the denominator, also called the *evidence*, is defined as $p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{z}) d\mathbf{z}$.

For many models of practical interest, this integral is unavailable in closed form, or prohibitively expensive to compute because of the high dimensionality of the latent space. The intractability of the posterior distribution makes inference hard and thus requires the use

[1] This chapter was co-authored with Kimia Nadjahi, a fellow intern at RIKEN.

of approximation schemes.

One important category of approximate inference methods gathers the *Monte Carlo Markov Chain* (MCMC) techniques, where the posterior is approximated through numerical sampling. Although this approach has been widely studied and applied, it has some major limitations in practice: sampling methods require expert level knowledge for finding an appropriate sampling scheme or diagnosing the convergence, and can be computationally demanding. By using optimization instead of sampling for approximating the posterior, Variational Inference offers a deterministic alternative to MCMC, with potentially biased but efficient algorithms whose convergence is easier to detect. Furthermore, these algorithms tend to be faster and scale better for large data sets, as pointed out in [Blei et al., 2017].

2.1.2 From inference to optimization

The main idea behind Variational Inference is to turn the inference problem into an optimization problem. Let $\mathcal{P}_\Lambda = \{q(\mathbf{z}|\boldsymbol{\lambda}) \mid \boldsymbol{\lambda} \in \Lambda\}$ be a family of densities over the latent variables $\mathbf{z}$ with $\boldsymbol{\lambda}$ the variational parameters. The goal is to find the distribution $q(\mathbf{z}|\boldsymbol{\lambda}^*) \in \mathcal{P}_\Lambda$ that approximates the best the intractable posterior distribution $p(\mathbf{z}|\mathbf{y})$. The similarity between the two densities is measured with the Kullback-Leibler (KL) divergence, so the optimal density $q(\mathbf{z}|\boldsymbol{\lambda}^*)$ is given by:

$$q(\mathbf{z}|\boldsymbol{\lambda}^*) = \arg\min_{q(Z|\lambda)\in\mathcal{P}_\Lambda} \mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, p(\mathbf{z}|\mathbf{y})] \tag{2.3}$$

where

$$\mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, p(\mathbf{z}|\mathbf{y})] = \int q(\mathbf{z}|\boldsymbol{\lambda}) \log \left\{ \frac{q(\mathbf{z}|\boldsymbol{\lambda})}{p(\mathbf{z}|\mathbf{y})} \right\}$$

The family of densities $\mathcal{P}_\Lambda$ should be flexible and rich enough to provide a good approximation to the exact posterior distribution, but also sufficiently simple so that the optimization problem could be efficiently solved. The choice of a set of variational parameters Λ is a way to restrict the family of approximating distributions.

However, the optimization problem 2.3 is intractable because we still need to compute $\log p(\mathbf{y})$. The dependence on $p(\mathbf{y})$ appears by re-writing the KL divergence and expanding the conditional as follows:

$$\begin{aligned}\mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, p(\mathbf{z}|\mathbf{y})] &= \mathbb{E}_q[\log q(\mathbf{z}|\boldsymbol{\lambda})] - \mathbb{E}_q[\log p(\mathbf{z}|\mathbf{y})] \\ &= \mathbb{E}_q[\log q(\mathbf{z}|\boldsymbol{\lambda})] - \mathbb{E}_q[\log p(\mathbf{z},\mathbf{y})] + \log p(\mathbf{y})\end{aligned}$$

Therefore, instead of minimizing the KL divergence, we want to maximize the *evidence lower bound* (ELBO), defined as:

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q[\log p(\mathbf{z},\mathbf{y})] - \mathbb{E}_q[\log q(\mathbf{z}|\boldsymbol{\lambda})] \tag{2.4}$$

Indeed, this alternative objective is equal to the KL divergence plus a term that doesn't depend on $q(\mathbf{z}|\,\boldsymbol{\lambda})$, so maximizing the ELBO in 2.4 with respect to $\boldsymbol{\lambda}$ is equivalent to minimizing the KL term.

The name of the ELBO comes from the fact that, since the KL divergence is non-negative:

$$\begin{aligned}\log p(\mathbf{y}) &= \mathcal{L}(\boldsymbol{\lambda}) + \mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, p(\mathbf{z}|\mathbf{y})] \\ &\geq \mathcal{L}(\boldsymbol{\lambda})\end{aligned}$$

2.1.3 The mean-field variational family

As mentioned above, the choice of the family of distributions $\mathcal{P}_\Lambda$ is crucial since it determines the complexity of the optimization and thus the quality of the approximation.

A flexible family is obtained by assuming that the latent variables $\mathbf{z}$ can be partitioned into independent ones $\mathbf{z}_j$, $j = 1, ..., N$, so that the variational distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ factorizes:

$$q(\mathbf{z}|\boldsymbol{\lambda}) = \prod_{j=1}^{N} q_j(\mathbf{z}_j|\boldsymbol{\lambda}_j) \tag{2.5}$$

Each latent variable $\mathbf{z}_j$ follows its own variational density $q_j(\mathbf{z}_j|\boldsymbol{\lambda}_j)$ characterized by the set of variational parameters $\boldsymbol{\lambda}_j$.

This case is known as the *mean-field approximation*, and the family of factorized variational distributions is called the *mean-field variational family*.

One common method to maximize the ELBO in this situation is the *Coordinate Ascent Variational Inference* (CAVI), which iteratively optimizes each factor of the variational distribution while holding the others fixed. By writing the objective as a function of $q_j(\mathbf{z}_j|\boldsymbol{\lambda}_j)$ and maximizing it with respect to this factor, we obtain a general expression for the optimal solution $q_j(\mathbf{z}_j|\boldsymbol{\lambda}_j^*)$:

$$q_j(\mathbf{z}_j|\boldsymbol{\lambda}_j^*) \propto \exp\{\mathbb{E}_{q_i, i\neq j}[\log p(\mathbf{y}, \mathbf{z})]\} \tag{2.6}$$

The algorithm thus initializes all of the factors q_j appropriately, and cycles through the factors at each iteration, updating each in turn with the formula given in Equation 2.6.

2.1.4 Gradient and stochastic gradient descents

The evidence lower bound can be maximized with the following gradient ascent algorithm:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \nabla_\theta \mathcal{L}(\boldsymbol{\lambda}_t) \tag{2.7}$$

where t is the iteration number, ρ_t the learning rate and $\nabla_\theta \mathcal{L}(\boldsymbol{\lambda}_t)$ the gradient of the ELBO with respect to $\boldsymbol{\lambda}$ evaluated at $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$.

This method can be very inefficient, especially when we have to deal with large datasets: evaluating the gradient on all samples of the training set can be expensive. To ensure scalability, we can use a stochastic-gradient method instead: suppose that $\mathcal{L}(\boldsymbol{\lambda})$ can be written as $\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_x[\ell(\boldsymbol{\lambda}, \mathbf{x})]$ where $\mathbf{x}$ is a random variable ; then, the stochastic-gradient ascent algorithm updates $\boldsymbol{\lambda}$ as follows:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \nabla_\theta \ell(\boldsymbol{\lambda}_t, \mathbf{x}_i) \tag{2.8}$$

where $\mathbf{x}_i$ is a random sample of $\mathbf{x}$.

This converges to a maximum of $\mathcal{L}(\boldsymbol{\lambda})$ when the learning rate ρ_t follows the Robbins-Monro conditions:

$$\sum_{t=1}^{\infty} \rho_t = \infty, \quad \sum_{t=1}^{\infty} \rho_t^2 < \infty$$

In this method, the gradient points in the direction of the steepest ascent, which implicitly depends on the Euclidean distance metric in the space of valid parameters Λ. However, this metric between two parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$ can be a poor measure of dissimilarity between the distributions they parametrize, $q(\mathbf{z}|\boldsymbol{\lambda})$ and $q(\mathbf{z}|\boldsymbol{\lambda}')$, as illustrated in [Hoffman et al., 2013]. This motivates the use of the natural gradient introduced in [Amari, 1998] to optimize the ELBO: distances are now defined with the KL divergence, which depends only on the properties of the distributions and not their parameters. The natural gradient then points in the direction of steepest ascent in Riemannian space instead of Euclidean, and as shown in [Amari, 1998], it can be obtained by multiplying the gradient of the objective function by the inverse of the Fisher information matrix of $q(\mathbf{z}|\boldsymbol{\lambda})$.

Because of their generality, both approaches can be applied to a wide range of inference problems as black-box methods. However, the stochastic-gradient method have some important limitations: it might not exploit the closed-form expressions that the ELBO potentially contains, and its efficiency highly depends on the parameterization of the variational distribution. We discuss these issues in more details in the next sections.

2.2 Conjugate vs Non-Conjugate Variational Inference

We introduce the notion of conjugacy and discuss the efficiency of Variational Inference in conjugate and non-conjugate models. Basically, conjugacy defines a posterior distribution that can be obtained through simple computations, extensively used to make inference. However, Variational Inference becomes computationally challenging with the presence of non-conjugate terms.

2.2.1 Conjugate models

Definition. *Conjugate models* refer to probabilistic graphical models where the prior distribution $p(\mathbf{z})$ is *conjugate* to the likelihood $p(\mathbf{y}|\mathbf{z})$.

We define *conjugacy* based on Chapter 2 of [Gelman et al., 2014]: suppose $\mathcal{F}$ is the class of data distributions $p(\mathbf{y}|\mathbf{z})$ parameterized by $\mathbf{z}$, and $\mathcal{P}$ is the class of prior distributions for $\mathbf{z}$, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if:

$$p(\mathbf{z}|\mathbf{y}) \in \mathcal{P}, \quad \forall p(\cdot|\boldsymbol{\theta}) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P} \tag{2.9}$$

Conjugate computations. Such models significantly simplify the computation of the posterior distribution. Indeed, the posterior has the same functional form as its prior distribution, so inference consists in learning the values of its parameters – and not the functional form. These parameters are available in closed-form and can easily be obtained through efficient and simple computations that we call *conjugate computations*.

Many Variational Inference techniques make use of these conjugate computations to approximate the posterior because of their high computational efficiency. One of these methods is called *Stochastic Variational Inference* (SVI) and is described later.

We give an example of one type of conjugate computations with the *conjugate exponential-family* model: we consider an exponential-family prior distribution, $p(\mathbf{z}) = h(\mathbf{z}) \exp\left[\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\eta} \rangle - A(\boldsymbol{\eta})\right]$ where $\boldsymbol{\eta}$ is the natural parameter, $\boldsymbol{\phi}$ the sufficient statistics, $\langle \cdot, \cdot \rangle$ an inner product, $h(\mathbf{z})$ the base measure, and $A(\boldsymbol{\eta})$ the log-partition function. In a conjugate model, the associated

likelihood also belongs to the exponential family:

$$p(\mathbf{y}|\mathbf{z}) = \exp\left[\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\eta}_{yz}(\mathbf{y})\rangle - f_y(\mathbf{y})\right]$$

where $\boldsymbol{\eta}_{yz}$ and f_y are functions of $\mathbf{y}$ only. The posterior distribution in such a model takes the following exponential form:

$$p(\mathbf{z}|\mathbf{y}) \propto h(\mathbf{z}) \exp\left[\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\eta} + \boldsymbol{\eta}_{yz}(\mathbf{y})\rangle\right] \tag{2.10}$$

We see that the computation of the posterior distribution is very easy since it is completely determined by its natural parameter, which is only the sum of the natural parameter of the prior and the likelihood's one.

Conjugate exponential-family models come up frequently in Bayesian statistics and statistical machine learning. Examples include Latent Dirichlet Allocation [Blei et al., 2003], Probabilistic Matrix Factorization Models [Mnih and Salakhutdinov, 2008], Probabilistic Principal Component Analysis [Bishop, 1999] among many others.

SVI. Stochastic Variational Inference [Hoffman et al., 2013] is a scalable algorithm for approximating the posterior distribution in conjugate exponential-family models with local and global hidden variables. The variational distribution is assumed to be in the mean-field family and each factor in the same exponential family as the associated conditional distributions, such that we can apply the coordinate ascent algorithm whose updates 2.6 are available in closed-form due to conjugacy. However, this algorithm becomes inefficient with large data sets in such models, since it has to optimize the local variational parameters for each data point before improving the global ones that are randomly initialized.

SVI solves this problem by performing stochastic natural gradient ascent to optimize the lower bound: at each iteration, the algorithm samples one data point from the training dataset, optimizes its local variational parameters, forms intermediate global parameters estimate using classical coordinate ascent updates, and sets the new global parameters to a convex combination of the old estimate and the intermediate ones. Since the parameters are updated using only a single data point or a small mini-batch and calculations from classical coordinate inference, SVI is an efficient and easy-to-configure algorithm. The choice of the natural gradient over the traditional one is motivated in the previous section: it accounts more for the differences between two densities rather than on how they are parameterized.

2.2.2 Non-conjugate models

Challenges with non-conjugacy. When the model also contains non-conjugate terms, methods that are specifically designed for conjugate models lose their computational efficiency. For example, the closed form coordinate updates required by SVI to compute noisy natural gradients are not valid anymore when the conditional distributions are not conjugate. The evidence lower bound might become intractable and thus hard to optimize because of the presence of non-conjugate terms.

This problem is for instance encountered in Gaussian Process (GP) models with a non-Gaussian likelihood: we consider N output pairs $\{y_n, \mathbf{x}_n\}$ indexed by n and the latent function $z_n = f(\mathbf{x}_n)$ drawn from a GP with mean 0 and covariance $\mathbf{K}$. The likelihood $p(y_n|z_n)$ is defined as non-Gaussian (e.g., logistic) and the joint distribution is given by:

$$p(\mathbf{y}, \mathbf{z}) = \left[\prod_{n=1}^{N} p(y_n|z_n)\right] \mathcal{N}(\mathbf{z}|0, \mathbf{K})$$

A common choice for the variational distribution used to approximate the posterior is a Gaussian distribution: $q(\mathbf{z}|\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})$, whose mean $\mathbf{m}$ and covariance $\mathbf{V}$ need to be estimated. In this case, the evidence lower bound is defined as:

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{n=1}^{N} \mathbb{E}_q[\log p(y_n|z_n)] - \mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, \mathcal{N}(\mathbf{z}|0, \mathbf{K})]$$

The difficulty stems from the likelihood terms. Indeed, while the KL divergence between two Gaussian distributions has an analytic expression, the expectation $\mathbb{E}_q[\log p(y_n|z_n)]$ doesn't have a closed-form expression for most non-Gaussian likelihoods. Hence, Variational Inference is computationally challenging because of the non-conjugate terms in the model.

Stochastic-gradient methods. To deal with this issue, many stochastic-gradient methods have been recently proposed ([Salimans and Knowles, 2013], [Ranganath et al., 2014], [Titsias and Lázaro-gredilla, 2014]). This approach relies on the fact that despite its intractability, the lower bound can still be optimized with a stochastic-gradient descent algorithm:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \widehat{\nabla}_\lambda \mathcal{L}(\boldsymbol{\lambda}_t)$$

where the notations were introduced in the section about Variational Inference. The main difference between the existing stochastic-gradient methods is in the way the noisy gradient $\widehat{\nabla}_\lambda \mathcal{L}(\boldsymbol{\lambda}_t)$ is being defined.

The key advantage of this approach is that it doesn't require significant model-specific analysis, unlike the methods used in conjugate models, which are based on a closed-form solution, or some Variational Inference algorithms designed for non-conjugate models, which require model specific computations ([Minka and Lafferty, 2002], [Wang and Blei, 2013]). In that sense, it can be applied to a wide-variety of models with little additional derivation.

Limitations. However, although the stochastic-gradient method provides generality and scalability, it has major limitations. First, it is not always computationally efficient and modular: the lower bound might contain conjugate terms with a closed-form expression, that the stochastic-gradient computation will approximate instead of directly exploiting it. Then, the efficiency and rate of convergence are sometimes dependent on the parameterization of the variational distribution, and choosing the parameterization that leads to the simplest updates is complicated.

In the next section, we derive the CVI algorithm, an alternative solution to make inference in models with both conjugate and non-conjugate terms and which doesn't suffer from these issues.

2.3 Conjugate-computation Variational Inference (CVI)

2.3.1 Motivation

We have seen that in many Bayesian models, the prior distribution is chosen to be conjugate to the likelihood, so as to guarantee that the posterior distribution can be obtained in closed form, or computed efficiently.

Other useful models from the statistics and machine learning literature contain both conjugate and non-conjugate terms. These are usually more expressive, since they use richer (non-conjugate) likelihoods despite their mathematical inconvenience. Examples include

Kalman Filters with non-Gaussian likelihoods [Rue and Held, 2005], Gaussian process classification [Kuss and Rasmussen, 2005], correlated topic models [Blei and Lafferty, 2007], or dynamic topic models [Blei and Lafferty, 2006]. In these models, Variational Inference can't rely on the efficiency of conjugate computations and becomes computationally challenging. In some cases, the non-conjugate terms can be further approached, for instance by using Taylor approximations or Laplace approximations, as in [Wang and Blei, 2013]. Nevertheless, such approximations can cause a loss of performance [Khan, 2012].
We have seen that another possible way of handling the non-conjugacy is stochastic-gradient methods, such as [Ranganath et al., 2014]. These do not require significant model-specific analysis, and therefore can be applied to various inference problems without much additional derivation. However, a naive application of the stochastic gradient ascent algorithm might lead to major issues, as it ignores the potential available closed-form expressions contained in the lower bound, and its efficiency and convergence might depend on the parameterization of the variational distribution.

The Conjugate-computation Variational Inference (CVI) algorithm introduced in [Khan and Lin, 2017] doesn't have these limitations as it combines the advantages of both types of inference: it uses conjugate computations for the conjugate terms to leverage its computational efficiency on one hand, while performing stochastic approximations with a gradient-ascent method for the non-conjugate terms on the other hand.
To achieve that, the evidence lower bound is maximized in the *mean-parameter space* with a stochastic *mirror-ascent method*, rather than in the *natural-parameter space* with a traditional *stochastic gradient-ascent algorithm* ; plus, the gradient steps are implemented using conjugate computations.

2.3.2 Assumptions

We give the two assumptions that CVI relies on. The first one concerns the form required for the variational distribution $q(\mathbf{z}|\boldsymbol{\lambda})$, while the second one concerns the expression of the joint distribution $p(\mathbf{y}, \mathbf{z})$. For that, we need to give a quick summary on exponential families.

Basics of Exponential Families We present a few relevant results about exponential families. We recall the definition of the exponential family, already given in the section about conjugate models. A member of the exponential family, say $q(\mathbf{z}|\boldsymbol{\lambda})$, takes the following exponential form:

$$q(\mathbf{z}|\boldsymbol{\lambda}) = h(\mathbf{z}) \exp\left\{\langle \boldsymbol{\lambda}, \boldsymbol{\phi}(\mathbf{z})\rangle - A(\boldsymbol{\lambda})\right\} \tag{2.11}$$

where $\boldsymbol{\phi} := [\phi_1, \phi_2, \ldots, \phi_M]$ is a vector of sufficient statistics, $\boldsymbol{\lambda} := [\lambda_1, \lambda_2, \ldots, \lambda_M]^T$ is a vector of natural parameters, $\langle \mathbf{a}, \mathbf{b}\rangle$ is an inner product, and $A(\boldsymbol{\lambda})$ is the log-partition function. The set of natural parameters is denoted by $\Omega := \{\boldsymbol{\lambda} \in \mathbb{R}^M | A(\boldsymbol{\lambda}) < \infty\}$.

The mean parameter associated with a sufficient statistic $\boldsymbol{\phi}_m$ is defined by:

$$\mu_m := \mathbb{E}_q\left[\phi_m(\mathbf{z})\right] \tag{2.12}$$

We denote by $\boldsymbol{\mu}$ the vector of all μ_m. The set of valid mean parameters is defined as:

$$\mathcal{M} := \{\boldsymbol{\mu} \in \mathbb{R}^M | \exists\ p\ s.t.\ \mathbb{E}_q[\phi_m(\mathbf{z})] = \mu_m,\ \forall m\} \tag{2.13}$$

The exponential representation 2.11 is called *minimal* when there does not exist a nonzero vector $\mathbf{a} \in \mathbb{R}^M$ such that the linear combination $\langle \mathbf{a}, \boldsymbol{\phi}\rangle$ is equal to a constant. This implies that each distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ has a unique natural parametrization $\boldsymbol{\lambda}$. It also means that there

is a one-to-one mapping between the mean parameter $\boldsymbol{\mu}$ and the natural parameter $\boldsymbol{\lambda}$.

We can now introduce the first assumption of CVI.

Assumption 1 [minimality] The variational distribution $q(\mathbf{z}|\boldsymbol{\lambda})$ is a minimal exponential-family distribution:

$$q(\mathbf{z}|\boldsymbol{\lambda}) = h(\mathbf{z}) \exp\left\{\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\lambda}\rangle - A(\boldsymbol{\lambda})\right\}, \tag{2.14}$$

with $\boldsymbol{\lambda}$ as its natural parameters.

This assumption allows us to express the ELBO as a function of $\boldsymbol{\mu} \in \mathcal{M}$ instead of $\boldsymbol{\lambda} \in \Lambda$ because of the one-to-one mapping between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ that the minimal representation implies. This reparametrized objective function is denoted by $\widetilde{\mathcal{L}}(\boldsymbol{\mu}) := \mathcal{L}(\boldsymbol{\lambda})$

The second assumption defines the class of joint distributions for which CVI can be applied.

Assumption 2 [conjugacy] We assume that the joint distribution contains some terms, collectively denoted by $\tilde{p}_c$, which take the same form as q with respect to $\mathbf{z}$, i.e. :

$$\tilde{p}_c(\mathbf{y}, \mathbf{z}) \propto h(\mathbf{z}) \exp\left\{\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\eta}\rangle\right\}, \tag{2.15}$$

where $\boldsymbol{\eta}$ is a known parameter vector. $\tilde{p}_c$ is the conjugate part of the model, while $\tilde{p}_{nc}$ refers to the non-conjugate part. This yields a partitioning of the joint distribution into conjugate and non-conjugate terms as follows: $p(\mathbf{y}, \mathbf{z}) \propto \tilde{p}_{nc}(\mathbf{y}, \mathbf{z})\tilde{p}_c(\mathbf{y}, \mathbf{z})$. These terms can be unnormalized with respect to $\mathbf{z}$.

2.3.3 Derivation of the updates

We can now explain the main ideas to derive the updates of the CVI algorithm.

In the stochastic-gradient methods introduced in the previous section, the evidence lower bound $\mathcal{L}$ is optimized with respect to the variational parameters $\boldsymbol{\lambda}$ by using the stochastic gradient descent algorithm. At iteration t, the algorithm implements the following step:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \widehat{\nabla}_\lambda \mathcal{L}(\boldsymbol{\lambda}_t) \tag{2.16}$$

where $\widehat{\nabla}_\lambda \mathcal{L}(\boldsymbol{\lambda}_t) := \widehat{\partial \mathcal{L}}/\partial \boldsymbol{\lambda}$ is a stochastic gradient of the lower bound evaluated at $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$. An equivalent formulation of the gradient step in equation 2.16 is given by:

$$\boldsymbol{\lambda}_{t+1} = \arg\max_{\lambda \in \Lambda} \left\langle \boldsymbol{\lambda}, \widehat{\nabla}_\lambda \mathcal{L}(\boldsymbol{\lambda}_t) \right\rangle - \frac{1}{2\rho_t} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_t\|_2^2 \tag{2.17}$$

where $||.||_2$ is the Euclidean norm.

The assumption that states a minimal representation for the variational distribution implies that 2.16 can also be viewed as a maximization problem over $\boldsymbol{\mu} \in \mathcal{M}$. The update in Equation 2.17 is therefore replaced by a stochastic mirror-descent update in the mean parameter space. The mirror-descent algorithm is a generalization of gradient-descent that induces non-Euclidean geometry by using the KL divergence instead of the Euclidean norm [Raskutti and

Mukherjee, 2015]. The gradient step becomes:

$$\boldsymbol{\mu}_{t+1} = \arg\max_{\mu \in \mathcal{M}} \left\langle \boldsymbol{\mu}, \widehat{\nabla}_{\mu} \widetilde{\mathcal{L}}(\boldsymbol{\mu}_t) \right\rangle - \frac{1}{\beta_t} \mathbb{D}_{KL}[q(\mathbf{z}|\boldsymbol{\lambda}) \,\|\, q(\mathbf{z}|\boldsymbol{\lambda}_t)] \tag{2.18}$$

where $\beta_t > 0$ is the step size.

[Khan and Lin, 2017] proved that 2.18 can be implemented by using a Bayesian inference in a conjugate model: the non-conjugate term of the model is approximated by an exponential-family whose natural parameter is a weighted sum of the gradients of the non-conjugate term.

Theorem Under Assumptions 1 and 2, the update 2.18 is equivalent to the Bayesian inference in the following conjugate model:

$$q(\mathbf{z}|\boldsymbol{\lambda}_{t+1}) \propto e^{\langle \phi(z), \widetilde{\lambda}_t \rangle} \widetilde{p}_c(\mathbf{y}, \mathbf{z}) \tag{2.19}$$

whose natural parameter can be obtained by conjugate computation: $\boldsymbol{\lambda}_{t+1} = \widetilde{\boldsymbol{\lambda}}_t + \boldsymbol{\eta}$ where $\widetilde{\boldsymbol{\lambda}}_t$ is the natural parameter of the exponential-family approximation to $\widetilde{p}_{nc}$ and can be obtained recursively as follows:

$$\widetilde{\boldsymbol{\lambda}}_t = (1 - \beta_t)\widetilde{\boldsymbol{\lambda}}_{t-1} + \beta_t \, \widehat{\nabla}_{\mu} \mathbb{E}_q[\log \widetilde{p}_{nc}] \Big|_{\mu=\mu_t} \tag{2.20}$$

with $\boldsymbol{\lambda}_0 = 0$ and $\boldsymbol{\lambda}_1 = \boldsymbol{\eta}$.

This results in the final algorithm shown in Algorithm 1.

Algorithm 1 CVI for exponential-family approximations.

1: Initialize $\widetilde{\boldsymbol{\lambda}}_0 = 0$ and $\boldsymbol{\lambda}_1 = \boldsymbol{\eta}$.
2: **for** $t = 1, 2, 3, \ldots,$ **do**
3: $\quad \widetilde{\boldsymbol{\lambda}}_t = (1 - \beta_t)\widetilde{\boldsymbol{\lambda}}_{t-1} + \beta_t \, \widehat{\nabla}_{\mu} \mathbb{E}_q \left[\log \tilde{p}_{nc}\right]|_{\mu=\mu_t}$.
4: $\quad \boldsymbol{\lambda}_{t+1} = \widetilde{\boldsymbol{\lambda}}_t + \boldsymbol{\eta}$.

We can see that as desired, the natural parameter $\boldsymbol{\lambda}_{t+1}$ of the variational distribution $q(\mathbf{z}|\boldsymbol{\lambda}_{t+1})$ is obtained by a simple conjugate computation: adding the natural parameter of $\tilde{p}_c(\mathbf{y}, \mathbf{z})$, $\boldsymbol{\eta}$, to the natural parameter of the exponential-family approximation to $\tilde{p}_{nc}$, $\widetilde{\boldsymbol{\lambda}}_t$. Stochastic gradients are only computed for the non-conjugate terms.

2.3.4 Extension to Mean-Field Approximation

[Khan and Lin, 2017] also extend CVI to Bayesian networks over $\mathbf{x} = \{\mathbf{y}, \mathbf{z}\}$ where $\mathbf{y}$ is the vector of observed nodes $\mathbf{y}_n$ for $n = 1, 2, \ldots, N$, and $\mathbf{z}$ is the vector of latent nodes $\mathbf{z}_i$ for $i = 1, 2, \ldots, M$. However, they restrict the posterior distribution to a mean-field approximation.

This CVI version is based on the two following assumptions.

Assumption 3 [mean-field + minimality] We assume that $q(\mathbf{z}) = \prod_i q_i(\mathbf{z}_i)$ with each factor being a minimal exponential-family distribution:

$$q_i(\mathbf{z}_i|\boldsymbol{\lambda}_i) := h_i(\mathbf{z}_i) \exp\left[\langle \boldsymbol{\phi}_i(\mathbf{z}_i), \boldsymbol{\lambda}_i \rangle - A_i(\boldsymbol{\lambda}_i)\right]. \tag{2.21}$$

Assumption 4 [conditional-conjugacy] For each node $\mathbf{z}_i$, we can split the following conditional distribution into a conjugate and a non-conjugate term as shown below:

$$\begin{aligned} p(\mathbf{z}_i|\mathbf{x}_{/i}) &\propto \prod_{a\in\mathbb{N}_i} p(\mathbf{x}_a|\mathbf{x}_{pa_a}) \\ &\propto h_i(\mathbf{z}_i) \prod_{a\in\mathbb{N}_i} \tilde{p}_{nc}^{a,i}(\mathbf{z}_i, \mathbf{x}_{a/i})\, e^{\left\{\langle \boldsymbol{\phi}_i(\mathbf{z}_i), \boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})\rangle\right\}} \end{aligned} \tag{2.22}$$

where $\mathbb{N}_i$ is the set containing the node $\mathbf{z}_i$ and all its children, $\tilde{p}_{nc}^{a,i}$ is the non-conjugate part, $\boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})$ is the natural parameter of the conjugate part for the factor a and $\mathbf{x}_{a/i}$ is the set of all nodes in the set $\mathbf{x}_a$ and their parents except $\mathbf{z}_i$.

The algorithm derivation is based on the ideas previously explained ; in this situation, the mirror-descent update in 2.18 can be rewritten as a sum over all nodes i, due to the mean-field approximation and the linearity of the dot product in the first term:

$$\boldsymbol{\mu}_{t+1} = \max_{\mu} \sum_{i=1}^{M} \left[\left\langle \boldsymbol{\mu}_i, \widehat{\nabla}_{\mu_i} \widetilde{\mathcal{L}}(\boldsymbol{\mu}_t) \right\rangle - \frac{1}{\beta_t} \mathbb{B}_{A^*}(\boldsymbol{\mu}_i \| \boldsymbol{\mu}_{i,t}) \right] \tag{2.23}$$

where $A^*(\boldsymbol{\mu})$ is the convex-conjugate of the log-partition function $A(\boldsymbol{\lambda})$ and $\mathbb{B}_{A^*}$ is the Bregman divergence defined by A^* over $\mathcal{M}$.

The final algorithm is shown in Algorithm 2 and its complete derivation is detailed in [Khan and Lin, 2017].

Algorithm 2 CVI for mean-field

1: Initialize $\boldsymbol{\lambda}_{i,0}$.
2: **for** $t = 0, 1, 2, 3, \ldots$, **do**
3: **for** for all node $\mathbf{z}_i$ (or a randomly sampled one) **do**
4: $\widetilde{\boldsymbol{\lambda}}_{i,t} = \sum_{a\in\mathbb{N}_i} \left[\tilde{\boldsymbol{\eta}}_{ai} + \widehat{\nabla}_{\mu_i} \mathbb{E}_{q_t}(\log \tilde{p}_{nc}^{a,i})|_{\boldsymbol{\mu}=\boldsymbol{\mu}_t} \right]$
5: $\boldsymbol{\lambda}_{i,t+1} = (1-\beta_t)\boldsymbol{\lambda}_{i,t} + \beta_t \widetilde{\boldsymbol{\lambda}}_{i,t}$.

We can see that the update of the natural parameter of $q_{i,t+1}$ at node i separates the conjugate computations from the non-conjugate ones: the first set of messages $\tilde{\boldsymbol{\eta}}_{a,i} = \mathbb{E}_{q_{/i,t}}\left[\boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})\right]$ is obtained from the conjugate parts by taking the expectation over their natural parameters $\boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})$, while the second set of messages is the stochastic-gradient of the non-conjugate term in factor a.

Finally, [Khan and Lin, 2017] also show that in the case of a conjugate model, the CVI updates reduce to SVI updates. Since mirror descent with a Bregman divergence (which corresponds to the CVI mean parameter update) is equivalent to the natural gradient descent algorithm along the dual Riemannian space [2] (which correponds to the SVI natural parameter update) [Raskutti and Mukherjee, 2015], CVI can be viewed as an extension of SVI to non-conjugate models.

[2] The space where local distance is defined by the KL divergence between probability distributions rather than the l_2 norm between their parameters.

Chapter 3

Topic Models

Topic modelling algorithms are a class of probabilistic algorithms aiming to discover and annotate large collections of documents with thematic information [Blei, 2012]. By analyzing the words contained in the texts, they uncover the hidden structure that pervades them, how they are linked to each other, and represent each document according to how it exhibits the inferred patterns. These hidden patterns often represent the underlying topical structure of the corpus. They can be used to explore and organize large collections of documents, for classification and annotation tasks, and for information retrieval.

In this chapter, we introduce latent Dirichlet Allocation the simplest topic model which serves as a building block for many other topic models. We present the approximate posterior inference for LDA with a variational approach, and then describe our model of interest, the correlated topic model.

3.1 Latent Dirichlet Allocation (LDA)

In the LDA model, a topic is defined as a distribution over a fixed vocabulary of terms. The intuition behind the model is to represent documents as generated from multiple topics. Across a collection, the documents share the same set of K topics, although each document exhibits them in its own proportion. This appears to be a sound assumption, as any document is usually a heterogeneous combination of some of main themes that run through the entire collection. In a collection of articles from a scientific journal, it is clear that each document exhibits the some of the fields to a different degree. A document might be about Biology and Statistics, while another one might be about Statistics and Engineering. They both share the statistics topic, while combining it with a different topic. The LDA model is closer to capturing the real composition of an article than a model that associates each document with a single topic.

The Goal of LDA, and more generally of topic models, is to automatically infer the topics in an corpus of documents that are not labeled with either topic or keywords. The documents are observed, while the topic structure is hidden. Indeed, LDA belongs to the wide category of *hidden variable models*, where we assume a hidden structured in the observed data, and then attempt to learn this structure by performing *probabilistic posterior inference*.

3.1.1 Notation

- **Words and Documents** words are organized into documents, and constitute the observations. The n-th word in the d-th document is denoted by w_{dn}, which is an element of the fixed *vocabulary* of V terms. $\mathbf{w}_d$ denotes the vector of N_d words associated withe document d. A corpus of D documents is denoted by $\mathbf{w}_{1...D}$.

- **Topics** A topic $\boldsymbol{\beta}_k$ is a distribution over the vocabulary, and each topic is a point of on the $V-1$ simplex (a positive vector of V elements that sums to one. The w-th entry in the the k-th topic is denoted by $\beta_k^{(w)}$.There are K topics $\boldsymbol{\beta}_{1\ldots K}$.
- **Topic Proportions** Each document is associated with a distribution over topic indices : the vector of *topic proportions* $\boldsymbol{\theta}_d$, which is a point on the $K-1$ simplex. It expresses the probability with which words are drawn from each topic in the corpus. The k-th entry of $\boldsymbol{\theta}_d$ is denoted by θ_{dk}.
- **Topic Assignments** Each word w_{dn} is drawn from a from a single topic, which is indexed by the *topic assignment* z_{dn}.

The only observed variables are the words. The topics, topic proportions and topic assignments constitute the latent variables.

3.1.2 Generative Process

LDA assumes that a collection of D documents arises from the following generative process:

1. For each topic $k \in \{1 \ldots K\}$:
   (a) Draw $\boldsymbol{\beta}_k \sim \text{Dir}(\eta)$
2. For each document $d \in \{1 \ldots D\}$:
   (a) Draw topic proportions $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha)$
   (b) For each word $n \in \{1 \ldots N\}$:
      i. Draw topic assignment $z_{dn} \sim \text{Mu}(\boldsymbol{\theta}_d)$
      ii. Draw word $w_{dn} \sim \text{Mu}(\boldsymbol{\beta}_{z_{dn}})$

This is presented as a graphical model in figure 3.1.

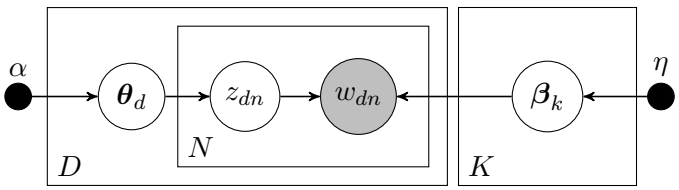


FIGURE 3.1: Graphical model representation of the LDA model

As is shown by the plates, there are three levels of representation: the parameters α and η are corpus-level parameters, sampled once when generating a corpus; $\boldsymbol{\theta}_d$ are document-level variables, sampled once for each document; and z_{dn} and w_{dn} are word-level variables that are sampled once per word and per document.

This structure differs from classical mixture models [Nigam et al., 2000], that only have two levels. For example, in a simple Dirichlet-Multinomial mixture model applied to text modeling, a Dirichlet is sampled once for the whole corpus, then each document is generated by first choosing a topic z, and the set of words in the documents is sampled independently from the multinomial $p(w|z)$. This implies that each document is restricted to only exhibit one topic, and therefore the Dirichlet-Multinomial mixture fails to model the topic heterogeneity of different documents within a corpus, that we have previously noted. Under LDA however, documents can be associated with different topics.

We note that LDA assumes Dirichlet priors for the topic proportions $\boldsymbol{\theta}_d$ and the topics $\boldsymbol{\beta}_k$. Although these distributions take K and V parameters respectively, we use exchangeable Dirichlet distributions, where all the components have the same scalar value η and α respectively.
The generative process that we have presented defines a joint probability distribution over the observed and hidden random variables, given the parameters α and η is:

$$p(\boldsymbol{\theta}_{1:D}, \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K}, \mathbf{w}_{1:D}|\alpha, \eta) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\eta) \prod_{d=1}^{d} \left(p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|\boldsymbol{\beta}_{1:K}, z_{dn}) \right)$$

The corresponding posterior distribution of the hidden variables given the observed documents $p(\boldsymbol{\theta}_{1:D}, \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K}|\mathbf{w}_{1:D}, \alpha, \eta)$ gives the hidden topic decomposition of a particular corpus. It can be thought of as reversing the generative process, i.e. finding the hidden structure that has generated the observed corpus.

3.1.3 Posterior Inference

Approximating the posterior distribution is essential to performing predictions, or corpus exploration and browsing tasks [Blei, 2012]. This distribution can be written as:

$$p(\boldsymbol{\theta}_{1:D}, \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K}|\mathbf{w}_{1:D}, \alpha, \eta) = \frac{p(\boldsymbol{\theta}_{1:D}, \mathbf{z}_{1:D}, \boldsymbol{\beta}_{1:K}\mathbf{w}_{1:D}|\alpha, \eta)}{p(\mathbf{w}_{1:D}|\alpha, \eta)}$$

And:

$$p(\mathbf{w}_{1:D}|\alpha, \eta) = \prod_{d=1}^{D} \int_{\boldsymbol{\theta}_d} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn}|\boldsymbol{\theta}_d) \int_{\boldsymbol{\beta}_{z_{dn}}} p(w_{dn}|\boldsymbol{\beta}_{z_{dn}}) p(\boldsymbol{\beta}_{z_{dn}}|\eta) \quad (3.1)$$

This posterior is intractable because the denominator in equation 3.1 involves a sum over all the possible configurations of the interdependent N topic assignments z_{dn} in each document d, which have K possible values. Several techniques have been developed to approximate this posterior: Gibbs sampling [Griffiths and Steyvers, 2004], expectation propagation [Minka and Lafferty, 2002] and mean field variational inference [Blei et al., 2003], which is the approach that we will briefly present in the next section.

3.1.4 Variational Inference

The coupling between the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ inside the summation over topic assignments is what causes the LDA posterior to be intractable. It's due to the edges between $\boldsymbol{\theta}$, $\mathbf{z}$ and the observed $\mathbf{w}$ nodes. In the mean-field variational distribution however, we drop these edges and the $\mathbf{w}$ nodes, obtaining a distribution where the variables are independent and governed by different variational parameters:

$$q(\boldsymbol{\theta}, \mathbf{z}, \beta|\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = \prod_{k=1}^{K} q(\boldsymbol{\beta}_k|\lambda_k) \prod_{d=1}^{D} \left(q(\boldsymbol{\theta}_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\boldsymbol{\phi}_{dn}) \right) \quad (3.2)$$

where:

$$q(\boldsymbol{\theta}_d) = \text{Dir}(\boldsymbol{\theta}_d|\gamma_d); \quad q(\boldsymbol{\beta}_k) = \text{Dir}(\boldsymbol{\beta}_k|\lambda_d); \quad q(z_{dn}) = \text{Mu}(z_{dn}|\boldsymbol{\phi}_{dn})$$

These parameters are fit to maximize the ELBO:

$$\log p(\mathbf{w}_{1:D}|\alpha,\eta) \geq \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi}) \tag{3.3}$$
$$= \mathbb{E}[\log(p(\mathbf{w},\mathbf{z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\eta)] - \mathbb{E}[\log q(\boldsymbol{\theta},\mathbf{z},\beta|\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})] \tag{3.4}$$

Since LDA is a conjugate exponential model, the optimisation of the ELBO $\mathcal{L}$ can be realized by a closed-form coordinate ascent over the variational parameters [Blei et al., 2003]. This leads to the variational Bayes procedure outlined in algorithm 3, which is guaranteed to converge to a stationary point of the ELBO [Hoffman et al., 2010].

Algorithm 3 Batch Variational Bayes for LDA

1: Initialize $\boldsymbol{\lambda}$ randomly
2: **while** relative improvement in $\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi}) > 0.001$ **do**
3: **for document** $d \in \{1 \dots D\}$ **do**
4: Initialize $\gamma_{dk} = 1$ for $k \in \{1 \dots K\}$
5: **repeat**
6: **for word** $n \in \{1 \dots N\}$ **do**

$$\phi_{dn}^{(k)} \propto \exp\left(\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_k^{(w_{dn})}]\right); \quad k \in \{1 \dots K\}$$

7: Set $\gamma_d = \alpha + \sum_{n=1}^{N} \boldsymbol{\phi}_{dn}$
8: **until** local parameters $\boldsymbol{\phi}_{dn}$ and γ_d converge
9: **for topic** $k \in \{1 \dots K\}$ **do**

$$\lambda_k = \eta + \sum_{d=1}^{D}\sum_{n=1}^{N} \phi_{dn}^{(k)} w_{dn}$$

3.2 The Correlated Topic Model (CTM)

It is reasonable to expect that in most document collections the occurrences of topics are highly correlated: an article about Calculus is more likely to also be about Physics than about Economics. However, LDA fails to account for this correlation, due to the use of a Dirichlet distribution for the topic proportions. The Dirichlet is computationally convenient because it is conjugate to the multinomial distribution over topic assignments; nevertheless, it implies that the components of the topic proportion vectors are nearly independent, but for the small negative correlation arising from the fact that they have to sum to one.

The correlated topic model [Blei and Lafferty, 2007], builds on the LDA model, but replaces the Dirichlet distribution for the topic proportions by a *logistic normal* distribution [Atchison and Shen, 1980], which is more flexible since it allows for a pattern of variability among the components. This is achieved by mapping a multivariate random variable from $\mathbb{R}^d$ to the $d-$simplex.
In the case of the correlated topic model, a sample is drawn from a multivariate Gaussian, then maps it to the $K-1$ simplex to obtain a multinomial parameter for the topic proportions. The covariance of the Gaussian engenders the dependencies between the components of the resulting multinomial.

The generative process for CTM is almost identical to that of LDA, except for the fact that topic proportions are drawn from a logistic normal rather than a Dirichlet. We use the notation introduced in 3.1.1. Given a $K-$mean vector $\boldsymbol{\mu}$, a$K \times K$ covariance matrix $\boldsymbol{\Sigma}$ and a $V-$ dimensional vector $\boldsymbol{\gamma}$:

1. For each topic $k \in \{1 \dots K\}$:
   (a) Draw $\boldsymbol{\beta}_k \sim \text{Dir}(\eta)$
2. For each document $d \in \{1 \dots D\}$:
   (a) Draw $\boldsymbol{\eta}_d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
   (b) For each word $n \in \{1 \dots N\}$:
      i. Draw topic assignment $z_{dn} \sim \text{Mu}(g(\boldsymbol{\eta}_d))$
      ii. Draw word $w_{dn} \sim \text{Mu}(\boldsymbol{\beta}_{zdn})$

where $g(\boldsymbol{\eta})$ maps a natural parametrization of the topic proportions $\boldsymbol{\mu}$ to the mean parametrization $\boldsymbol{\theta}$, i.e. to the $K-1$ simplex:

$$\boldsymbol{\theta} = g(\boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\eta})}{\sum_{k=1}^{K} \exp(\eta_k)} \tag{3.5}$$

This process is illustrated by the graphical model in figure 3.2.

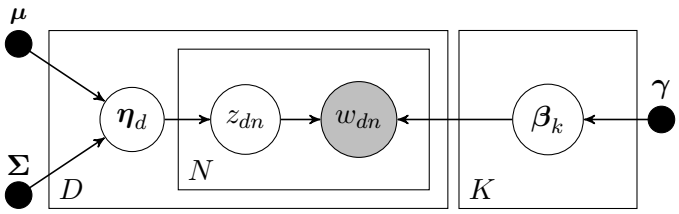


FIGURE 3.2: The graphical model representation of the correlated topic model

The correlated topic model is a more expressive representation of a corpus, as it removes the strong and unrealistic assumption that topic are not correlated, which is made in LDA. The structure provided by the covariance matrix can be used to explore and understand a large corpus. In [Blei and Lafferty, 2007], a topic graph is constructed from the covariance matrix estimated from the *Science* journal corpus, and is used to visualize the topics analyze the connections between them.

However, the extra expressiveness and flexibility comes at the cost of losing the conjugacy with mutlinomial variable for topic assignments. This will complicate the posterior inference for CTM, which we will present in the next chapter.

Chapter 4

CVI for the CTM

We have seen in the previous section that CTM provides the ability to model correlation between topics, while LDA does not. However, the conjugacy between the Dirichlet topic proportions and the multinomial topic assignments, which simplifies approximate posterior inference for LDA, is lost. Indeed, using Gibbs sampling [Griffiths and Steyvers, 2004] is no longer possible, since it was the conjugacy between latent variables that allowed the computation of the coordinate-wise posteriors analytically. An alternative would be applying a Metropolis-Hastings based MCMC sampling, but it would be prohibitive to scale it to large collections of documents.

We therefore approximate the posterior distribution using variational methods. In LDA, the coordinate ascent updates in the variational Bayes algorithm 3 are straightforward to derive [Blei et al., 2003]. This falls into a general approach to mean field variational inference, when the conditional distribution of each variable is in the exponential family. In this case, each variational parameter is updated in closed form with the expectation of the true posterior under the variational distribution. [Xing et al., 2002, Blei et al., 2006]. In the CTM however, the variational inference algorithm presented in [Blei and Lafferty, 2007] is not as simple or as fast, due to the non-conjugacy. In particular, the update for the mean and variance of the topic proportions is fit by gradient based optimization until the local parameters of each document converge.

As an alternative, we convert the problem to variational inference in conjugate models by applying CVI [Khan and Lin, 2017]. This allows us to exploit the efficiency of the existing conjugate models and combine them with (stochastic) gradient methods.

The rest of this chapter is organized as follows: first, we present the variational posterior inference approach to the correlated topic model, and we derive a batch and stochastic CVI algorithm for optimizing the variational lower bound. We then present briefly the coordinate ascent algorithm derived in [Blei and Lafferty, 2007], and compare it to the CVI algorithm.

4.1 Posterior Variational inference for CTM

As we have seen in the previous section, the main question in topic modeling is: given a collection of D documents, what are the underlying topics and how does each document exhibit them. We therefore wish to estimate the posterior distribution of the latent variables $\boldsymbol{\eta}, \mathbf{z}, \boldsymbol{\beta}_{1:K}$ given the data $\mathbf{w}$ and a model $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}\}$. It can be written as:

$$p(\boldsymbol{\eta}, \mathbf{z}, \boldsymbol{\beta}_{1:K} | \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}) = \frac{p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}_{1:K}, \mathbf{w} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})}{p(\mathbf{w} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})} \tag{4.1}$$

Let's develop the denominator, which is the marginal likelihood of a collection of D documents:

$$p(\mathbf{w}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{d=1}^{D} \int_{\boldsymbol{\eta}_d} p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \sum_{z_{dn}=1}^{K} p(z_{dn}|g(\boldsymbol{\eta}))p(w_{dn}|z_{dn}, \boldsymbol{\gamma})d\boldsymbol{\eta}_d \tag{4.2}$$

$$= \prod_{d=1}^{D} \int_{\boldsymbol{\eta}_d} p(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \sum_{z_{dn}=1}^{K} p(z_{dn}|g(\boldsymbol{\eta})) \int_{\boldsymbol{\beta}_{z_{dn}}} p(w_{dn}|\boldsymbol{\beta}_{z_{dn}})p(\boldsymbol{\beta}_{z_{dn}}|\boldsymbol{\gamma})d\boldsymbol{\eta}_d d\boldsymbol{\beta}_{z_{dn}} \tag{4.3}$$

The marginal probability of one document $\mathbf{w}_d$ is intractable both because we integrate over a combinatorial number of terms K^N, and because the distribution of topic proportions $p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is not conjugate to the distribution of topic assignments $p(z_{dn}|f(\boldsymbol{\eta}_d))$. We will use CVI to approximate this denominator and therefore approximate the posterior in 4.1.

We introduce the variational distribution over the latent variables:

$$q(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta}) = \prod_{d=1}^{D} q_d(\mathbf{z}_d, \boldsymbol{\eta}_d, \boldsymbol{\beta}|\Lambda) \tag{4.4}$$

We find $\mathcal{L}(\Lambda)$, a lower bound to the log-marginal likelihood of the corpus, and then maximize it with respect to the variational parameters Λ.

$$\begin{aligned}
\log p(\mathbf{w}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \log \int_{\boldsymbol{\eta}} \sum_{z_n=1}^{K} \int_{\boldsymbol{\beta}_{z_n}} p(\mathbf{w}, \mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{\eta}d\boldsymbol{\beta}_{z_n} \\
&= \mathbb{E}_q\left[\log \frac{p(\mathbf{w}, \mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{q(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta})}\right] \\
&\geq \mathbb{E}_q\left[\log \frac{p(\mathbf{w}, \mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{q(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta}|\Lambda)}\right] \\
&\quad \text{(by using Jensen's inequality and the concavity of the log)} \\
&= \mathbb{E}_q[\log p(\mathbf{w}, \mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma})] + \mathbb{H}(q) = \mathcal{L}(\Lambda)
\end{aligned} \tag{4.5}$$

where the second term $\mathbb{H}(q)$ is the entropy of the variational distribution. The first term is expected log joint distribution over the observed and latent variables given the hyperparameters of the model, which is:

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left[\prod_{d=1}^{D} \left(\mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \text{Mu}(z_n|g(\boldsymbol{\eta}))\text{Mu}(w_n|\boldsymbol{\beta}_{z_n})\right)\right] \prod_{k=1}^{K} \text{Dir}(\boldsymbol{\beta}_k|\boldsymbol{\gamma}) \tag{4.6}$$

For the variational distribution, we assume the following mean-field approximation (where each hidden variables independent and governed by a different distribution):

$$\begin{aligned}
q(\boldsymbol{\eta}, \mathbf{z}, \boldsymbol{\beta}|\Lambda) &= q(\boldsymbol{\eta}, \mathbf{z}, \boldsymbol{\beta}|m_{1:K}, v_{1:K}, \boldsymbol{\phi}_{1:N}, \boldsymbol{\alpha}_{1:K}) \\
&= \left[\prod_{d=1}^{D} \left(\prod_{k=1}^{K} \mathcal{N}(\eta_{dk}|m_{dk}, v_{dk}) \prod_{n=1}^{N} \text{Mu}(z_{dn}|\boldsymbol{\phi}_{dn})\right)\right] \prod_{k=1}^{K} \text{Dir}(\boldsymbol{\beta}_k|\boldsymbol{\alpha}_k)
\end{aligned} \tag{4.7}$$

Where:

- The variational distribution over $\boldsymbol{\eta}$ (the natural parameter of the multinomial representing the topic proportions) is a K-dimensional Gaussian. Similarly to [Blei and Lafferty, 2007], we use K independent univariate Gaussians parametrized by their mean and variance (m_k, v_k). Since the variational parameters are fit using one document, there is no need to use a non-diagonal covariance matrix for this distribution.

- $\boldsymbol{\phi}_n$ is the K-dimensional mean parameter of the of the multinomial variational distribution over the topic assignment z_n. Per the mean-field approximation, each observed word has a different variational distribution for its topic assignment, which allows different words to be associated with different topIcs.

- $\boldsymbol{\alpha}_i$ is the V-dimensional natural parameter of the dirichlet variational distribution over the topic distribution $\boldsymbol{\beta}_i$. Again, we use an exchangeable Dirichlet distribution, where all the components have the same scalar value α_i. This assumption is easy to relax, as pointed out in [Blei et al., 2003].

4.2 CVI

4.2.1 Batch CVI

We will now maximize the bound in 4.5 with repect to the variational parameters $m_{1:K}$, $v_{1:K}$, $\boldsymbol{\phi}_{1:N}$ and $\boldsymbol{\alpha}_{1:K}$ using CVI. We recall the CVI for mean-field update, where the natural parameter $\boldsymbol{\lambda}_{i,t+1}$ of $q_{i,t+1}$, the variational distribution of a latent node z_i is obtained by:

$$\widetilde{\boldsymbol{\lambda}}_{i,t} = \sum_{a \in \mathbb{N}_i} \left[\mathbb{E}_{q_{/i,t}}[\boldsymbol{\eta}_{a,i}(\mathbf{x}_{a/i})] + \nabla_{\mu_i} \mathbb{E}_{q_t}(\log \tilde{p}_{nc}^{a,i})|_{\boldsymbol{\mu}=\boldsymbol{\mu}_t} \right] \tag{4.8}$$

$$\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_{i,t} + \rho_t \widetilde{\boldsymbol{\lambda}}_{i,t} \tag{4.9}$$

where:

- $\mathbb{N}_i$ denotes the set containing the node z_i and all its children; and $\mathbf{x}_{a/i}$ the set of all the neighbors of a and their parents except z_i.

- For a node a in the neighborhood of z_i, $\boldsymbol{\eta}_{a,i}$ is the natural parameter of the conjugate part of the node a, and $\tilde{p}_{nc}^{a,i}$ is the non-conjugate part.

- q_t is the variational distribution at iteration t and $\rho_t > 0$ is the step-size.

As we have previously noted, for a document d, the topic proportions $\boldsymbol{\eta}_d$ and the topic assignments z_{dn} are local hidden variables. Indeed, they do not depend on variables from other documents; they only depend on the global variable representing the topics β, as well as variables in the same local context (the same document d). We can therefore drop the subscript d when deriving the updates for $\boldsymbol{\eta}_d$ and z_{dn}, since we will only be considering one document.

Update for the topic proportions For a latent node $\boldsymbol{\eta}$, we have:

- $\mathbb{N}_{\boldsymbol{\eta}} = \{z_n \; \forall n \leq N\}$
- $\tilde{p}_{nc}^{z_n,\boldsymbol{\eta}} = \text{Mu}(z_n | g(\boldsymbol{\eta})) \quad \forall z_n \in \mathbb{N}_{\boldsymbol{\eta}}$
- $\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \quad \frac{-1}{2}\boldsymbol{\Sigma}^{-1}\right]^T$ is the natural parameter of the conjugate distribution over the topic proportions $\mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The sufficient statistics and mean parameters of the variational distribution $q(\eta_k) = \mathcal{N}(\eta_k|m_k, v_k)$ are respectively:

$$\psi(\eta_k) = \begin{bmatrix}\eta_k & \eta_k^2\end{bmatrix}^T \tag{4.10}$$

$$\boldsymbol{\nu}_k = \begin{bmatrix}m_k & m_k^2 + v_k\end{bmatrix}^T \tag{4.11}$$

Therefore the variational distribution for the topic proportion $\boldsymbol{\eta}$ will be:

$$q_{t+1}(\boldsymbol{\eta}) \propto \left[\prod_{k=1}^{K} \exp\{\psi(\eta_k)^T \widetilde{\boldsymbol{\lambda}}_{k,t}\}\right] \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4.12}$$

where $\widetilde{\boldsymbol{\lambda}}_{k,t}$ are computed as follows:

$$\widetilde{\boldsymbol{\lambda}}_{k,t} = \begin{bmatrix}\tilde{\lambda}_{k,t}^{(1)} & \tilde{\lambda}_{k,t}^{(2)}\end{bmatrix}^T = \rho_t \sum_{n=1}^{N} \nabla_{\boldsymbol{\nu}} \mathbb{E}_{q_t}[\log \text{Mu}(z_n|\boldsymbol{\eta})]|_{\boldsymbol{\nu}=\boldsymbol{\nu}_{k,t}} + (1-\rho_t)\widetilde{\boldsymbol{\lambda}}_{k,t-1} \tag{4.13}$$

We proceed to compute $\nabla_{\boldsymbol{\nu}} \mathbb{E}_q[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))]$. Let's start by deriving the expression of the expectation.

$$\begin{aligned}
\mathbb{E}_q[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))] &= \mathbb{E}_q\left[\log \frac{\prod_{k=1}^{K} \exp(\eta_k)^{\mathbb{1}(z_n=k)}}{\sum_{k=1}^{K}\exp(\eta_k)}\right] \\
&= \mathbb{E}_q\left[\sum_{k=1}^{K} \mathbb{1}(z_n = k)\eta_k\right] - \mathbb{E}\left[\log \sum_{k=1}^{K}\exp(\eta_k)\right] \\
&= \sum_{k=1}^{K} m_k \phi_n^{(k)} - \mathbb{E}\left[\log \sum_{k=1}^{K}\exp(\eta_k)\right]
\end{aligned} \tag{4.14}$$

where we have used the fact that the expectation of an indicator is its probability.

The second term of equation 4.14 is hard to compute. In order to maintain an upper bound on the marginal likelihood, we find a lower bound to the second term using convex duality. The concave function $\log$ can be represented using its concave conjugate:

$$\begin{aligned}
\log(x) &= \min_{\xi}\{\xi x - \log^*(\xi)\} \\
&= \min_{\xi}\{\xi x - \log(\xi) - 1\} \\
&= \min_{\zeta}\{\zeta^{-1} x + \log(\zeta) - 1\}
\end{aligned} \tag{4.15}$$

where ζ is an additional variational parameter. Using 4.15 and the monotonicity of the expectation, we have:

$$\mathbb{E}\left[\log \sum_{k=1}^{K}\exp(\eta_k)\right] \leq \zeta^{-1} \sum_{k=1}^{K} \mathbb{E}_q[\exp(\eta_k)] + \log(\zeta) - 1 \tag{4.16}$$

And we have:

$$\begin{aligned}\mathbb{E}_q[\exp(\eta_k)] &= \int \exp(\eta_k)\mathcal{N}(\eta_k|m_k, v_k)d\eta_k \\ &= \int \frac{1}{\sqrt{2\pi v_k}} \exp\left(\frac{-1}{2v_k}(\eta_k - m_k)^2 + \eta_k\right)d\eta_k\end{aligned}$$

We complete the square in the exponent to recover the integral of the density of a Gaussian distribution:

$$\begin{aligned}&(\eta_k - m_k)^2 - 2v_k\eta_k \\ &= \eta_k^2 + m_k^2 - 2\eta_k m_k - 2v_k\eta_k \\ &= \eta_k^2 + m_k^2 - 2\eta_k(m_k + v_k) \\ &= [\eta_k^2 + (m_k + v_k)^2 - 2\eta_k(m_k + v_k)] + m_k^2 - (m_k + v_k)^2 \\ &= (\eta_k - m_k - v_k)^2 - v_k^2 - 2m_k v_k\end{aligned}$$

Replacing the above result back in the exponent, we get:

$$\mathbb{E}_q[\exp(\eta_k)] = \left[\int \frac{1}{\sqrt{2\pi v_k}} \exp\left(\frac{-1}{2v_k}(\eta_k - m_k - v_k)^2\right)d\eta_k\right] \exp\left(\frac{1}{2v_k}(v_k^2 + 2m_k v_k)\right)$$

The first term in the above equation being equal to 1, we finally obtain:

$$\mathbb{E}_q[\exp(\eta_k)] = \exp\left(m_k + \frac{v_k}{2}\right) \tag{4.17}$$

Using 4.17 and the bound in 4.16, equation 4.14 becomes:

$$\mathbb{E}_q[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))] \geq \sum_{k=1}^{K} m_k \phi_n^{(k)} - \zeta^{-1} \sum_{k=1}^{K} \exp\left(m_k + \frac{v_k}{2}\right) - \log(\zeta) + 1 \tag{4.18}$$

We drop the subscript k for simplicity, and take $f = \mathbb{E}_q[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))]$. The gradient of this expectation with respect to the mean parameter $\boldsymbol{\nu}$ of $q(\eta)$ is :

$$\nabla_{\boldsymbol{\nu}} f = \begin{bmatrix} \frac{\partial f}{\partial \nu^{(1)}} & \frac{\partial f}{\partial \nu^{(1)}} \end{bmatrix}^T \tag{4.19}$$

To compute it, we begin by expressing m and v as a function of the mean parameter $\boldsymbol{\nu}$, using 4.11:

$$m = \nu^{(1)} \tag{4.20}$$

$$v = \nu^{(2)} - (\nu^{(1)})^2 \tag{4.21}$$

Then, by using the chain rule, we express the gradients with respect to the mean parameter $\boldsymbol{\nu}$ as a function of the gradient with respect to the mean m and variance v:

$$\begin{aligned}\frac{\partial f}{\partial \nu^{(1)}} &= \frac{\partial f}{\partial m}\frac{\partial m}{\partial \nu^{(1)}} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial \nu^{(1)}} \\ &= \frac{\partial f}{\partial m} - 2\frac{\partial f}{\partial v}m\end{aligned} \tag{4.22}$$

$$\begin{aligned}\frac{\partial f}{\partial \nu^{(2)}} &= \frac{\partial f}{\partial m}\frac{\partial m}{\partial \nu^{(2)}} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial \nu^{(1)}} \\ &= \frac{\partial f}{\partial v}\end{aligned} \tag{4.23}$$

Reintroducing k, we now compute the derivatives of the expectation in 4.18 with respect to m_k and v_k:

$$\frac{\partial f}{\partial m_k} = \phi_n^{(k)} - \frac{1}{\zeta}\exp\left(m_k + \frac{v_k}{2}\right) \tag{4.24}$$

$$\frac{\partial f}{\partial v_k} = -\frac{1}{2\zeta}\exp\left(m_k + \frac{v_k}{2}\right) \tag{4.25}$$

We substituting back in 4.22 and 4.23, and get the desired gradient with respect to the mean parameter of $q(\eta_k)$:

$$\nabla_{\boldsymbol{\nu}}(f) = \nabla_{\boldsymbol{\nu}}\mathbb{E}_q[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))] = \left[\phi_n^{(k)} \quad -\tfrac{1}{2\zeta}\exp\left(m_k + \tfrac{v_k}{2}\right)\right]^T \tag{4.26}$$

We also maximize 4.18 with respect to ζ. The derivative is:

$$\frac{\partial f}{\partial \zeta} = \frac{1}{\zeta^2}\sum_{k=1}^{K}\exp(m_k + v_k/2) - \frac{1}{\zeta} \tag{4.27}$$

Therefore, $f(\zeta)$ is maximized by;

$$\hat{\zeta} = \sum_{k=1}^{K}\exp(m_k + \frac{v_k}{2}) \tag{4.28}$$

Using 4.26 and 4.28, as well as the mapping from the natural parameter to the moments of a gaussian, we can now write the updates for the natural parameter $\boldsymbol{\lambda}_k$, the moments (m_k, v_k) and the additional variational parameter ζ:

$$\boldsymbol{\lambda}_{k,t+1} = \rho_t \begin{bmatrix} \sum_{n=1}^{N}\boldsymbol{\phi}_n^{(k)} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{(k)} \\ \frac{-N}{2\zeta}\exp(m_{k,t} + \frac{v_{k,t}}{2}) + (\frac{-1}{2}\boldsymbol{\Sigma}^{-1})^{(k)} \end{bmatrix} + (1-\rho_t)\boldsymbol{\lambda}_{k,t} \tag{4.29}$$

$$m_{k,t+1} = \frac{-\boldsymbol{\lambda}_{k,t+1}^{(1)}}{2\boldsymbol{\lambda}_{k,t+1}^{(2)}} \qquad v_{k,t+1} = \frac{-1}{2\boldsymbol{\lambda}_{k,t+1}^{(2)}} \tag{4.30}$$

$$\zeta_{t+1} = \sum_{k=1}^{K}\exp\left(m_{dk,t+1} + \frac{v_{dk,t+1}}{2}\right) \tag{4.31}$$

Update for the topic assignments For a latent node z_n, we have:

- $\mathbb{N}_{z_n} = \{\boldsymbol{\beta}_k, \boldsymbol{\eta}\ \forall k \leq K\}$
- $\tilde{p}_{nc} = \text{Mu}(z_n|g(\boldsymbol{\eta}))$
- The conjugate term for z_n is the multinomial distribution over words $\text{Mu}(w_n|\boldsymbol{\beta}_{z_n})$

The sufficient statistics and mean parameters of the variational distribution $q(z_n) = \text{Mu}(z_n|\boldsymbol{\phi}_n)$ are respectively:

$$\psi(z_n) = \begin{bmatrix}\mathbb{1}(z_n = 1) & \dots & \mathbb{1}(z_n = K-1) & 0\end{bmatrix}^T \tag{4.32}$$

$$\boldsymbol{\nu}_n = \boldsymbol{\phi}_n \tag{4.33}$$

Therefore the variational distribution for the topic assignment z is:

$$q_{t+1}(z) \propto \prod_{n=1}^{N} \exp\{\psi(z_n)^T \widetilde{\boldsymbol{\kappa}}_{n,t}\} \text{Mu}(w_n|\boldsymbol{\beta}_{zn}) \tag{4.34}$$

$$= \prod_{n=1}^{N} \text{Mu}(z_n|\widetilde{\boldsymbol{\phi}}_{n,t}) \text{Mu}(w_n|\boldsymbol{\beta}_{zn}) \tag{4.35}$$

where $\widetilde{\boldsymbol{\kappa}}_n$ is updated by:

$$\widetilde{\boldsymbol{\kappa}}_{n,t} = \begin{bmatrix}\tilde{\kappa}_{n,t}^{(1)} & \dots & \tilde{\kappa}_{n,t}^{(K)}\end{bmatrix}^T = \rho_t \nabla_{\boldsymbol{\nu}} \mathbb{E}_{q_t}[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))]|_{\boldsymbol{\nu}=\boldsymbol{\nu}_{\boldsymbol{n,t}}} + (1-\rho_t)\widetilde{\boldsymbol{\kappa}}_{n,t-1} \tag{4.36}$$

Using the bound on $\mathbb{E}_q[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))]$ from equation 4.18, we differentiate it with respect to $\boldsymbol{\phi}_n$, the mean parameter of $q(z_n)$:

$$\nabla_{\boldsymbol{\phi_n}} \mathbb{E}_q[\log \text{Mu}(z_n|g(\boldsymbol{\eta}))] = \begin{bmatrix}m_1 & \dots & m_k & \dots m_K\end{bmatrix}^T = \mathbf{m} \tag{4.37}$$

where m_k is the mean of the variational distribution $q(\boldsymbol{\eta}_k) = \mathcal{N}(\boldsymbol{\eta}_k|m_k, v_k)$ for $k = 1 \dots K$.

Substituting 4.37 back into the update 4.36, we get the following updates for $\widetilde{\boldsymbol{\kappa}}_n$, and the corresponding mean parameter $\widetilde{\boldsymbol{\phi}}_n$:

$$\widetilde{\boldsymbol{\kappa}}_{n,t} = \rho_t \mathbf{m}_t + (1-\rho_t)\widetilde{\boldsymbol{\kappa}}_{n,t-1} \tag{4.38}$$

$$\widetilde{\boldsymbol{\phi}}_{n,t} = \begin{bmatrix}\frac{\exp(\tilde{\kappa}_{n,t}^{(1)})}{\sum_{k=1}^{K}\exp(\tilde{\kappa}_{n,t}^{(k)})} & \dots & \frac{\exp(\tilde{\kappa}_{n,t}^{(K)})}{\sum_{k=1}^{K}\exp(\tilde{\kappa}_{n,t}^{(k)})}\end{bmatrix}^T \tag{4.39}$$

Update for the topics Unlike the topic assignment and the topic proportion, the topic variable β is a global variable, as it depends on the words and the topic assignment for the entire collection. We therefore reintroduce the dependency on the document d in the notation.

The distribution over the topics is conjugate to the rest of the factors. It follows that for batch CVI, the update for the global nodes $\boldsymbol{\beta}_k$, will simply amount to computing the LDA coordinate ascent update and subtracting the current set of parameters. This is equivalent to taking a natural gradient step [Hoffman et al., 2013].
The variational distribution for the topics $\boldsymbol{\beta}_{1:K}$ is:

$$q_{t+1}(\boldsymbol{\beta}) = \prod_{i=1}^{k} Dir(\boldsymbol{\beta}_i|\boldsymbol{\alpha}_{i,t+1}) \tag{4.40}$$

where the variational concentration parameter follows the update:

$$\boldsymbol{\alpha}_{i,t+1} = (1-\rho_t)\boldsymbol{\alpha}_{i,t} + \rho_t\left(\gamma + \sum_{d=1}^{D}\sum_{n=1}^{N}\phi_{dn}^{(i)}w_{dn}\right) \tag{4.41}$$

We can see that this update depends on the variational topic assignment ϕ for every document.

Inference over conjugate models We can complete the squares in the exponential in 4.12 in order to write the update for the topic proportions as an inference in a bayesian linear model, represented in figure 4.1c.

$$q_{t+1}(\boldsymbol{\eta}) \propto \left[\prod_{k=1}^{K}\mathcal{N}(y_{k,t+1}|\eta_k, \sigma^2_{k,t+1})\right]\mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4.42}$$

where $y_{k,t+1} = \sigma^2_{k,t+1}\tilde{\lambda}^{(1)}_{k,t+1}$ and $\sigma^2_{k,t+1} = -1/(2\tilde{\lambda}^{(2)}_{k,t+1})$.

From 4.35, we can see that updates for the topics and the topic assignment can be written as inference over an LDA model characterized by the following joint distribution.

$$q(\mathbf{z}, \boldsymbol{\beta}) = \left[\prod_{d=1}^{D}\prod_{d=1}^{N}\left(\text{Mu}(z_{dn}|\widetilde{\boldsymbol{\phi}}_{dn})\text{Mu}(w_{dn}|\boldsymbol{\beta}_{z_{dn}})\right)\right]\prod_{i=1}^{K}\text{Dir}(\boldsymbol{\beta}_i|\boldsymbol{\alpha}_i) \tag{4.43}$$

where $\widetilde{\phi}$ plays the role of the topic proportion variable for the LDA model. The corresponding graphical model is shown in figure 4.1b.

Therefore, by using the coordinate ascent LDA updates seen in 3, we can compute the variational parameters for both $q(\boldsymbol{\beta}_i|\boldsymbol{\alpha}_i)$ and $q(z_n|\boldsymbol{\phi}_n)$:

$$\boldsymbol{\phi}^{(k)}_{dn,t+1} \propto \exp\{\widetilde{\boldsymbol{\kappa}}_{n,t} + \mathbb{E}_q[\log \boldsymbol{\beta}_k^{(w_{dn})}]\} \quad \text{for } n \in \{1\dots N\}, d \in \{1\dots D\} \tag{4.44}$$

$$\boldsymbol{\alpha}_{i,t+1} = (1-\rho_t)\boldsymbol{\alpha}_{i,t} + \rho_t\left(\gamma + \sum_{d=1}^{D}\sum_{n=1}^{N}\phi_n^{(i)}w_n\right) \quad \text{for } i \in \{1\dots K\} \tag{4.45}$$

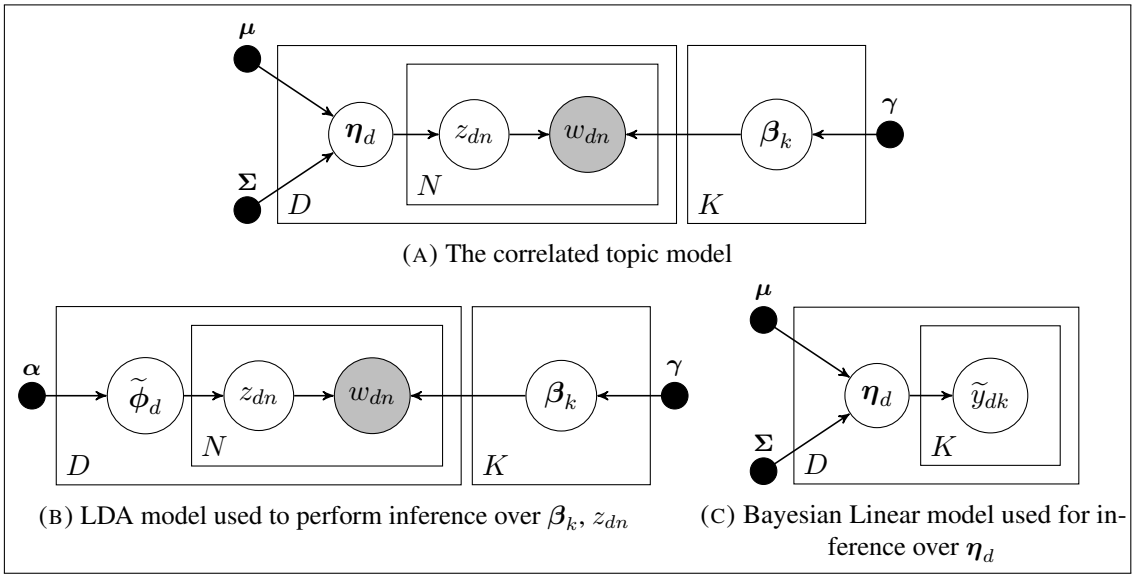
where :

$$\mathbb{E}_q[\log \boldsymbol{\beta}_k^{(w_n)}] = \Psi\big(\alpha_k^{(w_{dn})}\big) - \Psi\big(\sum_{v=1}^{V}\alpha_k^{(v)}\big)$$

where Ψ denotes the digamma function (the first derivative of the logarithm of the gamma function).

Using CVI, we have converted inference in the non-conjugate CTM model into inference over two conjugate models, LDA and a linear model. This simplifies the derivation of the updates for the variational parameters, as well as their implementation, since we can use already existing implementations of the simpler conjugate models.

Batch CVI for CTM algorithm Algorithm 4 outlines the procedure for Batch CVI for the CTM model. Batch inference is slow to apply for large corpora. Indeed, we compute the local variational parameters $\boldsymbol{\phi}_d$, $\mathbf{m}_d$ and $\mathbf{v}_d$ for every document before we can update the topics $\boldsymbol{\alpha}_{1:K}$, as the the update requires summing over the variational parameters for every word in



(A) The correlated topic model

(B) LDA model used to perform inference over $\boldsymbol{\beta}_k$, z_{dn}

(C) Bayesian Linear model used for inference over $\boldsymbol{\eta}_d$

FIGURE 4.1: Inference over the CTM model in 4.1a can be written as inference over the LDA model in 4.1b and the bayesian linear model in 4.1c

the document. This is especially inefficient when we have to perform the local inference over all the documents with a randomly initialized topic in the first iterations. Furthermore, it is not suited to "online" contexts, where new documents are constantly arriving.

4.2.2 Stochastic CVI

The CVI for CTM algorithm we have previously derived requires a full pass through the entire collection of documents at each iteration. In this section, we introduce a stochastic version to make posterior inference scalable, and adapted to online settings.

At each iteration t we sample one document d for the corpus. Since the local parameters do not depend on variables from other documents, the local inference phase remains almost the same as in Batch CVI, the only difference being that we only use the one randomly sampled document to update $\boldsymbol{\phi}_d$, $\mathbf{m}_d$ and $\mathbf{v}_d$.
For the variational topics $\boldsymbol{\alpha}$, we use the optimized local parameters to perform the SVI [Hoffman et al., 2013] update in the global phase, since the distribution over topics is conjugate to the rest of the factors. The update at iteration t is:

$$\boldsymbol{\alpha}_{i,t+1} = (1-\rho_t)\boldsymbol{\alpha}_{i,t} + \sigma_t\left(\boldsymbol{\gamma} + D\sum_{n=1}^{N}\phi_{dn}^{(i)}w_{dn}\right) \quad \text{for } i \in \{1\dots K\}$$

The general procedure for Stochastic CVI for CTM is presented in algorithm 5.

Step-size parametrization Similarly to [Hoffman et al., 2013] and [Hoffman et al., 2010], the step-size at iteration t is set as:

$$\sigma_t = (t+\tau_0)^{-\kappa} \tag{4.46}$$

The forgetting rate $\kappa \in (0.5, 1]$ controls the rate at which old information is forgotten while the delay $\tau_0 \geq 0$ slows down the early iterations of the algorithm.

Mini-batches In 5, we have presented a stochastic algorithm where only one document is sampled at a time. In order to reduce noise, stochastic learning algorithms usually use multiple observations. For the correlated topic model, this means sampling a mini-batch of S documents, computing the local variational parameters $\mathbf{m}_s$, $\mathbf{v}_s$ and $\boldsymbol{\phi}_s$ for each sample s in the mini-batch, and then computing the variational parameter for the topics as follows:

$$\boldsymbol{\alpha}_{i,t+1} = (1-\rho_t)\boldsymbol{\alpha}_{i,t} + \sigma_t\left(\boldsymbol{\gamma} + \frac{D}{S}\sum_{n=1}^{N}\phi_{dn}^{(i)}w_{dn}\right) \quad \text{for } i \in \{1\dots K\} \tag{4.47}$$

4.2.3 ELBO Derivation

In this section, we derive the ELBO that we will use to monitor the convergence of the CVI algorithm.

Using 4.5, we can easily verify that:

$$\log p(\mathbf{w}|\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{L}(m,v,\boldsymbol{\phi},\boldsymbol{\alpha}) + KL[q(\boldsymbol{\eta},\mathbf{z},\boldsymbol{\beta}|m,v,\boldsymbol{\phi},\boldsymbol{\alpha})||p(\boldsymbol{\eta},\mathbf{z},\boldsymbol{\beta}|\mathbf{w},m,v,\boldsymbol{\phi},\boldsymbol{\alpha})]$$

This shows that maximizing the lower bound to the log-marginal likelihood with respect to the variational parameters is equivalent to minimizing the KL divergence between the variational posterior $q(\boldsymbol{\eta},\mathbf{z},\boldsymbol{\beta}|m,v,\boldsymbol{\phi},\boldsymbol{\alpha})$ and the true posterior $p(\boldsymbol{\eta},\mathbf{z},\boldsymbol{\beta}|\mathbf{w},m,v,\boldsymbol{\phi},\boldsymbol{\alpha})$.

Using the factorizations of p in 4.6 and q in 4.7, we can write the ELBO as:

$$\begin{aligned}\mathcal{L}(m,v,\boldsymbol{\phi},\boldsymbol{\alpha}) \quad = \sum_{d=1}^{D}\Bigg[&\sum_{n=1}^{N} w_{dn}\mathbb{E}_q\{\log \text{Mu}(w_{dn}|\boldsymbol{\beta}_{z_n})\} \\ &+ \sum_{n=1}^{N}\mathbb{E}_q\{\log \text{Mu}(z_{dn}|g(\boldsymbol{\eta}))\} - \sum_{n=1}^{N}\mathbb{E}_q\{\log \text{Mu}(z_{dn}|\boldsymbol{\phi}_{dn})\} \\ &+ \mathbb{E}_q\{\log \mathcal{N}(\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma})\} - \sum_{k=1}^{K}\mathbb{E}_q\{\log \mathcal{N}(\eta_{dk}|m_{dk},v_{dk})\} \\ &+ \left(\sum_{k=1}^{K}\left[\mathbb{E}_q\{\text{Dir}(\boldsymbol{\beta}_k|\boldsymbol{\gamma})\} - \mathbb{E}_q\{Dir(\boldsymbol{\beta}_k|\boldsymbol{\alpha}_k\}\right]\right)/D\Bigg]\end{aligned} \tag{4.48}$$

Now we expand each expectation in term of the variational parameters:

The first expectation of 4.48 is:

$$\mathbb{E}_q\{\log \text{Mu}(w_{dn}|\boldsymbol{\beta}_{z_n})\} = \sum_{k=1}^{K}\boldsymbol{\phi}_{dn}^{(k)}\log\mathbb{E}_q[\boldsymbol{\beta}_k^{(w_{dn})}] = \sum_{k=1}^{K}\boldsymbol{\phi}_{dn}^{(k)}\left(\Psi(\alpha_k^{(w_{dn})}) - \Psi(\sum_{v=1}^{V}\alpha_k^{(v)})\right)$$

For the second expectation, we use the lower-bound on the expected log probability of a topic assignment that we have derived in equation 4.18:

$$\sum_{n=1}^{N}\mathbb{E}_q\{\log \text{Mu}(z_{dn}|g(\boldsymbol{\eta}))\} = \sum_{n=1}^{N}\left[\sum_{k=1}^{K}m_k\phi_n^{(k)} - \zeta^{-1}\sum_{k=1}^{K}\exp\left(m_k + \frac{v_k}{2}\right) - \log(\zeta) + 1\right]$$

The third term is:

$$\sum_{n=1}^{N}\mathbb{E}_q\{\log \text{Mu}(z_{dn}|\boldsymbol{\phi}_{dn})\} = \sum_{n=1}^{N}\sum_{k=1}^{K}\boldsymbol{\phi}_{dn}^{(k)}\log(\boldsymbol{\phi}_{dn}^{(k)})$$

The fourth term expands to:

$$\begin{aligned}&\mathbb{E}_q\{\log \mathcal{N}(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma})\}\\&= \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}| - \frac{K}{2}\log 2\pi - \frac{1}{2}\mathrm{Tr}(\mathrm{diag}(\mathbf{v}_d^2)\boldsymbol{\Sigma}^{-1}) + (\mathbf{m}_d - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}_d - \boldsymbol{\mu})\end{aligned}$$

The fifth term can be written as:

$$\sum_{k=1}^{K}\mathbb{E}_q\{\log\mathcal{N}(\eta_{dk}|m_{dk}, v_{dk})\} = \sum_{k=1}^{K}\frac{1}{2}(\log v_k^2 + \log 2\pi + 1))$$

Finally, we can write the last line of 4.48 as:

$$\begin{aligned}\sum_{k=1}^{K}\Bigg(&- \log\Gamma\Big(\sum_{v=1}^{V}\alpha_k^{(v)}\Big) + \sum_{v=1}^{V}(\gamma - \alpha_k^{(v)})\Big(\Psi\big(\alpha_k^{(v)}\big) - \Psi\Big(\sum_{i=1}^{V}\alpha_k^{(i)}\Big)\Big)\\&+ \log\Gamma\big(\alpha_k^{(v)}\big) + \log\Gamma(V\gamma) - V\log\Gamma(\gamma)\Bigg)/D\end{aligned}$$

4.3 Coordinate ascent optimization for CTM

We now give a rough outline of the variational inference algorithm presented in [Blei and Lafferty, 2007], which consists of a coordinate ascent algorithm that iteratively optimizes the bound 4.48 with respect to the variational parameters $m_{1:K}$, $v_{1:K}$, $\boldsymbol{\phi}_{1:N}$, $\boldsymbol{\alpha}_{1:K}$ and ζ.

We have already shown in 4.2.1 that the maximum for ζ is found at:

$$\hat{\zeta} = \sum_{i=1}^{K}\exp(m_i + \frac{v_i}{2}) \tag{4.49}$$

The maximization with respect to $\boldsymbol{\phi}_n$ gives a maximum at:

$$\boldsymbol{\phi}_n^{(i)} = \exp(m_i + \mathbb{E}[\log\beta_{iw_n}]) \qquad i \in \{1\dots K\}$$

The maximization with respect to $\boldsymbol{\alpha}_k$ gives the LDA coordinate ascent update for the topics:

$$\boldsymbol{\alpha}_k = \boldsymbol{\gamma} + \sum_{d=1}^{D}\sum_{n=1}^{N}\phi_{dn}^{(i)}w_{dn}$$

The maximization with respect to the parameters of the Gaussian variational distribution does not yield an analytic solution because of the non-conjugacy. The derivative with respect to $\mathbf{m}$ is:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{m}} = -\boldsymbol{\Sigma}^{-1}(\mathbf{m} - \boldsymbol{\mu}) + \sum_{n=1}^{N}\boldsymbol{\phi}_n - \frac{N}{\zeta}\exp(\mathbf{m} + \frac{\mathbf{v}}{2}) \tag{4.50}$$

And the derivative with respect to v_i is:

$$\frac{\partial\mathcal{L}}{\partial v_i} = -\frac{\boldsymbol{\Sigma}_{ii}^{-1}}{2} - \frac{N}{2\zeta}\exp(m_i + \frac{v_i}{2}) + \frac{1}{2v_i} \qquad i \in \{1\dots K\} \tag{4.51}$$

Iterating between the optimization of each parameter results in algorithm 6.

This algorithm requires the maximization of the ELBO both with respect to the mean and with respect to the variance of the variational distribution for topic proportions $q(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{v}_d)$, repeated for every document d until the local parameters for this document converge converge. Performing two full optimizations at each local iteration is computationally demanding and makes this variational inference algorithm slow to converge.
CVI however, avoids this inefficiency. At each local iteration, we take only one mirror descent step in the natural parameter space of $q(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{v}_d)$ (4.29), from which directly obtain the updates for the mean $\mathbf{m}_d$ and the variance $\mathbf{v}_d$ (4.30).

4.4 Summary

In this chapter, we have presented the main theoretical contribution of this work: deriving of the CVI algorithm for the correlated topic model. We also show that using CVI reduces inference in this non-conjugate models to in inference over two conjugate models: LDA and a linear model. This is a considerable advantage of CVI, as it will allow to use already existing software for the simpler models in our implementation.

We have then derived the stochastic CVI algorithm, solving the inefficiency of the batch version that performs a full pass through the entire corpus at each iteration. Stochastic CVI scales to large collections of documents, and is adapted to inference in online settings.

Algorithm 4 Batch CVI for CTM

Initialize $\alpha_{i,0}$ for topic $i \in \{1, \dots, K\}$
for document $d \in \{1 \dots D\}$ **do**
  Initialize $\boldsymbol{\lambda}_{dk,0}$, compute $m_{dk,0}$ and $v_{dk,0}$ for $k \in \{1 \dots K\}$, compute $\zeta_{d,0}$
  for word $w_{dn} \in \{1 \dots N\}$ **do**
    Initialize $\widetilde{\boldsymbol{\kappa}}_{dn,0}$, compute $\widetilde{\boldsymbol{\phi}}_{dn,0}$
while relative improvement in $\mathcal{L}(m, v, \boldsymbol{\phi}, \boldsymbol{\alpha}) > 0.001$ **do**
  for document $d \in \{1 \dots D\}$ **do**
    for $t = 0 \dots T$ **do**
      Update topic proportion distribution $q(\boldsymbol{\eta}_d | \mathbf{m}_d, \mathbf{v}_d)$
      Compute for $k \in \{1 \dots K\}$

$$\boldsymbol{\lambda}_{dk,t+1} = \begin{bmatrix} \lambda^{(1)}_{dk,t} \\ \\ \lambda^{(2)}_{dk,t} \end{bmatrix} = \rho_t \begin{bmatrix} \sum_{n=1}^{N} \boldsymbol{\phi}^{(k)}_{dn} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{(k)} \\ \\ \frac{-N}{2\zeta} \exp(m_{dk,t} + \frac{v_{dk,t}}{2}) + (\frac{-1}{2}\boldsymbol{\Sigma}^{-1})^{(k)} \end{bmatrix} + (1 - \rho_t)\boldsymbol{\lambda}_{dk,t}$$

$$m_{dk,t+1} = \frac{-\boldsymbol{\lambda}^{(1)}_{dk,t+1}}{2\boldsymbol{\lambda}^{(2)}_{dk,t+1}} \qquad v_{dk,t+1} = \frac{-1}{2\boldsymbol{\lambda}^{(2)}_{dk,t+1}}$$

$$\zeta_{t+1} = \sum_{k=1}^{K} \exp\left(m_{dk,t+1} + \frac{v_{dk,t+1}}{2}\right)$$

      for word $n \in \{1 \dots N\}$ **do**
        Update topic assignment distribution $q(z_{dn} | \boldsymbol{\phi}_{dn})$
        Compute

$$\widetilde{\boldsymbol{\kappa}}_{dn,t} = \rho_t \mathbf{m}_t + (1 - \rho_t)\widetilde{\boldsymbol{\kappa}}_{dn,t-1}$$

        Apply the coordinate ascent update for topic assignments in LDA:

$$\boldsymbol{\phi}^{(k)}_{dn,t+1} \propto \exp\{\widetilde{\boldsymbol{\kappa}}_{dn,t} + \mathbb{E}_q[\log \boldsymbol{\beta}^{(w_{dn})}_k]\}$$

$$\mathbb{E}_q[\log \boldsymbol{\beta}^{(w_{dn})}_k] = \Psi\big(\alpha^{(w_{dn})}_{k,t}\big) - \Psi\big(\sum_{v=1}^{V} \alpha^{(v)}_{k,t}\big)$$

  for topic $i \in \{1, \dots, K\}$ **do**
    Update the topic distribution $q(\boldsymbol{\beta}_k | \boldsymbol{\alpha}_k)$
    Apply the coordinate ascent update for topics in LDA:

$$\boldsymbol{\alpha}_{i,t+1} = (1 - \rho_t)\boldsymbol{\alpha}_{i,t} + \rho_t\left(\boldsymbol{\gamma} + \sum_{d=1}^{D}\sum_{n=1}^{N} \phi^{(i)}_{dn} w_{dn}\right)$$

Algorithm 5 Stochastic CVI for CTM

Initialize $\alpha_{i,0}$ for topic $i \in \{1, \dots, K\}$
for document $d \in \{1 \dots D\}$ **do**
  Initialize $\boldsymbol{\lambda}_{dk,0}$, compute $m_{dk,0}$ and $v_{dk,0}$ for $k \in \{1 \dots K\}$, compute $\zeta_{d,0}$
  for word $w_{dn} \in \{1 \dots N\}$ **do**
    Initialize $\widetilde{\boldsymbol{\kappa}}_{dn,0}$, compute $\widetilde{\boldsymbol{\phi}}_{dn,0}$
for $t = 0 \dots T$ **do**
  Uniformly sample a **document** $d \in \{1 \dots D\}$
  Update topic proportion distribution $q(\boldsymbol{\eta}_d | \mathbf{m}_d, \mathbf{v}_d)$
  Compute for $k \in \{1 \dots K\}$

$$\boldsymbol{\lambda}_{dk,t+1} = \begin{bmatrix} \lambda^{(1)}_{dk,t} \\ \lambda^{(2)}_{dk,t} \end{bmatrix} = \rho_t \begin{bmatrix} \sum_{n=1}^{N} \boldsymbol{\phi}^{(k)}_{dn} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{(k)} \\ \frac{-N}{2\zeta} \exp(m_{dk,t} + \frac{v_{dk,t}}{2}) + (\frac{-1}{2}\boldsymbol{\Sigma}^{-1})^{(k)} \end{bmatrix} + (1 - \rho_t)\boldsymbol{\lambda}_{dk,t}$$

$$m_{dk,t+1} = \frac{-\boldsymbol{\lambda}^{(1)}_{dk,t+1}}{2\boldsymbol{\lambda}^{(2)}_{dk,t+1}} \qquad v_{dk,t+1} = \frac{-1}{2\boldsymbol{\lambda}^{(2)}_{dk,t+1}}$$

$$\zeta_{t+1} = \sum_{k=1}^{K} \exp\left(m_{dk,t+1} + \frac{v_{dk,t+1}}{2}\right)$$

  for $n \in \{1 \dots N\}$ **do**
    Update topic assignment distribution $q(z_{dn} | \boldsymbol{\phi}_{dn})$
    Compute

$$\widetilde{\boldsymbol{\kappa}}_{dn,t} = \rho_t \mathbf{m}_t + (1 - \rho_t)\widetilde{\boldsymbol{\kappa}}_{dn,t-1}$$

    Apply the coordinate ascent update for topic assignments in LDA:

$$\boldsymbol{\phi}^{(k)}_{dn,t+1} \propto \exp\{\widetilde{\boldsymbol{\kappa}}_{dn,t} + \mathbb{E}_q[\log \boldsymbol{\beta}^{(w_{dn})}_k]\}$$

$$\mathbb{E}_q[\log \boldsymbol{\beta}^{(w_{dn})}_k] = \Psi\big(\alpha^{(w_{dn})}_{k,t}\big) - \Psi\big(\sum_{v=1}^{V} \alpha^{(v)}_{k,t}\big)$$

  for $i \in \{1, \dots, K\}$ **do**
    Update the topic distribution $q(\boldsymbol{\beta}_k | \boldsymbol{\alpha}_k)$
    Apply the SVI update for topics in LDA:

$$\boldsymbol{\alpha}_{i,t+1} = (1 - \rho_t)\boldsymbol{\alpha}_{i,t} + \sigma_t\left(\gamma + D\sum_{n=1}^{N} \phi^{(i)}_{dn} w_{dn}\right)$$

Algorithm 6 Coordinate ascent optimization for CTM

Initialize $m_{1:K}$, $v_{1:K}$, $\boldsymbol{\phi}_{1:N}$, $\boldsymbol{\alpha}_{1:K}$ and ζ
while relative improvement in $\mathcal{L} > 0.001$ **do**
  for document $d \in \{1 \dots D\}$ **do**
    repeat
      for word $n \in \{1 \dots N\}$ **do**
        Update topic assignment distribution $q(z_{dn}|\boldsymbol{\phi}_{dn})$

$$\boldsymbol{\phi}_{dn}^{(i)} = \exp(m_{di} + \mathbb{E}[\log \beta_{iw_n}]) \quad i \in \{1 \dots K\}$$

        Update ζ **with equation** 4.49
        Update topic proportion distribution $q(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{v}_d)$
          maximize $\mathcal{L}$ with respect to $\mathbf{m}_d$ using conjugate gradient with the derivative in equation 4.50
          maximize $\mathcal{L}$ with respect to v_{di} using L-BFGS-B with the constraint $v_{di} > 0$ and the derivative in equation 4.51
    until local parameters $\boldsymbol{\phi}_{dn}$ and γ_d converge
  for topic $k \in \{1 \dots K\}$ **do**
    Update the topic distribution $q(\boldsymbol{\beta}_k|\boldsymbol{\alpha}_k)$

$$\boldsymbol{\alpha}_k = \boldsymbol{\gamma} + \sum_{d=1}^{D} \sum_{n=1}^{N} \phi_{dn}^{(i)} w_{dn}$$

Chapter 5

Experiments

5.1 Setup

We adapt an available implementation of LDA, performing the batch variational bayes algorithm 3. In our experiments, we refer to this model as **LDA**. For CTM, we implement the variational inference algorithm of [Blei and Lafferty, 2007], presented in 4, in addition to the batch CVI algorithm in 6 and the stochastic CVI algorithm described in 5. These algorithms are labeled **CTM**, **CTM_CVI** and **CTM_STOCH_CVI** respectively.
All algorithms are implemented in *Python* using the standard scientific libraries *Numpy* and *Scipy*, and the implementations are as similar as possible.

We begin by comparing the model fit of the LDA and CTM models, as measured by the accuracy in a document classification task, and the predictive distributions over held out documents. We then compare the convergence speed of the traditional variational algorithm for CTM with the batch CVI algorithm. Finally, we evaluate the performance of the stochastic CVI algorithm.

5.1.1 Datasets

de-news de-news[1] is collection of daily news items between 1996 to 2000 in English. It contains 9756 documents, and 20000 distinct terms.

20Newsgroups 20Newsgroups[2] is a collection of news documents divided almost evenly across 20 different news groups. Each article is associated with a category label, that can be used as ground truth while performing document classification. We use a subset of 11314 documents and 60879 distinct terms.

Associated Press corpus (AP) The AP[3] corpus is constituted from Associated Press articles from the years 1988 through 1990. We use a subset of 2246 documents and 6786 unique terms.

In pre-processing all datasets, we removed a standard list of 50 stop words. From the 20Newsgroups, we removed indicative meta text such as headers and footers. This is done in order to force the document classification task, to rely only on the semantics of plain text.

5.1.2 Evaluation metrics

Heldout log-likelihood We use the average log-likelihood on held-out data as a measure of model fitness. We compute the average log probability of the held-out data given a model

[1] http://homepages.inf.ed.ac.uk/pkoehn/publications/de-news/
[2] http://scikit-learn.org/stable/datasets/twenty_newsgroups.html
[3] http://www.cs.columbia.edu/~blei/lda-c/

estimated from the training data. A model with better generalization performance will assign higher probability to the held out data.

$$\text{average log-likelihood}(D_{\text{test}}) = \frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}$$

Classification Accuracy is the accuracy of a classifier learned from the topic distribution of training documents and then applied to test documents. A higher accuracy means the unsupervised topic model better captures the underlying structure of the corpus.

5.1.3 Parameter setting

For LDA, the variational topics λ_k are fit using the training documents, and then held fixed for the testing documents. The per-document parameters γ_d and $\boldsymbol{\phi}_d$ for the topic proportion variational distribution $q(\boldsymbol{\theta}_d|\gamma_d)$ and the topic assignment variational distribution $q(z_{dn}|\boldsymbol{\phi}_{dn})$ are fit during the training phase, and estimated in the testing phase using the variational topics learned in the training phase. An identical procedure is used for both CTM algorithms.
All algorithms are trained with exactly the same convergence criteria: that the average change in the log-likelihood is less than 0.001.

For LDA, we set the Dirichlet prior hyper-parameters α and η are set to $\frac{1}{K}$. For CTM, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a K-dimensional gaussian with zero mean and unit variance; the Dirichlet prior hyper-parameter γ is set to $\frac{1}{K}$.

5.2 CTM and LDA comparison

5.2.1 Document Classification

Using LDA and CTM, we can reduce any document to its posterior topic distribution, which can be approximated using the variational topic distribution. Through document classification, we want to determine if accounting for topic correlations improves the quality of this low-dimensional representation.

We conduct this experiment on the 20Newsgroups dataset where the ground truth category labels are available for all the articles. We use 5-fold cross validation, and fit LDA and CTM models for each fold. A multi-class SVM classifier is then trained for each model based on the learned topic distribution of the training documents, and tested on the estimated topic distribution of the testing documents.

Figure 5.1, shows the mean classification accuracies obtained for varying numbers of topics, as well as the standard deviations over the 5 folds. In every case, CTM significantly outperforms LDA which treats topics independently, confirming that accurately modelling topic correlation is important for semantic modelling and representation.

5.2.2 Document Modelling

In this section, we wish to compare the relative generalization performance of the two models, CTM and LDA. For CTM, we use both the CVI algorithm and the coordinate ascent optimization algorithm.

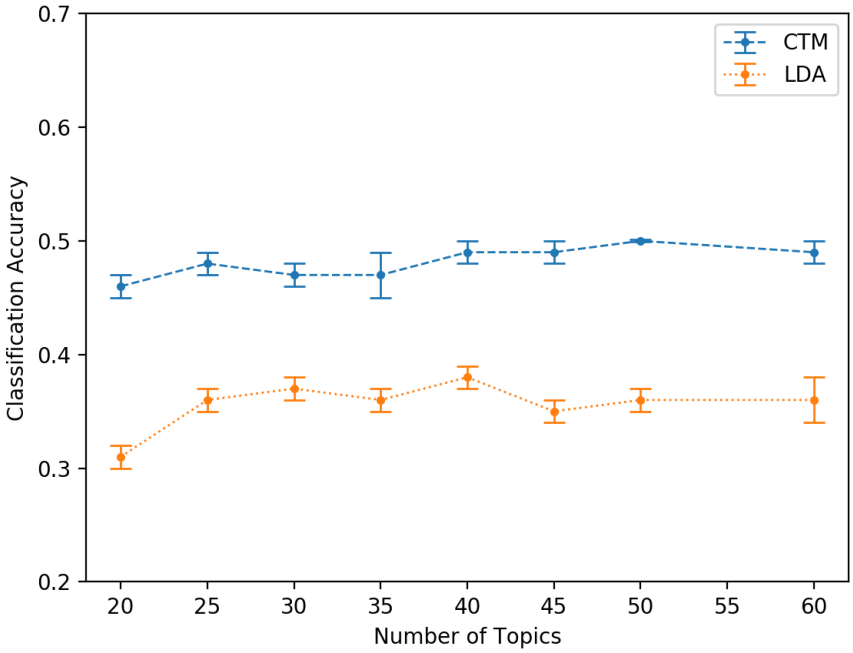


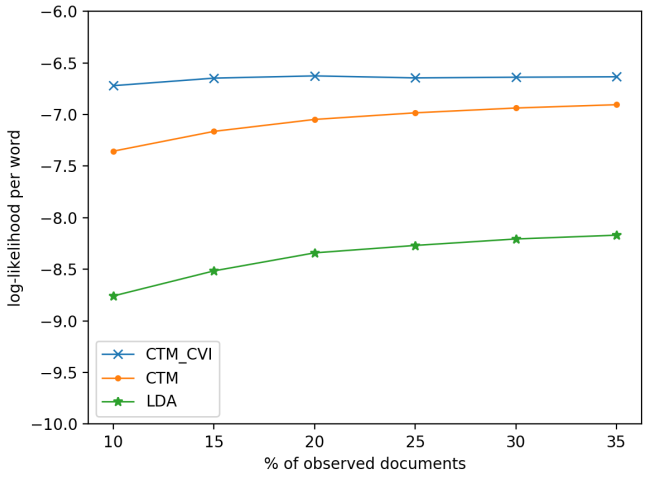
FIGURE 5.1: Mean Classification accuracy over the folds on 20Newsgroups

One quantitative evaluation for this is how well the models predict the remaining documents in a corpus after observing a portion of it [Blei and Lafferty, 2007]. We observe a randomly sampled subset S of the documents, and compare the predictive distributions $p(\mathbf{w}_d|\mathbf{w}_S)$ of the remaining documents under all algorithms, trained on the subset S with 10 topics. To compare these distributions, we use the average log-likelihood of the testing documents, defined above.
The results are shown in figure 5.2. Even with a relatively small number of observed documents, CTM has a significantly higher log-likelihood compared to LDA. The topic correlation is therefore useful: even after seeing few words in one topic, it allows to increase the probability of words in correlated topics.

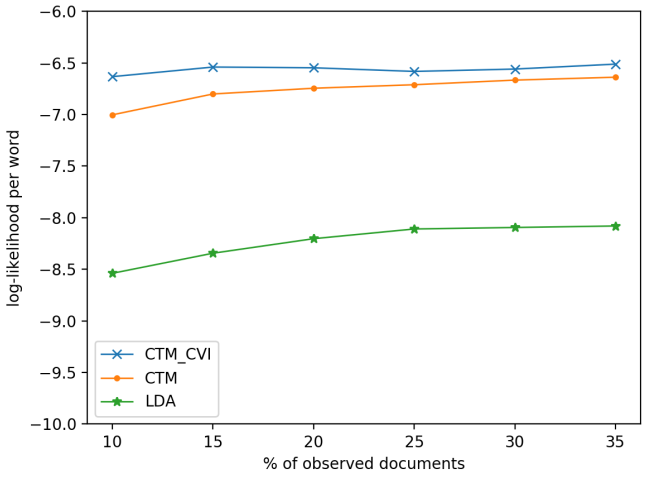
Next, we use 90% of the data for training and hold-out the rest for testing, and compute the average held out log probability as we vary the topics for each model. Figure 5.3 shows that the CTM model always gives a superior fit.
We also note that in both figure 5.3 and figure 5.2, the batch CVI algorithm nearly always converges to a better solution than the original coordinate ascent algorithm for CTM.

5.2.3 CVI performance evaluation

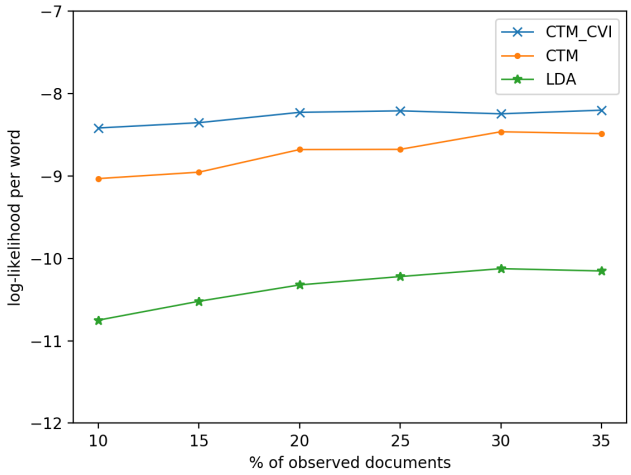
To compare the performance of the traditional CTM variational algorithm and the CVI algorithm for CTM, we evaluate the average held-out log-likelihood as a function of CPU time on the 20Newsgroups and de-news datasets. For every document, we run the optimization for the local variational parameters for 30 iterations, for both algorithms. The CVI step-size ρ_t is set to 0.7.
For both corpora, we use 70% of the data for training and the rest for testing. We set the number of topics $K = 20$ for the 20Newsgroups corpus and $K = 10$ for the de-news corpus.



(A) AP corpus



(B) de-news corpus



(C) 20Newsgroups corpus

FIGURE 5.2: Heldout average log-likelihood for different proportions of training data, on different corpora, $K = 10$

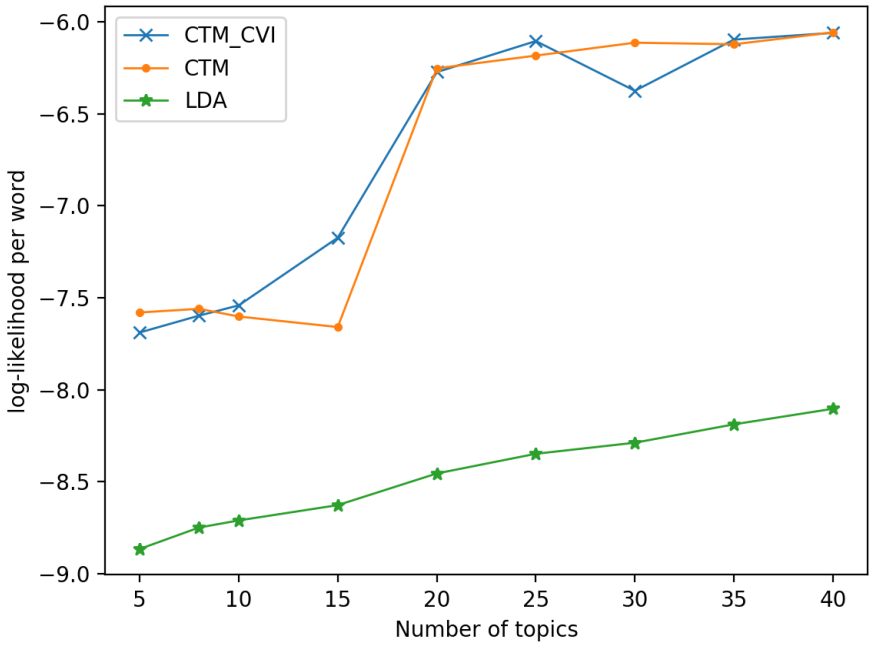
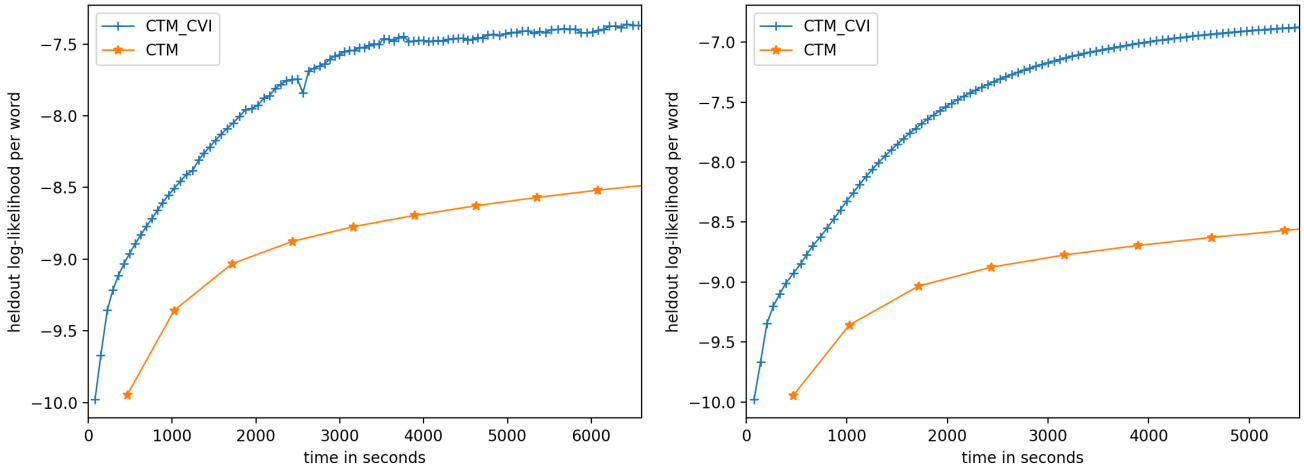


FIGURE 5.3: Heldout average log-likelihood for different number of topics, on the de-news dataset

Figure 5.4 presents the results. On the two datasets, CVI finds a better solution with much less computation compared to the traditional CTM algorithm. This significant speed-up is due to the fact that CVI avoids the costly optimization of the ELBO to obtain the update for the mean m_d and the variance v_d of $q(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{v}_d)$, the variational distribution for topic proportions. As we have seen in the previous section, the CTM algorithm performs two full gradient-based optimizations for every document d until the local parameters converge. Instead, the CVI algorithm updates the natural parameter of $q(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{v}_d)$ by taking only one mirror descent step, from which the updates for the mean are directly obtained in closed-form.

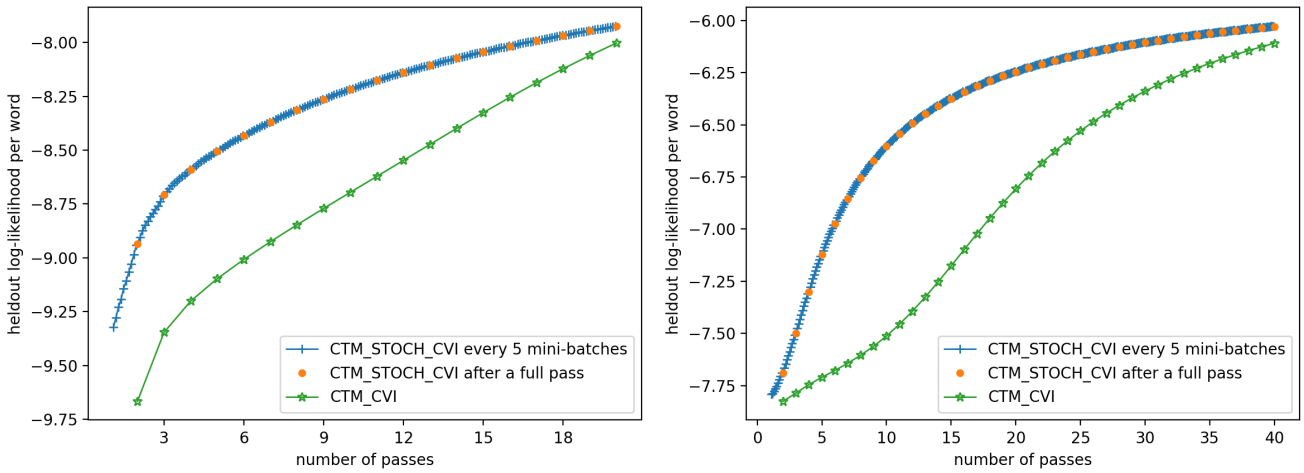
5.2.4 Stochastic CVI performance evaluation

We also compare the performances of the batch and stochastic CVI algorithms. We use the same datasets and the same partitioning as in the previous experiment. For the 20Newsgroups corpus, we set the the forgetting rate $\kappa = 0.7$, the delay $\tau = 10$ and the mini-batch size $S = 150$. For de-news, $\kappa = 0.6$, $\tau = 64$ and $S = 140$. For both algorithms, the step-size for the local variational parameters β_t is set to 0.7.
For each iteration of the stochastic algorithm, we evaluate the heldout log-likelihood after seeing 5 mini-batches, and after seeing the entire collection of training documents. Figure 5.5 shows that stochastic CVI converges much faster to similar solutions as batch CVI.



(A) 20Newsgroups corpus, $K = 20$ (B) de-news corpus, $K = 10$

FIGURE 5.4: Heldout average log-likelihood obtained on the 20Newsgroups (left) and de-news (right) corpora as a function of CPU time



(A) 20Newsgroups corpus, $K = 20$ (B) de-news corpus, $K = 10$

FIGURE 5.5: Heldout average log-likelihood obtained on the 20Newsgroups (left) and de-news (right) corpora as a function of the number of full passes through the data (epochs)

Conclusion

We have studied the Correlated Topic Model, which is an extension of LDA, using a non-conjugate logistic-normal prior, in order to model topic correlation patterns with a Gaussian covariance matrix. The CTM posterior is intractable, and the logistic-normal prior distribution induces a non-conjugate term in the model.

We derive and implement Conjugate-Computation variational inference algorithm for the CTM. CVI performs a stochastic mirror descent update in mean-parameter space, and provides an efficient and modular method for variational inference in non-conjugate models. Using CVI, we convert posterior inference in the CTM into inference over two conjugate models : LDA and a linear model. This allows us to leverage already-existing implementations of the two simpler models.
We evaluate the algorithms on different corpora. We validate that accounting for the correlations between topics improves the predictive performance of the model and yields a more accurate prediction in a classification task. We show that the CVI algorithm converges to similar or better solutions compared with the traditional variational inference method for CTM, with much less computation. Furthermore, we derive a stochastic version of the CVI algorithm that converges faster than the batch CVI algorithm.

The CTM eliminates the assumption made in LDA regarding the independence of topic occurrences. However, both LDA and the CTM assume that documents are exchangeable within the corpus. This assumption is ill-suited for corpora with evolving content with respect to time. The Dynamic Topic Model (DTM) [Blei and Lafferty, 2006] is a more sophisticated extension of LDA which captures the evolution of topics when the collections of documents are organized sequentially. The data is divided by time-slices. The documents of each time-slice are modeled by a topic model, and the topics evolve between the slices. A logistic normal distribution is used to model uncertainty about the time-series topics, and induces non-conjugate terms. Deriving a CVI approach for posterior approximation in the DTM could be an interesting next step.

Bibliography

[Amari, 1998] Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

[Atchison and Shen, 1980] Atchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272.

[Bishop, 1999] Bishop, C. M. (1999). Bayesian pca. In *Advances in neural information processing systems*, pages 382–388.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

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

[Blei et al., 2006] Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

[Blei et al., 2016] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational inference: A review for statisticians. *CoRR*, abs/1601.00670.

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted).

[Blei and Lafferty, 2006] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

[Blei and Lafferty, 2007] Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

[Blei and Lafferty, 2009] Blei, D. M. and Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications*, 10(71):34.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Gelman et al., 2014] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

[Griffiths and Steyvers, 2004] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

[Hoffman et al., 2010] Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

[Hoffman et al., 2013] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

[Khan, 2012] Khan, M. E. (2012). *Variational Learning for Latent Gaussian Models of Discrete Data*. PhD thesis, University of British Columbia.

[Khan and Lin, 2017] Khan, M. E. and Lin, W. (2017). Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *arXiv preprint arXiv:1703.04265*.

[Kuss and Rasmussen, 2005] Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary gaussian process classification. *Journal of machine learning research*, 6(Oct):1679–1704.

[Minka and Lafferty, 2002] Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc.

[Mnih and Salakhutdinov, 2008] Mnih, A. and Salakhutdinov, R. R. (2008). Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264.

[Nigam et al., 2000] Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134.

[Ranganath et al., 2014] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

[Raskutti and Mukherjee, 2015] Raskutti, G. and Mukherjee, S. (2015). The information geometry of mirror descent. In *International Conference on Networked Geometric Science of Information*, pages 359–368. Springer.

[Rue and Held, 2005] Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

[Salimans and Knowles, 2013] Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Anal.*, 8(4):837–882.

[Titsias and Lázaro-gredilla, 2014] Titsias, M. and Lázaro-gredilla, M. (2014). Doubly stochastic variational bayes for non-conjugate inference. In Jebara, T. and Xing, E. P., editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979. JMLR Workshop and Conference Proceedings.

[Wang and Blei, 2013] Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031.

[Wei and Croft, 2006] Wei, X. and Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM.

[Xing et al., 2002] Xing, E. P., Jordan, M. I., and Russell, S. (2002). A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 583–591. Morgan Kaufmann Publishers Inc.