Journal of Machine Learning Research (Submitted) Submitted ; Published

Stochastic Variational Inference

Matt Hoffman
Department of Statistics
Columbia University
MDH2002@COLUMBIA.EDU

David M. Blei
Department of Computer Science
Princeton University
BLEI@CS.PRINCETON.EDU

Chong Wang
Department of Computer Science
Princeton University
CHONGW@CS.PRINCETON.EDU

John Paisley
Department of Computer Science
University of California, Berkeley
JPAISLEY@BERKELEY.EDU

Editor: (Under submission)

Abstract

We develop stochastic variational inference, a scalable algorithm for approximating posterior distributions. We develop this technique for a large class of probabilistic models and we demonstrate it with two probabilistic topic models, latent Dirichlet allocation and the hierarchical Dirichlet process topic model. Using stochastic variational inference, we analyze several large collections of documents: 300K articles from *Nature*, 1.8M articles from *The New York Times*, and 3.8M articles from *Wikipedia*. Stochastic inference can handle the full data, and outperforms traditional variational inference on a subset. (Further, we show that the Bayesian nonparametric topic model outperforms its parametric counterpart.) Stochastic variational inference lets us apply complex Bayesian models to very large data sets.[1]

[1] This paper is under submission. We welcome comments and suggestions.

1. Introduction

Here are several examples of modern data analysis problems. (1) We have an archive of two million books, scanned and stored online. We want to organize these books by subject and build a navigator for users to explore our collection. (2) We have data from an online shopping website containing millions of users' purchase histories as well as descriptions of each item in the catalog. We want to recommend items to users based on this information. (3) We are continuously collecting data from an online feed of photographs. We want to build an classifier from these data. (4) We have measured the gene sequences and many traits of a large population. We want to make hypotheses about connections between genes and traits.

These problems illustrate some of the modern challenges that we face when developing methods for analyzing data. Our data are complex and high-dimensional; we have assumptions to make—from science, intuition, or other data analyses—and these assumptions often take the form of the *hidden structure*, structure that we believe exists in the data but that we cannot directly observe; and finally our data sets are large, possibly even arriving in a never-ending stream.

Statistical machine learning research has addressed some of these challenges by developing the field of probabilistic modeling, a field that provides an elegant approach to developing new methods for analyzing data (Pearl, 1988; Jordan, 1999; Bishop, 2006; Koller and Friedman, 2009). In particular, *probabilistic graphical models* give us a visual language for expressing assumptions about data, and the corresponding *posterior inference algorithms* let us analyze data under those assumptions. With the results of this field, we now enjoy a powerful suite of probability models to connect and combine. And we have general-purpose computational strategies for fitting our models and for estimating the quantities needed to use them.

The problem we face is scale. Inference algorithms of the 1990s and 2000s used to be considered scalable, but they cannot easily handle the amount of data that we described in the four examples above. This is the problem we address in this paper. We present an approach to approximate posterior inference in a large class of probability models that is appropriate for massive data sets, data that might not fit in memory or even be stored locally. Our method does not require clusters of computers or specialized hardware, though it can be further sped up with these amenities.

Our algorithm builds on variational inference, a method that turns complex inference problems into high-dimensional optimization problems (Jordan et al., 1999). Traditionally, this optimization has been solved with a coordinate ascent algorithm, iterating between reanalyzing every data point in the data set and re-estimating parameters that summarize its hidden structure. However, this is inefficient for the large data sets we hope to consider. In this paper we will derive a more efficient algorithm by using stochastic optimization (Robbins and Monro, 1951), an optimization technique that follows noisy estimates of the gradient of the objective. In the context of variational inference, this gives an efficient algorithm that iterates between subsampling the data and adjusting the parameters based only on the subsample. We call our method *stochastic variational inference.*

We will derive stochastic variational inference for a large class of graphical models. We will study its performance on two kinds of probabilistic topic models, which are probabilistic models of text that can uncover the hidden thematic structure in large collections of documents. (For example, topic models are used in problem #1 above.) In particular, we demonstrate stochastic variational inference on latent Dirichlet allocation (Blei et al., 2003), a simple topic model, and the hierarchical Dirichlet process topic model (Teh et al., 2006a), a more flexible model where the number of discovered topics grows with the data. (This latter application demonstrates how to use stochastic variational inference in a variety of Bayesian nonparametric settings.) Stochastic variational inference can efficiently analyze massive data sets with complex probabilistic models.

Technical summary. We now turn to the technical context of our method. In probabilistic modeling, we use latent variables $\boldsymbol{z}$ to encode hidden structure in observed data $\boldsymbol{x}$; we articulate the relationship between the hidden and observed variables with a factorized probability distribution (i.e., a graphical model); and we use inference algorithms to estimate the posterior distribution of the hidden structure given the observations $p(\boldsymbol{z} \,|\, \boldsymbol{x})$. In descriptive tasks, like problems #1 and #4, the posterior provides a new way to explore the data—the organization of books or the connections between genes and traits—with the hidden structure probabilistically "filled in." In predictive tasks, like problems #2 and #3, we use the posterior to form the predictive distribution of a new observation. In particular, the predictive distribution marginalizes out the hidden structure via the posterior and can then be used, for example, to make a recommendation to a user or to classify a new image.

Consider a graphical model of hidden and observed random variables for which we want to compute the posterior distribution $p(\boldsymbol{z} \,|\, \boldsymbol{x})$. For many models of interest, this posterior is not tractable to compute and we must appeal to approximate methods. The two most prominent strategies in statistics and machine learning are Markov chain Monte Carlo (MCMC) sampling and variational inference. In MCMC sampling, we construct a Markov chain over the hidden variables whose stationary distribution is the posterior of interest (Robert and Casella, 2004). We run the chain until it has (hopefully) reached equilibrium and collect samples to approximate the posterior. In variational inference, we define a flexible family of distributions over the hidden variables, indexed by free parameters (Jordan et al., 1999). We then find the setting of the parameters (i.e., the member of the family) that is closest to the posterior. Thus the inference problem turns into an optimization problem.

Neither MCMC nor variational inference scales to the kinds of settings described in the first paragraph. Researchers have proposed speed-ups of both approaches, but these usually are tailored to specific models or compromise the correctness of the algorithm (or both). Here, we develop a general variational method that scales.

As we mentioned above, the main idea is to use stochastic optimization (Robbins and Monro, 1951; Spall, 2003). In stochastic optimization, we find the maximum of an objective function by following noisy estimates of its gradient. If the expectation of the noisy gradient equals the true gradient then stochastic optimization provably converges to an optimum of the objective. Stochastic optimization is particularly attractive when the objective (and therefore its gradient) is a sum of many terms that can be computed independently. In that setting, we can repeatedly subsample the terms to give noisy gradients that are cheaper to compute than the true gradient.

Variational inference is amenable to stochastic optimization because the variational objective decomposes into a sum of terms, one for each data point in the analysis. We can cheaply obtain noisy estimates of the gradient by subsampling the data and computing a scaled gradient on the subsample. If we sample independently then the expectation of this noisy gradient is equal to the true gradient. With one more detail—the idea of a natural gradient (Amari, 1998), a type of gradient that is particularly simple when applied to the variational objective—stochastic variational inference has an attractive form:

1. Subsample one or more data points from the dataset.
2. Analyze the subsample using the current variational parameters.
3. Implement a closed form update of the variational parameters.
4. Repeat.

While traditional algorithms require repeatedly analyzing the whole data set before updating the variational parameters, this algorithm only requires that we analyze randomly sampled subsets. We will show how to use this algorithm for a large class of graphical models.

Related work. Scalable posterior inference has received much attention from the machine learning community. Most research has focused on parallelizing existing algorithms to allow the application of large-scale computation to large-scale data analysis. The topic modeling community has been particularly interested in these approaches, producing parallel variational inference (Zhai et al., 2012) and parallel MCMC (Newman et al., 2009; Smola and Narayanamurthy, 2010; Ahmed et al., 2012).

An alternative approach to large-scale learning uses stochastic optimization. These techniques have a long history in the machine learning literature, going back to the early days of neural networks (Rosenblatt, 1958; Widrow and Hoff, 1960). Interest in stochastic optimization algorithms for machine learning has blossomed again in recent years as the theoretical efficiency of stochastic gradient algorithms has become better understood (LeCun et al., 1998; Bottou and LeCun, 2004; Bottou and Bousquet, 2011). As examples, these algorithms have been successfully adapted to fit support vector machines (Shalev-Shwartz et al., 2007) and perform dictionary learning (Mairal et al., 2010). This work has shown both empirically and theoretically that using stochastic learning can dramatically accelerate learning speed in the presence of large amounts of data, without relying on the availability of vast computational resources.

The work discussed above uses stochastic optimization in non-probabilistic contexts. Sato and Ishii (2000) developed a stochastic variant on the expectation-maximization (EM) algorithm of Dempster et al. (1977) that finds maximum-marginal-likelihood estimates of parameters to probabilistic models, integrating over a set of hidden variables. Neal and Hinton (1999) proposed a similar algorithm, though not couched in the theory of stochastic optimization. Finally, Sato (2001) developed an online variational Bayes algorithm using stochastic optimization, although it can only be applied to the fairly limited set of models that can be fit using the classic EM algorithm. In contrast to Sato's algorithm, stochastic variational inference generalizes to a much wider set of probabilistic models, in particular those that are amenable to closed-form coordinate ascent inference (Bishop et al., 2003; Xing et al., 2003; Beal, 2003).

The organization of this paper. In Section 2, we review variational inference for graphical models and then derive stochastic variational inference. In Section 3, we review probabilistic topic models and Bayesian nonparametric models and then derive the stochastic variational inference algorithms in these settings. In Section 4, we study stochastic variational inference on several large text data sets.

2. Stochastic Variational Inference

We derive *stochastic variational inference*, a stochastic optimization algorithm for mean-field variational inference. Our algorithm approximates the posterior distribution of a probabilistic model with hidden variables, and can handle large (or even streaming) data sets of observations.

We divide this section into four parts.

1. We define the class of models to which our algorithm applies. We define *local* and *global* hidden variables, and requirements on the conditional distributions within the model.

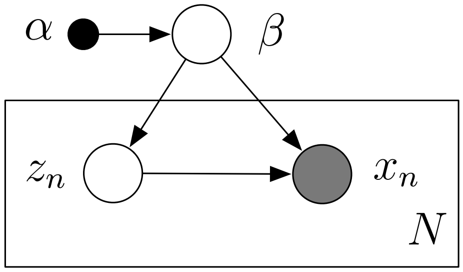


Figure 1: A graphical model with observations $x_{1:N}$, local hidden variables $z_{1:N}$ and global hidden variables β. The distribution of each observation x_n only depends on its corresponding local variable z_n and the global variables β. (Though not pictured, each hidden variable z_n, observation x_n, and global variable β may be a vector consisting of multiple random variables.)

2. We review *mean-field variational inference*, an approximate inference strategy that seeks to find a tractable distribution over the hidden variables that is close to the posterior distribution. We derive the traditional variational inference algorithm in our setting, which is a coordinate ascent algorithm.

3. We review the *natural gradient*, and derive the natural gradient of the variational objective function. The natural gradient closely relates to the coordinate ascent variational inference algorithm.

4. We review *stochastic optimization*, a technique that uses noisy estimates of a gradient to optimize an objective function, and apply it to variational inference. Specifically, we use stochastic optimization with noisy estimates of the natural gradient of the variational objective. These estimates arise from repeatedly subsampling the data set. We show how stochastic variational inference easily builds on traditional variational inference algorithms, but can handle much larger data sets.

2.1 Models with local and global hidden variables

We study models involving observations, global hidden variables, local hidden variables, and fixed parameters. The N observations are $\boldsymbol{x} = x_{1:N}$; the vector of global hidden variables is β; the N local hidden variables are $\boldsymbol{z} = z_{1:N}$, each of which is a collection of J variables $z_n = z_{n,1:J}$; the vector of fixed parameters is α.[2]

[2] The fixed parameters α can contain parameters that partly govern any of the random variables (e.g., fixed parts of the conditional distribution of observations). However, to keep notation simple, we consider them only to govern the global hidden variables.

The distinction between local and global hidden variables is determined by the conditional dependencies that define the model. In particular, the nth observation x_n and the nth local hidden variable z_n are conditionally independent (given β) of all other observations and local hidden variables x_{-n} and z_{-n}. (The notation x_{-n} refers to the set of variables except the nth.) That is, the joint distribution factorizes into a global term and a product of local terms,

$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{n=1}^{N} p(z_n, x_n \,|\, \beta). \tag{1}$$

Figure 1 illustrates the graphical model. Our goal is to approximate the posterior distribution of the hidden variables given the observations, $p(\boldsymbol{z}, \beta \,|\, \boldsymbol{x})$.

This kind of model frequently arises in Bayesian statistics. In that setting, the global variables β are parameters endowed with a prior $p(\beta)$ and each local variable z_n contains the hidden structure that governs the nth observation. For example, consider a Bayesian mixture of Gaussians. The global variables are the mixture proportions and the means and variances of the mixture components; the local variable z_n is the hidden cluster label for the nth observation x_n.

We have described the independence assumptions of the hidden variables. We make further assumptions about the *complete conditionals* in the model. A complete conditional is the conditional distribution of a hidden variable given the other hidden variables and the observations. We assume that these distributions are in the exponential family,

$$p(\beta \,|\, \boldsymbol{x}, \boldsymbol{z}) = h(\beta) \exp\{\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha)^\top t(\beta) - a_g(\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha))\} \tag{2}$$

$$p(z_{n,j} \,|\, x_n, z_{n,-j}, \beta) = h(z_{n,j}) \exp\{\eta_{\ell,j}(x_n, z_{n,-j}, \beta)^\top t(z_{n,j}) - a_{\ell,j}(\eta_{\ell,j}(x_n, z_{n,-j}, \beta))\}. \tag{3}$$

The scalar functions $h(\cdot)$ and $a(\cdot)$ are respectively the *base measure* and *log-normalizer*; the vector functions $\eta(\cdot)$ and $t(\cdot)$ are respectively the *natural parameter* and *sufficient statistics*.[3] These are conditional distributions, so the natural parameter is a function on the variables that are being conditioned on. (The subscripts on the natural parameter η indicate complete conditionals for local or global variables.) Notice that for the local variables $z_{n,j}$, the complete conditional distribution is determined by the global variables β and the other local variables in the nth context, i.e., the nth data point x_n and the local variables $z_{n,-j}$. This follows from the factorization in Equation 1.

These assumptions on the complete conditionals imply a conjugacy relationship between the global variables β and the local contexts (z_n, x_n); and this relationship

[3] We use overloaded notation for the functions $h(\cdot)$ and $t(\cdot)$ so that they depend on the names of their arguments; for example, $h(z_{n,j})$ can be thought of as a shorthand for the more formal (but more cluttered) notation $h_{z_{n,j}}(z_{n,j})$.

implies a specific form of the complete conditional for β. Specifically, the distribution of the local context given the global variables is in an exponential family,

$$p(x_n, z_n|\beta) = h(x_n, z_n) \exp\{g(\beta)^\top t(x_n, z_n) - a_\ell(g(\beta))\}. \tag{4}$$

Note that the global variables are passed through a function $g(\beta)$ to form the natural parameter. The prior distribution $p(\beta)$ is also in an exponential family,

$$p(\beta) = h(\beta) \exp\{\alpha^\top t(\beta) - a_g(\alpha)\}, \tag{5}$$

where the sufficient statistics are $t(\beta) = \langle g(\beta), -a_\ell(g(\beta))\rangle$ and thus the hyperparameter α has two components $\alpha = \langle \alpha_1, \alpha_2\rangle$. The first component is a vector of the same dimension as $g(\beta)$; the second is a scalar.

Equations 4 and 5 imply that the complete conditional for the global variable in Equation 2 is in the same exponential family as the prior with parameter

$$\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha) = (\alpha_1 + \textstyle\sum_{n=1}^N t(z_n, x_n), \alpha_2 + N). \tag{6}$$

This form will be important when we derive stochastic variational inference in Section 2.4. For a general discussion of conjugacy and the exponential family, see Bernardo and Smith (1994).

This general family of distributions—those with local and global variables, and where the complete conditionals are in the exponential family—contains many useful statistical models from the machine learning and statistics literature. Examples include Bayesian mixture models (Attias, 2000), latent Dirichlet allocation (Blei et al., 2003), hidden Markov models (and many variants) (Rabiner, 1989; Fine et al., 1998; Fox et al., 2011b; Paisley and Carin, 2009a), Kalman filters (and many variants) (Kalman, 1960; Fox et al., 2011a), factorial models (Ghahramani and Jordan, 1997), hierarchical linear regression models (Gelman and Hill, 2007), hierarchical probit classification models (McCullagh and Nelder, 1989; Girolami and Rogers, 2006), probabilistic matrix factorization models (Tipping and Bishop, 1999; Collins et al., 2002; Wang, 2006; Salakhutdinov and Mnih, 2008; Paisley and Carin, 2009b; Hoffman et al., 2010b), and certain Bayesian nonparametric mixture models (Antoniak, 1974; Escobar and West, 1995; Teh et al., 2006a).[4]

Analyzing data with one of these models amounts to computing the posterior distribution of the hidden variables given the observations:

$$p(\boldsymbol{z}, \beta \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z}, \beta)}{\int_\beta \int_z p(\boldsymbol{x}, \boldsymbol{z}, \beta)}. \tag{7}$$

[4] We note that our assumptions can be relaxed to the case where the full conditional $p(\beta|x, z)$ is not tractable, but each partial conditional $p(\beta_k|x, z, \beta_{-k})$ associated with the global variable β_k is in a tractable exponential family. The topic models of the next section do not require this complexity, so we chose to keep the derivation a little simpler.

We then use this posterior to explore the hidden structure of our data or to make predictions about future data. Unfortunately, for many models (such as the examples listed above) we cannot directly compute the posterior because the denominator is an intractable integral. Thus we resort to approximate posterior inference, a problem that has been a focus of modern Bayesian statistics. We now turn to mean-field variational inference, which roots our strategy for scalable approximate inference.

2.2 Mean-field variational inference

Variational inference turns the inference problem into an optimization problem. We introduce a family of distributions over the hidden variables that is indexed by a set of free parameters, and then optimize those parameters to find the member of the family that is closest to the posterior of interest. (Closeness is measured with Kullback-Leibler divergence.) We use the resulting distribution, called the *variational distribution*, to approximate the posterior.

In this section we review mean-field variational inference, which uses a variational family where each hidden variable is independent. We describe the variational objective function, discuss the mean-field variational family, and derive a coordinate ascent algorithm for fitting the variational parameters. This coordinate-ascent algorithm will be a stepping stone to stochastic variational inference.

The evidence lower bound. Variational inference minimizes the Kullback-Leibler (KL) divergence from the variational distribution to the posterior distribution by maximizing the *evidence lower bound* (ELBO), a lower bound on the logarithm of the marginal probability of the observations $\log p(\boldsymbol{x})$. The ELBO is equal to the KL divergence up to an additive constant.

We derive the ELBO by introducing a distribution over the hidden variables $q(\boldsymbol{z}, \beta)$ and using Jensen's inequality,[5]

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int_{\boldsymbol{z},\beta} p(\boldsymbol{x}, \boldsymbol{z}, \beta) && (8)\\
&= \log \int_{\boldsymbol{z},\beta} p(\boldsymbol{x}, \boldsymbol{z}, \beta) \frac{q(\boldsymbol{z}, \beta)}{q(\boldsymbol{z}, \beta)} && (9)\\
&= \log \left(\mathbb{E}_q \left[\frac{p(\boldsymbol{x}, \boldsymbol{z}, \beta)}{q(\boldsymbol{z}, \beta)} \right] \right) && (10)\\
&\geq \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z}, \beta)] - \mathbb{E}_q[\log q(\boldsymbol{z}, \beta)] && (11)\\
&\triangleq \mathcal{L}(q). && (12)
\end{aligned}$$

[5] Recall that Jensen's inequality implies that the logarithm of a function's expected value is greater than or equal to the expected value of the logarithm.

The ELBO contains two terms. The first term is the expected log joint, $\mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{z}, \beta)]$. The second term is the entropy of the variational distribution, $-\mathbb{E}_q[\log q(\boldsymbol{z}, \beta)]$. Both these terms depend on $q(\boldsymbol{z}, \beta)$, a distribution of the hidden variables.

We restrict $q(\boldsymbol{z}, \beta)$ to be in a family that is tractable, i.e., one for which the expectations in the ELBO can be efficiently computed, and we try to find the member of the family that maximizes the ELBO. Solving this maximization problem is equivalent to finding the member of the family that is closest in KL divergence to the posterior (Jordan et al., 1999; Wainwright and Jordan, 2008),

$$\begin{aligned}
\mathrm{KL}(q(\boldsymbol{z}, \beta)||p(\boldsymbol{z}, \beta|\boldsymbol{x})) &= \mathbb{E}_q\left[\log q(\boldsymbol{z}, \beta)\right] - \mathbb{E}_q\left[p(\boldsymbol{z}, \beta \,|\, x)\right] \\
&= \mathbb{E}_q\left[\log q(\boldsymbol{z}, \beta)\right] - \mathbb{E}_q\left[p(\boldsymbol{x}, \boldsymbol{z}, \beta)\right] + \log p(\boldsymbol{x}) \\
&= -\mathcal{L}(q) + \text{const.}
\end{aligned}$$

Note we absorbed $\log p(\boldsymbol{x})$ into a constant because it does not depend on q. We use the optimized distribution $q^*(\boldsymbol{z}, \beta)$ as a proxy for the posterior.

The mean-field variational family. The simplest variational family of distributions is the *mean-field family*. In this family, each hidden variable is independent and governed by its own parameter,

$$q(\boldsymbol{z}, \beta) = q(\beta \,|\, \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{n,j} \,|\, \phi_{n,j}). \tag{13}$$

The global parameters λ govern the global variables; the local parameters ϕ_n govern the local variables in the nth context. The ELBO is a function of these parameters.

Equation 13 gives the factorization of the variational family, but does not specify its form. We set $q(\beta|\lambda)$ and $q(z_{n,j}|\phi_{n,j})$ to be in the same exponential family as the complete conditional distributions $p(\beta|\boldsymbol{x}, \boldsymbol{z})$ and $p(z_{n,j}|x_n, z_{n,-j}, \beta)$ (from Equations 2 and 3) and since we focus on conjugate exponential family models these conditionals are in the same family as the prior. The variational parameters λ and $\phi_{n,j}$ are the natural parameters to those families,

$$q(\beta \,|\, \lambda) = h(\beta) \exp\{\lambda^\top t(\beta) - a_g(\lambda)\} \tag{14}$$

$$q(z_{n,j} \,|\, \phi_{n,j}) = h(z_{n,j}) \exp\{\phi_{n,j}^\top t(z_{n,j}) - a_{\ell,j}(\phi_{n,j})\}. \tag{15}$$

Again, we overload the base measures $h(\cdot)$ and sufficient statistics $t(\cdot)$. Assuming that these exponential families are the same as their corresponding conditionals means that $t(\cdot)$ and $h(\cdot)$ in Equation 14 are the same functions as $t(\cdot)$ and $h(\cdot)$ in Equation 2. Likewise, $t(\cdot)$ and $h(\cdot)$ in Equation 15 are the same as in Equation 3. We will sometimes suppress the explicit dependence on ϕ and λ, substituting $q(z_{n,j})$ for $q(z_{n,j}|\phi_{n,j})$ and $q(\beta)$ for $q(\beta|\lambda)$.

The mean-field family has several computational advantages. At the outset, the entropy term decomposes,

$$-\mathbb{E}_q[\log q(\boldsymbol{z}, \beta)] = -\mathbb{E}_\lambda[\log q(\beta)] - \sum_{n=1}^{N} \sum_{j=1}^{J} \mathbb{E}_{\phi_{n,j}}[\log q(z_{n,j})], \tag{16}$$

where $\mathbb{E}_{\phi_{n,j}}[\cdot]$ denotes an expectation with respect to $q(z_{n,j} \,|\, \phi_{n,j})$ and $\mathbb{E}_\lambda[\cdot]$ denotes an expectation with respect to $q(\beta \,|\, \lambda)$.

The gradient of the ELBO and coordinate ascent inference. We have defined the objective function in Equation 11 and the variational family in Equation 13. Our goal is to optimize the objective with respect to the variational parameters.

In traditional mean-field variational inference, we optimize Equation 11 with coordinate ascent. We iteratively optimize each variational parameter, holding the other parameters fixed. With the assumptions that we have made about the model and variational distribution—that each conditional is in an exponential family and that the corresponding variational distribution is in the same exponential family—we can optimize each coordinate in closed form.

We first derive the coordinate update for the parameter to the variational distribution of the global variables $q(\beta \,|\, \lambda)$. As a function of λ, we can rewrite the objective as

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta \,|\, \boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_q[\log q(\beta)] + \text{const.} \tag{17}$$

The first two terms are expectations that involve β; the third term is constant with respect to λ. The constant absorbs quantities that relate to the other hidden variables. Those quantities do not depend on $q(\beta \,|\, \lambda)$ because all variables are independent in the mean-field family.

Equation 17 reproduces the full ELBO in Equation 11. The second term of Equation 17 is the entropy of the global variational distribution. The first term derives from the expected log joint likelihood, where we use the chain rule to separate terms that depend on the variable β from terms that do not,

$$\mathbb{E}_q[\log p(\beta, \boldsymbol{z}, \boldsymbol{x})] = \mathbb{E}_q[\log p(\boldsymbol{z}, \boldsymbol{x})] + \mathbb{E}_q[\log p(\beta \,|\, \boldsymbol{z}, \boldsymbol{x})]. \tag{18}$$

The constant absorbs $\mathbb{E}_q[\log p(\boldsymbol{z}, \boldsymbol{x})]$, leaving the expected log conditional $\mathbb{E}_q[\log p(\beta \,|\, \boldsymbol{z}, \boldsymbol{x})]$.

Finally, we substitute the form of $q(\beta \,|\, \lambda)$ in Equation 14 to obtain the final expression for the ELBO as a function of λ,

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(\boldsymbol{z}, \boldsymbol{x}, \alpha)]^\top \nabla_\lambda a(\lambda) - \lambda^\top \nabla_\lambda a(\lambda) + a(\lambda) + \text{const.} \tag{19}$$

In the first term of this expression, we used the exponential family identity that the expectation of the sufficient statistics is the gradient of the log normalizer, $\mathbb{E}_\lambda[t(\beta)] =$

$\nabla_\lambda a(\lambda)$. The constant has further absorbed the expected log normalizer of the conditional distribution $-\mathbb{E}_q[a(\eta_g(\boldsymbol{z}, \boldsymbol{x}, \alpha))]$.

Equation 19 simplifies the ELBO as a function of the global variational parameter. To derive the coordinate ascent update, we take its gradient,

$$\nabla_\lambda \mathcal{L} = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_\phi[\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha)] - \lambda). \tag{20}$$

We can zero this gradient by setting

$$\lambda = \mathbb{E}_\phi[\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha)]. \tag{21}$$

This sets the global variational parameter equal to the expected natural parameter of its complete conditional distribution. Implementing this update, holding all other variational parameters fixed, optimizes the ELBO over λ. Notice that the mean-field assumption plays an important role. The update is the expected conditional parameter $\mathbb{E}_\phi[\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha)]$, which is an expectation of a function of the other random variables and observations. Thanks to the mean-field assumption, this expectation is only a function of the local variational parameters and does not depend on λ.[6]

The gradient for each local parameter $\phi_{n,j}$ is nearly identical to the global case,

$$\nabla_{\phi_{n,j}} \mathcal{L} = \nabla^2_{\phi_{n,j}} a_{\ell,j}(\phi_{n,j})(\mathbb{E}_{\lambda,\phi_{n,-j}}[\eta_{\ell,j}(x_n, z_{n,-j}, \beta)] - \phi_{n,j}). \tag{22}$$

It equals zero when

$$\phi_{n,j} = \mathbb{E}_{\lambda,\phi_{n,-j}}[\eta_\ell(x_n, z_{n,-j}, \beta)]. \tag{23}$$

Mirroring the global update, this expectation does not depend on $\phi_{n,j}$. However, while the global update depends on all the local variational parameters, this update only depends on the global parameters and the other parameters associated with the nth context.

The updates in Equations 21 and 23 form the algorithm for coordinate ascent variational inference, iterating between updating each local parameter and the global parameters. The full algorithm is in Figure 2. It is guaranteed to find a stationary point of the ELBO. It is the “classical” variational inference algorithm, used in many settings.

As an aside, these updates reveal a connection between mean-field variational inference and Gibbs sampling (Gelfand and Smith, 1990). In Gibbs sampling, we iteratively sample from each complete conditional. In variational inference, we take variational expectations of the natural parameters of the same distributions.

[6] Computing this expectation for specific models is easy for directed graphical models with tractable complete conditionals. The details use the relationship between the sufficient statistics and log-normalizers. In Section 3, we show that these updates are tractable for many topic models.

```
1: Initialize $\lambda^{(0)}$ randomly.
2: repeat
3:     for each local variational parameter $\phi_{n,j}$ do
4:         Update $\phi_{n,j}$, $\phi_{n,j}^{(t)} = \mathbb{E}_{\lambda^{(t-1)}}[\eta_{\ell,j}(x_n, z_{n,-j}, \beta)]$.
5:     end for
6:     Update the global variational parameters, $\lambda^{(t)} = \mathbb{E}_{\phi^{(t)}}[\eta_g(z_{1:N}, x_{1:N})]$.
7: until the ELBO converges
```

Figure 2: Coordinate ascent mean-field variational inference.

The local steps (Steps 3 and 4 in Figure 2) are trivially parallelizable. The data can be distributed across many machines and the local variational updates can be implemented in parallel. These results can then be aggregated in Step 6 to find the new global variational parameters.

However, the local steps also reveal an inefficiency in the algorithm. The algorithm begins by initializing the global parameters λ randomly—the initial value of λ does not reflect any regularity in the data. But before completing even one iteration, the algorithm must analyze every data point using these initial (random) values. This is wasteful, especially if we expect that we can learn something about the global variational parameters from only a subset of the data. Further, if the data are "infinite", i.e., if they represent a data source where information arrives in a constant stream, then this algorithm can never complete even one iteration.

We will solve this problem with stochastic optimization. With stochastic optimization we can handle massive and streaming data sets, continually improving our estimates of the global variational parameters as we analyze more observations. The efficiency of our stochastic optimization algorithm hinges on using the *natural gradients* of the variational objective, and on the relationship between the natural gradients and the coordinate updates that we derived here. We will next discuss natural gradients and their role in mean-field variational inference.

2.3 The natural gradient of the ELBO

The natural gradient of a function accounts for the geometry of its parameter space, using a Riemannian metric to adjust the direction of the traditional gradient. Amari (1998) discusses natural gradients for maximum-likelihood estimation, which give faster convergence than standard gradients. In this section we describe Riemannian metrics for probability distributions and the natural gradient of the ELBO.

Gradients and probability distributions. The classical gradient method for maximization tries to find a maximum of a function $f(\lambda)$ by taking a series of steps of size ρ in the direction of the gradient,

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho \nabla_\lambda f(\lambda^{(t)}). \tag{24}$$

The gradient points in the direction of steepest ascent. That is, the gradient $\nabla_\lambda f(\lambda)$ points in the same direction as

$$\arg \max_{d\lambda} f(\lambda + d\lambda) \quad \text{subject to } ||d\lambda||^2 < \epsilon^2 \tag{25}$$

for sufficiently small ϵ, assuming that f is differentiable at λ. In words, Equation 25 implies that if we could only move a tiny distance ϵ away from λ then we should move in the direction of the gradient. Initially this seems reasonable, but there is a complication. The gradient direction implicitly depends on the Euclidean distance metric associated with the space in which λ lives. There is no reason to think that the Euclidean metric captures an inherently meaningful notion of distance between settings of λ; for example, it is sensitive to rescalings of the elements of λ, a sensitivity that leads to moves in suboptimal directions.

The problem with Euclidean distance is especially clear in our setting, where we are trying to optimize the ELBO with respect to the parameter of a probability distribution. For the moment, we focus on the variational distribution of the global variable $q(\beta|\lambda)$. When optimizing over a probability distribution, the Euclidean distance between two parameter vectors λ and λ' is often not the best way of measuring the dissimilarity of the distributions $q(\beta \,|\, \lambda)$ and $q(\beta \,|\, \lambda')$. For example, consider $q(\beta)$ to be a univariate normal and λ to be the mean μ and scale σ of that distribution. The distributions $\mathcal{N}(0, 10000)$ and $\mathcal{N}(10, 10000)$ are almost indistinguishable, and the Euclidean distance between their parameter vectors is 10. By contrast, the distributions $\mathcal{N}(0, 0.01)$ and $\mathcal{N}(0.1, 0.01)$ barely overlap, but this is not reflected in the Euclidean distance between their parameter vectors, which is only 0.1.

Natural gradients and probability distributions. A natural measure of dissimilarity between probability distributions is the symmetrized KL divergence

$$D^{\text{sym}}_{KL}(\lambda, \lambda') = \mathbb{E}_\lambda \left[\log \tfrac{q(\beta \,|\, \lambda)}{q(\beta \,|\, \lambda')} \right] + \mathbb{E}_{\lambda'} \left[\log \tfrac{q(\beta \,|\, \lambda')}{q(\beta \,|\, \lambda)} \right]. \tag{26}$$

Symmetrized KL depends on the distributions themselves, rather than on how they are parameterized; it is invariant to parameter transformations.

With distances defined using symmetrized KL, we find the direction of steepest ascent in the same way as for gradient methods,

$$\arg \max_{d\lambda} f(\lambda + d\lambda) \quad \text{subject to } D^{\text{sym}}_{KL}(\lambda, \lambda + d\lambda) < \epsilon. \tag{27}$$

As $\epsilon \to 0$, the solution to this problem points in the same direction as the *natural gradient.* While the Euclidean gradient points in the direction of steepest ascent in Euclidean space, the natural gradient points in the direction of steepest ascent in the Riemannian space, i.e., the space where local distance is defined by KL divergence rather than the L^2 norm.

We manage the more complicated constraint in Equation 27 with a Riemannian metric $G(\lambda)$. This metric defines linear transformations of λ under which the squared Euclidean distance between λ and a nearby vector $\lambda + d\lambda$ is the KL between $q(\beta|\lambda)$ and $q(\beta|\lambda + d\lambda)$,

$$d\lambda^T G(\lambda) d\lambda = D_{KL}^{\text{sym}}(\lambda, \lambda + d\lambda), \tag{28}$$

and note that the transformation can be a function of λ. For a distance measure (such as symmetrized KL) and its corresponding metric $G(\lambda)$, we find the natural gradient by premultiplying the gradient by $G^{-1}(\lambda)$ (Amari, 1998),

$$\hat{\nabla}_\lambda f(\lambda) \triangleq G(\lambda)^{-1} \nabla_\lambda f(\lambda). \tag{29}$$

When the distance is the symmetrized KL divergence of Equation 26, the Riemannian metric is the Fisher information matrix G (Amari, 1982; Kullback and Leibler, 1951),

$$G(\lambda) = \mathbb{E}_\lambda \left[(\nabla_\lambda \log q(\beta \,|\, \lambda))(\nabla_\lambda \log q(\beta \,|\, \lambda))^\top \right]. \tag{30}$$

We can show that Equation 30 satisfies Equation 28 by approximating $\log q(\beta \,|\, \lambda + d\lambda)$ using a first-order Taylor approximation of $\log q$ about λ and plugging the result into Equation 26. When $q(\beta \,|\, \lambda)$ is in the exponential family (Equation 14) the metric is the second derivative of the log normalizer,

$$\begin{aligned} G(\lambda) &= \mathbb{E}_\lambda \left[(\nabla_\lambda \log p(\beta \,|\, \lambda))(\nabla_\lambda \log p(\beta \,|\, \lambda))^\top \right] \\ &= \mathbb{E}_\lambda \left[(t(\beta) - \mathbb{E}_\lambda[t(\beta)])(t(\beta) - \mathbb{E}_\lambda[t(\beta)])^\top \right] \\ &= \nabla_\lambda^2 a(\lambda). \end{aligned} \tag{31}$$

This is due to the exponential family identity that the Hessian of the log normalizer function a with respect to the natural parameter λ is the covariance matrix of the sufficient statistic vector $t(\beta)$.

Natural gradients and mean field variational inference. We now return to variational inference, and compute the natural gradient of the ELBO with respect to the variational parameters. Researchers have used the natural gradient in variational inference for nonlinear state space models (Honkela et al., 2008) and Bayesian mixtures (Sato, 2001).[7]

[7] Our work here—using the natural gradient in a stochastic optimization algorithm—is closest to that of Sato (2001), though we develop the algorithm via a different path and that paper does not address models for which the joint conditional $p(z_n|\beta, x_n)$ is not tractable.

Consider the global variational parameter λ. The gradient of the ELBO with respect to λ is in Equation 20. Since λ is a natural parameter to an exponential family distribution, the Fisher metric defined by $q(\beta)$ is $\nabla^2_\lambda a(\lambda)$. Note that the Fisher metric is the first term in Equation 20. We premultiply the gradient by the inverse Fisher information to find the natural gradient. This reveals that the natural gradient has the following simple form,

$$\hat{\nabla}_\lambda \mathcal{L} = \mathbb{E}_\phi[\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha)] - \lambda. \tag{32}$$

An analogous computation goes through for the local variational parameters,

$$\hat{\nabla}_{\phi_{n,j}} \mathcal{L} = \mathbb{E}_{\lambda, \phi_{n,-j}}[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{n,j}. \tag{33}$$

The natural gradients are closely related to the coordinate ascent updates. Consider a full set of variational parameters λ and $\boldsymbol{\phi}$. We can compute the natural gradient by computing the coordinate updates in parallel, i.e., the first terms in Equation 32 and Equation 33, and subtracting the current setting of the parameters. The classical coordinate ascent algorithm of Figure 2 was not derived as a natural gradient algorithm, but it can be interpreted as a projected natural gradient algorithm; updating any single parameter λ or $\phi_{n,j}$ by taking a natural gradient step of length 1 is equivalent to performing the update in Equation 21 or Equation 23.

We have motivated natural gradients by mathematical reasoning around the geometry of the parameter space. Even more important in this setting, however, is that natural gradients are easier to compute than classical gradients, and this opens the door to efficient gradient-based algorithms for variational inference. They are easier to compute because premultiplying by the Fisher information matrix—which we must do to compute the classical gradient but which is not needed to compute the natural gradient—is prohibitively expensive for variational parameters with many components. As we will see in the next section, being able to efficiently compute gradients lets us develop scalable variational inference algorithms.

2.4 Stochastic variational inference

We have reviewed mean-field variational inference in models with exponential family conditionals and showed that the natural gradient of the variational objective function is easy to compute. We now discuss *stochastic optimization*, optimization with repeated noisy estimates of the gradient, which is the basis for scalable variational inference.

Stochastic optimization. Stochastic optimization algorithms follow noisy estimates of the gradient with a decreasing step size. Noisy estimates of a gradient are often cheaper to compute than the true gradient, and following such estimates can allow algorithms to escape shallow local optima of complex objective functions. In statistical estimation problems, including variational inference of the global parameters, the gradient can be written as a sum of terms for each data point and we can compute a fast noisy approximation by subsampling the data. With certain conditions on the step-size schedule, these algorithms provably converge to an optimum (Spall, 2003; Robbins and Monro, 1951). Spall (2003) gives an overview of stochastic optimization; Bottou (2003) gives an overview of its role in machine learning.

Consider an objective function $f(\lambda)$ and a random function $B(\lambda)$ that has expectation equal to the gradient $\mathbb{E}_q[B(\lambda)] = \nabla_\lambda f(\lambda)$. The stochastic gradient algorithm, which is a type of stochastic optimization, optimizes $f(\lambda)$ by iteratively following realizations of $B(\lambda)$. At iteration t, the update for λ is

$$\lambda^{(t)} = \lambda^{(t-1)} + \rho_t b_t(\lambda^{(t-1)}), \tag{34}$$

where b_t is an independent draw from the noisy gradient B. If the sequence of step sizes ρ_t satisfies

$$\begin{aligned} \textstyle\sum \rho_t &= \infty \\ \textstyle\sum \rho_t^2 &< \infty \end{aligned} \tag{35}$$

then λ^t will converge to the optimal λ^* (if f is convex) or a local optimum of f (if not convex).[8] The same results apply if we premultiply the noisy gradients b_t by a sequence of positive-definite matrices G_t^{-1} (whose eigenvalues are bounded). The resulting algorithm is

$$\lambda^t = \lambda^{t-1} + \rho_t G_t^{-1} b_t(\lambda^{t-1}). \tag{36}$$

As our notation suggests, we will use the Fisher metric for G_t, replacing stochastic Euclidean gradients with stochastic natural gradients.

Stochastic variational inference. The coordinate ascent algorithm in Figure 2 is inefficient for large data sets because we must optimize the local variational parameters for each data point before re-estimating the global variational parameters. Stochastic variational inference uses stochastic optimization to fit the global variational parameters. We repeatedly subsample the data to form noisy estimates of the natural gradient of the ELBO, and we follow these estimates with a decreasing step-size.

The resulting algorithm is in Figure 3. At each iteration we have a current setting of the global variational parameters. We repeat the following steps:

1. Sample a data point from the set; compute its local variational parameters.

[8] To find a local optimum, f must be three-times differentiable and meet a few other technical requirements (Bottou, 2003). The variational objective satisfies these criteria.

2. Form intermediate global variational parameters, as though we were running classical coordinate ascent and the sampled data point were repeated N times to form the collection.

3. Update the global variational parameters to be a weighted average of the intermediate parameters and their current setting.

We now show that this algorithm is stochastic natural gradient ascent of the global variational parameters.

Our goal is to find a setting of the global variational parameters λ that maximizes the ELBO. We define $\mathcal{L}(\lambda)$ to be the ELBO when λ is held fixed and the local variational parameters ϕ are set to their optimal value,

$$\mathcal{L}(\lambda) \triangleq \max_{\phi} \mathcal{L}(\lambda, \phi). \tag{37}$$

Holding the global parameters λ fixed, let $\phi(\lambda)$ be all the optimal local parameters.

We can compute the (natural) gradient of $\mathcal{L}(\lambda)$ by first finding the corresponding optimal local parameters $\phi(\lambda)$ and then computing the (natural) gradient of $\mathcal{L}(\lambda, \phi(\lambda))$, holding $\phi(\lambda)$ fixed. The reason is that the gradient of $\mathcal{L}(\lambda)$ with respect to λ is the same as the gradient of the two-parameter ELBO $\mathcal{L}(\lambda, \phi(\lambda))$,

$$\begin{aligned} \nabla_\lambda \mathcal{L}(\lambda) &= \nabla_\lambda \mathcal{L}(\lambda, \phi(\lambda)) + (\nabla_\lambda \phi(\lambda))^\top \nabla_\phi \mathcal{L}(\lambda, \phi(\lambda)) && (38) \\ &= \nabla_\lambda \mathcal{L}(\lambda, \phi(\lambda)), && (39) \end{aligned}$$

where $\nabla_\lambda \phi(\lambda)$ is the Jacobian of $\phi(\lambda)$ and we use the fact that the gradient of $\mathcal{L}(\lambda, \phi)$ with respect to ϕ is zero at $\phi(\lambda)$.

Stochastic variational inference optimizes the maximized ELBO $\mathcal{L}(\lambda)$ by subsampling the data to form noisy estimates of the natural gradient. First, we decompose $\mathcal{L}(\lambda)$ into a global term and a sum of local terms,

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + \sum_{n=1}^{N} \max_{\phi_n} (\mathbb{E}_q[\log p(z_n, x_n \,|\, \beta)] - \mathbb{E}_q[\log q(z_n)]). \tag{40}$$

Now consider a variable that chooses an index of the data uniformly at random, $I \sim \text{Unif}(1, \ldots, N)$. Define $\mathcal{L}_I(\lambda)$ to be the following random function of the variational parameters,

$$\mathcal{L}_I(\lambda) \triangleq \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + N \max_{\phi_I} \left(\mathbb{E}_q[\log p(z_I, x_I \,|\, \beta) - \mathbb{E}_q[\log q(z_I)]\right). \tag{41}$$

The expectation of $\mathcal{L}_I$ is equal to the objective in Equation 40. Therefore, the natural gradient of $\mathcal{L}_I$ with respect to each global variational parameter λ is a noisy but unbiased estimate of the natural gradient of the variational objective. This process—sampling a data point and then computing the natural gradient of $\mathcal{L}_I$—will provide cheaply computed noisy gradients for stochastic optimization.

We now compute the noisy gradient. Suppose we have sampled the ith data point. Notice that Equation 41 is equivalent to the full objective of Equation 40 where the ith data point is observed N times. Thus the natural gradient of Equation 41—which is a noisy natural gradient of the ELBO—can be found using Equation 32,

$$\hat{\nabla}\mathcal{L}_i = \mathbb{E}_q[\eta_g(z_i^{(N)}, x_i^{(N)}, \alpha)] - \lambda, \tag{42}$$

where $\{x_i^{(N)}, z_i^{(N)}\}$ are a data set formed by N replicates of observation x_n and hidden variables z_n.

We compute this expression in more detail. Recall the complete conditional $\eta_g(\boldsymbol{x}, \boldsymbol{z}, \alpha)$ from Equation 6. From this equation, we can compute the conditional natural parameter for the global parameter given N replicates of x_n,

$$\eta_g(z_i^{(N)}, x_i^{(N)}, \alpha) = \alpha + N(t(z_n, x_n), 1). \tag{43}$$

Substituting this into the natural gradient of the ELBO from Equation 32 gives the noisy natural gradient,

$$\hat{\nabla}_\lambda \mathcal{L}_i = \alpha + N\mathbb{E}_{\phi_i(\lambda)}[(t(z_i, x_i), 1)] - \lambda, \tag{44}$$

where $\phi_i(\lambda) = \arg\max_{\phi_i} \mathcal{L}_i$. While the full natural gradient would use the local variational parameters for the whole data set, the noisy natural gradient only considers the local parameters for one randomly sampled data point. These noisy gradients are cheaper to compute.

Finally, we use these natural gradients in a Robbins-Monro algorithm to optimize the ELBO. At each iteration we update the global variational parameter with a noisy gradient. The step-size at iteration t is ρ_t. The update is

$$\begin{aligned} \lambda^{(t)} &= \lambda^{(t-1)} + \rho_t(\mathbb{E}_{\phi_I(\lambda^{(t-1)})}[\eta_g(z_I^{(N)}, x_I^{(N)}, \alpha)] - \lambda^{(t-1)}) && (45) \\ &= (1-\rho_t)\lambda^{(t-1)} + \rho_t \mathbb{E}_{\phi_I(\lambda^{(t-1)}}[\eta_g(z_I^{(N)}, x_I^{(N)}, \alpha)]. && (46) \end{aligned}$$

This is a weighted average of the previous estimate of λ and the estimate of λ that we would obtain if the sampled data point was replicated N times.

Figure 3 presents the full algorithm. At each iteration, the algorithm has an estimate of the global variational parameter $\lambda^{(t-1)}$. It samples a single data point from the data and computes the "single data-point" optimal global variational parameter, i.e., the next value of λ if the data set contained N replicates of the sampled point. We

```
Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule $\rho_t$ appropriately.
**repeat**
    Sample a data point $x_t$ uniformly from the data set.
    Compute its local variational parameter,
        $\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_t^{(N)}, z_t^{(N)})].$
    Compute intermediate global parameters as though $x_t$ is replicated $N$ times,
        $\hat{\lambda} = \mathbb{E}_{\phi}[\eta_g(z_t^{(N)}, x_t^{(N)})].$
    Update the current estimate of the global variational parameters,
        $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}.$
**until** forever
```

Figure 3: Stochastic variational inference.

emphasize that this is much cheaper than iteratively optimizing the local variational parameters for each data point (as in the algorithm of Figure 2).

Finally, it sets the new estimate of the global variational parameter to be a weighted average of the previous estimate and the single-data-point optimal. We set the step-size at iteration t as follows,

$$\rho_t = (t + \tau)^{-\kappa}. \tag{47}$$

This satisfies the conditions in Equation 35. The *forgetting rate* $\kappa \in (0.5, 1]$ controls how quickly old information is forgotten; the *delay* $\tau \geq 0$ downweights early iterations. In Section 4 we fix the delay to be one and explore a variety of forgetting rates. Note that this is just one way to parameterize the learning rate. As long as the step size conditions in Equation 35 are satisfied, this iterative algorithm converges to a local optimum of the ELBO.

2.5 Extensions

We now describe two extensions of the basic stochastic inference algorithm in Figure 3: the use of multiple samples ("minibatches") to improve the algorithm's stability, and methods for hyperparameter estimation.

Minibatches. So far, we have considered stochastic variational inference algorithms where only one observation x_n is analyzed at a time. Many stochastic optimization algorithms benefit from "minibatches," i.e., several examples at a time (Bottou and Bousquet, 2008; Liang et al., 2009; Mairal et al., 2010). In stochastic variational inference, we can sample a set of S examples at each iteration $x_{t,1:S}$ (with or without replacement), compute the local variational parameters $\phi_s(\lambda^{(t-1)})$ for each data point, compute the intermediate global parameters $\hat{\lambda}_s$ for each data point $x_{t,s}$, and finally average the $\hat{\lambda}_s$ variables in the update

$$\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \frac{\rho_t}{S} \sum_s \hat{\lambda}_s. \tag{48}$$

The stochastic natural gradients associated with each point x_s have expected value equal to the gradient. Therefore, the average of these stochastic natural gradients has the same expectation and the algorithm remains valid.

There are two reasons to use minibatches. The first reason is to amortize any computational expenses associated with updating the global parameters across more data points; for example, if the expected sufficient statistics of β are expensive to compute, using minibatches allows us to incur that expense less frequently. The second reason is that it may help the algorithm to find better local optima. Although stochastic variational inference is guaranteed to converge to a local optimum, it may be that taking large steps on the basis of very few data points leads the algorithm to a poor local optimum. As we will see in Section 4, using more of the data per update can help the algorithm avoid such pathological local optima.

Empirical Bayes estimation of hyperparameters. In some cases one may want to both estimate the posterior of the hidden random variables β and $\boldsymbol{z}$ and obtain a point estimate of the values of the hyperparameters α. One approach to fitting α is maximum marginal likelihood, also known as empirical Bayes (Maritz and Lwin, 1989). Ideally, the hyperparameters α would be set to maximize the marginal likelihood of the data $p(\boldsymbol{x} \,|\, \alpha)$. Since we cannot compute $p(\boldsymbol{x} \,|\, \alpha)$ exactly, an alternative is to instead maximize the lower bound $\mathcal{L}$ over α. In the non-stochastic setting, α can be optimized by interleaving the coordinate ascent updates in Figure 2 with an update for α that increases the ELBO with the variational parameters λ and $\phi_{1:N}$ held fixed. This is called "variational expectation-maximization."

In the stochastic setting, we update α simultaneously with λ. We can take a step in the direction of the gradient of the replicated ELBO $\mathcal{L}_t$ with respect to α, scaled by the step-size ρ_t,

$$\alpha^{(t)} = \alpha^{(t-1)} + \rho_t \nabla_\alpha \mathcal{L}_t(\lambda^{(t-1)}, \phi, \alpha).$$

Here $\lambda^{(t-1)}$ are the global parameters from the previous iteration and ϕ are the optimized local parameters for the currently sampled data point.

3. Stochastic Variational Inference in Topic Models

We derived stochastic variational inference, a scalable inference algorithm that can be applied to a large class of hierarchical Bayesian models. In this section we show how to use the general algorithm of Section 2 to derive stochastic variational inference for two probabilistic topic models: latent Dirichlet allocation (LDA) (Blei et al., 2003) and its Bayesian nonparametric counterpart, the hierarchical Dirichlet process (HDP) topic model (Teh et al., 2006a).

Topic models are probabilistic models of document collections that use latent variables to encode recurring patterns of word use (Blei, 2012). Topic modeling algorithms are inference algorithms; they uncover a set of patterns that pervade a collection and represent each document according to how it exhibits them. (These patterns tend to be thematically coherent, which is why the models are called "topic models.") Topic models are used for both descriptive tasks, such as to build thematic navigators of large collections of documents, and for predictive tasks, such as to aid document classification. Topic models have been extended and applied in many domains.

Topic models assume that the words of each document arise from a mixture of multinomials. Across a collection, the documents share the same mixture components (called *topics*). Each document, however, is associated with its own mixture proportions (called *topic proportions*). In this way, topic models represent documents heterogenously—the documents share the same set of topics, but each exhibits them to a different degree. For example, a document about sports and health will be associated with the sports and health topics; a document about sports and business will be associated with the sports and business topics. They both share the sports topic, but each combines sports with a different topic. More generally, this kind of model is known as a *mixed-membership model.*

The central computational problem in topic modeling is posterior inference: Given a collection of documents, what are the topics that it exhibits and how does each document exhibit them? Most posterior inference algorithms for topic models are based on Markov chain Monte Carlo (MCMC) sampling or variational inference. In practical applications of topic models, scale is important. Topic models promise an unsupervised approach to organizing large collections of text (and, with simple adaptations, images, sound, and other data). Thus they are a good testbed for stochastic variational inference.

This section illustrates how to use the results from Section 2 to develop algorithms for specific models. We will derive the algorithms in several steps: (1) we specify the model assumptions; (2) we derive the complete conditional distributions of the latent variables; (3) we form the mean-field variational family; (4) we derive the corresponding stochastic inference algorithm. In Section 4, we will report our empirical study of stochastic variational inference with these models.

3.1 Notation

We follow the notation of Blei et al. (2003).

- Observations are *words*, organized into documents. The nth word in the dth document is $w_{d,n}$. Each word is an element in a fixed vocabulary of V terms.
- A *topic* β_k is a distribution over the vocabulary. Each topic is a point on the $V-1$ simplex, a positive vector of length V that sums to one. In LDA there are K topics; in the HDP topic model there are an infinite number of topics.
- Each document in the collection is associated with a vector of *topic proportions* θ_d, which is a distribution over topics. In LDA θ_d is a point on the $K-1$ simplex. In the HDP topic model, θ_d is a point on the infinite simplex. (We give details about this below in Section 3.3.)
- Each word in each document is assumed to have been drawn from a topic. The *topic assignment* $z_{d,n}$ indexes the topic from which $w_{d,n}$ is drawn.

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

3.2 Latent Dirichlet allocation

LDA is the simplest topic model. It assumes that each document exhibits K topics with different proportions. The generative process is

1. Draw topics $\beta_k \sim \mathrm{Dir}_V(\eta)$ for $k \in \{1, \ldots, K\}$.
2. For each document $d \in \{1, \ldots, D\}$:
   (a) Draw topic proportions $\theta \sim \mathrm{Dir}_K(\alpha)$.
   (b) For each word $w \in \{1, \ldots, N\}$:
      i. Draw topic assignment $z_{d,n} \sim \mathrm{Mult}(\theta_d)$.
      ii. Draw word $w_{d,n} \sim \mathrm{Mult}(\beta_{z_{dn}})$.

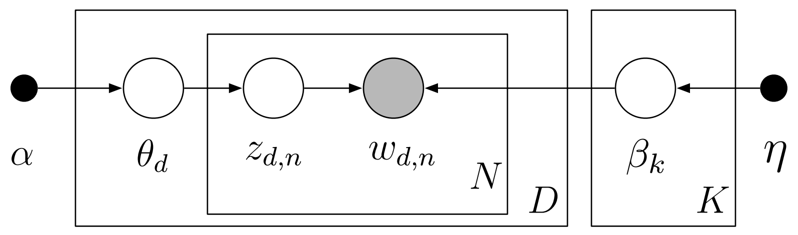
Figure 4 illustrates LDA as a graphical model.[9]

In LDA, each document exhibits the same shared topics but with different proportions. We have assumed exchangeable Dirichlet priors, which take a scalar parameter. For example, the prior on topics is $\mathrm{Dir}_V(\eta)$. It produces V-vectors on the $V-1$ simplex from a Dirichlet distribution with all parameters set to η. The prior on topic proportions is analagous, but produces K-vectors on the $K-1$ simplex. (We note that Blei et al. (2003) and Wallach et al. (2009) found improved empirical performance with non-exchangeable Dirichlet priors.)

LDA models an observed collection of documents $\boldsymbol{w} = w_{1:D}$, where each w_d is a collection of words $w_{d,1:N}$. Analyzing the documents amounts to posterior inference of $p(\beta, \boldsymbol{\theta}, \boldsymbol{z} \,|\, \boldsymbol{w})$. Conditioned on the documents, the posterior distribution captures the topics that describe them $\beta = \beta_{1:K}$, how each document exhibits those topics $\boldsymbol{\theta} = \theta_{1:D}$, and which topics each word was assigned to $\boldsymbol{z} = z_{1:D,1:N}$. We can use the posterior to explore large collections of documents.

The posterior is intractable to compute (Blei et al., 2003). Approximating the posterior in LDA is a central computational problem for topic modeling. Researchers have developed many methods, including Markov chain Monte Carlo methods (Griffiths and Steyvers, 2004), expectation propagation (Minka and Lafferty, 2002), and variational inference (Blei et al., 2003; Teh et al., 2006b; Asuncion et al., 2009). Here we use

[9] Each document d may not have the same number of words N. However, to keep the notation simple, we suppress N's dependence on the document index d.



Var	Type	Conditional	Param	Relevant Expectations
$z_{d,n}$	Multinomial	$\log \theta_{d,k} + \log \beta_{k,w_{d,n}}$	$\phi_{d,n}$	$\mathbb{E}[Z^k_{d,n}] = \phi^k_{d,n}$
θ_d	Dirichlet	$\alpha + \sum_{n=1}^N z_{d,n}$	γ_d	$\mathbb{E}[\log \theta_{d,k}] = \Psi(\gamma_{d,k}) - \sum_{j=1}^K \Psi(\gamma_{d,j})$
β_k	Dirichlet	$\eta + \sum_{d=1}^D \sum_{n=1}^N z^k_{d,n} w_{d,n}$	λ_k	$\mathbb{E}[\log \beta_{k,v}] = \Psi(\lambda_{k,v}) - \sum_{y=1}^V \Psi(\lambda_{k,y})$

Figure 4: (Top) The graphical model representation of Latent Dirichlet allocation. (Bottom) In LDA: hidden variables, complete conditionals, variational parameters, and expected sufficient statistics.

the results of Section 2 to develop stochastic inference for LDA. This scales the original variational algorithm for LDA in Blei et al. (2003) to massive collections of documents.[10]

Indicator vectors and Dirichlet distributions. Before deriving the algorithm, we discuss two mathematical details.

First, we represent categorical variables like the topic assignments $z_{d,n}$ and observed words $w_{d,n}$ with *indicator vectors*. An indicator vector is a binary vector with a single one. For example, the topic assignment $z_{d,n}$ can take on one of K values (one for each topic). Thus, it is represented as a K-vector with a one in the component corresponding to the value of the variable: if $z^k_{d,n} = 1$ then the nth word is assigned to the kth topic.

Second, we review the Dirichlet distribution. As we described above, a K-dimensional Dirichlet is a distribution on the $K-1$ simplex, i.e., positive vectors over K elements that sum to one. It is parameterized by a positive K-vector γ,

$$p(\theta \,|\, \gamma) = \frac{\Gamma\left(\sum_{i=1}^K \gamma_i\right)}{\prod_{i=1}^K \Gamma(\gamma_i)} \prod_{i=1}^K \theta^{\gamma_i - 1}, \tag{49}$$

where $\Gamma(\cdot)$ is the Gamma function, which is a real-valued generalization of the factorial function. The expection of the Dirichlet is its normalized parameter,

$$\mathbb{E}[\theta_k \,|\, \gamma] = \frac{\gamma_k}{\sum_{i=1}^K \gamma_i}. \tag{50}$$

The expectation of its log uses $\Psi(\cdot)$, which is the first derivative of the log Gamma function,

$$\mathbb{E}[\log \theta_k \,|\, \gamma] = \Psi(\gamma_k) - \Psi\left(\sum_{i=1}^K \gamma_i\right). \tag{51}$$

This can be derived by putting the Dirichlet in its exponential family form, noticing that $\log \theta_k$ are the sufficient statistics, and taking the derivative of the log-normalizer to find their expectation.

Complete conditionals and variational distributions. We specify the global and local variables of LDA to place it in the stochastic variational inference setting of Section 2. In topic modeling, the local context is a document. The local observations are its observed words $w_{d,1:N}$. The local hidden variables are the topic proportions θ_d and the topic assignments $z_{d,1:N}$. The global hidden variables are the topics $\beta_{1:K}$.

[10] The algorithm we present was originally developed in Hoffman et al. (2010a), which is a special case of the stochastic variational inference algorithm we developed in Section 2.

Recall from Section 2 that the complete conditional is the conditional distribution of a variable given all of the other variables, hidden and observed. In mean-field variational inference, the variational distributions of each variable are in the same family as the complete conditional.

We begin with the topic assignment $z_{d,n}$. The complete conditional of the topic assignment is a multinomial,

$$p(z_{d,i} = k \mid \theta_d, \beta_{1:K}, w_{d,i}) \propto \exp\{\log \theta_{d,k} + \log \beta_{k,w_{d,i}}\}. \tag{52}$$

Thus its variational distribution is a multinomial $q(z_{d,i}) = \text{Mult}(\phi_{d,i})$, where the variational parameter $\phi_{d,i}$ is a distribution over the K topics. Per the mean-field assumptions, each observed word is endowed with a different variational distribution for its topic assignment. These variational distributions can capture that some observed words are likely to be from from one topic; others are likely to be from another.

The complete conditional of the topic proportions is a posterior Dirichlet,

$$p(\theta_d \mid \beta_{1:K}, \boldsymbol{z}_d) = \text{Dir}\left(\alpha + \textstyle\sum_{n=1}^{N} z_{d,n}\right). \tag{53}$$

Since $z_{d,n}$ is an indicator vector, the parameter to this Dirichlet is the usual posterior parameter—for each topic k, its posterior Dirichlet parameter is the sum of the hyperparameter α and the number of words assigned to topic k. Note that while α is a scalar parameter to an exchangeable Dirichlet, the posterior Dirichlet is a full distribution with K different parameter values.

With this conditional, the variational distribution of the topic proportions is also Dirichlet $q(\theta_d) = \text{Dir}(\gamma_d)$, where γ_d is a K-vector Dirichlet parameter. As for the topic assignments, we emphasize that there is a different variational Dirichlet parameter for each document. These variational distributions capture that different documents reflect the topics with different proportions.

Notice that these are local hidden variables. The complete conditionals only depended on other variables in the local context (i.e., the document) and the global variables.

Finally, the complete conditional for the topic β_k is also a posterior Dirichlet,

$$p(\beta_k \mid \boldsymbol{z}, \boldsymbol{w}) = \text{Dir}\left(\eta + \textstyle\sum_{d=1}^{D} \sum_{n=1}^{N} z_{d,n}^{k} w_{d,n}\right). \tag{54}$$

For a particular term and topic k, its posterior Dirichlet parameter is the sum of the hyperparameter η and the number of times that the term was assigned to topic k. This is a global variable—its complete conditional depends on the words and topic assignments of the entire collection.

The variational distribution for each topic is a V-dimensional Dirichlet,

$$q(\beta_k) = \text{Dir}(\lambda_k). \tag{55}$$

As we will see in the next section, the traditional variational inference algorithm for LDA is inefficient with large collections of documents. The root of this inefficiency is the update for the topic parameter λ_k, which (from Equation 54) requires summing over variational parameters for every word in the collection.

Batch variational inference. With the complete conditionals in hand, we now derive the coordinate ascent variational inference algorithm, i.e., batch inference. We form each coordinate update by taking the expectation of the natural parameter of the complete conditional. This is the stepping stone to stochastic variational inference.

The variational parameters are the global per-topic Dirichlets $\lambda_{1:K}$, local per-document Dirichlets $\gamma_{1:D}$, and local per-word multinomials $\phi_{1:D,1:N}$. Coordinate ascent variational inference iterates between updating all of the local variational parameters (Equation 23) and updating the global variational parameters (Equation 21).

We update each document's local variational in a local coordinate ascent routine, iterating between updating each word's topic assignment and the per-document topic proportions,

$$\begin{aligned} \phi_{d,n} &\propto \exp\left\{\Psi(\gamma_d) + \Psi(\lambda_{\cdot,w_{d,n}}) - \Psi\left(\textstyle\sum_v \lambda_{\cdot,v}\right)\right\} \quad \text{for } n \in \{1,\ldots,N\} && (56) \\ \gamma_d &= \alpha + \textstyle\sum_{n=1}^N \phi_{d,n}. && (57) \end{aligned}$$

These updates derive from taking the expectations of the natural parameters of the complete conditionals in Equation 52 and Equation 53. (We then map back to the usual parameterizations of the multinomial and Dirichlet.) For the update on the topic assignment, we have used the Dirichlet expectations in Equation 51. For the update on the topic proportions, we have used that the expectation of an indicator is its probability, $\mathbb{E}_q\left[z_{d,n}^k\right] = \phi_{d,n}^k$.

After finding variational parameters for each document, we update the variational Dirichlet for each topic,

$$\lambda_k = \eta + \textstyle\sum_{d=1}^D \sum_{n=1}^N \phi_{d,n}^k w_{d,n}. \qquad (58)$$

This update depends on the variational parameters for every document.

Batch inference is inefficient for large collections of documents. Before updating the topics $\lambda_{1:K}$, we compute the local variational parameters for every document. This is particularly wasteful in the beginning of the algorithm when, before completing the first iteration, we must analyze every document with randomly initialized topics.

Stochastic variational inference Stochastic variational inference provides a scalable method for approximate posterior inference in LDA. The global variational parameters are the topic Dirichlets λ_k; the local variational parameters are the per-document topic proportion Dirichlets γ_d and the per-word topic assignment multinomials $\phi_{d,n}$.

```
Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule $\rho_t$ appropriately.
repeat
    Sample a document $w_d$ uniformly from the data set.
    Initialize $\gamma_{d,k} = 1$, for $k \in \{1, \ldots, K\}$.
    repeat
        For $n \in \{1, \ldots, N\}$ set
            $\phi^k_{d,n} \propto \exp\left\{\mathbb{E}[\log \theta_{d,k}] + \mathbb{E}[\log \beta_{k,w_{d,n}}]\right\}, \; k \in \{1, \ldots, K\}.$
        Set $\gamma_d = \alpha + \sum_n \phi_{d,n}$.
    until local parameters $\phi_{d,n}$ and $\gamma_d$ converge.
    For $k \in \{1, \ldots, K\}$ set intermediate topics
        $\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi^k_{d,n} w_{d,n}.$
    Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
until forever
```

Figure 5: Stochastic variational inference for LDA. The relevant expectations for each update are found in Figure 4.

We follow the general algorithm of Figure 3. Let $\lambda^{(t)}$ be the topics at iteration t. At each iteration we sample a document d from the collection. In the local phase, we compute optimal variational parameters by iterating between updating the per-document topic Dirichlet γ_d (Equation 57) and per-word topic assignment multinomial $\phi_{d,1:N}$ (Equation 56). This is the same subroutine as in batch inference, though here we only analyze one subsampled document.

In the global phase we use these fitted local variational parameters to form intermediate topics,

$$\hat{\lambda}_k = \eta + D \textstyle\sum_{n=1}^{N} \phi_{d,n}^{k} w_{d,n}. \tag{59}$$

This comes from Equation 58 for a corpus containing D replicates of document d. We then set the topics at the next iteration to be a weighted combination of the intermediate topics and current topics,

$$\lambda_k^{(t+1)} = (1 - \rho_t)\lambda_k^{(t)} + \rho_t \hat{\lambda}_k. \tag{60}$$

The algorithm for stochastic variational inference for LDA is in Figure 5.[11]

Stochastic inference versus batch inference for LDA. Figure 6 illustrates the performance of 100-topic LDA on three large collections—*Nature* contains 350K documents, *New York Times* contains 1.8M documents, and *Wikipedia* contains 3.8M documents. (Section 4 describes the complete study, including the details of the performance measure and corpora.) We compare stochastic inference for LDA on the full collection to batch inference on a 100K document subset, for which batch inference can run in reasonable speed. We see that stochastic variational inference converges faster and to a better model.

3.3 Bayesian nonparametric topic models with the HDP

Stochastic inference for LDA lets us analyze large collections of documents. One limitation of LDA, however, is that the number of topics is fixed in advance. Typically, researchers find the "best" number of topics with cross-validation (Blei et al., 2003). However, for very large data (or streaming data) this approach is not practical. We address this issue with a Bayesian nonparametric topic model.

[11] This algorithm, as well as the algorithm for the HDP, specifies that we initialize the topics λ_k randomly. There are many ways to initialize the topics. We use a Gamma distribution,

$$\lambda_{k,v} \sim \text{Gamma}(1,1) * D * 100/(KV) + \eta.$$

This relates to—but is not exactly the same as—having a corpus of size D with 100 words per document, and allocating those words randomly to different topics.

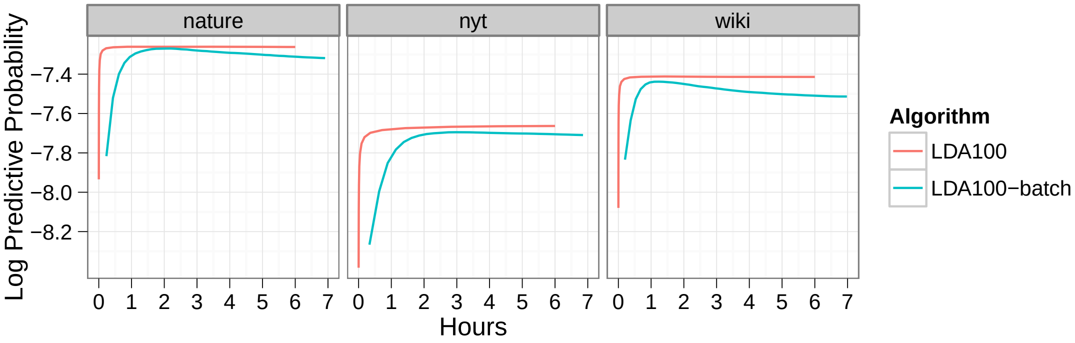


Figure 6: The per-word predictive log likelihood for a 100-topic LDA model on three large corpora. Stochastic variational inference on the full data converges faster and to a better place than batch variational inference on a reasonably sized subset. Section 4 gives the details of our empirical study.

We derive stochastic variational inference for the Bayesian nonparametric variant of LDA, the hierarchical Dirichlet process (HDP) topic model. Like LDA, the HDP topic model is a mixed-membership model of text collections. However, the HDP assumes an "infinite" number of topics. Given a collection of documents, the posterior distribution of the hidden structure determines how many topics are needed to describe them. Further, the HDP is flexible in that it allows future data to exhibit new and previously unseen topics.

More broadly, stochastic variational inference for the HDP topic model demonstrates the possibilities of stochastic inference in the context of Bayesian nonparametric statistics. Bayesian nonparametrics gives us a collection of flexible models—mixture models, mixed-membership models, factor models, and models with more complex structure—which grow and expand with data. Flexible and expanding models are particularly apt for analyzing large data sets, where searching for a specific latent structure (such as a number of topics or a tree structure of components) is prohibitive.

This section is organized as follows. We first give some background on the Dirichlet process and its definition via Sethuraman's stick breaking construction, which is a distribution on the infinite simplex. We then show how to use this construction to form the HDP topic model and how to use stochastic variational inference to approximate the posterior.[12]

[12] This algorithm first appeared in Wang et al. (2011). Here we place it in the more general context of Section 2 and relate it to stochastic inference for LDA.

The stick-breaking construction of the Dirichlet process. Bayesian nonparametric (BNP) methods use *distributions of distributions*, placing flexible priors on the shape of the data-generating density function. BNP models draw a distribution from that prior and then independently draw data from that random distribution. Data analysis proceeds by evaluating the posterior distribution of the (random) distribution from which the data were drawn. Because of the flexible prior, that posterior can potentially have mass on a wide variety of distribution shapes. For a reviews of BNP methods, see the edited volume of Hjort et al. (2010) and the tutorial of Gershman and Blei (2011).

The most common BNP prior is the *Dirichlet process* (DP). The Dirichlet process is parameterized by a *base distribution* G_0 and a scaling factor α. These are used to form a distribution of discrete distributions, i.e., those that place their mass on a countably infinite set of atoms. The locations of the atoms are independently drawn from the base distribution G_0 (which need not be discrete) and the closeness of the probabilities to G_0 is determined by the scaling factor α. When α is small, more mass is placed on fewer atoms, and the draw will likely look very different from G_0; when α is large, the mass is spread around many atoms, and the draw will more closely resemble the base distribution.

There are several representations of the Dirichlet process. For example, it is a normalized gamma process (Ferguson, 1973), and its marginalization gives the Chinese restaurant process (Pitman, 2002). We will focus on its definition via Sethuraman's stick breaking construction (Sethuraman, 1994). The stick-breaking construction explicitly defines the distribution of the probabilities that make up a random discrete distribution. It is the gateway to variational inference in Bayesian nonparametric models (Blei and Jordan, 2005).

Let $G \sim \mathrm{DP}(\alpha, G_0)$ be drawn from a Dirichlet process prior. It is a discrete distribution with mass on an infinite set of atoms. Let β_k be the atoms in this distribution and σ_k be their corresponding probabilities. We can write G as

$$G = \sum_{k=1}^{\infty} \sigma_k \delta_{\beta_k}. \tag{61}$$

The atoms are drawn independently from G_0. The stick-breaking construction specifies the distribution of their probabilities.

The stick-breaking construction uses an infinite collection of beta-distributed random variables. Recall that the beta is a distribution on $(0, 1)$ and define the following collection,

$$v_i \sim \mathrm{Beta}(1, \alpha) \quad i \in \{1, 2, 3, \ldots\}. \tag{62}$$

These variables combine to form a point on the infinite simplex. Imagine a stick of unit length. Break off the proportion of the stick given by v_1, call it σ_1, and set it aside. From the remainder (of length $1 - \sigma_1$) break off the proportion given by v_2, call it σ_2, and set it aside. The remainder of the stick is now $1 - \sigma_2 - \sigma_1 = (1 - v_1)(1 - v_2)$. Repeat this process for the infinite set of v_i. The resulting stick lengths σ_i will sum to one.

More formally, we define the function σ_i to take the collection of realized v_i variables and to return the stick length of the ith component,

$$\sigma_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1 - v_j), \tag{63}$$

and note that $\sum_{i=1}^{\infty} \sigma_i(\mathbf{v}) = 1$. We call v_i the ith *breaking proportion.*

Combining these steps, we form the distribution G according to the following process,

$$\begin{aligned} \beta_i &\sim G_0 \quad i \in \{1, 2, 3, \ldots\} \\ v_i &\sim \text{Beta}(1, \alpha) \quad i \in \{1, 2, 3, \ldots\} \\ G &= \textstyle\sum_{i=1}^{\infty} \sigma_i(\mathbf{v}) \delta_{\beta_i}. \end{aligned}$$

In the random distribution G the ith atom β_i is an independent draw from G_0 and it has probability given by the ith stick length $\sigma_i(\mathbf{v})$. Sethuraman (1994) showed that the distribution of G is $\text{DP}(\alpha, G_0)$.

The most important property of G is the “clustering” property. Even though G places mass on a countably infinite set of atoms, N draws from G will tend to exhibit only a small number of them. (How many depends on the scalar α, as we described above.) Formally, this is most easily seen via other perspectives on the DP (Ferguson, 1973; Blackwell and MacQueen, 1973; Pitman, 2002), though it can be seen intuitively with the stick-breaking construction. The intuition is that as α gets smaller more of the stick is absorbed in the first break locations because the breaking proportions are drawn from $\text{Beta}(1, \alpha)$. Thus, those atoms associated with the first breaks of ths stick will have larger mass in the distribution G, and that in turn encourages draws from the distribution to realize fewer individual atoms. In general, the first break locations tend to be larger than the later break locations. This property is called *size biasedness.*

The HDP topic model. We now construct a Bayesian nonparametric topic model that has an “infinite” number of topics. The hierarchical Dirichlet process topic model (Teh et al., 2006a) is a two-level Dirichlet process. The base distribution H of the top-level DP is a symmetric Dirichlet over the vocabulary simplex—its atoms are topics. We draw once from this DP, $G_0 \sim \text{DP}(\omega, H)$. In the second level, we use G_0 (drawn from the top-level DP) as a base measure to a document-level DP, $G_d \sim \text{DP}(\alpha, G_0)$. We draw the words of each document from topics from G_d. The consequence of this two-level construction is that all documents share the same collection of topics but exhibit them with different proportions.

We construct the HDP topic model using a stick-breaking construction at each level—one at the document level and one at the corpus level.[13] The generative process of the HDP topic model is as follows.

1. Draw an infinite number of topics, $\beta_i \sim \text{Dir}_V(\eta)$ for $k \in \{1, 2, 3, \ldots\}$.
2. Draw corpus breaking proportions, $v_k \sim \text{Beta}(1, \omega)$ for $k \in \{1, 2, 3, \ldots\}$.
3. For each document d:
    (a) Draw document-level topic indices, $c_{d,i} \sim \text{Mult}(\sigma(\mathbf{v}))$ for $i \in \{1, 2, 3, \ldots\}$.
    (b) Draw document breaking proportions, $\pi_{d,i} \sim \text{Beta}(1, \alpha)$ for $i \in \{1, 2, 3, \ldots\}$.
    (c) For each word n:
        i. Draw topic assignment $z_{d,n} \sim \text{Mult}(\sigma(\boldsymbol{\pi}_d))$.
        ii. Draw word $w_n \sim \text{Mult}(\beta_{c_{d,z_{d,n}}})$.

Figure 7 illustrates this process as a graphical model.

In this construction, topics β_k are drawn as in LDA (Step 1). Corpus-level breaking proportions $\boldsymbol{v}$ (Step 2) define a probability distribution on these topics, which indicates their relative prevalence in the corpus. At the document level, breaking proportions $\pi_{d,i}$ create a set of probabilities (Step 3b) and topic indices $c_{d,i}$, drawn from $\sigma(\boldsymbol{v})$, attach each document-level stick length to a topic (Step 3a). This creates a document-level distribution over topics, and words are then drawn as for LDA (Step 3c).

The posterior distribution of the HDP topic model gives a mixed-membership decomposition of a corpus where the number of topics is unknown in advance and unbounded. However, it is not possible to compute the posterior. Approximate posterior inference for BNP models in general is an active field of research (Escobar and West, 1995; Neal, 2000; Blei and Jordan, 2005; Teh et al., 2007).

The main advantage of our construction is that it meets the conditions of Section 2. All the complete conditionals are in exponential families in closed form, and it neatly separates global variables from local variables. The global variables are topics and corpus-level breaking proportions; the local variables are document-level topic indices and breaking proportions. Following the same procedure as for LDA, we now derive stochastic variational inference for the HDP topic model.

[13] See the original HDP paper of Teh et al. (2006a) for other constructions of the HDP—the random measure construction, the construction by the Chinese restaurant franchise, and an alternative stick-breaking construction. This construction was alluded to by Fox et al. (2008). It was formally proposed and first used for the HDP by Wang et al. (2011).

Complete conditionals and variational distributions. We form the complete conditional distributions of all variables in the HDP topic model. We begin with the latent indicator variables,

$$\begin{aligned} p(z_{d,n} = k|\boldsymbol{\pi}_d, \beta_{1:K}, w_{d,n}, \boldsymbol{c}_d) &\propto \exp\{\log \sigma_k(\boldsymbol{\pi}_d) + \textstyle\sum_{i=1}^{\infty} c_{d,k}^i \log \beta_{i,w_{d,n}}\}, & (64) \\ p(c_{d,i} = k|\mathbf{v}, \beta_{1:K}, \boldsymbol{w}_d, \boldsymbol{z}_d) &\propto \exp\{\log \sigma_k(\mathbf{v}) + \textstyle\sum_{n=1}^{N} z_{d,n}^k \log \beta_{k,w_{d,n}}\}. & (65) \end{aligned}$$

Note the interaction between the two levels of latent indicators. In LDA the ith component of the topic proportions points to the ith topic. Here we must account for the topic index $c_{d,i}$, which is a random variable that points to one of the topics.

This interaction between indicators is also seen in the conditionals for the topics,

$$p(\beta_k|\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{w}) = \mathrm{Dir}\left(\eta + \textstyle\sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{d,i}^k \sum_{n=1}^{N} z_{d,n}^i w_{d,n}\right). \qquad (66)$$

The innermost sum collects the sufficient statistics for words in the dth document that are allocated to the ith local component index. However, these statistics are only kept when the ith topic index $c_{d,i}$ points to the kth global topic.

The full conditionals for the breaking proportions follow those of a standard stick-breaking construction (Blei and Jordan, 2005),

$$\begin{aligned} p(v_k|\boldsymbol{c}) &= \mathrm{Beta}\left(1 + \textstyle\sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{d,i}^k \,,\, \omega + \sum_{d=1}^{D} \sum_{i=1}^{\infty} \sum_{j>k} c_{d,i}^j\right), & (67) \\ p(\pi_{d,i}|z_d) &= \mathrm{Beta}\left(1 + \textstyle\sum_{n=1}^{N} z_{d,n}^i \,,\, \alpha + \sum_{n=1}^{N} \sum_{j>i} z_{d,n}^j\right). & (68) \end{aligned}$$

The complete conditionals for all the latent variables are all in the same family as their corresponding distributions in the generative process. Accordingly, we will define the variational distributions to be in the same family. However, the main difference between BNP models and parametric models is that BNP models contain an infinite number of hidden variables. These cannot be completely represented in the variational distribution as this would require optimizing an infinite number of variational parameters. We solve this problem by truncating the variational distribution (Blei and Jordan, 2005). At the corpus level, we truncate at K, fitting posteriors to K breaking points, K topics, and allowing the topic pointer variables to take on one of K values. At the document level we truncate at T, fitting T breaking proportions, T topic pointers, and letting the topic assignment variable take on one of T values. Thus the variational family is,

$$q(\beta, v, \boldsymbol{z}, \boldsymbol{\pi}) = \left(\prod_{k=1}^{K} q(\beta_k \mid \lambda_k) q(v_k \mid a_k)\right) \left(\prod_{d=1}^{D} \prod_{i=1}^{T} q(c_{d,i} \mid \zeta_{d,i}) q(\pi_{d,i} \mid \gamma_{d,i}) \prod_{n=1}^{N} q(z_{d,n} \mid \phi_{d,n})\right)$$

We emphasize that this is not a finite model. With truncation levels set high enough, the variational posterior will use as many topics as the posterior needs, but will not necessarily use all K topics to explain the observations. (If K is set too small then the truncated variational distribution will use all of the topics, but this problem can be easily diagnosed and corrected.) Further, a particular advantage of this two-level stick-breaking distribution is that the document truncation T can be much smaller than K. Though there may be hundreds of topics in a large corpus, we expect each document will only exhibit a small subset of them.

Stochastic variational inference for HDP topic models. From the complete conditionals, batch variational inference proceeds by updating each variational parameter using the expectation of its conditional distribution's natural parameter. In stochastic inference, we sample a data point, update its local parameters as for batch inference, and then update the global variables.

To update the global topic parameters, we again form intermediate topics with the sampled document's optimized local parameters,

$$\hat{\lambda}_k = \eta + D \textstyle\sum_{i=1}^{T} \mathbb{E}_q[c^k_{d,i}] \sum_{n=1}^{N} \mathbb{E}_q[z^i_{d,n}] w_{d,n}. \tag{69}$$

We then update the global variational parameters by taking a step in the direction of the stochastic natural gradient

$$\lambda^{(t+1)} = (1 - \rho_t)\lambda^{(t)} + \rho_t \hat{\lambda}_k. \tag{70}$$

This parallels the update for LDA.

The other global variables in the HDP are the corpus-level breaking proportions v_k, each of which is associated with a set of beta parameters $a_k = \langle a_k^{(1)}, a_k^{(2)} \rangle$ for its variational distribution. Using the same randomly selected document and optimized variational parameters as above, first construct the two-dimensional vector

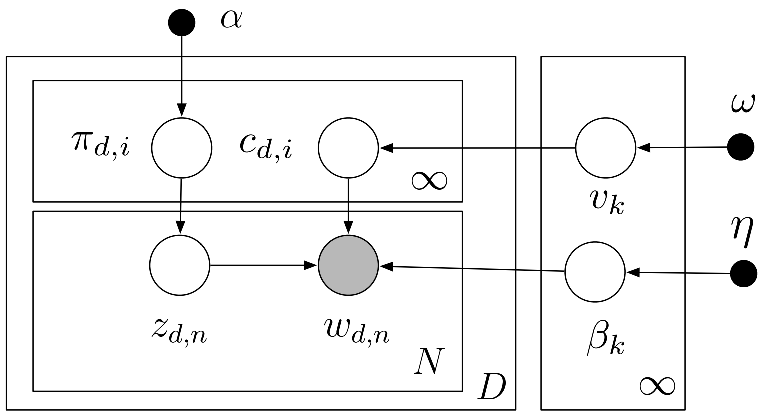
$$\hat{a}_k = \left\langle 1 + D \textstyle\sum_{i=1}^{T} \mathbb{E}_q[c^k_{d,i}]\, , \, \omega + D \sum_{i=1}^{T} \sum_{j=k+1}^{K} \mathbb{E}_q[c^j_{d,i}] \right\rangle. \tag{71}$$

Then, update the parameters

$$a_k^{(t+1)} = (1 - \rho_t) a_k^{(t)} + \rho_t \hat{a}_k. \tag{72}$$

Note that we use the truncations K and T. Figure 7 summarizes the complete conditionals, variational parameters, and relevant expectations for the full algorithm. Figure 8 gives the stochastic variational inference algorithm for the HDP topic model.

Stochastic inference versus batch inference for the HDP. Figure 9 illustrates the performance of the HDP topic model on the same three large collections as in Figure 6. As for LDA, stochastic variational inference for the HDP converges faster and to a better model.



Var	Type	Conditional	Param	Relevant expectation
$z_{d,n}$	Multinomial	$\log \sigma_i(\boldsymbol{\pi}_d) + \sum_{k=1}^{\infty} c_{d,i}^k \log \beta_{k,w_{d,n}}$	$\phi_{d,n}$	$\mathbb{E}[Z_{d,n}^i] = \phi_{d,n}^i$
$\pi_{d,i}$	Beta	$\langle 1 + \sum_{n=1}^N z_{d,n}^i,\ \alpha + \sum_{n=1}^N \sum_{j=i+1}^{\infty} z_{d,n}^j \rangle$	$\langle \gamma_{d,i}^{(1)}, \gamma_{d,i}^{(2)} \rangle$	(Expectations are similar to those for v_k.)
$c_{d,i}$	Multinomial	$\log \sigma_k(\mathbf{V}) + \sum_{n=1}^N z_{d,n}^i \log \beta_{k,w_{d,n}}$	$\zeta_{d,i}$	$\mathbb{E}[c_{d,i}^k] = \zeta_{d,i}^k$
v_k	Beta	$\langle 1 + \sum_d \sum_i c_{d,i}^k,\ \omega + \sum_d \sum_i \sum_{\ell=k+1}^{\infty} c_{d,i}^{\ell} \rangle$	$\langle a_k^{(1)}, a_k^{(2)} \rangle$	$\mathbb{E}[\log V_k] = \Psi(a_k) - \Psi(a_k + b_k)$ $\mathbb{E}[\log(1 - V_k)] = \Psi(b_k) - \Psi(a_k + b_k)$ $\mathbb{E}[\log \sigma_k(\mathbf{V})] = \mathbb{E}[\log V_k] + \sum_{\ell=1}^{k-1} \mathbb{E}[\log(1 - V_\ell)]$
β_k	Dirichlet	$\eta + \sum_{d=1}^D \sum_{i=1}^{\infty} c_{d,i}^k \sum_{n=1}^N z_{d,n}^i w_{d,n}$	λ_k	$\mathbb{E}[\log \beta_{k,v}] = \Psi(\lambda_{k,v}) - \Psi\left(\sum_{v'} \lambda_{k,v'}\right)$

Figure 7: A graphical model for the HDP topic model, and a summary of its variational inference algorithm.

```
Initialize $\lambda^{(0)}$ randomly. Set $a^{(0)} = 1$ and $b^{(0)} = \omega$.
Set the step-size schedule $\rho_t$ appropriately.
repeat
    Sample a document $w_d$ uniformly from the data set.
    For $i \in \{1, \ldots, T\}$ initialize
        $\zeta_{d,i}^{k} \propto \exp\{\sum_{n=1}^{N} \mathbb{E}[\log \beta_{k,w_{d,n}}]\}, \ k \in \{1, \ldots, K\}.$
    For $n \in \{1, \ldots, N\}$ initialize
        $\phi_{d,n}^{i} \propto \exp\left\{\sum_{k=1}^{K} \zeta_{d,i}^{k} \mathbb{E}[\log \beta_{k,w_{d,n}}]\right\}, \ i \in \{1, \ldots, T\}.$
    repeat
        For $i \in \{1, \ldots, T\}$ set
            $\gamma_{d,i}^{(1)} = 1 + \sum_{n=1}^{N} \phi_{d,n}^{i},$
            $\gamma_{d,i}^{(2)} = \alpha + \sum_{n=1}^{N} \sum_{j=i+1}^{T} \phi_{d,n}^{j}$
            $\zeta_{d,i}^{k} \propto \exp\left\{\mathbb{E}[\log \sigma_k(\mathbf{V})] + \sum_{n=1}^{N} \phi_{d,n}^{i} \mathbb{E}[\log \beta_{k,w_{d,n}}]\right\}, \ k \in \{1, \ldots, K\}.$
        For $n \in \{1, \ldots, N\}$ set
            $\phi_{d,n}^{i} \propto \exp\left\{\mathbb{E}[\log \sigma_i(\boldsymbol{\pi}_d)] + \sum_{k=1}^{K} \zeta_{d,i}^{k} \mathbb{E}[\log \beta_{k,w_{d,n}}]\right\}, \ i \in \{1, \ldots, T\}.$
    until local parameters converge.
    For $k \in \{1, \ldots, K\}$ set intermediate topics
        $\hat{\lambda}_{k,v} = \eta + D \sum_{i=1}^{T} \zeta_{d,i}^{k} \sum_{n=1}^{N} \phi_{d,n}^{i} w_{d,n}$
        $\hat{a}_k = 1 + D \sum_{i=1}^{T} \zeta_{d,i}^{k}$
        $\hat{b}_k = \omega + D \sum_{i=1}^{T} \sum_{\ell=k+1}^{K} \zeta_{d,i}^{\ell}.$
    Set
        $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$
        $a^{(t)} = (1 - \rho_t)a^{(t-1)} + \rho_t \hat{a}$
        $b^{(t)} = (1 - \rho_t)b^{(t-1)} + \rho_t \hat{b}.$
until forever
```

Figure 8: Stochastic variational inference for the HDP topic model. The corpus-level truncation is K; the document-level truncation as T. Relevant expectations are found in Figure 7.

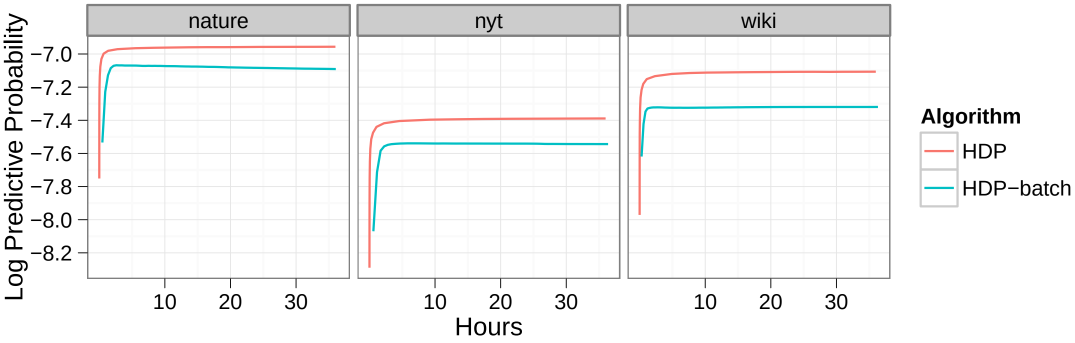


Figure 9: The per-word predictive log likelihood for an HDP model on three large corpora. As for LDA, stochastic variational inference on the full data converges faster and to a better place than batch variational inference on a reasonably sized subset. Section 4 gives the details of our empirical study.

4. Empirical Study

We studied stochastic variational inference for latent Dirichlet allocation (LDA) and the hierarchical Dirichlet process (HDP) topic model. With these algorithms, we can apply and compare these models with very large collections of documents. We also investigated how the forgetting rate κ and mini-batch size S influenced the algorithms. Finally, we compared stochastic variational inference to the traditional batch variational inference algorithm.[14]

Data. We evaluated our algorithms on three collections of documents. For each collection, we computed a vocabulary by removing stop words, rare words, and very frequent words. The data are as follows.

- *Nature*: This collection contains 350,000 documents from the journal *Nature* (spanning the years 1869–2008). After processing, it contains 58M observed words from a vocabulary of 4,200 terms.

- *New York Times*: This collection contains 1.8M documents from the *New York Times* (spanning the years 1987–2007). After processing, this data contains 461M observed words from a vocabulary of 8,000 terms.

[14] We implemented all algorithms in Python using Numpy, making the implementations as similar as possible. All of our code is available on the web.

- *Wikipedia*: This collections contains 3.8M documents from Wikipedia. After processing, it contains 482M observed words from a vocabulary of 7,700 terms.

For each collection, we set aside a test set of $10,000$ documents for evaluating model fitness; these test sets were not given to the algorithms for training.

Evaluating model fitness. We evaluate how well a model fits the data with the *predictive distribution.* We are given a corpus and estimate its topics. We then are given part of a test document, which we combine with the topics to form a predictive distribution of words. Under this predictive distribution, a better model will assign higher probability to the held-out words.

In more detail, we divide each test document w into a set of observed words w_{obs} and held-out words w_{ho}. We approximate the posterior distribution of topics β from the training data $\mathcal{D}$, and then to use that approximate posterior to estimate the predictive distribution of words $p(w \,|\, w_{\mathrm{obs}}, \mathcal{D})$. Finally, we evaluate the log probability of the words in w_{ho} under this distribution.

This metric was used in Teh et al. (2007) and Asuncion et al. (2009). Unlike previous methods, like held-out perplexity in Blei et al. (2003), evaluating the predictive distribution avoids comparing bounds or forming approximations of the evaluation metric. It rewards a good predictive distribution, however it is computed.

Operationally, we use the training data to compute variational Dirichlet parameters for the topics. We then use these parameters with the observed test wods w_{obs} to compute the variational distribution of the topic proportions. Taking the inner product of the expected topics and the expected topic proportions gives the predictive distribution.

To see this is a valid approximation, note the following for a K-topic LDA model,

$$\begin{aligned} p(w \,|\, \mathcal{D}, w_{\mathrm{obs}}) &= \int_{\boldsymbol{\beta}} \int_{\theta} \left(\textstyle\sum_{k=1}^{K} \theta_k \beta_{k,w} \right) p(\theta \,|\, w_{\mathrm{obs}}, \beta) p(\beta \,|\, \mathcal{D}) && (73) \\ &\approx \int_{\beta} \int_{\theta} \left(\textstyle\sum_{k=1}^{K} \theta_k \beta_{k,w} \right) q(\theta) q(\beta) && (74) \\ &= \mathbb{E}_q[\theta \,|\, w_{\mathrm{obs}}]^{\top} \mathbb{E}_q[\beta_{\cdot,w} \,|\, \mathcal{D}]. && (75) \end{aligned}$$

We have conditioned the expectations to make clear with which data the q distribution is fit. The metric independentally evaluates each held out word under this distribution. In the HDP, the reasoning is identical. The differences are that the topic proportions are computed via the two-level variational stick-breaking distribution and K is the truncation level of the approximate posterior.

Learning parameters. Stochastic variational inference introduces several parameters in setting the learning rate schedule (see Equation 47). The forgetting rate $\kappa \in (0.5, 1]$ controls how quickly old information is forgotten; the delay $\tau \geq 0$ downweights early iterations; and the mini-batch size S is how many documents are subsampled and analyzed in each iteration. Although stochastic variational inference algorithm converges to a stationary point for any valid κ, τ, and S, the quality of this stationary point and the speed of convergence may depend on how these parameters are set.

We set $\tau = 1$ and explored the following forgetting rates and minibatch sizes:[15]

- Forgetting rate $\kappa \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- Minibatch size $S \in \{10, 50, 100, 500, 1000\}$

We periodically paused each run to compute predictive likelihoods from the test data.

Results on LDA and HDP topic models. We studied LDA and the HDP. In LDA, we varied the number of topics K to be 25, 50, 100, 200 and 300; we set the Dirichlet hyperparameters $\alpha = 1/K$. In the HDP, we set both concentration parameters γ and α equal to 1; we set the top-level truncation $K = 300$ and the second level truncation $T = 20$. (Here $T \ll K$ because we do not expect documents to exhibit very many unique topics.) In both models, we set the topic Dirichlet parameter $\eta = 0.01$. Figure 10 shows example topics from the HDP (on *New York Times*).

Figure 11 gives the average predictive log likelihood for both models. We report the value for a forgetting rate $\kappa = 0.9$ and a batch size of 500. Stochastic inference lets us perform a large scale comparison of these models. The HDP gives consistently better performance. For larger numbers of topics, LDA overfits the data. As the modeling assumptions promise, the HDP stays robust to overfitting.[16]

We now turn to the sensitivity of stochastic inference to its learning parameters. First, we consider the HDP (the algorithm presented in Figure 8). We fixed the batch size to 500 and explored the forgetting rate.[17] Figure 12 shows the results on all three corpora. All three fits were sensitive to the forgetting rate; we see that a higher value (i.e., close to one) leads to convergence to a better optimum.

[15] We also explored various values of the delay τ, but found that the algorithms were not sensitive. To make this presentaton simpler, we fixed $\tau = 1$ in our report of the empirical study.

[16] Though not illustrated, we note that using the traditional measure of fit, held-out perplexity, does *not* reveal this overfitting (though the HDP still outperforms LDA with that metric as well). We feel that the predictive distribution is a better metric for model fitness.

[17] We fit distributions using the entire grid of parameters described above. However, to simplify presenting results we will hold one of the parameters fixed and vary the other.

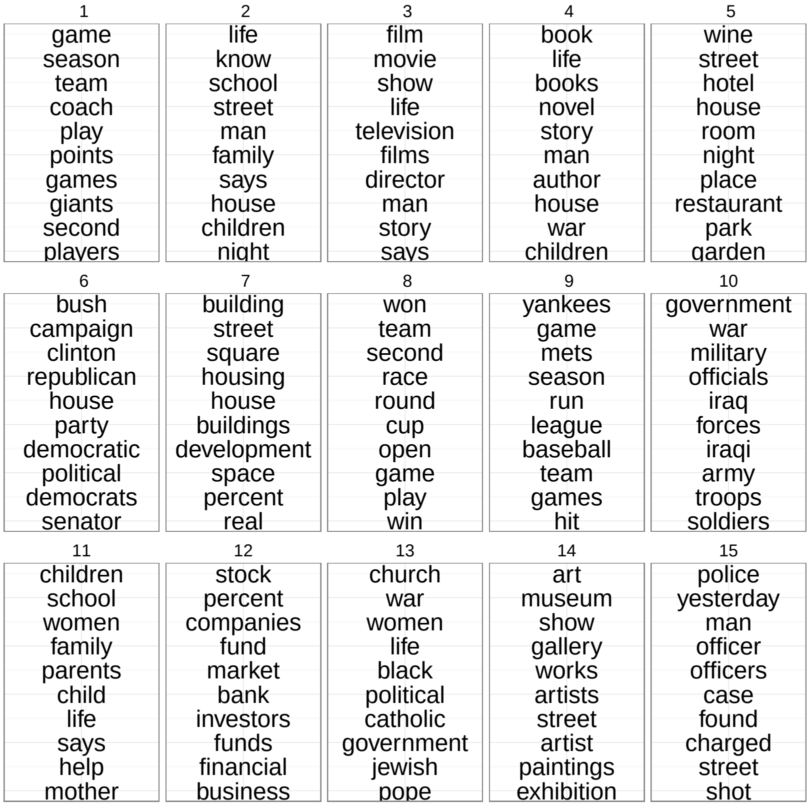


Figure 10: The 15 most frequent topics from the HDP posterior on the *New York Times*. Each topic plot illustrates the topic's most frequent words.

	Nature	*New York Times*	*Wikipedia*
LDA 25	-7.24	-7.73	-7.44
LDA 50	-7.23	-7.68	-7.43
LDA 100	-7.26	-7.66	-7.41
LDA 200	-7.50	-7.78	-7.64
LDA 300	-7.86	-7.98	-7.74
HDP	**-6.97**	**-7.38**	**-7.07**

Figure 11: Stochastic inference lets us compare performance on several large data sets. We fixed the forgetting rate $\kappa = 0.9$ and the batch size to 500 documents. We find that LDA is sensitive to the number of topics; the HDP gives consistently better predictive performance. Traditional variational inference (on subsets of each corpus) did not perform as well as stochastic inference.

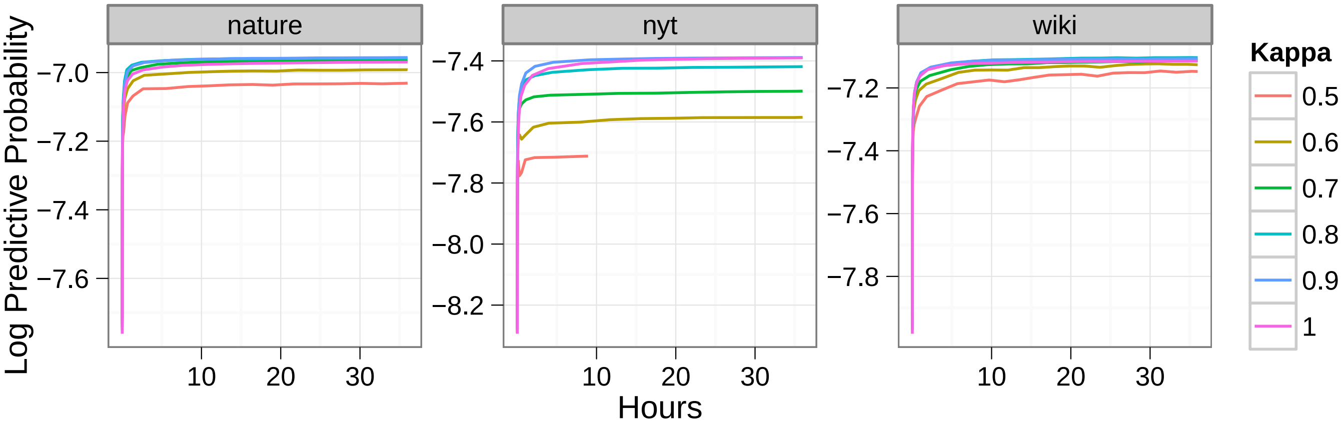


Figure 12: HDP inference: Holding the batch size fixed at 500, we varied the forgetting rate κ. Slower forgetting rates are preferred.

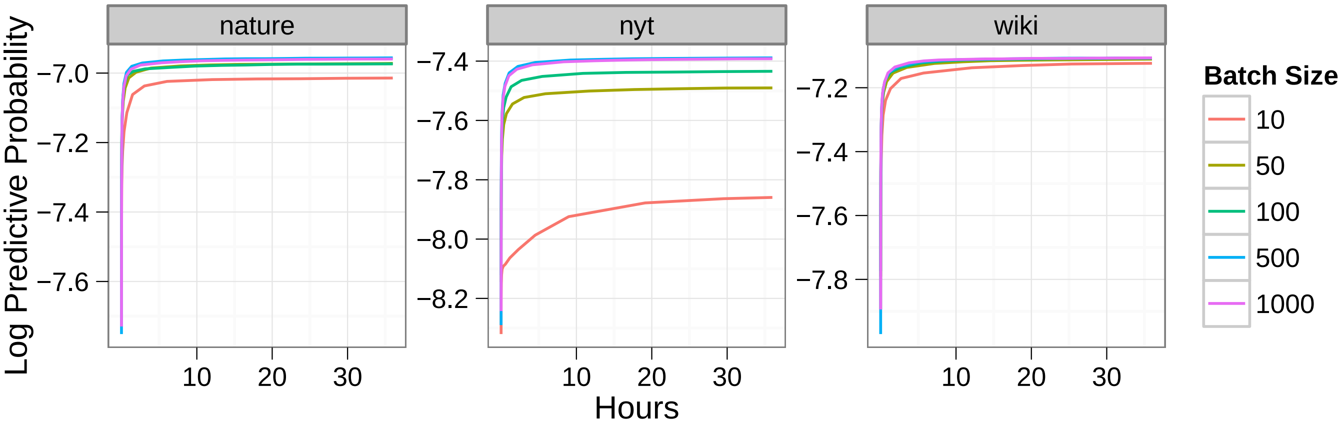


Figure 13: HDP inference: Holding the forgetting rate κ fixed at 0.9, we varied the batch size. Batch sizes may be set too small (e.g., ten documents) but the difference in performance is small once set high enough.

Fixing the forgetting rate to 0.9, we explored various mini-batch sizes. Figure 13 shows the results on all three corpora. Batch sizes that are too small (e.g., ten documents) can affect performance; larger batch sizes are preferred. That said, there was not a big difference between batch sizes of 500 and 1,000. The *New York Times* corpus was most sensitive to batch size; the *Wikipedia* corpus was least sensitive.

Figure 15 illustrate LDA's sensitivity to the forgetting rate and batch size, respectively. Again, we find that large learning rates and batch sizes perform well.

5. Discussion

We developed stochastic variational inference, a scalable variational inference algorithm that lets us analyze very large data sets with complex probabilistic models. The main idea is to use stochastic optimization to optimize the variational objective, following noisy esimates of the natural gradient where the noise arises by repeatedly subsampling the data. We illustrated this approach with two probabilistic topic models, latent Dirichlet allocation and the hierarchical Dirichlet process topic model. With stochastic variational inference, we can easily apply topic modeling to collections of millions documents. More importantly, this algorithm generalizes to many settings.

Stochastic variational inference opens the door to several promising research directions.

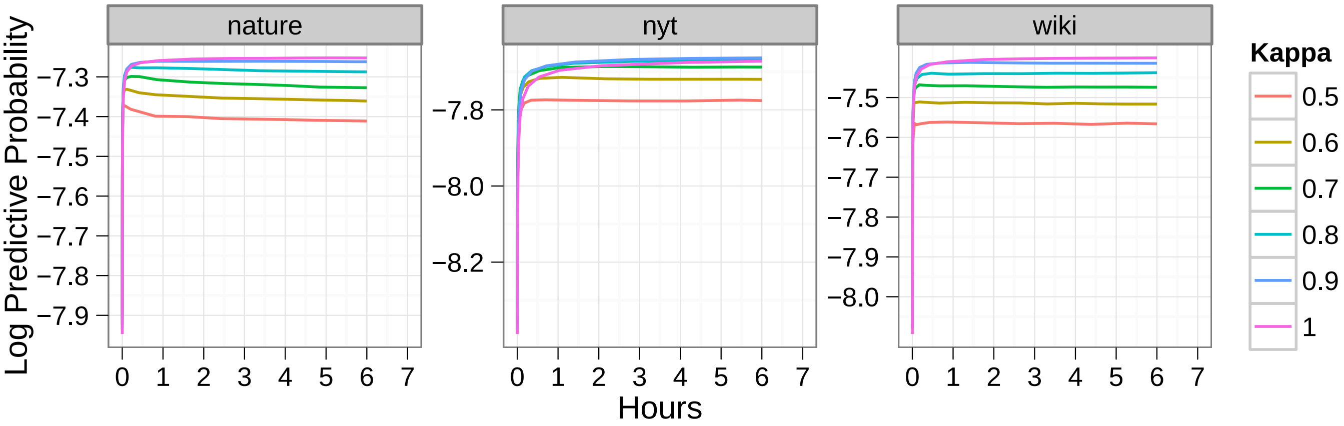


Figure 14: 100-topic LDA inference: Holding the batch size fixed at 500, we varied the forgetting rate κ. Slower forgetting rates are preferred.

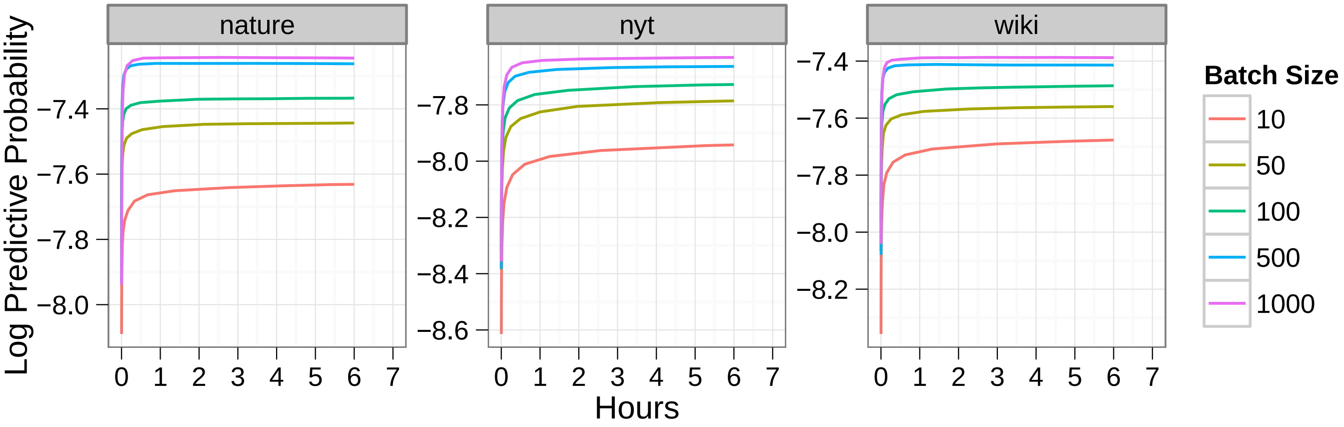


Figure 15: 100-topic LDA inference: Holding the learning rate κ fixed at 0.9, we varied the batch size. Bigger batch sizes are preferred.

We developed our algorithm with conjugate exponential family models. This class of models is expressive, but nonconugate models—models where a richer prior is used at the expense of mathematical convenience—have expanded the suite of probabilistic tools at our disposal. For example, nonconjugate models can capture correlations between topics (Blei and Lafferty, 2007) or topics changing over time (Blei and Lafferty, 2006; Wang et al., 2008), and the algorithm presented here cannot be used in these settings. (In other work, Paisley et al. (2012b) developed a stochastic variational inference algorithm for a specific nonconjugate Bayesian nonparametric model.) Recent research has developed general methods for non-conjugate models (Knowles and Minka, 2011; Gershman et al., 2012; Paisley et al., 2012a). Can these be scaled up with stochastic optimization?

We developed our algorithm with mean-field variational inference and closed form coordinate updates. Another promising direction is to use stochastic optimization to scale up recent advances in variational inference, moving beyond closed form updates and fully factorized approximate posteriors. As one example, collapsed variational inference (Teh et al., 2006b, 2007) marginalizes out some of the hidden variables, trading closed-form updates for a more focused posterior. In related work, we developed a stochastic variational algorithm for fitting topic models—combining MCMC at a local level and variational inference at a global level (Mimno et al., 2012)—but could this be generalized to the full setting considered here? As another example, structured variational distributions let us handle more complex posteriors, such as those arising in time-series models (Ghahramani and Jordan, 1997; Blei and Lafferty, 2006). Being able to relax the mean-field assumption would expand the repertoire of models that we can use on large analysis problems.

Finally, our algorithm lets us potentially connect innovations in stochastic optimization to better methods for approximate posterior inference. For example, Wahabzada and Kersting (2011) sample from data non-uniformly to better focus on more informative data points. We might also consider data whose distribution changes over time, such as when we want to model an infinite stream of data but to "forget" data from the far past in a current estimate of the model. Last, we can analyze our estimates of the gradient. How is it affected by sparse data? Are there ways to reduce its variance, but maintain its unbiasedness?

References

A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. Smola. Scalable inference in latent variable models. In *WSDM*, pages 123–132, New York, NY, USA, 2012. ACM.

S. Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 1982.

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

C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

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

H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, 2000.

M Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

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

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

C. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 777–784. MIT Press, Cambridge, MA, 2003.

D. Blackwell and J. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

D. Blei. Introduction to probabilistic topic modeling. *Communications of the ACM*, 55(4):77–84, 2012.

D. Blei and M. Jordan. Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144, 2005.

D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

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

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

L. Bottou. Stochastic learning. In *Advanced lectures on machine learning*, pages 146–168. Springer, 2003.

L. Bottou and O. Bousquet. Learning using large datasets. In *Mining Massive DataSets for Security*, NATO ASI Workshop Series. IOS Press, Amsterdam, 2008.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*, pages 351–368. MIT Press, 2011.

L. Bottou and Y. LeCun. Large scale online learning. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. In *Neural Information Processing Systems*, 2002.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32:41–62, 1998.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. An HDP-HMM for systems with state persistence. In *International Conference on Machine Learning*, 2008.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *IEEE Transactions on Signal Processing*, 59 (4):1569–1585, 2011a.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011b.

A. Gelfand and A. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

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

S. Gershman and D. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 2011.

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

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

Mark Girolami and Simon Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8), 2006.

T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Science*, 2004.

N. Hjort, C. Holmes, P. Muller, and S. Walker, editors. *Bayesian nonparametrics*. Cambridge University Press, 2010.

M. Hoffman, D. Blei, and F. Bach. On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems*, 2010a.

M. Hoffman, D. Blei, and P. Cook. Bayesian nonparametric matrix factorization for recorded music. In *International Conference on Machine Learning*, 2010b.

A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In *Neural Information Processing*, pages 305–314. Springer, 2008.

M. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

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

R. Kalman. A new approach to linear filtering and prediction problems a new approach to linear filtering and prediction problems,". *Transaction of the AMSE: Journal of Basic Engineering*, 82:35–45, 1960.

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

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009. ISBN 0262013193.

S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Y. LeCun, L. Bottou, G. Orr, and K. Müller. Efficient backprop. *Neural networks: Tricks of the trade, LNCS 1524*, pages 9–50, 1998.

P. Liang, M. Jordan, and D. Klein. Learning semantic correspondences with less supervision. In *Association of Computational Linguisitics*, 2009.

J. Mairal, J. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

J. Maritz and T. Lwin. *Empirical Bayes methods*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1989.

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

D. Mimno, M. Hoffman, and D. Blei. Sparse stochastic inference for latent Dirichlet allocation. In *International Conference on Machine Learning*, 2012.

T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Uncertainty in Artificial Intelligence (UAI)*, 2002.

R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. MIT Press, 1999.

D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *JMLR*, 10:1801–1828, 2009.

J. Paisley and L. Carin. Hidden Markov models with stick-breaking priors. *IEEE Transactions on Signal Processing*, 57(10):3905–3917, 2009a.

J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *International Conference on Machine Learning*, 2009b.

J. Paisley, D. Blei, and M. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012a.

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

J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988. ISBN 1558604790.

J. Pitman. *Combinatorial Stochastic Processes*. Lecture Notes for St. Flour Summer School. Springer-Verlag, New York, NY, 2002.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, NY, 2004.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008.

M Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.

M. Sato and S. Ishii. On-line EM algorithm for the normalized gaussian network. *Neural Computation*, 12(2):407–432, 2000.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th international conference on Machine learning*, pages 807–814. ACM, 2007.

A. Smola and S. Narayanamurthy. An architecture for parallel topic models. In *VLDB*, 2010.

J. Spall. *Introduction to stochastic search and optimization: Estimation, simulation, and control.* John Wiley and Sons, 2003.

Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006a.

Y. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Neural Information Processing Systems*, 2006b.

Y. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *Neural Information Processing Systems*, 2007.

M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

M. Wahabzada and K. Kersting. Larger residuals, less work: Active document scheduling for latent dirichlet allocation. In *Proceedings of ECML PKDD*, 2011.

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

H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

C. Wang. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 2006.

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

C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2011.

B. Widrow and M. Hoff. Adaptive switching circuits. *IRE WESCON Conv. Record, Part 4*, pages 96–104, 1960.

E. Xing, M. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, 2003.

K. Zhai, J. Boyd-Graber, N. Asadi, and M. Alkhouja. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In *ACM International Conference on the World Wide Web*, 2012.