Structured Stochastic Variational Inference

Matthew D. Hoffman
Adobe Research

David M. Blei
Departments of Statistics and Computer Science
Columbia University

Abstract

Stochastic variational inference makes it possible to approximate posterior distributions induced by large datasets quickly using stochastic optimization. The algorithm relies on the use of fully factorized variational distributions. However, this "mean-field" independence approximation limits the fidelity of the posterior approximation, and introduces local optima. We show how to relax the mean-field approximation to allow arbitrary dependencies between global parameters and local hidden variables, producing better parameter estimates by reducing bias, sensitivity to local optima, and sensitivity to hyperparameters.

1 Introduction

Hierarchical Bayesian modeling is a powerful framework for learning from rich data sources. Unfortunately, the intractability of the posteriors of rich models drives practitioners to resort to approximate inference algorithms such as mean-field variational inference or Markov chain Monte Carlo (MCMC). These classes of methods have complementary strengths and weaknesses—MCMC methods have strong asymptotic guarantees of unbiasedness but are often slow, while mean-field variational inference is often faster but tends to misrepresent important qualities of the posterior of interest and is more vulnerable to local optima. Incremental versions of both methods based on stochastic optimization have been developed that are applicable to large datasets (Welling and Teh, 2011; Hoffman et al., 2013).

In this paper we focus on variational inference. In particular, we are interested in estimating the parameters of high-dimensional Bayesian models with highly multimodal posteriors, such as mixture models, topic models, and factor models. We focus less on uncertainty estimates, which are difficult to trust and interpret in this setting.

Mean-field variational inference approximates the intractable posterior distribution implied by the model and data with a factorized approximating distribution in which all parameters are independent. This mean-field distribution is then tuned to minimize its Kullback-Leibler divergence to the posterior, which is equivalent to maximizing a lower bound on the marginal probability of the data. The restriction to factorized distributions makes the problem tractable, but reduces the fidelity of the approximation and introduces local optima (Wainwright and Jordan, 2008).

A partial remedy is to weaken the mean-field factorization by restoring some dependencies, resulting in "structured" mean-field approximations (Saul and Jordan, 1996). The applicability, speed, effectiveness, and ease-of-implementation of standard structured mean-field algorithms is limited because the lower bound implied by the structured distribution must be available in closed form.

More recent work manages these intractable variational lower bounds using stochastic optimization, which allows one to optimize functions that can only be computed approximately. For example, Ji et al. (2010) use mean-field approximations to the posteriors of "collapsed" models where some parameters have been analytically marginalized out, Salimans and Knowles (2013) apply a structured approximation to the posterior of a stochastic volatility model, and Mimno et al. (2012) use a structured approximation to the posterior of a collapsed model.

In parallel, Hoffman et al. (2013) proposed the stochastic variational inference (SVI) framework, which uses stochastic optimization to apply mean-field variational inference to massive datasets. SVI splits the unobserved variables in a hierarchical model into global parameters β (which are shared across all observations) and groups of local hidden variables $z_1, \ldots, z_N$ (each of which is specific to a small group of observations y_n). The goal is to minimize the Kullback-Leibler (KL) divergence between a tractable approximating distribution $q(z, \beta)$ over the local and global parameters and the true posterior $p(z, \beta|y)$ over those parameters. SVI approximates posteriors much more quickly than traditional batch variational inference algorithms.

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

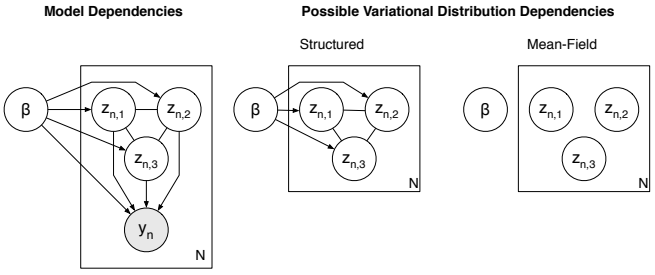

Figure 1: Left: Graphical model showing the independence structure of models in our framework. We make no assumptions on dependencies between the elements of z_n, so the elements of z_n are all connected. Here we show only three local z variables, but there could be more or fewer. Middle: Graphical model of structured variational distributions $q(\beta, z)$ that are possible in our framework. Right: Graphical model of mean-field variational distributions.

Like batch variational inference, SVI relies on the mean-field approximation, which requires that q factorize as $(\prod_k q(\beta_k)) \prod_{n,m} q(z_{n,m})$; that is, SVI approximates the joint posterior $p(z, \beta|y)$ with a distribution q that cannot represent any dependencies between random variables. Mimno et al. (2012) proposed a variant for latent Dirichlet allocation that restores dependencies between sets of local hidden variables $z_{n,1:M}$ so that the approximating distribution has the form $q(z, \beta) = (\prod_k q(\beta_k)) \prod_n q(z_{n,1:M})$, improving q's ability to approximate the posterior $p(z, \beta|y)$. But their method still breaks the dependence between the global and local variables β and z.

In this paper, we introduce structured stochastic variational inference (SSVI), a generalization of the SVI framework that can restore the dependence between global and local variables. SSVI approximates the posteriors $p(z, \beta|y)$ of a wide class of models with distributions of the form $q(z, \beta) = (\prod_k q(\beta_k)) \prod_n q(z_n|\beta)$, allowing for arbitrary dependencies between β and z_n.

In experiments on three models and datasets, we find that in all cases restoring these dependencies allows SSVI to find qualitatively and quantitatively better parameter estimates, avoiding the local optima and sensitivity to hyperparameters that plague mean-field variational inference.

2 Structured Stochastic Variational Inference

In this section, we will present two SSVI algorithms. We first review the class of models to which SSVI can be applied and the variational distributions that it employs.

2.1 Model Assumptions

As in SVI (Hoffman et al., 2013), we assume we have N groups of observations $y_{1:N}$ and a probability model that factorizes as $p(y, z, \beta) = p(\beta) \prod_n p(y_n, z_n|\beta)$. The independence structure of such a model is visualized in figure 1. The global parameters β are shared across all observations, and the local hidden variables $z_{1:N}$ are conditionally independent of one another given the global parameters β.

We will restrict our attention to conditionally conjugate models. Specifically, we assume that

$$\begin{aligned} p(\beta) &= h(\beta) \exp\{\eta \cdot t(\beta) - A(\eta)\} \\ p(y_n, z_n|\beta) &= \exp\{t(\beta) \cdot \eta_n(y_n, z_n) + g_n(y_n, z_n)\}, \end{aligned} \tag{1}$$

where the base measure h and log-normalizer A are scalar-valued functions, η is a vector of natural parameters, $t(\beta)$ is a vector-valued sufficient statistic function, g_n is a scalar-valued function and η_n is a vector-valued function.

This form for $p(y_n, z_n, \beta)$ includes all conjugate pairs of distributions $p(\beta)$, $p(y_n, z_n|\beta)$ (Gelman et al., 2013); that is, it is the most general family of distributions for which the conditional $p(\beta|y, z)$ is in the same family as the prior $p(\beta)$. This conditional is

$$\begin{aligned} p(\beta|y, z) = h(\beta) \exp\{(\eta + \textstyle\sum_n \eta_n(y_n, z_n)) \cdot t(\beta) - \\ A(\eta + \textstyle\sum_n \eta_n(y_n, z_n))\}. \end{aligned} \tag{2}$$

These restrictions are a weaker version of those imposed by Hoffman et al. (2013); the difference is that we make no assumptions about the tractability of the conditional distributions $p(z_n|y_n, \beta)$ or $p(z_{n,m}|y_n, z_{n,\setminus m}, \beta)$. This work is therefore applicable to any model that fits in the SVI framework, including mixture models, LDA, hidden Markov models (HMMs), factorial HMMs, Kalman filters, factor analyzers, probabilistic matrix factorizations, hierarchical linear regression, hierarchical probit regression, and many other hierarchical models. Unlike SVI, it can also address models without tractable local conditionals, such as multilevel logistic regressions (Gelman and Hill, 2007) or the correlated topic model (Blei and Lafferty, 2006).

The conjugacy assumptions above simplify the form of the SSVI updates in algorithm 1, but they could be relaxed. However, the simpler SSVI-A updates in algorithm 1 depend strongly on these assumptions.

2.2 Approximating Distribution

Our goal is to approximate the intractable posterior $p(z, \beta|y)$ with a distribution $q(z, \beta)$ in some restricted, tractable family. We will choose a q distribution from this family by solving an optimization problem, minimizing the Kullback-Leibler (KL) divergence between $q(z, \beta)$ and the posterior $p(z, \beta|y)$.

The simplest approach is to make the mean-field approximation, restricting q to factorize so that $q(z, \beta) = q(\beta) \prod_n \prod_m q(z_{n,m})$. This restriction dramatically simplifies the form of the KL-divergence between q and the

posterior, but this simplicity comes at a price. Every dependence that we break to make q easier to work with makes q less able to closely approximate the posterior $p(z, \beta|y)$. Breaking dependencies may also introduce additional local minima into the KL divergence between q and the posterior; imposing independence assumptions places nonconvex constraints on the dual of the solution space, which may block the path from a bad solution to a good one (Wainwright and Jordan, 2008).

Structured mean-field partially relaxes the mean-field independence restriction (Saul and Jordan, 1996). Traditional structured mean-field algorithms require the practitioner to identify and exploit some model-specific structure; for example, Ghahramani and Jordan (1997) exploited dynamic programming algorithms for HMMs to derive a structured mean-field algorithm for factorial HMMs.

Mimno et al. (2012) proposed a structured stochastic variational inference algorithm for latent Dirichlet allocation that depends less on model-specific structure, placing no restrictions on the joint distributions $q(z_{n,1:M})$ so that $q(z, \beta)$ factorizes as $q(z, \beta) = (\prod_k q(\beta_k)) \prod_n q(z_n)$. The optimal $q(z_n)$ may not be tractable to normalize, but it can still be sampled from using Markov chain Monte Carlo (MCMC), which is all that is necessary to generate a stochastic natural gradient for a stochastic variational inference algorithm. The result was a significant improvement in the quality of the inference algorithm's ability to obtain high-quality estimates of model parameters. However, the approximate posterior of Mimno et al. (2012) still breaks the dependence between global parameters β and local hidden variables z.

We introduce a framework for *structured stochastic variational inference* (SSVI) algorithms that restore the dependence between β and z. Our variational distribution q is of the form

$$q(z, \beta) = (\textstyle\prod_k q(\beta_k)) \prod_n q(z_n|\beta). \tag{3}$$

The only remaining factorization we impose is between the elements of β; the conditional independence between the z_ns given β is implied by the model in equation 2.

We will restrict $q(\beta)$ to be in the same exponential family as the prior $p(\beta)$, so that $q(\beta) = h(\beta) \exp\{\lambda \cdot t(\beta) - A(\lambda)\}$. λ is a vector of free parameters that controls $q(\beta)$. We also require that any dependence under q between z_n and β be mediated by some vector-valued function $\gamma_n(\beta)$, so that we may write $q(z_n|\beta) = q(z_n|\gamma_n(\beta))$.

This form for q allows for rich dependencies between nearly all model variables. This comes at a cost, however. In mean-field variational inference, we maximize a lower bound on the marginal probability of the data; this is equivalent to minimizing the KL divergence from q to the posterior (Bishop, 2006). However, this lower bound contains expectations that become impossible to compute when we allow z_n to depend on β in q. This issue may seem insurmountable, but even though we cannot compute the bound, we can still optimize it using stochastic optimization.

2.3 The Structured Variational Objective

Our goal is to find a distribution $q(\beta, z)$ that has low KL divergence to the posterior $p(\beta, z|y)$. The KL divergence between q and the full posterior is

$$\begin{aligned}\mathrm{KL}(q_{z,\beta}||p_{z,\beta|y}) = -\mathbb{E}_q[\log p(y, z, \beta)] + \\ \mathbb{E}_q[\log q(z, \beta)] + \log p(y).\end{aligned} \tag{4}$$

Because the KL divergence must be non-negative, this yields the evidence lower bound (ELBO)

$$\begin{aligned}\mathcal{L} &\equiv \mathbb{E}_q[\log p(y, z, \beta)] - \mathbb{E}_q[\log q(z, \beta)] \\ &= \mathbb{E}_q[\log \tfrac{p(\beta)}{q(\beta)}] + \textstyle\sum_n \mathbb{E}_q[\log \frac{p(y_n, z_n|\beta)}{q(z_n|\beta)}] \\ &= \textstyle\int_\beta q(\beta) \big(\log \frac{p(\beta)}{q(\beta)} + \\ &\quad + \textstyle\sum_n \int_{z_n} q(z_n|\beta) \log \frac{p(y_n, z_n|\beta)}{q(z_n|\beta)} dz_n \big) d\beta \le \log p(y),\end{aligned} \tag{5}$$

We used the conditional independence structure assumed in section 2.1 to break $\log p(y, z|\beta)$ into a sum over n. Our goal is to maximize the ELBO subject to some restrictions on q.

Before describing the form we choose for $\gamma_n(\beta)$, we first note that the second integral in equation 5 is itself a lower bound on the marginal probability of the nth group of observations:

$$\begin{aligned}&\textstyle\int_{z_n} q(z_n|\beta) \log \frac{p(y_n, z_n|\beta)}{q(z_n|\beta)} dz_n \\ &= -\mathrm{KL}(q_{z_n|\beta}||p_{z_n|y_n,\beta}) + \log p(y_n|\beta) \le \log p(y_n|\beta).\end{aligned} \tag{6}$$

Thus, for any particular value of β we can maximize the global ELBO over $q(z_n|\beta)$ by minimizing the KL divergence between $q(z_n|\beta)$ and $p(z_n|y_n, \beta)$. We will assume that the function $\gamma_n(\beta)$ that controls $q(z_n|\beta) = q(z_n|\gamma_n(\beta))$ is defined to do just that, so that $\gamma_n(\beta)$ is at a local maximum of this "local ELBO", i.e.,

$$\nabla_{\gamma_n} \textstyle\int_{z_n} q(z_n|\gamma_n(\beta)) \log \frac{p(y_n, z_n|\beta)}{q(z_n|\gamma_n(\beta))} dz_n = 0. \tag{7}$$

The function $\gamma_n(\beta)$ may be implicit; e.g., it might be evaluated by solving an optimization problem.

2.4 Algorithms

Our algorithm for optimizing the ELBO from equation 5 is summarized in algorithm 1. Each iteration, we sample the global parameters β from $q(\beta)$. We then compute the parameters $\gamma_n(\beta)$ to each local variational distribution $q(z_n|\gamma_n(\beta))$ that maximize the local ELBO $\mathbb{E}_q[\log p(y_n, z_n|\beta) - \log q(z_n|\beta)|\beta]$ and compute an unbiased estimate $\hat{\eta}_n$ of $\mathbb{E}_q[\eta_n(y_n, z)|\beta]$, the expected value

of $\eta_n(y_n, z)$ with respect to the maximized local ELBO. Finally, we update λ with some step size ρ so that:

$$\lambda' = (1-\rho)\lambda + \rho(\eta + V(\beta,\lambda)\textstyle\sum_n \hat{\eta}_n). \tag{8}$$

where $V(\beta, \lambda)$ is a matrix that is defined in terms of the cumulative distribution functions (CDFs) quantile functions (inverse-CDFs) of $q(\beta)$. Defining the CDF $Q_k(\beta_k) \equiv \int_{-\infty}^{\beta_k} q(\beta'_k)d\beta'_k$ and the quantile function $R_k(Q_k(\beta_k)) \equiv \beta_k$, $V(\beta, \lambda)$ is defined as the product of two matrices: the inverse of the second derivative of the log-normalizer A of q, and the Jacobian of $t(R(Q(\beta)))$ with respect to λ:

$$V(\beta,\lambda) \equiv \nabla^2_\lambda A(\lambda)^{-1} \nabla_\lambda R(Q(\beta)) \nabla_\beta t(\beta)^\top. \tag{9}$$

Equation 8 is derived in appendix A as a stochastic natural gradient update of the SSVI ELBO. It resembles the standard SVI update, with two differences: we use a sample from $q(\beta)$ instead of $\mathbb{E}_q[t(\beta)]$ when estimating $\hat{\eta}_n$, and we multiply $\hat{\eta}_n$ by the matrix $V(\beta, \lambda)$.

The matrix V appears because of the dependence between β and $\hat{\eta}$. V's expected value is the identity matrix, since $\mathbb{E}_q[\nabla_\lambda t(R(Q(\beta)))] = \nabla_\lambda \mathbb{E}_q[t(\beta)] = \nabla^2_\lambda A(\lambda)$. (The last identity follows from $q(\beta)$ being in an exponential family.) So we can decompose $V\hat{\eta}$ as

$$V\hat{\eta} = (V - \mathbb{E}_q[V])(\hat{\eta} - \mathbb{E}_q[\hat{\eta}]) + \hat{\eta}. \tag{10}$$

If the covariance-like expression $(V - \mathbb{E}_q[V])(\hat{\eta} - \mathbb{E}_q[\hat{\eta}])$ is small (either because the variance of V or $\hat{\eta}$ is small or because V and $\hat{\eta}$ do not depend strongly on one another), we can neglect it without introducing much bias. This suggests the simpler update (labeled SSVI-A in algorithm 1)

$$\lambda' = (1-\rho)\lambda + \rho(\eta + \textstyle\sum_n \hat{\eta}_n), \tag{11}$$

which avoids the difficulty and (modest) expense of computing derivatives of quantiles and log-normalizers.

2.5 A matrix of approaches

We are free to choose any form for $q(z_n|\beta)$ and any unbiased estimator $\hat{\eta}$ that we want; different choices have different properties:

Exact conditional: $q(z_n|\beta) = p(z_n|y_n, \beta)$. Setting $q(z_n|\beta)$ to the local conditional distribution gives the best possible local ELBO. But this conditional may be intractable, in which case one can use MCMC to estimate $\mathbb{E}_q[\eta_n(y_n, z_n)|\beta]$. With this choice SSVI minimizes the KL divergence between $q(\beta)$ and the *marginal* posterior $p(\beta|y)$, integrating out the "nuisance parameters" z.

Mean-field: $q(z_n|\beta) = \prod_m q(z_{n,m}|\beta)$. This choice may be computationally cheaper, and it will often make it possible to compute $\mathbb{E}_q[\eta_n(y_n, z_n|\beta)]$ in closed form, but it reduces the algorithm's ability to approximate the marginal posterior $p(\beta|y)$. That said, this is still better than a full mean-field approach where one also breaks the dependencies between z and β.

Algorithm 1 Structured stochastic variational inference (SSVI).

Initialize $t = 1$, initialize $\lambda^{(0)}$ randomly.
repeat
  Compute step size $\rho^{(t)} = st^{-\kappa}$.
  Set $N_t = \min\{t, N\}$.
  Sample global parameters $\beta^{(t)}$ from $q(\beta)$.
  Compute the local variational parameters $\gamma_n(\beta^{(t)})$.
  Compute an estimate $\hat{\eta}_n$ such that $\mathbb{E}[\hat{\eta}_n] = \int_{z_n} q(z_n|\gamma_n(\beta^{(t)}))\eta_n(y_n, z_n)dz_n$.
  Option 1: SSVI.
  Set $\lambda^{(t)} = (1-\rho^{(t)})\lambda^{(t-1)} + \rho^{(t)}(\eta + V(\beta^{(t)}, \lambda^{(t)})\sum_n \hat{\eta}_n)$. (See equation 9 for definition of $V(\beta, \lambda)$.)
  Option 2: SSVI-A.
  Set $\lambda^{(t)} = (1-\rho^{(t)})\lambda^{(t-1)} + \rho^{(t)}(\eta + \sum_n \hat{\eta}_n)$.
until convergence

Non-bound-preserving approaches: Although we derived SSVI assuming that $q(z_n|\beta)$ is chosen to maximize the local ELBO $\mathcal{L}_n$, one could obtain a distribution over z_n in other ways. For example, for latent Dirichlet allocation one could adapt the CVB0 method of Asuncion et al. (2009) to use a fixed value of β, resulting in an algorithm akin to that of Foulds et al. (2013).

All of these choices of local variational distribution could also be used in a traditional mean-field setup where $q(\beta, z) = q(\beta)q(z)$, or as part of a variational maximum a posteriori (MAP) estimation algorithm. So for any model that enjoys conditional conjugacy we have a matrix of possible variational inference algorithms: we can match any "E-step" (e.g. mean-field or sampling from the exact conditional) used to approximate $p(z_n|y_n, \beta)$ with any "M-step" (e.g. MAP, mean-field, SSVI, SSVI-A) used to update our approximation to $p(\beta|y)$.

2.6 Extensions

There are several ways in which the basic algorithms presented above can be extended:

Subsampling the data. As in SVI, we can compute an unbiased estimate of the sum over n in equation 8 by only computing $\hat{\eta}_n$ for some randomly sampled subset of S observations, resulting in the update

$$\lambda' = (1-\rho)\lambda + \rho(\eta + V(\beta,\lambda)\tfrac{N}{S}\textstyle\sum_n \hat{\eta}_n). \tag{12}$$

For large datasets, the reduced computational effort of only looking at a fraction of the data far outweighs the noise that

this subsampling introduces. Taking a cue from the recent work of Broderick et al. (2013) and Wang and Blei (2012), we suggest gradually ramping up the multiplier N over the course of the first sweep over the dataset.

Hyperparameter updates and parameter hierarchies. As in the mean-field stochastic variational inference framework of Hoffman et al. (2013), we can optimize any hyperparameters in our model by taking steps in the direction of the gradient of the ELBO with respect to those hyperparameters. We can also extend the framework developed in this paper to models with hierarchies of global parameters as in appendix A of (Hoffman et al., 2013).

3 Related Work

The idea of sampling from global variational distributions to optimize intractable variational inference problems has been proposed previously in several contexts. Ji et al. (2010); Nott et al. (2012); Gerrish (2013); Paisley et al. (2012), and Ranganath et al. (2014) proposed sampling without a change of variables as a way of coping with non-conjugacy. Kingma and Welling (2014) and Titsias and Lázaro-Gredilla (2014) proposed methods that do use a change of variables, although their methods focus more on speed and/or dealing with nonconjugacy than on improving the accuracy of the variational approximation. Salimans and Knowles (2013) also used a change of variables to dramatically reduce the variance of their stochastic gradients.

Although some of the above methods use stochastic optimization to improve the quality of the mean-field approximation, there are major differences between these methods and SSVI. Ji et al. (2010) apply their method to models where some parameters have been analytically marginalized out, but do not consider explicitly structured variational distributions. Also, in our informal experiments we found that the variance of their gradient estimator was unacceptably high for high-dimensional problems.

Salimans and Knowles (2013) apply their stochastic linear regression method to structured variational distributions in which the natural parameters of lower-level variational distributions are explicit functions of draws from higher-level variational distributions. In comparison, SSVI's implicit form for the conditional distribution $q(z|\beta)$ allows for more complicated dependencies between β and z.

A final difference is that the papers mentioned above do not exploit conjugacy relationships, which are central to the ease of implementation and efficiency of SSVI-A.

Two other related algorithms are due to Mimno et al. (2012) (which SSVI generalizes) and Wang and Blei (2012). The algorithm of Wang and Blei (2012) also uses sampling and conjugacy relationships to attempt to restore the dependencies broken in mean-field algorithms, although their method lacks guarantees of convergence or correctness.

4 Experiments

In this section we empirically evaluate SSVI and SSVI-A's ability to estimate parameters for three hierarchical Bayesian models. In each case, relaxing the mean-field approximation allows SSVI and SSVI-A to find significantly better parameter estimates than mean-field. We also find evidence suggesting that this superior performance is primarily due to SSVI/SSVI-A's ability to avoid local optima.

4.1 Latent Dirichlet allocation

We evaluated the quality of parameter estimates from SSVI and SSVI-A on the latent Dirichlet allocation (LDA) topic model fit to the 3,800,000-document Wikipedia dataset from (Hoffman et al., 2013). We compared with full mean-field stochastic variational inference (Blei et al., 2003; Hoffman et al., 2010a), a mean-field M-step with a Gibbs sampling E-step (Mimno et al., 2012), SSVI with Gibbs, and SSVI-A with Gibbs. Results for other E-step/M-step combinations are in the supplement.

To speed up learning, each update we subsample a mini-batch of 1,000 documents rather than analyzing the whole dataset each iteration. We also experimented with various settings of the hyperparameters α and η, which mean-field variational inference for LDA is known to be quite sensitive to (Asuncion et al., 2009). For all algorithms we used a step size schedule $\rho^{(t)} = t^{-0.75}$.

We held out a test set of 10,000 documents, and periodically evaluated the average per-word marginal log probability assigned by the model to each test document, using the expected value under the variational distribution as a point estimate of the topics. We estimated marginal log probabilities with a Chib-style estimator (Wallach et al., 2009).

Figure 2 summarizes the results for $\alpha = 0.1$, which yielded the best results for all algorithms. The method of Mimno et al. (2012) outperforms the online LDA algorithm of Hoffman et al. (2010a), but both methods are very sensitive to hyperparameter selection. SSVI achieves good results regardless of hyperparameter choice. SSVI-A's performance is very slightly worse than that of SSVI.

The approximating Dirichlet distributions found by SSVI were more diffuse than those found by SVI or SSVI-A, having lower overall parameter values λ. Many elements of λ were smaller than the prior hyperparameter η, a solution that is not stable under the SVI and SSVI-A updates.

We also evaluated the stochastic gradient Riemannian Langevin dynamics (SGRLD) algorithm for LDA, which Patterson and Teh (2013) found outperformed the Gibbs-within-SVI method of Mimno et al. (2012). We experimented with various hyperparameter settings, including

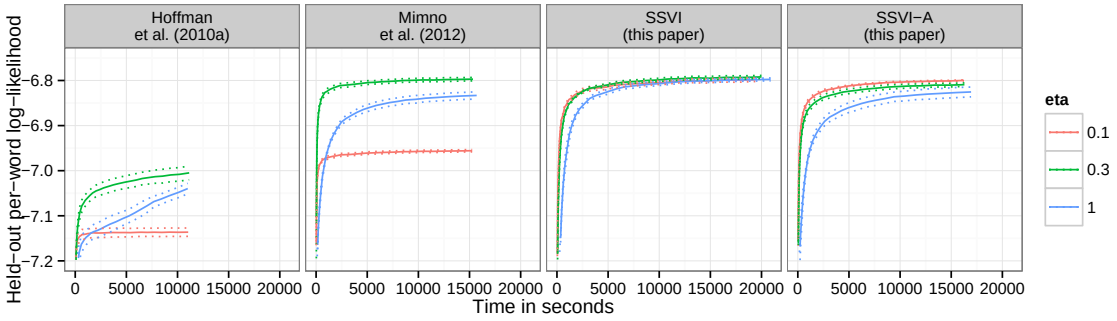

Figure 2: Predictive accuracy for various algorithms and hyperparameter settings as a function of wallclock time when fitting LDA to 3.8 million Wikipedia articles. Solid lines show average performance across five runs, dotted lines are drawn one standard deviation above and below the mean. Each algorithm ran for two sweeps over the corpus.

the optimal values reported by Patterson and Teh (2013). The best per-word marginal log probability achieved by SGRLD was -6.83. SGRLD thus outperforms standard SVI regardless of hyperparameters, but only outperforms Gibbs-within-SVI for some hyperparameter settings, and achieves performance comparable to that of SSVI and SSVI-A. This suggests that the poor performance of Gibbs-within-SVI in the experiments of Patterson and Teh (2013) may be due to their setting $\eta = 0.1$.

Computational costs: With mini-batches of 1000 documents the computational costs of SSVI and SSVI-A were comparable to those of Gibbs-within-SVI: SSVI took about 30% longer than SSVI to analyze the same number of documents, and SSVI-A's speed is very close to SSVI. Using Gibbs sampling in the "E-step" rather than variational inference costs about 50% more than a mean-field E-step. We found that SGLRD was significantly slower per document than either SSVI or SSVI-A, although this may be due to implementation differences; in principle the methods should have comparable costs.

Local optima: The improved performance of SSVI and SSVI-A over SVI might be because SSVI/SSVI-A optimize an objective function that more closely approximates the KL divergence between $q(\beta)$ and $p(\beta|y)$ than the MF objective does. But it could also be because the MF objective includes undesirable local optima that the SSVI objective does not, and so SVI cannot help getting stuck in these local optima. Our experiments above show that SSVI and SSVI-A consistently find better parameter estimates than MF even with multiple restarts, but this does not rule out local optima as an explanation—even with many restarts it might be very difficult for MF to find a good local optimum.

To test this question, we intialized mean-field SVI with the variational parameters found by SVI, Gibbs-within-SVI, SSVI, and SSVI-A, and did another sweep through the Wikipedia dataset. Figure 3 plots the mean-field ELBOs

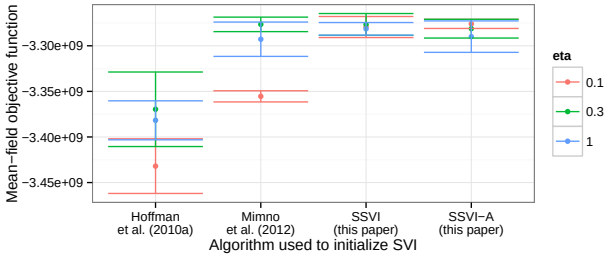

Figure 3: ELBOs obtained by initializing SVI with the variational parameters obtained by four algorithms. Points denote ELBOs averaged over five runs, error bars denote the range of ELBOs obtained.

obtained by running SVI from each initialization. SVI initialized with the result from the structured algorithms finds a much better local optimum of the ELBO on the training set than SVI initialized randomly.

This result suggests that the main weakness of mean-field methods may not be the inability of factorized distributions to adequately approximate the posterior, but the difficulty of finding a good local optimum of the ELBO (at least in high-dimensional, multimodal problems). Conversely, the improved performance of the structured methods may be due to the relative lack of nasty local optima in the structured ELBOs. That the structured methods find variational distributions that are also good in the mean-field setting suggests that the structured ELBO may resemble a smoother version of the mean-field ELBO.

4.2 Dirichlet process mixture of Bernoullis

We used a synthetic data experiment to compare mean-field and SSVI-A's ability to correctly approximate the posterior of (a finite approximation to) a Dirichlet process mixture of Bernoullis model. The data were generated according to the process

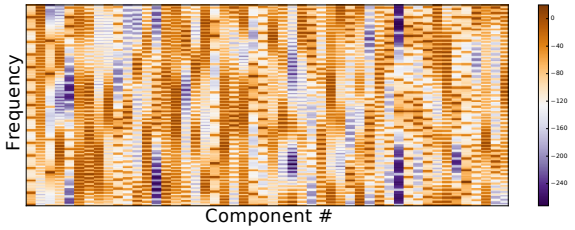

Figure 4: Dictionary of synthetic spectra used in the GaP-KL-NMF experiment. Magnitudes are shown in dB.

$\pi \sim \text{Dirichlet}(\frac{\alpha}{K})$; $\phi_{k,d} \sim \text{Beta}(1,1)$; $z_n \sim \text{Multinomial}(\pi)$; $y_{n,d} \sim \text{Bernoulli}(\phi_{z_n,d})$, where the size of π is $K = 100$ and the hyperparameter $\alpha = 20$. Each observation y_n is a 100-dimensional binary vector, and 1000 such vectors were sampled. The model ultimately used only 56 of the 100 mixture components to generate data. Given the correct hyperparameters, we used mean-field and SSVI-A[1] (using the full dataset as a "minibatch") to approximate the posterior $p(z, \pi, \phi|y)$. We also applied collapsed Gibbs sampling (CGS) (Neal, 2000), which yields samples from the posterior that are asymptotically unbiased.

SSVI-A's performance closely mirrored that of CGS; both methods were more accurate than mean-field. Mean-field only discovered 17 of the 56 mixture components; the rest were not significantly associated with data. By contrast, SSVI-A and CGS used 54 and 55 components respectively. We also estimated (using Monte Carlo) the KL divergence between the true data-generating distribution $p(y|\pi, \phi)$ and $p(y|\hat{\pi}, \hat{\phi})$, where $\hat{\pi}$ and $\hat{\phi}$ are the estimates of the posterior means of π and ϕ obtained by mean-field, SSVI-A, and CGS. Mean-field achieved a KL divergence of 5.23, while CGS and SSVI-A achieved much lower KL divergences of 1.9 and 1.94, respectively.

4.3 Bayesian nonparametric nonnegative matrix factorization

We also evaluated the ability of SSVI-A to determine an appropriate number of active components in a Bayesian nonparametric model of audio magnitude spectrograms proposed by Nakano et al. (2011) as a variant on the GaP-NMF model of Hoffman et al. (2010b). The model, which we will call GaP-KL-NMF, assumes that a quantized magnitude spectrogram matrix Y is sampled according to the following generative process:

$$\begin{aligned} W_{f,k} \sim \text{Gamma}(a, a); \quad H_{k,t} \sim \text{Gamma}(b, b) \\ \theta_k \sim \text{Gamma}(\alpha/K, \alpha c); \\ Y_{f,t} \sim \text{Poisson}(\textstyle\sum_k \theta_k W_{f,k} H_{k,t}), \end{aligned} \tag{13}$$

[1]We did not compare with SSVI, since its performance on the LDA experiment was so similar to that of SSVI-A.

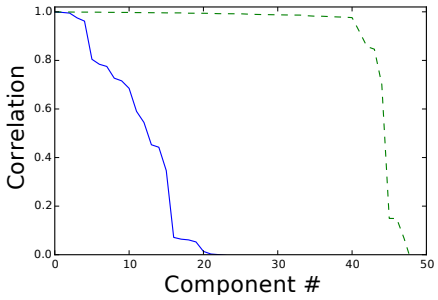

Figure 5: Sorted correlations between elements of the true dictionary of bases used to generate a synthetic spectrogram and the dictionaries learned by mean-field (solid blue line) and SSVI-A (dashed green line) on GaP-KL-NMF.

so that $Y \approx W\text{diag}(\theta)H$. As K gets large, the prior on θ approximates a gamma process, in which most elements of θ are expected to be very near 0 but a few may be much larger than 0. This prior allows the model to determine how many components it needs to explain the data.

We performed an experiment to compare how well mean-field variational inference and SSVI-A can accomplish the task of determining how many latent components actually generated a synthetic spectrogram. We randomly generated a 257-by-50 dictionary matrix W whose columns resemble the magnitude spectra of speech sounds (shown in figure 4), sampled a 50-by-1000 activation matrix H of independent draws from a $\text{Gamma}(0.2, 0.2)$ distribution, and generated a spectrogram Y such that $Y_{f,t} \sim \text{Poisson}(\sum_k W_{f,k} H_{k,t})$. We then used MF and SSVI-A to approximate the posterior $p(W, H, \theta|Y)$ under the GaP-KL-NMF model with hyperparameters $a = 0.25$, $b = 0.2$, $\alpha = 1$, $c = 1$.

Figure 5 plots the correlation between each column of the true dictionary W and its closest match among the columns of the estimated dictionaries $\mathbb{E}_q[W]$ obtained using MF and SSVI-A. These correlations are sorted in decreasing order. SSVI-A recovers good approximations of nearly all of the bases used to generate the data, whereas MF performs quite poorly. This is because MF grossly underestimates the number of components used to generate the data—the components with near-zero correlations are not used to explain the data. As in section 4.2, MF quickly gets stuck in a bad local optimum from which it cannot recover.

5 Discussion

We have presented stochastic structured variational inference (SSVI), an algorithmic framework that uses stochastic variational inference to restore the dependencies between global and local unobserved variables that mean-field variational inference breaks. Experiments suggest that both SSVI-A and SSVI can fit models in a way that outperforms

previously existing variational inference algorithms.

A Algorithm derivation

The convergence proofs that our stochastic variational inference algorithm relies on require that the ELBO be twice continuously differentiable in λ (Hoffman et al., 2013). We also require that $q(\beta)$ and $p(y, z, \beta)$ be continuously differentiable in β for any y and z. Finally, we require that $q(z_n|\gamma_n)$ be continuously differentiable in γ_n, that $\gamma_n(\beta)$ be continuously differentiable in β, and that it be possible to compute an unbiased estimate $\hat{\eta}_n$ of the expectation $\int_{z_n} q(z_n|\beta)\eta_n(y_n, z_n)dz_n$ for any β (for example, using Markov chain Monte Carlo to sample from $q(z_n|\beta)$).

Our derivation will rely heavily on the concept of *sampling by inversion*, a general method for sampling from univariate distributions. Considering for the moment the case where each β_k is a scalar[2], we can obtain a draw of β from q by first sampling K uniform random variables $u_k \sim \text{Uniform}([0, 1])$ and then passing each through the quantile function R; i.e., $\beta \sim q \Leftrightarrow u_k \sim \text{Uniform}([0, 1]), \beta_k = R_k(u_k)$. We will use the vector-valued functions $R(u) = (R_1(u_1), \ldots, R_K(u_K))^\top$ $Q(u) = (Q_1(u_1), \ldots, Q_K(u_K))^\top$ to denote the concatenations of the outputs of the K quantile functions and CDFs. R and Q depend on both u and on the parameters λ that control $q(\beta)$, but we suppress the dependence on λ to avoid clutter. Below, we will need the derivative of R with respect to λ; this can be obtained using an identity derived in the supplemental material.

We begin by writing the bound in equation 5 as a function of λ. We define $\mathcal{L}(\lambda)$ as the ELBO achieved by setting $q(z_n|\beta) = q(z_n|\gamma_n(\beta))$ for all values of β:

$$\mathcal{L}(\lambda) \equiv \int_\beta q(\beta)(\log \tfrac{p(\beta)}{q(\beta)} + \textstyle\sum_n \mathcal{L}_n(\beta, \gamma_n(\beta)))d\beta; \quad (15)$$

$$\mathcal{L}_n(\beta, \gamma) \equiv \int_{z_n} q(z_n|\gamma(\beta)) \log \tfrac{p(y_n, z_n|\beta)}{q(z_n|\gamma_n(\beta))} dz_n \quad (16)$$

where we define $\mathcal{L}_n(\beta, \gamma_n)$ as a shorthand for the part of the ELBO that depends on y_n and z_n.

[2]We can also sample by inversion from multivariate distributions such as the Dirichlet or multivariate normal by generating a set of independent random variables by inversion and then transforming them, so that

$$R(u) = T(\hat{R}(u)), \quad (14)$$

where $\hat{R}$ is a vector of K univariate quantile functions and T is a transformation that introduces dependencies. For example, if q is a multivariate normal distribution with mean 0 and covariance Σ, we could make each $\hat{R}_k$ the quantile function of a standard normal and let $T(\hat{R}) \equiv \sqrt{\Sigma}R$, so that if u is a vector of uniform random numbers then $T(\hat{R}(u))$ would be a draw from q. Or, to sample from a Dirichlet we could use the relation $\hat{\beta}_k \sim \text{Gamma}(\lambda_k, 1) \Rightarrow \hat{\beta}/\sum_l \hat{\beta}_l \sim \text{Dirichlet}(\lambda)$, set $\hat{R}_k(u_k)$ to be the $\text{Gamma}(\lambda_k, 1)$ quantile function, and set $T(\hat{R}(u)) = \hat{R}(u)/\sum_k \hat{R}_k(u)$.

We will optimize this bound using stochastic optimization. To do so, we need to consider the derivative of $\mathcal{L}(\lambda)$ with respect to λ. The derivative of the first term in the expectation simplifies:

$$\begin{aligned} &\nabla_\lambda \textstyle\int_\beta q(\beta) \log \frac{p(\beta)}{q(\beta)} d\beta \\ &= \nabla_\lambda((\eta - \lambda)^\top \nabla_\lambda A(\lambda) - A(\eta) + A(\lambda)) \\ &= \nabla^2_\lambda A(\lambda)(\eta - \lambda). \end{aligned} \quad (17)$$

This is a consequence of the exponential-family identity $\mathbb{E}_q[t(\beta)] = \nabla_\lambda A(\lambda)$. The derivatives of the remaining terms do not simplify so easily. It will be easier to simplify them by rewriting the expectation of $\mathcal{L}_n$ in terms of the quantile function $R(u)$ we defined in section 2.3:

$$\textstyle\int_\beta q(\beta)\mathcal{L}_n(\beta, \gamma_n(\beta))d\beta = \int_u \mathcal{L}_n(R(u), \gamma_n(R(u)))du. \quad (18)$$

Now, taking the derivative of $\mathcal{L}_n$ with respect to λ using the chain rule yields

$$\begin{aligned} &\nabla_\lambda \mathcal{L}_n(R(u), \gamma_n(R(u))) \\ &= (\nabla_\lambda R(u))\Big(\nabla_\beta \mathcal{L}_n(R(u), \gamma_n(R(u))) \\ &\qquad + (\nabla_\beta \gamma_n(R(u)))\nabla_{\gamma_n} \mathcal{L}_n\big(R(u), \gamma_n(R(u))\big)\Big). \end{aligned} \quad (19)$$

The second term vanishes because $\nabla_{\gamma_n} \mathcal{L}_n(\beta, \gamma_n(\beta)) = 0$ by the definition in equation 7, and so the derivative of equation 18 simplifies to

$$\begin{aligned} &\nabla_\lambda \textstyle\int_u \mathcal{L}_n(R(u), \gamma_n(R(u)))du \\ &= \textstyle\int_u (\nabla_\lambda R(u))\nabla_\beta \mathcal{L}_n(R(u), \gamma_n(R(u)))du \\ &= \textstyle\int_{u,z_n} q(z_n|\gamma_n(R(u)))(\nabla_\lambda R(u))(\nabla_\beta t(R(u)))^\top \\ &\qquad\qquad \eta_n(y_n, z_n)dz_n du \end{aligned} \quad (20)$$

Now, if we sample the global parameters $\beta \sim q_\beta$ and compute an unbiased estimate $\hat{\eta}_n$ of the expectation $\int_{z_n} q(z_n|\gamma_n(\beta))\eta_n(y_n, z_n)dz_n$, then we can compute a random vector g whose expectation is the true gradient:

$$\begin{aligned} g &\equiv (\nabla^2_\lambda A(\lambda))(\eta - \lambda) \\ &\quad + (\nabla_\lambda R(Q(\beta)))(\nabla_\beta t(\beta))^\top \textstyle\sum_n \hat{\eta}_n; \quad \mathbb{E}[g] = \nabla_\lambda \mathcal{L}. \end{aligned} \quad (21)$$

A stochastic natural gradient can be obtained by preconditioning this stochastic gradient with the inverse of the Fisher information matrix of $q(\beta)$:

$$g^{\text{nat}} \equiv -\lambda + \eta + V(\beta, \lambda) \textstyle\sum_n \hat{\eta}_n, \quad (22)$$

where $V(\beta, \lambda)$ is defined as in equation 9 as

$$V(\beta, \lambda) \equiv (\nabla^2_\lambda A(\lambda))^{-1}(\nabla_\lambda R(Q(\beta)))(\nabla_\beta t(\beta))^\top \quad (23)$$

It is easy to verify that $\mathbb{E}[g^{\text{nat}}] = (\nabla^2_\lambda A(\lambda))^{-1}\nabla_\lambda \mathcal{L}$. Because the Fisher information matrix of $q(\beta)$ is the second derivative of A with respect to λ, $\mathbb{E}[g^{\text{nat}}]$ is the natural gradient of $\mathcal{L}$ with respect to λ (Sato, 2001). We can therefore optimize $\mathcal{L}$ with respect to λ by repeatedly sampling values of g^{nat} and plugging them into a standard Robbins-Monro stochastic approximation algorithm (Hoffman et al., 2013). Such an approach is summarized in algorithm 1 (SSVI).

References

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

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

Blei, D. and Lafferty, J. (2006). Correlated topic models. In *Advances in Neural Information Processing Systems 18 (NIPS) 18*, pages 147–154. MIT Press.

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

Broderick, T., Boyd, N., Wibisono, A., Wilson, A., , and Jordan, M. (2013). Streaming variational Bayes. In *Advances in Neural Information Processing Systems (NIPS)*.

Foulds, J., Boyles, L., DuBois, C., Smyth, P., and Welling, M. (2013). Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2013). *Bayesian Data Analysis*. Chapman & Hall, London.

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

Gerrish, S. (2013). *Applications of Latent Variable Models in Modeling Influence and Decision Making*. PhD thesis, Princeton University.

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

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

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

Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Ji, C., Shen, H., and West, M. (2010). Bounded approximations for marginal likelihoods. Technical report, Tech. report, Duke University.

Kingma, D. and Welling, M. (2014). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

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

Nakano, M., Le Roux, J., Kameoka, H., Nakamura, T., Ono, N., and Sagayama, S. (2011). Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

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

Nott, D., Tan, S., Villani, M., and Kohn, R. (2012). Regression density estimation with variational methods and stochastic approximation. *Journal of Computational and Graphical Statistics*, 21(3):797–820.

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

Patterson, S. and Teh, Y.-W. (2013). Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems (NIPS)*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Proc. Artificial Intelligence and Statistics (AISTATS)*.

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

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

Saul, L. and Jordan, M. (1996). Exploiting tractable substructures in intractable networks. *Neural Information Processing Systems*.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational bayes for non-conjugate inference. In *International Conference on Machine Learning*.

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

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM New York, NY, USA.

Wang, C. and Blei, D. (2012). Truncation-free stochastic variational inference for Bayesian nonparametric models. In *Neural Information Processing*.

Welling, M. and Teh, Y. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*.