Stochastic Structured Variational Inference—Supplement

Matthew D. Hoffman
Adobe Research

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

1 Derivatives of Quantile Functions

One definition of the quantile function is as the inverse of the cumulative distribution function (CDF) $Q_k(\beta_k, \lambda) = q(\beta_k < \beta_k)$. Writing down the definition of an inverse function and Differentiating both sides of this definition shows that

$$\begin{aligned} R_k(Q_k(\beta_k, \lambda), \lambda) &= \beta_k \\ \tfrac{\partial R_k}{\partial u_k}|_{Q_k(\beta_k,\lambda),\lambda} \tfrac{\partial Q_k}{\partial \lambda}|_{\beta_k,\lambda} + \tfrac{\partial R_k}{\partial \lambda}|_{Q_k(\beta_k,\lambda),\lambda} &= 0 \\ \tfrac{\partial R_k}{\partial \lambda}|_{Q_k(\beta_k,\lambda),\lambda} &= -(\tfrac{\partial Q_k}{\partial \beta_k}|_{\beta_k,\lambda})^{-1} \tfrac{\partial Q_k}{\partial \lambda}|_{\beta_k,\lambda} = -q(\beta_k)^{-1} \tfrac{\partial Q_k}{\partial \lambda}|_{\beta_k,\lambda}, \end{aligned} \tag{1}$$

where we use the identities that the derivative of a function's inverse is one over the derivative of that function and that the derivative of a CDF with respect to the random variable is the corresponding probability distribution function (PDF). The derivative of Q_k with respect to λ_k can be obtained numerically using finite differences or automatic differentiation. (The same is true of R_k, but CDFs are often much cheaper to compute than quantile functions.) For multivariate distributions defined as in equation 7 (main text) we can compute $\frac{\partial R}{\partial \lambda}$ as

$$\tfrac{\partial R}{\partial \lambda}|_{u,\lambda} = \tfrac{\partial T}{\partial \hat{R}}|_{\hat{R}(u,\lambda)} \tfrac{\partial \hat{R}}{\partial \lambda}|_{u,\lambda}; \qquad \tfrac{\partial \hat{R}_k}{\partial \lambda} = -\hat{q}_k(\hat{R}(u_k, \lambda))^{-1} \tfrac{\partial \hat{Q}_k}{\partial \lambda}, \tag{2}$$

where $\hat{q}_k$ is the PDF of the kth random variable obtained via the kth univariate quantile function $\hat{R}_k$ and $\hat{Q}_k$ is the CDF that is the inverse of $\hat{R}_k$.

2 SSVI for Latent Dirichlet Allocation

In this section we demonstrate how to use SSVI to do approximate posterior inference on the popular topic model latent Dirichlet allocation (LDA) (Blei et al., 2003). LDA is a generative model of text that assumes that the words in a corpus of documents are generated according to the process

$$\begin{aligned} \beta_k &\sim \text{Dirichlet}(\eta, \ldots, \eta); \quad \theta_d \sim \text{Dirichlet}(\alpha, \ldots, \alpha); \\ z_{d,n} &\sim \text{Multinomial}(\theta_d); \quad w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}}), \end{aligned} \tag{3}$$

where $w_{n,m} \in \{1, \ldots, V\}$ is the index into the vocabulary of the mth word in the nth document, $z_{n,m} \in \{1, \ldots, K\}$ indicates which topic is responsible for $w_{n,m}$, $\theta_{n,k}$ is the prior probability of a word in document n coming from topic k, and $\beta_{k,v}$ is the probability of drawing the word index v from topic k. For simplicity we use symmetric Dirichlet priors.

LDA fits into the SSVI framework; the random variables β, θ, z, and w can be broken into global variables (β) and N sets of local variables (θ_n, z_n, and w_n) that are conditionally independent given the global variables, and the posterior over β given w, z, and θ is in the same tractable exponential family as the prior (i.e., a Dirichlet):

$$p(\beta|w, z, \theta) = \textstyle\prod_k \text{Dirichlet}(\beta_k; \eta + c_k); \quad c_{k,v} \equiv \sum_{n,m} \mathbb{I}[w_{n,m} = v]\mathbb{I}[z_{n,m} = k], \tag{4}$$

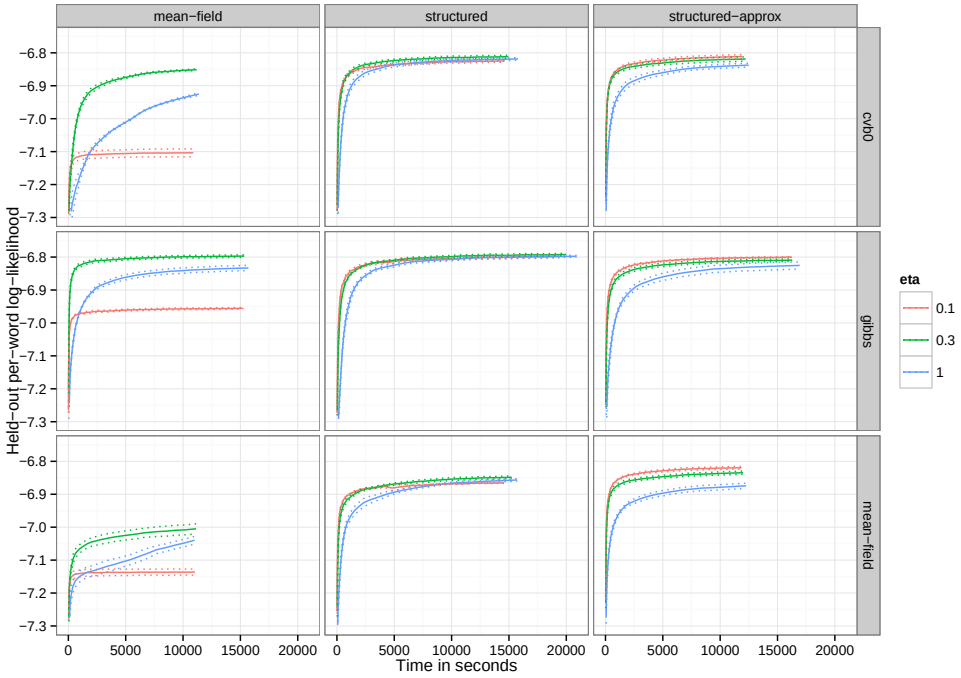

Figure 1: Predictive accuracy for various algorithms as a function of wallclock time when fitting LDA to 3.8 million Wikipedia articles. Each algorithm ran two sweeps over the dataset. Solid lines show average performance across five runs, dotted lines are drawn one standard deviation above and below the mean. The algorithms used to update the global parameters and local conditional distributions vary horizontally and vertically, respectively.

where $c_{k,v}$ counts the number of times that the word v is associated with topic k. Our goal will be to approximate the marginal posterior $p(\beta|w) \propto \int_{\theta,z} p(w,z,\theta,\beta)dzd\theta$ with a product of Dirichlet distributions $q(\beta) = \prod_k \text{Dirichlet}(\beta_k;\lambda_k)$. Algorithm 1 requires that we be able to sample from q_β by inversion, but the Dirichlet distribution lacks a well-defined quantile function. However, a Dirichlet random variable can be constructed from a set of independent gamma random variables:

$$\beta_k \sim \text{Dirichlet}(\lambda_{k,1},\ldots,\lambda_{k,v}) \Leftrightarrow \beta'_{k,v} \sim \text{Gamma}(\lambda_{k,v},1); \beta_{k,v} = \frac{\beta'_{k,v}}{\sum_i \beta'_{k,i}}. \tag{5}$$

So we could sample from q_β by sampling KV independent uniform random variables $u_{k,v}$, passing each through the gamma quantile function $\hat{R}$ to get $\beta'_{k,v} \equiv \hat{R}(u_{k,v},\lambda_{k,v},1)$, and letting $\beta_{k,v} = \beta'_{k,v}/\sum_i \beta'_{k,i}$ so that we have $R(u_{k,v},\lambda_{k,v}) \equiv \hat{R}(u_{k,v},\lambda_{k,v},1)/\sum_i \hat{R}(u_{k,i},\lambda_{k,i},1)$.

To compute the update for λ in algorithm 1 we need to know $(\frac{\partial^2 A}{\partial\lambda\partial\lambda^\top})^{-1}(\frac{\partial t}{\partial\beta}|_{\beta^{(t)}} \frac{\partial R}{\partial\lambda}|_{u^{(t)},\lambda^{(t)}})^\top \eta_n(w_{n^{(t)}},z^{(t)})$. Since each β_k is independent of all of the other topic vectors under q, we need only consider a single β_k at a time. The sufficient statistic vector for the Dirichlet distribution is $t(\beta_k) = \log\beta_k$, so we have

$$\frac{\partial t}{\partial\beta_k}|_{\beta_k^{(t)}} \frac{\partial R}{\partial\lambda_k}|_{u_k^{(t)},\lambda_k^{(t)}} = \text{diag}(\beta_k)^{-1}(\textstyle\sum_v \beta'_{k,v})^{-1}(I - \beta 1^\top)\frac{\partial\hat{R}}{\partial\lambda_k}|_{u_k,\lambda_k}. \tag{6}$$

$\frac{\partial\hat{R}}{\partial\lambda_k}$ can be evaluated using equation 1. $\eta_n(w,z)$ is simply a matrix counting how many times each unique word is associated with each topic: $\eta_n(w,z)_{k,v} = \sum_m \mathbb{I}[w_m = v]\mathbb{I}[z_m = k]$. Finally, the log-normalizer for the Dirichlet is $A(\lambda_k) = -\log\Gamma(\sum_v \lambda_{k,v}) + \sum_v \log\Gamma(\lambda_{k,v})$, and the Fisher matrix $\frac{\partial^2 A}{\partial\lambda_k\partial\lambda_k^\top}$ is a diagonal matrix plus a rank-one matrix: $\frac{\partial^2 A}{\partial\lambda_k\partial\lambda_k^\top} = \text{diag}(\Psi'(\lambda_k)) - \Psi'(\sum_v \lambda_{k,v})11^\top$, where 1 is a column vector of ones and Ψ' is the second derivative of the logarithm of the gamma function. The product of the inverse of the Fisher matrix and a vector can therefore be computed in $O(V)$ time using the matrix inversion lemma (Minka, 2000).

We now have everything we need to apply algorithm 1 to LDA.

3 Full Matrix of LDA Results

We tested various combinations of E-steps and M-steps for latent Dirichlet allocation with 100 topics on the 3,800,000-document Wikipedia dataset from (Hoffman et al., 2013). To update the global variational distributions, we used traditional mean-field updates, SSVI updates, and SSVI-A updates. For the local variational distributions, we used the traditional mean-field approximation (Blei et al., 2003), the CVB0 algorithm of Asuncion et al. (2009), and Gibbs sampling as in (Mimno et al., 2012). 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 mini-batches of 1000 documents and a step size schedule $\rho^{(t)} = t^{-0.75}$.

Figure 1 summarizes the results for $\alpha = 0.1$, which yielded the best results for all variational algorithms. Using traditional mean-field inference (bottom row) to approximate $p(z_n|y_n, \beta)$ degrades performance, but the CVB0 approximation (top row) works almost as well as Gibbs sampling (middle row) for the two SSVI algorithms. CVB0 is outperformed by Gibbs when using the mean-field M-step. The two SSVI algorithms perform comparably well, but the mean-field M-step (left column) is very sensitive to hyperparameter selection compared to SSVI and SSVI-A.

References

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

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

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

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

Minka, T. (2000). Estimating a Dirichlet distribution. Technical report, M.I.T.