Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation

Indraneel Mukherjee **David M. Blei**

Department of Computer Science
Princeton University
35 Olden Street
Princeton, NJ 08540
{imukherj,blei}@cs.princeton.edu

Appendix

Proof of Lemma 1: By choice q minimizes the collapsed variational free energy, so that by equation (2) of our paper,

$$\texttt{CVB}(\vec{x}) = \mathcal{F}(\vec{x}, q) = \mathbb{E}_{q(\vec{z})} \log \frac{q(\vec{z})}{p(\vec{z}, \vec{x})}. \tag{1}$$

From a result due to [1], [2],[3], the choice of parameters $\vec{\gamma}$ that minimize the variational free energy $\mathcal{F}(\vec{x}, q, \vec{\gamma})$ (defined in Section 2 of the paper) are given by

$$\gamma_j \quad \triangleq \quad \sum_i q_i(z), \forall j \in [k]. \tag{2}$$

If VB chose variational parameters as $\phi = q$ and $\vec{\gamma}$ as in (2), then it approximates the posterior $p(\vec{z}, \theta|\vec{x})$ by $q(\vec{z})q(\theta)$ where $q(\theta)$ is the Dirichlet prior with parameters $\vec{\gamma}$, and maybe written as

$$q(\theta) \triangleq \frac{1}{B(\vec{\gamma} + \vec{\alpha})} \prod_z \theta_z^{\gamma_z + \alpha_z - 1}, \tag{3}$$

Here $B(\vec{\gamma} + \vec{\alpha})$ is the normalization constant. In general, for Dirichlet parameters $\vec{\nu}$, $B(\vec{\nu})$ is

$$B(\vec{\nu}) = \frac{\prod_j \Gamma(\nu_j)}{\Gamma(\sum_j \nu_j)}. \tag{4}$$

Since VB chooses variational parameters to minimize its free energy, we have by equation (1) of our paper,

$$\texttt{VB}(\vec{x}) \leq \mathcal{F}(\vec{x}, q, \vec{\gamma}) = \mathbb{E}_{q(\vec{z})q(\theta)} \log \frac{q(\vec{z})q(\theta)}{p(\vec{z}, \vec{x}, \theta)}.$$

Expanding the above expression by the chain rule for computing relative entropy [4],

$$\texttt{VB}(\vec{x}) \leq \mathbb{E}_{q(\vec{z})} \log \frac{q(\vec{z})}{p(\vec{z}, \vec{x})} + \mathbb{E}_{q(\vec{z})} \mathbb{E}_{q(\theta)} \log \frac{q(\theta)}{p(\theta|\vec{z}, \vec{x})}$$

Combining (1) with the above

$$\texttt{VB}(\vec{x}) - \texttt{CVB}(\vec{x}) \leq \mathbb{E}_{q(\vec{z})} \mathbb{E}_{q(\theta)} \log \frac{q(\theta)}{p(\theta|\vec{z}, \vec{x})}. \tag{5}$$

We will bound the right side of (5). From the conjugacy of the multinomial and Dirichlet distributions, we know that $p(\theta|\vec{z}, \vec{x})$ is given by

$$p(\theta|\vec{z}, \vec{x}) = \frac{1}{B(\vec{m} + \vec{\alpha})} \prod_j \theta_j^{m_j + \alpha_j - 1},$$

where $m_j = m_j(\vec{z})$ denotes the number of occurrences of topic j in the collection of topics $\vec{z}$. We denote the collection $(m_1, \ldots, m_k)$ by the vector $\vec{m}$. Plugging the above expression for $p(\theta|\vec{z}, \vec{x})$, and (3) for $q(\theta)$, into the right side of (5),

$$\mathbb{E}_{q(\vec{z})}\mathbb{E}_{q(\theta)} \log \frac{q(\theta)}{p(\theta|\vec{x}, \vec{z})} = \mathbb{E}_{q(\vec{z})}\left[\log \frac{B(\vec{m}+\vec{\alpha})}{B(\vec{\gamma}+\vec{\alpha})}\right] + \mathbb{E}_{q(\vec{z})}\left[\sum_j (\gamma_j - m_j)\left(\mathbb{E}_{q(\theta)} \log \theta_z\right)\right].$$

We will show that

$$\mathbb{E}_{q(\vec{z})}\left[\sum_z (\gamma_z - m_z)\left(\mathbb{E}_{q(\theta)} \log \theta_z\right)\right] = 0 \tag{6}$$

and

$$\mathbb{E}_{q(\vec{z})}\left[\log \frac{B(\vec{m}+\vec{\alpha})}{B(\vec{\gamma}+\vec{\alpha})}\right] = \sum_j \left(\mathbb{E}_{q(\vec{z})}[\log\Gamma(m_j+\alpha_j)] - \log\Gamma(\gamma_j+\alpha_j)\right), \tag{7}$$

which will imply

$$\mathbb{E}_{q(\vec{z})}\mathbb{E}_{q(\theta)}\left[\log \frac{q(\theta)}{p(\theta|\vec{x}, \vec{z})}\right] = \sum_j \left(\mathbb{E}_{q(\vec{z})}[\log\Gamma(m_j+\alpha_j)] - \log\Gamma(\gamma_j+\alpha_j)\right),$$

and complete the proof, by (5).

To see (6), observe that $m_z = \sum_i \mathbf{1}[z_i = z]$ so that $\mathbb{E}_{q(\vec{z})} m_z = \sum_i q_i(z) = \gamma_z$. Hence, when we take expectation under $q(\vec{z})$, each summand disappears. By linearity of expectation, the entire sum is zero.

$$\begin{aligned}
&\mathbb{E}_{q(\vec{z})}[(\gamma_z - m_z)\mathbb{E}_{q(\theta)} \log \theta_z] \\
&\quad= \left(\mathbb{E}_{q(\theta)} \log \theta_z\right)\left(\mathbb{E}_{q(\vec{z})}(\gamma_z - m_z)\right) \\
&\quad= \left(\mathbb{E}_{q(\theta)} \log \theta_z\right)\left(\gamma_z - \mathbb{E}_{q(\vec{z})} m_z\right) \\
&\quad= 0.
\end{aligned}$$

To show (7), we first use (4) to evaluate

$$\begin{aligned}
\log \frac{B(\vec{m}+\vec{\alpha})}{B(\vec{\gamma}+\vec{\alpha})} &= \sum_j (\log\Gamma(m_j+\alpha_j) - \log\Gamma(\gamma_j+\alpha_j)) \\
&+ \log\Gamma\left(\sum_j \gamma_j + \sum_j \alpha_j\right) - \log\Gamma\left(\sum_j m_j + \sum_j \alpha_j\right).
\end{aligned}$$

From the definitions of γ_j (2), and m_j, we have $\sum_j \gamma_j = \sum_j m_j = m$ so that the last two terms in the above expression disappear. Taking expectations over $q(\vec{z})$ now yields (7). $\square$

Proof of Lemma 2: Let $F(q_1, \ldots, q_m)$ denote the expectation $\mathbb{E}[f(X_1 + \ldots + X_m)]$. Using iterated expectation we may rewrite

$$\begin{aligned}
F(q_1, \ldots, q_m) &= \sum_{X_1, X_2} \Pr(X_1, X_2)\mathbb{E}[f(X_1 + \ldots + X_m)|X_1, X_2] \\
&= \sum_{x_1, x_2} \Pr(X_1 = x_1, X_2 = x_2)\mathbb{E}[f(X_3 + \ldots + X_m)].
\end{aligned}$$

By independence, $\Pr(X_1 = x_1, X_2 = x_2) = \Pr(X_1 = x_1)\Pr(X_2 = x_2)$. Also, each random variable X_i takes $0, 1$ values, and the probabilities are given by $\Pr(X_i = 1) = q_i, \Pr(X_i = 0) = 1 - q_i$. Define random variable $Y = X_3 + \ldots + X_m$. Using these facts in the previous equation,

$$\begin{aligned}
F(q_1, \ldots, q_m) &= (1-q_1)(1-q_2)\mathbb{E}[f(Y)] + q_1 q_2 \mathbb{E}[f(2+Y)] \\
&\quad +\mathbb{E}[f(1+Y)](q_1(1-q_2) + q_2(1-q_1)) \\
&= (q_1+q_2)(\mathbb{E}[f(1+Y)] - \mathbb{E}[f(Y)]) \\
&\quad +q_1 q_2(\mathbb{E}[f(Y)] + \mathbb{E}[f(2+Y)] - 2\mathbb{E}[f(1+Y)]).
\end{aligned}$$

Fix $q_3, \ldots, q_m$. Since the q_i's sum to a fixed value, fixing implies $q_1 + q_2$ is a constant. Note $\mathbb{E}[f(Y)], \mathbb{E}[f(Y+1)], \mathbb{E}[f(Y+2)]$ are constants independent of q_1, q_2. Maximizing F is now equivalent to maximizing the second term of the right side of the last equation. By linearity of expectation, $\mathbb{E}[f(Y)] + \mathbb{E}[f(2+Y)] - 2\mathbb{E}[f(1+Y)] = \mathbb{E}[f(Y) + f(2+Y) - 2f(1+Y)]$. Since f is convex, the previous term is non-negative. Thus we need to maximize $q_1 q_2$ under the constraint that their sum is fixed. The optimum choice is $q_1 = q_2$.

Starting from a choice of $q_1, \ldots, q_m$ that maximizes F, we may, by our arguments, set the minimum and maximum of the q_i's, say q_1 and q_2, to a common value $(q_1+q_2)/2$, without decreasing F. This decreases the potential $\Phi(q_1, \ldots, q_m) \triangleq \sum_{i,j} |q_i - q_j|$ of the optimal solution by a factor of at least $(1 - \frac{1}{m^2})$. By repeating this process, we can find optimal solutions with arbitrarily small potential. Continuity of F now implies a solution with zero potential is optimal. We end by observing that zero potential is achieved only by $q_1 = \ldots = q_m = \frac{\gamma}{m}$.

□

Proof of Lemma 3: Assume without loss of generality $q \neq 0$. Let $\mu = mq$ be the mean. Define

$$f(c) \triangleq \mathbb{E}\left[\log\Gamma(X) | (X-\mu) \in [-c\sqrt{m}, c\sqrt{m}]\right].$$

Using the following concentration bound

$$\Pr[|X-\mu| > r] < 2e^{-r^2/2m} \tag{8}$$

we have

$$\mathbb{E}\left[\log\Gamma(X+a)\right] \leq f(c) + 2e^{-c^2/2}\log\Gamma(m+a). \tag{9}$$

Now

$$f(c) = \log\Gamma(\mu+a) + \sum_{i=1}^{c\sqrt{m}} \{\Pr(\mu-i)\log\Gamma(\mu-i+a) + \Pr(\mu+i)\log\Gamma(\mu+i+a)\}. \tag{10}$$

We will first obtain bounds on each summand term. Using $\Gamma(x+1) = x\Gamma(x)$, we get

$$\begin{aligned}\log\Gamma(\mu+a+i) &= \log\Gamma(\mu+a) + \sum_{r=0}^{i-1}\log(\mu+a+r) \\ &= \log\Gamma(\mu+a) + i\log(\mu+a) + \sum_{r=0}^{i-1}\log(1+\frac{r}{\mu+a}).\end{aligned}$$

From $1 + x \leq \exp(x)$, we may upper-bound the last summation by

$$\sum_{r=0}^{i-1}\frac{r}{\mu+a} \leq \frac{i^2}{2(\mu+a)}.$$

Therefore we get

$$\log\Gamma(\mu+a+i) \leq \log\Gamma(\mu+a) + i\log(\mu+a) + \frac{i^2}{2(\mu+a)}. \tag{11}$$

Similarly, we can get

$$\log\Gamma(\mu+a-i) = \log\Gamma(\mu+a) - i\log(\mu+a) - \sum_{r=1}^{i}\log(1-\frac{r}{\mu+a}).$$

This time we will use $-\log(1-x) \leq \log(1+2x) \leq 2x$; but this only holds in the range $x \in [0, \frac{1}{2})$. Since in our case $x \leq \frac{c\sqrt{m}}{mq+a}$, for this to be applicable it suffices if m is at least $\frac{4c^2}{q^2}$. We can now bound

$$\log\Gamma(\mu+a-i) \leq \log\Gamma(\mu+a) - i\log(\mu+a) + \frac{i^2}{\mu+a} + \frac{i}{\mu+a}. \tag{12}$$

Using (11) and (12), we can upper-bound each summand in (10) by

$$\begin{aligned}
&\log\Gamma(\mu+a)\left\{\Pr(\mu+i)+\Pr(\mu-i)\right\} \\
&+\quad \log(\mu+a)\left\{\Pr(\mu+i)i-\Pr(\mu-i)i\right\}+\frac{3i^2+i}{\mu+a}\left\{\Pr(\mu+i)+\Pr(\mu-i)\right\}.
\end{aligned}$$

Summing up the first term over i we get at most $\log\Gamma(\mu+a)$. The second term becomes at most $\log(\mu+a)m\Pr(|X-\mu| > c\sqrt{m})$ since the mean is μ. Finally, the third term is at most three times the sum of the variance, $\mu(1-q)$, and a term smaller than the variance, divided by $\mu+a$. Combining, and using the concentration bound in (8) we get

$$f(c) \le \log\Gamma(\mu+a)+O(1-q)+\frac{c}{\sqrt{m}}+2e^{-c^2/2}m\log m.$$

Choosing $c=2\sqrt{\log m}$ and plugging into (9), we get

$$\mathbb{E}\left[\log\Gamma(X+a)\right] \le \log\Gamma(\mu+a)+O(1-q)+o(1)$$

with $o(1)=O(\sqrt{\frac{\log m}{m}})$. With this choice of c, the required lower bound on m is $1/q^{2+o(1)}$.

$\square$

References

[1] M. Beal. *Variational Algorithms for approximate Bayesian inference*. PhD thesis, University College London, 2003.

[2] 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.

[3] C. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS 15*, pages 777–784. MIT Press, Cambridge, MA, 2003.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, August 1991.