Two Useful Bounds for Variational Inference

John Paisley

Department of Computer Science
Princeton University, Princeton, NJ

jpaisley@princeton.edu

Abstract

We review and derive two lower bounds on the expectation of the log-sum function in the context of variational inference. The first bound relies on the first-order Taylor expansion about an auxiliary parameter. The second bound relies on an auxiliary probability distribution. We show how these auxiliary parameters can be removed entirely from the model. We then discuss the advantage of keeping these parameters in certain cases, giving an example of a likelihood/prior pair for each case.

1 Introduction

Variational inference methods (Wainwright and Jordan, 2008) are useful for approximate posterior inference in hierarchical models. They also provide an approximate value of the log-evidence of a model. For example, let $\mathbf{Y} = Y_1, \dots, Y_n$ be a set of observed data and $\mathbf{X} = X_1, \dots, X_K$ a set of latent parameters. Let $p(\mathbf{Y}, \mathbf{X})$ be the joint likelihood of a given model and $Q(\mathbf{X})$ be any probability distribution. Then by Jensen's inequality,

$$\begin{aligned} \ln \int_{\Omega_{\mathbf{X}}} p(\mathbf{Y}|\mathbf{X}) p(\mathbf{X}) d\mathbf{X} &= \ln \int_{\Omega_{\mathbf{X}}} \frac{p(\mathbf{Y}, \mathbf{X})}{Q(\mathbf{X})} Q(\mathbf{X}) d\mathbf{X} \\ &\geq \int_{\Omega_{\mathbf{X}}} Q(\mathbf{X}) \ln \left\{ \frac{p(\mathbf{Y}, \mathbf{X})}{Q(\mathbf{X})} \right\} d\mathbf{X} \qquad (1) \end{aligned}$$

In variational methods, this Q distribution is given a factorized form, $Q(\mathbf{X}) := \prod_{k=1}^{K} q(X_k)$, and is predefined. It can be shown that maximizing this lower bound on the log-evidence with respect to $Q(\mathbf{X})$ is equivalent to minimizing the Kullback-Leibler divergence between this Q distribution and the true posterior $p(\mathbf{X}|\mathbf{Y})$. Therefore, the factorized Q distribution can be interpreted as an approximation to the true posterior.

It is often the case that an integral in (1) is intractable, meaning that the expression

$$\int_{\Omega_{X_{k \in \mathcal{I}}}} \prod_{k \in \mathcal{I}} Q(X_k) \ln p(\mathbf{Y}|X_{k \in \mathcal{I}}) dX_{k \in \mathcal{I}} \qquad (2)$$

cannot be analytically solved for some set of parameters indexed by $k \in \mathcal{I}$. In this case, approximating functions can be used to lower bound the terms of interest and provide analytical tractability. In this technical report, we review two simple lower bounds that are often useful for variational inference.

2 Two Lower Bounds Without Auxiliary Parameters

In this section, we present the two bounds of interest in this technical report and show how the auxiliary parameters used to derive each bound can be set such that they disappear from the model. The result is a function that depends only on the original parameters of interest and also represents the tightest lower bound on the original function given the functional form of the selected lower bound.

2.1 A lower bound on $-\mathbb{E}_Q\left[\ln \sum_{k=1}^K X_k\right]$

The tightest lower bound on $-\mathbb{E}_Q\left[\ln \sum_{k=1}^K X_k\right]$ using a first-order Taylor expansion is

$$-\mathbb{E}_Q\left[\ln \sum_{k=1}^K X_k\right] \geq -\ln \sum_{k=1}^K \mathbb{E}_Q\left[X_k\right] \tag{3}$$

Derivation: The function $-\ln(\cdot)$ is convex. Therefore, a first-order Taylor expansion about the point $\omega > 0$ produces the inequality

$$-\mathbb{E}_Q\left[\ln \sum_{k=1}^K X_k\right] \geq -\ln \omega - \frac{\sum_k \mathbb{E}_Q[X_k] - \omega}{\omega} \tag{4}$$

Taking the derivative with respect to ω and setting to zero will give the value of ω that maximizes this function for a given value of $\sum_k \mathbb{E}_Q[X_k]$. The derivative of this function is

$$\frac{df}{d\omega} = -\frac{1}{\omega} + \frac{1}{\omega^2}\sum_{k=1}^K \mathbb{E}_Q[X_k] \tag{5}$$

Setting to zero and solving for ω gives $\omega = \sum_k \mathbb{E}_Q[X_k]$. Plugging this value back into (4) produces (3). Therefore, the parameter ω can be removed and all instances of $-\mathbb{E}_Q\left[\ln \sum_{k=1}^K X_k\right]$ can be replaced with $-\ln \sum_{k=1}^K \mathbb{E}_Q\left[X_k\right]$.

2.2 A lower bound on $\mathbb{E}_Q\left[\ln\sum_{k=1}^K X_k\right]$

The tightest lower bound on $\mathbb{E}_Q\left[\ln\sum_k X_k\right]$ using an auxiliary probability distribution is

$$\mathbb{E}_Q\left[\ln\sum_{k=1}^K X_k\right] \geq \ln\sum_{k=1}^K \mathrm{e}^{\mathbb{E}_Q[\ln X_k]} \tag{6}$$

Derivation: The function $\ln(\cdot)$ is concave. Therefore, using an auxiliary probability vector, $(p_1,\ldots,p_K)$, where $p_k > 0$ and $\sum_k p_k = 1$, it follows from Jensen's inequality that

$$\begin{aligned}\mathbb{E}_Q\left[\ln\sum_{k=1}^K X_k\right] &= \mathbb{E}_Q\left[\ln\sum_{k=1}^K p_k\frac{X_k}{p_k}\right]\\ &\geq \sum_{k=1}^K p_k\mathbb{E}_Q\left[\ln\frac{X_k}{p_k}\right]\\ &= \sum_{k=1}^K p_k\mathbb{E}_Q\left[\ln X_k\right] - \sum_{k=1}^K p_k\ln p_k \qquad (7)\end{aligned}$$

Taking the derivative with respect to p_k, finding it's proportionality and normalizing produces

$$p_k = \frac{\mathrm{e}^{\mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln X_j]}} \tag{8}$$

Replacing $p_1,\ldots,p_K$ in (7) with these values gives

$$\mathbb{E}_Q\left[\ln\sum_{k=1}^K X_k\right] \geq \sum_{k=1}^K \frac{\mathrm{e}^{\mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln X_j]}}\mathbb{E}_Q[\ln X_k] - \sum_{k=1}^K \frac{\mathrm{e}^{\mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln X_j]}}\ln\frac{\mathrm{e}^{\mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln X_j]}} \tag{9}$$

The rightmost term can be expanded as

$$\sum_{k=1}^K \frac{\mathrm{e}^{\mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln X_j]}}\mathbb{E}_Q[\ln X_k] - \sum_{k=1}^K \frac{\mathrm{e}^{\mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln X_j]}}\ln\sum_{j=1}^K \mathrm{e}^{\mathbb{E}_Q[\ln X_j]} \tag{10}$$

The first term in (10) cancels the first term on the right of the inequality in (9) to give

$$\mathbb{E}_Q\left[\ln\sum_{k=1}^K X_k\right] \geq \sum_{k=1}^K \frac{\mathrm{e}^{\mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln X_j]}}\ln\sum_{j=1}^K \mathrm{e}^{\mathbb{E}_Q[\ln X_j]} \tag{11}$$

The logarithm on the right can be brought outside of the summation, which then sums to one. This produces (6). As with the inequality of the previous section, the auxiliary parameters do not remain in the final bound; all instances of $\mathbb{E}_Q\left[\ln\sum_{k=1}^K X_k\right]$ can simply be replaced with $\ln\sum_{k=1}^K \mathrm{e}^{\mathbb{E}_Q[\ln X_k]}$.

3 The Advantage of Retaining Auxiliary Parameters

Given certain likelihood/prior combinations, retaining the auxiliary parameters ω and $(p_1, \ldots, p_K)$ may significantly simplify model learning. This simplification comes in the form of analytical posterior updates for certain variational parameters, while using the tightest lower bounds given in Section 2 require gradient methods.

Gradient methods introduce potential issues: steepest ascent vs. Newton-type methods; setting vs. learning step sizes; checking for feasibility; assessing convergence vs. fixing the number of gradient steps when updating a parameter. In most of these concerns, there is usually a trade-off between accuracy and computation time. Analytical updates, on the other hand, are fast and move to the optimal value – i.e., the value that locally maximizes the lower bound – in one step.

In this section, we discuss two cases where keeping the auxiliary parameters ω and $(p_1, \ldots, p_K)$ can be beneficial. In both cases, the lower bound approximation results in analytical updates when the variational q distribution is chosen to be of the same form as the prior. Therefore, the effect of each approximation can be viewed as altering the likelihood such that it is conjugate to the prior.

3.1 Working with $-\mathbb{E}_Q\left[\ln \sum_{k=1}^K X_k\right] \geq -\ln\omega - \frac{\sum_k \mathbb{E}_Q[X_k]-\omega}{\omega}$

Consider the following hierarchical process.

$$Z_n \overset{iid}{\sim} \sum_{k=1}^{K} \frac{X_k}{\sum_j X_j}\delta_k, \qquad X_k \overset{ind}{\sim} \text{Gamma}(a_k, b_k) \tag{12}$$

This process could appear in a mixture modeling setting, where $X_1, \ldots, X_K$ are random variables that construct a discrete probability distribution and Z_n is a random index value that selects parameters, θ_{Z_n}, from a set. This would be followed by an observation, $Y_n \sim F(\theta_{Z_n})$. When $b_k = b_{k'}$ for all $k \neq k'$, this is equivalent to a Dirichlet prior. If $K \to \infty$ and $a_1, a_2, \ldots$ are scaled weights from a Dirichlet process (Ferguson, 1973; Sethuraman, 1994), this is the hierarchical Dirichlet process (Teh et al., 2006). Furthermore, when the values $b_1, b_2, \ldots$ are log-normal, this the infinite logistic normal distribution (Paisley et al., 2010). For some of these models, we note that other representations are available that do not require the approximation discussed in this section.

The posterior of $X_{1:K}$ in this model is proportional to

$$p(X_{1:K}|Z_{1:N}, a_{1:K}, b_{1:K}) \propto \left[\prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{X_k}{\sum_j X_j}\right)^{\mathbb{I}(Z_n=k)}\right]\left[\prod_{k=1}^{K} X_k^{a_k-1}\mathrm{e}^{-b_kX_k}\right] \tag{13}$$

Under a factorized Q distribution, the variational lower bound at nodes $X_1, \ldots, X_K$ is

$$\begin{aligned}\mathbb{E}_Q[\ln p(X_{1:K}|Z_{1:N}, a_{1:K}, b_{1:K})] + \mathbb{H}[Q] \;\;\propto\;\; & \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{P}_Q(Z_n = k)\mathbb{E}_Q[\ln X_k] - N\mathbb{E}_Q\left[\ln \sum_{k=1}^{K} X_k\right] \\ & + \sum_{k=1}^{K}(\mathbb{E}_Q[a_k] - 1)\mathbb{E}_Q[\ln X_k] - \sum_{k=1}^{K}\mathbb{E}_Q[b_k]\mathbb{E}_Q[X_k] \\ & + \sum_{k=1}^{K}\mathbb{H}[Q(X_k)] \end{aligned} \qquad (14)$$

Analytical calculation of all expectations in (14) is not possible because of the term $-N\mathbb{E}_Q[\ln \sum_{k=1}^{K} X_k]$. Using the lower bound given in Section 2.1, this term is replaced with $-N \ln \omega - N(\sum_{k=1}^{K}\mathbb{E}_Q[X_k] - \omega)/\omega$, where $\omega > 0$.

Before learning can proceed for $X_1, \ldots, X_K$, the form of their variational q distributions must be defined; we assume q distributions have been selected for all other parameters as well. As shown by Winn and Bishop (2005), the procedure for finding the optimal form and parameterization of a given q distribution is to exponentiate the variational lower bound with all expectations involving the parameter of interest not taken. When two distributions are conjugate, this will always produce an analytical result that is of the same form as the prior. Following this procedure with respect to the lower bound approximation of (14),

$$\begin{aligned} q(X_k) \;\;&\propto\;\; \mathrm{e}^{\mathbb{E}_{Q-X_k}[\ln p(X_k|Z_{1:N}, a_{1:K}, b_{1:K})]} \\ &\propto\;\; X_k^{\mathbb{E}_Q[a_k] + \sum_{n=1}^{N}\mathbb{P}_Q(Z_n = k) - 1}\;\mathrm{e}^{-(\mathbb{E}_Q[b_k] + N/\omega)X_k} \end{aligned} \qquad (15)$$

Therefore, the optimal q distribution for X_k is $q(X_k) = \text{Gamma}(X_k|a'_k, b'_k)$, with

$$a'_k \;=\; \mathbb{E}_Q[a_k] + \sum_{n=1}^{N}\mathbb{P}_Q(Z_n = k) \qquad (16)$$

$$b'_k \;=\; \mathbb{E}_Q[b_k] + N/\omega \qquad (17)$$

The impact of the Taylor approximation is to alter the likelihood, making it conjugate to the prior. As motivated in the introduction, this requires setting the auxiliary parameter, ω. For a given iteration, let $\dot{a}'_k$ and $\dot{b}'_k$ be the variational parameters resulting from the previous iteration. Then the auxiliary parameter ω can be set to

$$\omega = \sum_{k=1}^{K}\dot{a}'_k/\dot{b}'_k \qquad (18)$$

which, as motivated in Section 2.1, creates the tightest lower bound for the approximation. Of course, other options are available as well, such as incrementally updating ω after updating each (a'_k, b'_k), rather than after updating the set of all K.

3.2 Working with $\mathbb{E}_Q\left[\ln \sum_{k=1}^{K} X_k\right] \geq \sum_{k=1}^{K} p_k \mathbb{E}_Q\left[\ln X_k\right] - \sum_{k=1}^{K} p_k \ln p_k$

Consider the following hierarchical process.

$$N_m \overset{ind}{\sim} \text{Poisson}\left(\sum_{k=1}^{K} c_{k,m} X_k\right), \qquad X_k \overset{iid}{\sim} \text{Gamma}(a_k, b_k) \tag{19}$$

An example of where this process arises is in non-negative matrix factorization. For example, this could be part of the generative process of a matrix of non-negative integers, $Y \in \mathbb{N}^{D \times M}$, and N_m would be the value in the m^{th} column of a particular row of Y (not indexed). In this framework, the data could be counts of word occurrence in documents, with $Y_{d,m}$ equal to the number of times word d appears in document m.

The posterior of $X_{1:K}$ in this model is proportional to

$$p(X_{1:K}|\boldsymbol{N}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{b}) \propto \left[\prod_{m=1}^{M}\left(\sum_{k=1}^{K} c_{k,m} X_k\right)^{N_m} \mathrm{e}^{-\sum_{k=1}^{K} c_{k,m} X_k}\right]\left[\prod_{k=1}^{K} X_k^{a_k - 1} \mathrm{e}^{-b_k X_k}\right] \tag{20}$$

Under a factorized Q distribution, the variational lower bound at nodes $X_1, \ldots, X_K$ is

$$\begin{aligned} \mathbb{E}_Q[\ln p(X_{1:K}|\boldsymbol{N}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{b})] + \mathbb{H}[Q] &= \sum_{m=1}^{M} N_m \mathbb{E}_Q\left[\ln \sum_{k=1}^{K} c_{k,m} X_k\right] - \sum_{m=1}^{M}\sum_{k=1}^{K} \mathbb{E}_Q[c_{k,m}]\mathbb{E}_Q[X_k] \\ &\quad + \sum_{k=1}^{K}(\mathbb{E}_Q[a_k] - 1)\mathbb{E}_Q[\ln X_k] - \sum_{k=1}^{K} \mathbb{E}_Q[b_k]\mathbb{E}_Q[X_k] \\ &\quad + \sum_{k=1}^{K} \mathbb{H}[Q(X_k)] \end{aligned} \tag{21}$$

Analytical calculation of all expectations in (21) is not possible because of the term $N_m \mathbb{E}_Q\left[\ln \sum_{k=1}^{K} c_{k,m} X_k\right]$. Using the lower bound given in Section 2.2, this term is replaced with $N_m \sum_{k=1}^{K} p_k^{(m)} \mathbb{E}_Q\left[\ln c_{k,m} X_k\right] - N_m \sum_{k=1}^{K} p_k^{(m)} \ln p_k^{(m)}$, where $\boldsymbol{p}^{(m)} \in \Delta_K$.

As in Section 3.1, q distributions must be defined for $X_1, \ldots, X_K$ before learning can continue. We again follow the procedure outlined in (Winn and Bishop, 2005) for finding these distributions and their parameter settings.

$$\begin{aligned} q(X_k) &\propto \mathrm{e}^{\mathbb{E}_{Q_{-X_k}}[\ln p(X_k|\boldsymbol{N}, \boldsymbol{c}, \boldsymbol{a}, \boldsymbol{b})]} \\ &\propto X_k^{\mathbb{E}_Q[a_k] + \sum_{m=1}^{M} N_m p_k^{(m)} - 1}\, \mathrm{e}^{-(\mathbb{E}_Q[b_k] + \sum_{m=1}^{M} \mathbb{E}_Q[c_{k,m}])X_k} \end{aligned} \tag{22}$$

Therefore, the optimal q distribution for X_k is $q(X_k) = \text{Gamma}(X_k|a'_k, b'_k)$, with

$$a'_k = \mathbb{E}_Q[a_k] + \sum_{m=1}^{M} N_m p_k^{(m)} \tag{23}$$

$$b'_k = \mathbb{E}_Q[b_k] + \sum_{m=1}^{M} \mathbb{E}_Q[c_{k,m}] \tag{24}$$

Again, the impact of the approximation is to modify the likelihood to a form that is conjugate to the prior. This requires values for $p_k^{(m)}$ for $k = 1, \dots, K$ and $m = 1, \dots, M$. Following the same line of thought as in Section 3.1, let $\dot{a}'_k$ and $\dot{b}'_k$ be the variational parameters following a given iteration. Prior to the next iteration, the auxiliary distributions can be set to

$$p_k^{(m)} = \frac{\mathrm{e}^{\mathbb{E}_Q[\ln c_{k,m}] + \mathbb{E}_Q[\ln X_k]}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln c_{j,m}] + \mathbb{E}_Q[\ln X_j]}} = \frac{\mathrm{e}^{\mathbb{E}_Q[\ln c_{k,m}] + \psi(\dot{a}'_k) - \ln \dot{b}'_k}}{\sum_j \mathrm{e}^{\mathbb{E}_Q[\ln c_{j,m}] + \psi(\dot{a}'_j) - \ln \dot{b}'_j}} \tag{25}$$

This tightens the approximation for the current iteration. As in Section 3.1, the vector $\boldsymbol{p}^{(m)}$ can be updated at other points in the inference process, such as after updating each (a'_k, b'_k).

4 Summary

In this technical report, we have reviewed two simple lower bounds on the log-sum function with a focus on variational inference applications. We showed how retaining auxiliary parameters can aid the inference procedure in the form of analytical parameter updates. Hoffman et al. (2010) give another instance of this advantage in a matrix factorization model where the emissions are exponentially distributed. Their model includes the approximation in Section 2.1, and an approximation similar to Section 2.2 that uses an auxiliary probability distribution, but where the concave function is the negative inverse-sum function.

References

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

Hoffman, M., Blei, D. M., and Cook, P. (2010). Bayesian nonparametric matrix factorization for recorded music. In *ICML*.

Paisley, J., Wang, C., and Blei, D. M. (2010). The infinite logistic normal distribution for nonparametric correlated topic modeling. *In Progress*.

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

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

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

Winn, J. and Bishop, C. (2005). Variational message passing. *Journal of Machine Learning Research*, 6:661–694.