Fast Variational Inference in the Conjugate Exponential Family

James Hensman[*]
Department of Computer Science
The University of Sheffield
james.hensman@sheffield.ac.uk

Magnus Rattray
Faculty of Life Science
The University of Manchester
magnus.rattray@manchester.ac.uk

Neil D. Lawrence[*]
Department of Computer Science
The University of Sheffield
n.lawrence@sheffield.ac.uk

Abstract

We present a general method for deriving collapsed variational inference algorithms for probabilistic models in the conjugate exponential family. Our method unifies many existing approaches to collapsed variational inference. Our collapsed variational inference leads to a new lower bound on the marginal likelihood. We exploit the information geometry of the bound to derive much faster optimization methods based on conjugate gradients for these models. Our approach is very general and is easily applied to any model where the mean field update equations have been derived. Empirically we show significant speed-ups for probabilistic inference using our bound.

1 Introduction

Variational bounds provide a convenient approach to approximate inference in a range of intractable models [Ghahramani and Beal, 2001]. Classical variational optimization is achieved through coordinate ascent which can be slow to converge. A popular solution [King and Lawrence, 2006, Teh et al., 2007, Kurihara et al., 2007, Sung et al., 2008, Lázaro-Gredilla and Titsias, 2011, Lázaro-Gredilla et al., 2011] is to marginalize analytically a portion of the variational approximating distribution, removing this from the optimization. In this paper we provide a unifying framework for collapsed inference in the general class of models composed of conjugate-exponential graphs (CEGs).

First we review the body of earlier work with a succinct and unifying derivation of the collapsed bounds. We describe how the applicability of the collapsed bound to any particular CEG can be determined with a simple d-separation test. Standard variational inference via *coordinate ascent* turns out to be *steepest ascent* with a unit step length on our unifying bound. This motivates us to consider natural gradients and conjugate gradients for fast optimization of these models. We apply our unifying approach to a range of models from the literature obtaining, often, an order of magnitude or more increase in convergence speed. Our unifying view allows collapsed variational methods to be integrated into general inference tools like infer.net [Minka et al., 2010].

[*]also at Sheffield Institute for Translational Neuroscience, SITraN

2 The Marginalised Variational Bound

The advantages to marginalising analytically a subset of variables in variational bounds seem to be well understood: several different approaches have been suggested in the context of specific models. In Dirichlet process mixture models Kurihara et al. [2007] proposed a collapsed approach using both truncated stick-breaking and symmetric priors. Sung et al. [2008] proposed 'latent space variational Bayes' where both the cluster-parameters and mixing weights were marginalised, again with some approximations. Teh et al. [2007] proposed a collapsed inference procedure for latent Dirichlet allocation (LDA). In this paper we unify all these results from the perspective of the 'KL corrected bound' [King and Lawrence, 2006]. This lower bound on the model evidence is also an upper bound on the original variational bound, the difference between the two bounds is given by a Kullback Leibler divergence. The approach has also been referred to as the *marginalised variational bound* by Lázaro-Gredilla et al. [2011], Lázaro-Gredilla and Titsias [2011]. The connection between the KL corrected bound and the collapsed bounds is not immediately obvious. The key difference between the frameworks is the order in which the marginalisation and variational approximation are applied. However, for CEGs this order turns out to be irrelevant. Our framework leads to a more succinct derivation of the collapsed approximations. The resulting bound can then be optimised without recourse to approximations in either the bound's evaluation or its optimization.

2.1 Variational Inference

Assume we have a probabilistic model for data, $\mathcal{D}$, given parameters (and/or latent variables), $\mathbf{X}$, $\mathbf{Z}$, of the form $p(\mathcal{D}, \mathbf{X}, \mathbf{Z}) = p(\mathcal{D} \,|\, \mathbf{Z}, \mathbf{X})p(\mathbf{Z} \,|\, \mathbf{X})p(\mathbf{X})$. In variational Bayes (see e.g. Bishop [2006]) we approximate the posterior $p(\mathbf{Z}, \mathbf{X}|\mathcal{D})$ by a distribution $q(\mathbf{Z}, \mathbf{X})$. We use Jensen's inequality to derive a lower bound on the model evidence $\mathcal{L}$, which serves as an objective function in the variational optimisation:

$$p(\mathcal{D}) \geq \mathcal{L} = \int q(\mathbf{Z}, \mathbf{X}) \ln \frac{p(\mathcal{D}, \mathbf{Z}, \mathbf{X})}{q(\mathbf{Z}, \mathbf{X})} \mathrm{d}\mathbf{Z}\, \mathrm{d}\mathbf{X}. \tag{1}$$

For tractability the mean field (MF) approach assumes q factorises across its variables, $q(\mathbf{Z}, \mathbf{X}) = q(\mathbf{Z})q(\mathbf{X})$. It is then possible to implement an optimisation scheme which analytically optimises each factor alternately, with the optimal distribution given by

$$q^\star(\mathbf{X}) \propto \exp\left\{ \int q(\mathbf{Z}) \ln p(\mathcal{D}, \mathbf{X}|\mathbf{Z}) \, \mathrm{d}\mathbf{Z} \right\}, \tag{2}$$

and similarly for $\mathbf{Z}$: these are often referred to as VBE and VBM steps. King and Lawrence [2006] substituted the expression for the optimal distribution (for example $q^\star(\mathbf{X})$) back into the bound (1), eliminating one set of parameters from the optimisation, an approach that has been reused by Lázaro-Gredilla et al. [2011], Lázaro-Gredilla and Titsias [2011]. The resulting bound is not dependent on $q(\mathbf{X})$. King and Lawrence [2006] referred to this new bound as 'the KL corrected bound'. The difference between the bound, which we denote $\mathcal{L}_{\text{KL}}$, and a standard mean field approximation $\mathcal{L}_{\text{MF}}$, is the Kullback Leibler divergence between the optimal form of $q^*(\mathbf{X})$ and the current $q(\mathbf{X})$.

We re-derive their bound by first using Jensen's inequality to construct the variational lower bound on the *conditional* distribution,

$$\ln p(\mathcal{D}|\mathbf{X}) \geq \int q(\mathbf{Z}) \ln \frac{p(\mathcal{D}, \mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \, \mathrm{d}\mathbf{Z} \triangleq \mathcal{L}_1. \tag{3}$$

This object turns out to be of central importance in computing the final KL-corrected bound and also in computing gradients, curvatures and the distribution of the collapsed variables $q^\star(\mathbf{X})$. It is easy to see that it is a function of $\mathbf{X}$ which lower-bounds the log likelihood $p(\mathcal{D} \,|\, \mathbf{X})$, and indeed our derivation treats it as such. We now marginalize the conditioned variable from this expression,

$$\ln p(\mathcal{D}) \geq \ln \int p(\mathbf{X}) \exp\{\mathcal{L}_1\} \, \mathrm{d}\mathbf{X} \triangleq \mathcal{L}_{\text{KL}}, \tag{4}$$

giving us the bound of King and Lawrence [2006] & Lázaro-Gredilla et al. [2011]. Note that one set of parameters was marginalised *after* the variational approximation was made.

Using (2), this expression also provides the approximate posterior for the marginalised variables $\mathbf{X}$:

$$q^\star(\mathbf{X}) = p(\mathbf{X}) e^{\mathcal{L}_1 - \mathcal{L}_{\text{KL}}} \tag{5}$$

and $e^{\mathcal{L}_{\text{KL}}}$ appears as the constant of proportionality in the mean-field update equation (2).

3 Partial Equivalence of the Bounds

We can recover $\mathcal{L}_{\text{MF}}$ from $\mathcal{L}_{\text{KL}}$ by again applying Jensen's inequality,

$$\mathcal{L}_{\text{KL}} = \ln \int q(\mathbf{X}) \frac{p(\mathbf{X})}{q(\mathbf{X})} \exp\{\mathcal{L}_1\} \, \mathrm{d}\mathbf{X} \geq \int q(\mathbf{X}) \ln \left\{ \frac{p(\mathbf{X})}{q(\mathbf{X})} \exp\{\mathcal{L}_1\} \right\} \mathrm{d}\mathbf{X}, \tag{6}$$

which can be re-arranged to give the mean-field bound,

$$\mathcal{L}_{\text{KL}} \geq \int q(\mathbf{X}) q(\mathbf{Z}) \ln \left\{ \frac{p(\mathcal{D}|\mathbf{Z}, \mathbf{X}) p(\mathbf{Z}) p(\mathbf{X})}{q(\mathbf{Z}) q(\mathbf{X})} \right\} \mathrm{d}\mathbf{X} \, \mathrm{d}\mathbf{Z}, \tag{7}$$

and it follows that $\mathcal{L}_{\text{KL}} = \mathcal{L}_{\text{MF}} + \text{KL}(q^*(\mathbf{X})||q(\mathbf{X}))$ and[1] $\mathcal{L}_{\text{KL}} \geq \mathcal{L}_{\text{MF}}$. For a given $q(\mathbf{Z})$, the bounds are equal after $q(\mathbf{X})$ is updated via the mean field method: the approximations are ultimately the same. The advantage of the new bound is to reduce the number of parameters in the optimisation. It is particularly useful when variational parameters are optimised by gradient methods. Since VBEM is equivalent to a steepest descent gradient method with a fixed step size, there appears to be a lot to gain by combining the KLC bound with more sophisticated optimization techniques.

3.1 Gradients

Consider the gradient of the KL corrected bound with respect to the parameters of $q(\mathbf{Z})$:

$$\frac{\partial \mathcal{L}_{\text{KL}}}{\partial \theta_z} = \exp\{-\mathcal{L}_{\text{KL}}\} \frac{\partial}{\partial \theta_z} \int \exp\{\mathcal{L}_1\} p(\mathbf{X}) \, \mathrm{d}\mathbf{X} = \mathbb{E}_{q^\star(\mathbf{X})} \Big[\frac{\partial \mathcal{L}_1}{\partial \theta_z} \Big], \tag{8}$$

where we have used the relation (5). To find the gradient of the mean-field bound we note that it can be written in terms of our conditional bound (3) as $\mathcal{L}_{\text{MF}} = \mathbb{E}_{q(\mathbf{X})} \Big[\mathcal{L}_1 + \ln p(\mathbf{X}) - \ln q(\mathbf{X}) \Big]$ giving

$$\frac{\partial \mathcal{L}_{\text{MF}}}{\partial \theta_z} = \mathbb{E}_{q(\mathbf{X})} \Big[\frac{\partial \mathcal{L}_1}{\partial \theta_z} \Big] \tag{9}$$

thus setting $q(\mathbf{X}) = q^\star(\mathbf{X})$ not only makes the bounds equal, $\mathcal{L}_{\text{MF}} = \mathcal{L}_{\text{KL}}$, but also their *gradients* with respect to θ_Z.

Sato [2001] has shown that the variational update equation can be interpreted as a *gradient* method, where each update is also a step in the steepest direction in the canonical parameters of $q(\mathbf{Z})$. We can combine this important insight with the above result to realize that we have a simple method for computing the gradients of the KL corrected bound: we only need to look at the update expressions for the mean-field method. This result also reveals the weakness of standard variational Bayesian expectation maximization (VBEM): it is a steepest ascent algorithm. Honkela et al. [2010] looked to rectify this weakness by applying a conjugate gradient algorithm to the mean field bound. However, they didn't obtain a significant improvement in convergence speed. Our suggestion is to apply conjugate gradients to the KLC bound. Whilst the value and gradient of the MF bound matches that of the KLC bound after an update of the collapsed variables, the *curvature* is always greater. In practise this means that much larger steps (which we compute using conjugate gradient methods) can be taken when optimizing the KLC bound than for the MF bound leading to more rapid convergence.

3.2 Curvature of the Bounds

King and Lawrence [2006] showed empirically that the KLC bound could lead to faster convergence because the bounds differ in their curvature: the curvature of the KLC bound enables larger steps to be taken by an optimizer. We now derive analytical expressions for the curvature of both bounds. For the mean field bound we have

$$\frac{\partial^2 \mathcal{L}_{\text{MF}}}{\partial \theta_z^2} = \mathbb{E}_{q(\mathbf{X})} \Big[\frac{\partial^2 \mathcal{L}_1}{\partial \theta_z^2} \Big], \tag{10}$$

[1] We use $\text{KL}(\cdot||\cdot)$ to denote the Kullback Leibler divergence between two distributions.

and for the KLC bound, with some manipulation of (4) and using (5):

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_{\text{KL}}}{\partial \theta_z^{[i]} \partial \theta_z^{[j]}} &= e^{-\mathcal{L}_{\text{KL}}} \frac{\partial^2 e^{\mathcal{L}_{\text{KL}}}}{\partial \theta_z^{[i]} \partial \theta_z^{[j]}} - e^{-2\mathcal{L}_{\text{KL}}} \left\{ \frac{\partial e^{\mathcal{L}_{\text{KL}}}}{\partial \theta_z^{[i]}} \right\} \left\{ \frac{\partial e^{\mathcal{L}_{\text{KL}}}}{\partial \theta_z^{[j]}} \right\} \\ &= \mathbb{E}_{q^\star(\mathbf{X})} \left[\frac{\partial^2 \mathcal{L}_1}{\partial \theta_z^{[i]} \partial \theta_z^{[j]}} \right] + \mathbb{E}_{q^\star(\mathbf{X})} \left[\frac{\partial \mathcal{L}_1}{\partial \theta_z^{[i]}} \frac{\partial \mathcal{L}_1}{\partial \theta_z^{[j]}} \right] - \left\{ \mathbb{E}_{q^\star(\mathbf{X})} \left[\frac{\partial \mathcal{L}_1}{\partial \theta_z^{[i]}} \right] \right\} \left\{ \mathbb{E}_{q^\star(\mathbf{X})} \left[\frac{\partial \mathcal{L}_1}{\partial \theta_z^{[j]}} \right] \right\}. \end{aligned} \tag{11}$$

In this result the first term is equal to (10), and the second two terms combine to be always positive semi-definite, proving King and Lawrence [2006]'s intuition about the curvature of the bound. When curvature is negative definite (e.g. near a maximum), the KLC bound's curvature is less negative definite, enabling larger steps to be taken in optimization. Figure 1(b) illustrates the effect of this as well as the bound's similarities.

3.3 Relationship to Collapsed VB

In *collapsed inference* some parameters are marginalized *before* applying the variational bound. For example, Sung et al. [2008] proposed a latent variable model where the model parameters were marginalised, and Teh et al. [2007] proposed a non-parametric topic model where the document proportions were collapsed. These procedures lead to improved inference, or faster convergence.

The KLC bound derivation we have provided also marginalises parameters, but *after* a variational approximation is made. The difference between the two approaches is distilled in these expressions:

$$\ln \mathbb{E}_{p(\mathbf{X})} \left[\exp \left\{ \mathbb{E}_{q(\mathbf{Z})} \left[\ln p(\mathcal{D}|\mathbf{X}, \mathbf{Z}) \right] \right\} \right] \qquad \mathbb{E}_{q(\mathbf{Z})} \left[\ln \left\{ \mathbb{E}_{p(\mathbf{X})} \left[p(\mathcal{D}|\mathbf{X}, \mathbf{Z}) \right] \right\} \right] \tag{12}$$

where the left expression appears in the KLC bound, and the right expression appears in the bound for collapsed variational Bayes, with the remainder of the bounds being equal. Whilst appropriately conjugate formulation of the model will always ensure that the KLC expression is analytically tractable, the expectation in the collapsed VB expression is not. Sung et al. [2008] propose a first order approximation to the expectation of the form $\mathbb{E}_{q(\mathbf{Z})} \Big[f(\mathbf{Z}) \Big] \approx f(\mathbb{E}_{q(\mathbf{Z})} \Big[\mathbf{Z} \Big])$, which reduces the right expression to the that on the left. Under this approximation[2] the KL corrected approach is equivalent to the collapsed variational approach.

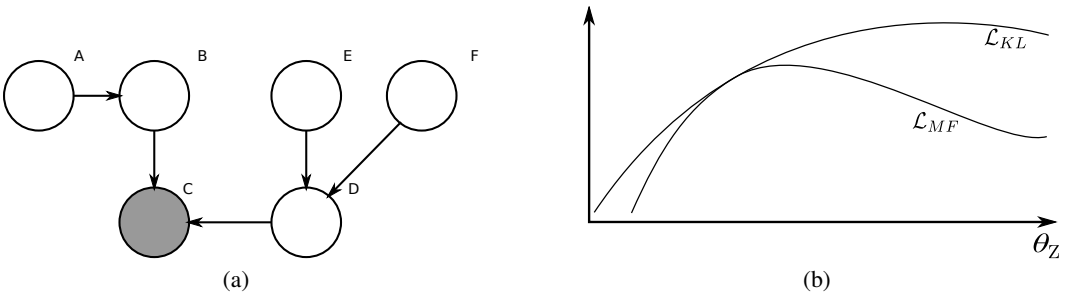
3.4 Applicability

To apply the KLC bound we need to specify a subset, $\mathbf{X}$, of variables to marginalize. We select the variables that break the dependency structure of the graph to enable the analytic computation of the integral in (4). Assuming the appropriate conjugate exponential structure for the model we are left with the requirement to select a sub-set that induces the appropriate factorisation. These induced factorisations are discussed in some detail in Bishop [2006]. They are factorisations in the approximate posterior which arise from the form of the variational approximation and from the structure of the model. These factorisations allow application of KLC bound, and can be identified using a simple d-separation test as Bishop discusses.

The d-separation test involves checking for independence amongst the marginalised variables ($\mathbf{X}$ in the above) conditioned on the observed data $\mathcal{D}$ and the approximated variables ($\mathbf{Z}$ in the above). The requirement is to select a sufficient set of variables, $\mathbf{Z}$, such that the effective likelihood for $\mathbf{X}$, given by (3) becomes conjugate to the prior. Figure 1(a) illustrates the d-separation test with application to the KLC bound.

For latent variable models, it is often sufficient to select the latent variables for $\mathbf{X}$ whilst collapsing the model variables. For example, in the specific case of mixture models and topic models, approximating the component labels allows for the marginalisation of the cluster parameters (topics

[2] Kurihara et al. [2007] and Teh et al. [2007] suggest a further second order correction and assume that that $q(\mathbf{Z})$ is Gaussian to obtain tractability. This leads to additional correction terms that augment KLC bound. The form of these corrections would need to be determined on a case by case basis, and has in fact been shown to be less effective than those methods unified here [Asuncion et al., 2012].



(a) (b)

Figure 1: (a) An example directed graphical model on which we could use the KLC bound. Given the observed node C, the nodes A,F d-separate given nodes B,D,E. Thus we could make an explicit variational approximation for A,F, whilst marginalising B,D,E. Alternatively, we could select B,D,E for a parameterised approximate distribution, whilst marginalising A,F. (b) A sketch of the KLC and MF bounds. At the point where the mean field method has $q(\mathbf{X}) = q^\star(\mathbf{X})$, the bounds are equal in value as well as in gradient. Away from the this point, the different between the bounds is the Kullback Leibler divergence between the current MF approximation for $\mathbf{X}$ and the implicit distribution $q^\star(\mathbf{X})$ of the KLC bound.

allocations) and mixing proportions. This allowed Sung et al. [2008] to derive a general form for latent variable models, though our formulation is general to any conjugate exponential graph.

4 Riemannian Gradient Based Optimisation

Sato [2001] and Hoffman et al. [2012] showed that the VBEM procedure performs gradient ascent in the space of the natural parameters. Using the KLC bound to collapse the problem, gradient methods seem a natural choice for optimisation, since there are fewer parameters to deal with, and we have shown that computation of the gradients is straightforward (the variational update equations contain the model gradients). It turns out that the KLC bound is particularly amenable to *Riemannian* or *natural gradient* methods, because the information geometry of the exponential family distrubution(s), over which we are optimising, leads to a simple expression for the natural gradient. Previous investigations of natural gradients for variational Bayes [Honkela et al., 2010, Kuusela et al., 2009] required the inversion of the Fisher information at every step (ours does not), and also used VBEM steps for *some* parameters and Riemannian optimisation for other variables. The collapsed nature of the KLC bound means that these VBEM steps are unnecessary: the bound can be computed by parameterizing the distribution of only one set of variables ($q(\mathbf{Z})$) whilst the implicit distribution of the other variables is given in terms of the first distribution and the data by equation (5).

We optimize the lower bound $\mathcal{L}_{\text{KL}}$ with respect to the parameters of the approximating distribution of the non-collapsed variables. We showed in section 2 that the gradient of the KLC bound is given by the gradient of the standard MF variational bound, after an update of the *collapsed* variables. It is clear from their definition that the same is true of the natural gradients.

4.1 Variable Transformations

We can compute the natural gradient of our collapsed bound by considering the update equations of the non-collapsed problem as described above. However, if we wish to make use of more powerful optimisation methods like conjugate gradient ascent, it is helpful to re-parameterize the natural parameters in an unconstrained fashion. The natural gradient is given by [Amari and Nagaoka, 2007]:

$$\widetilde{\mathbf{g}}(\boldsymbol{\theta}) = G(\boldsymbol{\theta})^{-1} \frac{\partial \mathcal{L}_{\text{KL}}}{\partial \boldsymbol{\theta}} \tag{13}$$

where $G(\boldsymbol{\theta})$ is the Fisher information matrix whose i,j^{th} element is given by

$$G(\boldsymbol{\theta})_{[i,j]} = -\mathbb{E}_{q(\mathbf{X} \mid \boldsymbol{\theta})}\Big[\frac{\partial^2 \ln q(\mathbf{X} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{[i]} \partial \boldsymbol{\theta}^{[j]}}\Big]. \tag{14}$$

For exponential family distributions, this reduces to $\nabla^2_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})$, where ψ is the log-normaliser. Further, for exponential family distributions, the Fisher information in the canonical parameters ($\boldsymbol{\theta}$) and that in the *expectation* parameters ($\boldsymbol{\eta}$) are reciprocal, and we also have $G(\boldsymbol{\theta}) = \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}$. This means that the natural gradient in $\boldsymbol{\theta}$ is given by

$$\widetilde{\mathbf{g}}(\boldsymbol{\theta}) = G(\boldsymbol{\theta})^{-1}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\theta}}\frac{\partial\mathcal{L}_{\text{KL}}}{\partial\boldsymbol{\eta}} = \frac{\partial\mathcal{L}_{\text{KL}}}{\partial\boldsymbol{\eta}} \text{ and } \widetilde{\mathbf{g}}(\boldsymbol{\eta}) = \frac{\partial\mathcal{L}_{\text{KL}}}{\partial\boldsymbol{\theta}}. \tag{15}$$

The gradient in one set of parameters provides the natural gradient in the other. Thus when our approximating distribution q is exponential family, we can compute the natural gradient *without* the expensive matrix inverse.

4.2 Steepest Ascent is Coordinate Ascent

Sato [2001] showed that the VBEM algorithm was a gradient based algorithm. In fact, VBEM consists of taking *unit* steps in the direction of the natural gradient of the canonical parameters. From equation (9) and the work of Sato [2001], we see that the gradient of the KLC bound can be obtained by considering the standard mean-field update for the non-collapsed parameter $\mathbf{Z}$. We confirm these relationships for the models studied in the next section in the supplementary material.

Having confirmed that the VB-E step is equivalent to steepest-gradient ascent we now explore whether the procedure could be improved by the use of conjugate gradients.

4.3 Conjugate Gradient Optimization

One idea for solving some of the problems associated with steepest ascent is to ensure each gradient step is conjugate (geometrically) to the previous. Honkela et al. [2010] applied conjugate gradients to the standard mean field bound, we expect much faster convergence for the KLC bound due to its differing curvature. Since VBEM uses a step length of 1 to optimize,[3] we also used this step length in conjugate gradients. In the natural conjugate gradient method, the search direction at the i^{th} iteration is given by $\mathbf{s}_i = -\widetilde{\mathbf{g}}_i + \beta\mathbf{s}_{i-1}$. Empirically the Fletcher-Reeves method for estimating β worked well for us:

$$\beta_{FR} = \frac{\langle\widetilde{\mathbf{g}}_i, \widetilde{\mathbf{g}}_i\rangle_i}{\langle\widetilde{\mathbf{g}}_{i-1}, \widetilde{\mathbf{g}}_{i-1}\rangle_{i-1}} \tag{16}$$

where $\langle\cdot,\cdot\rangle_i$ denotes the inner product in Riemannian geometry, which is given by $\widetilde{\mathbf{g}}^\top G(\rho)\widetilde{\mathbf{g}}$. We note from Kuusela et al. [2009] that this can be simplified since $\widetilde{\mathbf{g}}^\top G\widetilde{\mathbf{g}} = \widetilde{\mathbf{g}}^\top GG^{-1}\mathbf{g} = \widetilde{\mathbf{g}}^\top\mathbf{g}$, and other conjugate methods, defined in the supplementary material, can be applied similarly.

5 Experiments

For empirical investigation of the potential speed ups we selected a range of probabilistic models. We provide derivations of the bound and fuller explanations of the models in the supplementary material. In each experiment, the algorithm was considered to have converged when the change in the bound or the Riemannian gradient reached below 10^{-6}. Comparisons between optimisation procedures always used the same initial conditions (or set of initial conditions) for each method. First we recreate the mixture of Gaussians example described by Honkela et al. [2010].

5.1 Mixtures of Gaussians

For a mixture of Gaussians, using the d-separation rule, we select for $\mathbf{X}$ the cluster allocation (latent) variables. These are parameterised through the softmax function for unconstrained optimisation. Our model includes a fully Bayesian treatment of the cluster parameters and the mixing proportions, whose approximate posterior distributions appear as (5). Full details of the algorithm derivation are given in the supplementary material. A neat feature is that we can make use of the discussion above to derive an expression for the natural gradient without a matrix inverse.

[3] We empirically evaluated a line-search procedure, but found that in most cases that Wolfe-Powell conditions were met after a single step of unit length.

Table 1: Iterations to convergence for the mixture of Gaussians problem, with varying overlap (R). This table reports the average number of iterations taken to reach (within 10 nats of) the best known solution. For the more difficult scenarios (with more overlap in the clusters) the VBEM method failed to reach the optimum solution within 500 restarts

CG. method	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$
Polack-Ribiére	$3,100.37$	$15,698.57$	$5,767.12$	$1,613.09$	$3,046.25$
Hestenes-Stiefel	$1,371.55$	$5,501.25$	$5,922.4$	$\mathbf{358.03}$	$\mathbf{172.39}$
Fletcher-Reeves	$\mathbf{416.18}$	$\mathbf{1,161.35}$	$\mathbf{5,091.0}$	792.10	494.24
VBEM	∞	∞	∞	992.07	429.57

Table 2: Time and iterations taken to run LDA on the NIPS 2011 corpus, $\pm$ one standard deviation, for two conjugate methods and VBEM. The Fletcher-Reeves conjugate algorithm is almost ten times as fast as VBEM. The value of the bound at the optimum was largely the same: deviations are likely just due to the choice of initialisations, of which we used 12.

Method	Time (minutes)	Iterations	Bound
Hestenes-Steifel	56.4 ± 18.5	644.3 ± 214.5	$-1,998,780 \pm 201$
Fletcher-Reeves	$\mathbf{38.5 \pm 8.7}$	$\mathbf{447.8 \pm 100.5}$	$-1,998,743 \pm 194$
VBEM	370 ± 105	$4,459 \pm 1,296$	$-1,998,732 \pm 241$

In Honkela et al. [2010] data are drawn from a mixture of five two-dimensional Gaussians with equal weights, each with unit spherical covariance. The centers of the components are at $(0, 0)$ and $(\pm R, \pm R)$. R is varied from 1 (almost completely overlapping) to 5 (completely separate). The model is initialised with eight components with an uninformative prior over the mixing proportions: the optimisation procedure is left to select an appropriate number of components.

Sung et al. [2008] reported that their collapsed method led to improved convergence over VBEM. Since our objective is identical, though our optimisation procedure different, we devised a metric for measuring the efficacy of our algorithms which also accounts for their propensity to fall into local minima. Using many randomised restarts, we measured the average number of iterations taken to reach the *best-known optimum*. If the algorithm converged at a lesser optimum, those iterations were included in the denomiator, but we didn't increment the numerator when computing the average. We compared three different conjugate gradient approaches and standard VBEM (which is also steepest ascent on the KLC bound) using 500 restarts.

Table 1 shows the number of iterations required (on average) to come within 10 nats of the best known solution for three different conjugate-gradient methods and VBEM. VBEM sometimes failed to find the optimum in any of the 500 restarts. Even relaxing the stringency of our selection to 100 nats, the VBEM method was always at least twice as slow as the best conjugate method.

5.2 Topic Models

Latent Dirichlet allocation (LDA) [Blei et al., 2003] is a popular approach for extracting topics from documents. To demonstrate the KLC bound we applied it to 200 papers from the 2011 NIPS conference. The PDFs were preprocessed with `pdftotext`, removing non-alphabetical characters and coarsely filtering words by popularity to form a vocabulary size of 2000.[4] We selected the latent topic-assignment variables for parameterisation, collapsing the topics and the document proportions. Conjugate gradient optimization was compared to the standard VBEM approach.

We used twelve random initializations, starting each algorithm from each initial condition. Topic and document distributions where treated with fixed, uninformative priors. On average, the Hestenes-Steifel algorithm was almost ten times as fast as standard VB, as shown in Table 2, whilst the final bound varied little between approaches.

[4] Some extracted topics are presented in the supplementary material.

5.3 RNA-seq alignment

An emerging problem in computational biology is inference of transcript structure and expression levels using next-generation sequencing technology (RNA-Seq). Several models have been proposed. The BitSeq method [Glaus et al., 2012] is based on a probabilistic model and uses Gibbs sampling for approximate inference. The sampler can suffer from particularly slow convergence due to the large size of the problem, which has six million latent variables for the data considered here. We implemented a variational version of their model and optimised it using VBEM and our collapsed Riemannian method. We applied the model to data described in Xu et al. [2010], a study of human microRNA. The model was initialised using four random initial conditions, and optimised using standard VBEM and the conjugate gradient versions of the algorithm. The Polack-Ribiére conjugate method performed very poorly for this problem, often giving negative conjugation: we omit it here. The solutions found for the other algorithms were all fairly close, with bounds coming within 60 nats. The VBEM method was dramatically outperformed by the Fletcher-Reeves and Hestenes-Steifel methods: it took 4600 ± 20 iterations to converge, whilst the conjugate methods took only 268 ± 4 and 265 ± 1 iterations to converge. At about 8 seconds per iteration, our collapsed Riemannian method requires around forty minutes, whilst VBEM takes almost eleven hours. All the variational approaches represent an improvement over a Gibbs sampler, which takes approximately one week to run for this data [Glaus et al., 2012].

6 Discussion

Under very general conditions (conjugate exponential family) we have shown the equivalence of collapsed variational bounds and marginalized variational bounds using the KL corrected perspective of King and Lawrence [2006]. We have provided a succinct derivation of these bounds, unifying several strands of work and laying the foundations for much wider application of this approach.

When the collapsed variables are updated in the standard MF bound the KLC bound is identical to the MF bound in value and gradient. Sato [2001] has shown that coordinate ascent of the MF bound (as proscribed by VBEM updates) is equivalent to steepest ascent of the MF bound using natural gradients. This implies that standard variational inference is also performing steepest ascent on the KLC bound. This equivalence between natural gradients and the VBEM update equations means our method is quickly implementable for any model where the mean field update equations have been computed. It is only necessary to determine which variables to collapse using a d-separation test. Importantly this implies our approach can readily be incorporated in automated inference engines such as that provided by infer.net [Minka et al., 2010]. We'd like to emphasise the ease with which the method can be applied: we have provided derivations of equivalencies of the bounds and gradients which should enable collapsed conjugate optimisation of *any* existing mean field algorithm, with minimal changes to the software. Indeed our own implementations (see supplementary material) use just a few lines of code to switch between the VBEM and conjugate methods.

The improved performance arises from the curvature of the KLC bound. We have shown that it is always less negative than that of the original variational bound allowing much larger steps in the variational parameters as King and Lawrence [2006] suggested. This also provides a gateway to second-order optimisation, which could prove even faster.

We provided empirical evidence of the performance increases that are possible using our method in three models. In a thorough exploration of the convergence properties of a mixture of Gaussians model, we concluded that a conjugate Riemannian algorithm can find solutions that are not found with standard VBEM. In a large LDA model, we found that performance can be improved many times over that of the VBEM method. In the BitSeq model for differential expression of genes transcripts we showed that very large improvements in performance are possible for models with huge numbers of latent variables.

Acknowledgements

The authors would like to thank Michalis Titsias for helpful commentary on a previous draft and Peter Glaus for help with a C++ implementation of the RNAseq alignment algorithm. This work was funded by EU FP7-KBBE Project Ref 289434 and BBSRC grant number BB/1004769/1.

References

S. Amari and H. Nagaoka. *Methods of information geometry*. AMS, 2007.

A. Asuncion, M. Welling, P. Smyth, and Y. Teh. On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662*, 2012.

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

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

Z. Ghahramani and M. Beal. Propagation algorithms for variational Bayesian learning. *Advances in neural information processing systems*, pages 507–513, 2001.

P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 2012. doi: 10.1093/bioinformatics/bts260. Advance Access.

M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *arXiv preprint arXiv:1206.7051*, 2012.

A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *The Journal of Machine Learning Research*, 9999:3235–3268, 2010.

N. King and N. D. Lawrence. Fast variational inference for Gaussian process models through KL-correction. *Machine Learning: ECML 2006*, pages 270–281, 2006.

K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 20, page 19, 2007.

M. Kuusela, T. Raiko, A. Honkela, and J. Karhunen. A gradient-based algorithm competitive with variational Bayesian EM for mixture of Gaussians. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 1688–1695. IEEE, 2009.

M. Lázaro-Gredilla and M. K. Titsias. Variational heteroscedastic Gaussian process regression. In *Proceedings of the International Conference on Machine Learning (ICML), 2011*, 2011.

M. Lázaro-Gredilla, S. Van Vaerenbergh, and N. Lawrence. Overlapping mixtures of Gaussian processes for the data association problem. *Pattern Recognition*, 2011.

T. P. Minka, J. M. Winn, J. P. Guiver, and D. A. Knowles. Infer .NET 2.4. Microsoft Research Cambridge, 2010.

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

J. Sung, Z. Ghahramani, and S. Bang. Latent-space variational Bayes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2236–2242, 2008.

Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in neural information processing systems*, 19:1353, 2007.

G. Xu et al. Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. *RNA*, pages 1610–1622, June 2010. ISSN 1355-8382. doi: 10.1261/rna.2194910. URL `http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.2194910`.