Supplementary Material for: Fast Variational Inference in the Conjugate Exponential Family

James Hensman, Magnus Rattray and Neil D. Lawrence

November 7, 2012

This supplementary material accompanies the NIPS paper on Fast Variational Inference in the Conjugate Exponential Family. Its purpose is to provide details of how our very general framework applies in the case of the specific models described in the paper. First we briefly mention the form of the three conjugate gradient algorithms we used in optimization.

1 Conjugate gradient algorithms

There are several different methods for approximating the parameter β in the conjugate gradient algorithm. We used the Polack-Ribière, Fletch-Reeves or Hestenes-Stiefel methods:

$$\begin{aligned}\beta_{PR} &= \frac{\langle \widetilde{\mathbf{g}}_i, \widetilde{\mathbf{g}}_i - \widetilde{\mathbf{g}}_{i-1}\rangle_i}{\langle \widetilde{\mathbf{g}}_{i-1}, \widetilde{\mathbf{g}}_{i-1}\rangle_{i-1}} \\ \beta_{FR} &= \frac{\langle \widetilde{\mathbf{g}}_i, \widetilde{\mathbf{g}}_i\rangle_i}{\langle \widetilde{\mathbf{g}}_{i-1}, \widetilde{\mathbf{g}}_{i-1}\rangle_{i-1}} \\ \beta_{HS} &= \frac{\langle \widetilde{\mathbf{g}}_i, \widetilde{\mathbf{g}}_i - \widetilde{\mathbf{g}}_{i-1}\rangle_i}{\langle \widetilde{\mathbf{g}}_{i-1}, \widetilde{\mathbf{g}}_i - \widetilde{\mathbf{g}}_{i-1}\rangle_{i-1}}\end{aligned} \tag{1}$$

where $\langle \cdot, \cdot \rangle_i$ denotes the inner product in Riemannian geometry, which is given by $\widetilde{\mathbf{g}}^\top G(\rho)\widetilde{\mathbf{g}}$

2 Mixture of Gaussians

A MoG model is defined as follows. We have a set of N D-dimensional vectors $\mathbf{Y} = \{\mathbf{y}_n\}_{n=1}^N$. The likelihood is

$$p(\mathbf{Y}|\boldsymbol{\eta}, \mathbf{L}) = \prod_{k=1}^K \prod_{n=1}^N \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{\ell_{nk}} \tag{2}$$

where $\mathbf{L}$ is a collection of binary latent variables indicating cluster membership, $\mathbf{L} = \{\{\boldsymbol{\ell}_{nk}\}_{n=1}^N\}_{k=1}^K$ and $\boldsymbol{\eta}$ is a collection of cluster parameters, $\boldsymbol{\eta} = \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\}_{k=1}^K$

The prior over $\mathbf{L}$ is given by a multinomial distribution with components $\boldsymbol{\pi}$, which in turn have a Dirichlet prior with uniform concentrations for simplicity:

$$p(\mathbf{L}|\boldsymbol{\pi}) = \prod_{k=1}^K \prod_{n=1}^N \pi_k^{\ell_{nk}}, \quad p(\boldsymbol{\pi}) = R_D(\boldsymbol{\alpha}) \prod_{k=1}^K \pi_k^{\alpha-1} \tag{3}$$

with $\boldsymbol{\alpha}$ representing a K dimensional vector with elements α, and R_D being the normalising constant for the Dirichlet distribution, $R_D(\boldsymbol{\alpha}) = \Gamma(K\alpha)\Gamma(\alpha)^{-K}$.

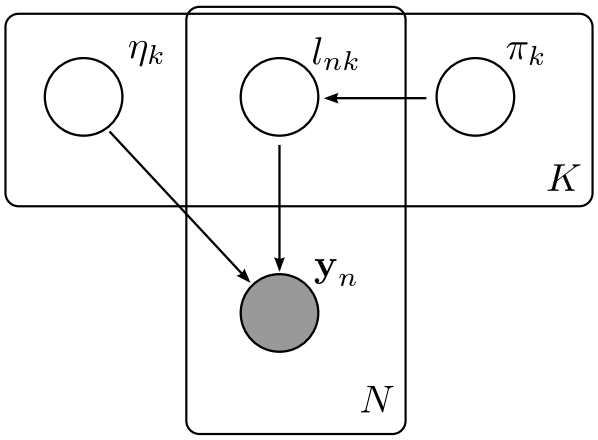

Figure 1: A graphical model representation of the MoG model. A d-separation test quickly shows that it is possible to marginalise $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$ given a variational parameterisation of $\mathbf{L}$.

Finally we choose a conjugate Gaussian-Wishart prior for the cluster parameters which can be written

$$\begin{aligned}\ln p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) &= \ln R_{GW}(\mathbf{S}_0, \nu_0, \kappa_0) + \frac{\nu_0 - D}{2} \ln |\boldsymbol{\Lambda}_k| \\ -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Lambda}_k \left(\kappa_0 \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top + \kappa_0 \mathbf{m}_0 \mathbf{m}_0^\top - 2\kappa_0 \mathbf{m}_0 \boldsymbol{\mu}_k^\top + \mathbf{S}_0\right)\right)\end{aligned} \tag{4}$$

where R_{GW} is the normalising constant, and is given by

$$R_{GW}(\mathbf{S}, \nu, \kappa) = |\mathbf{S}|^{\frac{\nu}{2}} 2^{-\frac{(\nu+1)D}{2}} \pi^{-\frac{D(D+1)}{4}} \kappa^{\frac{D}{2}} \Big(\prod_{d=1}^{D} \Gamma((\nu + 1 - d)/2)\Big)^{-1}.$$

2.1 Applying the KLC bound

The first task in applying the KLC bound is to select which variables to parameterise and which to marginalise. From the graphical model representation of the MoG problem in Figure 1, we can see that we can select the latent variables $\mathbf{Z} = \{\mathbf{L}\}$ for parameterisation, whilst marginalising the mixing proportions and cluster parameters ($\mathbf{X} = \{\boldsymbol{\pi}, \boldsymbol{\eta}\}$). We note that it is possible to select the variables the other way around: parameterising $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$ and marginalising $\mathbf{L}$, but parameterisation of the latent variables makes implementation a little simpler.

We use a factorised multinomial distribution $q(\mathbf{L})$ to approximate the posterior for $p(\mathbf{L}|\mathbf{Y})$, parameterised using the softmax functions so

$$q(\mathbf{L}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{\boldsymbol{\ell}_{nk}}, \quad r_{nk} = \frac{e^{\rho_{nk}}}{\sum_{i=1}^{K} e^{\rho_{ni}}}. \tag{5}$$

We are now ready to apply the procedure described above to derive the KLC bound. First,

$$\ln p(\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\pi}) \geq \int q(\mathbf{L}) \{\ln p(\mathbf{Y}|\boldsymbol{\eta}, \mathbf{L}) + \ln p(\mathbf{L}|\boldsymbol{\pi})\} \, \mathrm{d}\mathbf{L} + H_L, \tag{6}$$

where H_L is the entropy of the distribution $q(\mathbf{L})$. We expand to give

$$\begin{aligned}\mathcal{L}_1 = \tfrac{1}{2} \sum_{k=1}^{K} \Big\{ &- \mathrm{tr}\big(\boldsymbol{\Lambda}_k \big(\widehat{r}_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top + C_k - 2\boldsymbol{\mu}_k \bar{\mathbf{y}}_k\big)\big) \\ &+ \widehat{r}_k \ln \pi_k + \widehat{r}_k \ln |\boldsymbol{\Lambda}_k| \Big\} + H_L - \tfrac{ND}{2} \ln(2\pi)\end{aligned} \tag{7}$$

where $\widehat{r}_k = \sum_{n=1}^N r_{nk}$, $C_k = \sum_{n=1}^N r_{nk}\mathbf{y}_n\mathbf{y}_n^\top$, and $\bar{\mathbf{y}}_k = \sum_{n=1}^N r_{nk}\mathbf{y}_n$. The conjugacy between the intermediate bound $\mathcal{L}_1$ and the prior now emerges, making the second integral in the KLC bound tractable.

After exponentiating this expression and multiplying by the prior, $p(\boldsymbol{\eta})p(\boldsymbol{\pi})$, we find that the integrals with respect to both $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$ are tractable. This result means that the only variational parameters needed are those of $q(\mathbf{L})$. The integrals result in

$$\begin{aligned}\mathcal{L}_{\text{KL}} = -\frac{ND}{2}\ln(2\pi) + \ln R_{Di}(\boldsymbol{\alpha}) - \ln R_{Di}(\boldsymbol{\alpha}') \\ +K\ln R_{GW}(\mathbf{S}_0,\nu_0,\kappa_0) - \sum_{k=1}^K \ln R_{GW}(\mathbf{S}_k,\nu_k,\kappa_k) + H_L\end{aligned} \tag{8}$$

where we have defined

$$\begin{aligned}\alpha_k &= \alpha + \widehat{r}_k & \kappa_k &= \kappa_0 + \widehat{r}_k \\ \mathbf{m}_k &= (\kappa_0\mathbf{m}_0 + \bar{\mathbf{y}}_k)/\kappa_k & \nu_k &= \nu_0 + \widehat{r}_k \\ \mathbf{S}_k &= S_0 + C_k + \kappa_0\mathbf{m}_0\mathbf{m}_0^\top - \kappa_k\mathbf{m}_k\mathbf{m}_k^\top\end{aligned} \tag{9}$$

and $\boldsymbol{\alpha}'$ represents a vector containing each α_k. Some simplification of (8) leads to

$$\begin{aligned}\mathcal{L}_{\text{KL}} = \sum_{k=1}^K \Big\{ &\ln\Gamma(\alpha_k) - \tfrac{D}{2}\ln\kappa_k - \tfrac{\nu_k}{2}\ln|S_k| \\ &+ \sum_{d=1}^D \ln\Gamma((\nu_k+1-d)/2)\Big\} + H_L + \text{const.}\end{aligned} \tag{10}$$

where const. contains terms independent of $\mathbf{r}$.

Equations (9) are similar to the update equations for the approximating distributions in the VBEM methodology [see e.g. Bishop, 2006]. However, for our model they are simply intermediate variables, representing combinations of the true variational parameters $\mathbf{r}$, the data, and the model prior parameters. When optimizing the model with respect to the variational parameters, the dependency of these intermediate variables on $\mathbf{r}$ is not ignored as it would be in MF variational approach.

The gradient of the MV bound (10) with respect to the parameters $\mathbf{r}$ is given by

$$\begin{aligned}\frac{\partial\mathcal{L}_{\text{KL}}}{\partial r_{nk}} = &-\tfrac{D}{2}\kappa_k^{-1} - \tfrac{1}{2}\ln|S_k| + \psi(\alpha_k) - \ln r_{nk} \\ &- \tfrac{\nu_k}{2}(y_n - \mathbf{m}_k)^\top S_k^{-1}(y_n - \mathbf{m}_k)) \\ &+ \tfrac{1}{2}\sum_{d=1}^D \psi((\nu_k+1-d)/2) - 1.\end{aligned} \tag{11}$$

Taking a step in this direction (in the valiables γ) yields exactly the VB-E step associated with the mean-field bound. the gradient in r is the natural gradient in γ (see paper section 4.1).

3 Latent Dirichlet Allocation

Latent Dirichlet allocation is a popular topic model. See Blei et al. [2003] for a thorough introduction.

Suppose we have D documents, K topics and a vocabulary of size V. The d^{th} document contains N_d words $W_d = \{w_{dn}\}_{n=1}^{N_d}$, and each word is represented as a binary vector $w_{dn} \in \{0,1\}^V$. Each word is associated with a latent variable $\boldsymbol{\ell}_{dn}$, which assigns the word to a topic, thus $\boldsymbol{\ell}_{dn} \in \{0,1\}^K$. We'll use W to represent the colletion of all words, $W = \{W_d\}_{d=1}^D$, and $\mathbf{L}$ to represent the collection of all latent variables $\mathbf{L} = \{\{\boldsymbol{\ell}_{dn}\}_{n=1}^{N_d}\}_{d=1}^D$.

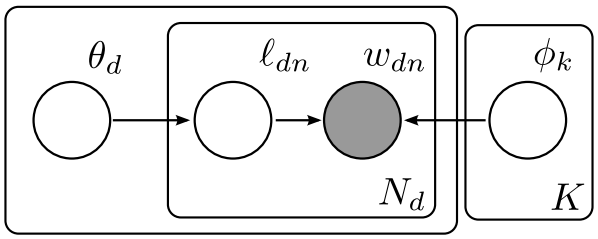

Figure 2: A graphical model representation of Latent Dirichlet allocation. A d-separation test quickly shows that it is possible to marginalise θ and ϕ given a variational parameterisation of $\mathbf{L}$.

Each document has an associated vector of topic proportions, $\theta_d \in [0,1]^K$, and each topic is represented by a vector of word proportions $\phi_k \in [0,1]^V$. We assume a symmetrical prior distribution over topics in each document $p(\theta_d) = \textit{Dir}(\theta_d|\alpha)$, and similarly for words within topics. $p(\phi_k) = \textit{Dir}(\phi_k|\beta)$.

The LDA generative model states that for each word, first the associated topic is drawn from the topic proportions for the document, and then the word is drawn from the selected topic.

$$\begin{aligned} p(\ell_{dn}|\theta_d) &= \prod_{k=1}^{K} \theta_{dk}^{\boldsymbol{\ell}_{dnk}} \\ p(w_{dn}|\boldsymbol{\ell}_{dn}, \phi) &= \prod_{k=1}^{K}\prod_{v=1}^{V} \phi_{kv}^{w_{dnv}\boldsymbol{\ell}_{dnk}} \end{aligned} \tag{12}$$

3.1 The collapsed bound

To derive the colapsed bound, we use a similar d-separation test as for the mixture model to select the latent variables as the parameteriser (non-collapsed) nodes. See Figure 2.

To proceed we assume a factorising multinomial posterior for $\mathbf{L}$:

$$q(\mathbf{L}) = \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K} r_{dnk}^{\boldsymbol{\ell}_{dnk}} \tag{13}$$

subject to the constraint $\sum_{k=1}^{K} \boldsymbol{\ell}_{dnk} = 1$, which we enforce through a softmax reparameterisation

$$r_{dnk} = \frac{e^{\rho_{dnk}}}{\sum_{k'=1}^{K} e^{\rho_{dnk'}}}. \tag{14}$$

We proceed by deriving the conditional bound

$$\ln p(W \mid \theta, \phi) \geq \mathcal{L}_1 = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\sum_{v=1}^{V}(w_{dnv} r_{dnk}) \ln \phi_{kv} + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}(r_{dnk} \ln \theta_{dk}) + H[q(\mathbf{L})]. \tag{15}$$

To marginalise the variables θ, ϕ, we exponentiate this bound and take the expectation under the priors. This

results in

$$p(W) \geq \int \exp\{\mathcal{L}_1\} p(\theta) p(\phi)\, \mathrm{d}\theta\, \mathrm{d}\phi = \int \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{\sum_{d=1}^{d} \sum_{n=1}^{n_d} (w_{dnv} r_{dnk})} \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{(\sum_{n=1}^{N_d} r_{dnk})} \prod_{k=1}^{K} R_{Di}(\beta) \prod_{v=1}^{V} \phi_{kv}^{\beta-1} \prod_{d=1}^{D} R_{Di}(\alpha) \prod_{k=1}^{K} \theta_{dk}^{\alpha-1}\, \mathrm{d}\theta\, \mathrm{d}\phi \exp\{H[q(\mathbf{L})]\}. \tag{16}$$

Careful inspection of the above reveals that the two integrals separate as expected, and result in the normalizers for each of the independent Dirichlet approximations. Taking the logarithm reults in

$$\mathcal{L}_{\mathrm{KL}} = D \ln R_{Di}(\alpha) - \sum_{d=1}^{D} \ln R_{Di}(\alpha'_d) + K \ln R_{Di}(\beta) - \sum_{k=1}^{K} \ln R_{Di}(\beta'_k) + H[q(\mathbf{L})] \tag{17}$$

where we have defined $\alpha'_{dk} = \alpha + \sum_{n=1}^{N_d} r_{dnk}$ and $\beta'_{kv} = \beta + \sum_{d=1}^{d} \sum_{n=1}^{n_d} w_{dnv} r_{dnk}$.

3.2 Topics found by LDA

For completeness we show here some topics found by LDA on the NIPS conference data.

Table 1: some topics found using LDA on papers from the 2011 NIPS conference.

neural	training	distribution	data	features	model
input	feature	gaussian	points	image	models
neurons	classification	inference	point	object	variables
network	class	process	clustering	images	parameters
fig	tree	prior	distance	objects	structure
estimate	prediction	sampling	dataset	scene	variable
neuron	label	likelihood	similarity	recognition	markov
visual	accuracy	posterior	cluster	reference	observed
nonlinear	labels	distributions	manifold	detection	graphical
linear	classifier	bayesian	spectral	part	hidden

4 BitSeq Model

The generative model for an RNA-seq assay is as follows. We assume that the experiment consists of a pile of RNA fragments, where the abundance of fragments from transcript T_m in the assay is θ_m. The sequencer then selects a fragment at random from the pile, such that the probability of picking a fragment corresponding to transcript T_m is θ_m. Introducing a convenient membership vector $\boldsymbol{\ell}_n$ for each read, we can write

$$p(\mathbf{L}|\boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{m=1}^{M} \theta_m^{\boldsymbol{\ell}_{nm}} \tag{18}$$

where $\ell_{nm} \in \{0, 1\}$ is a binary variable which indicates whether the nth fragment came from the mth transcript ($\ell_{nm} = 1$) and is subject to $\sum_{m=1}^{M} \ell_{nm} = 1$. We use $\mathbf{L}$ to represent the collection of all alignment variables.

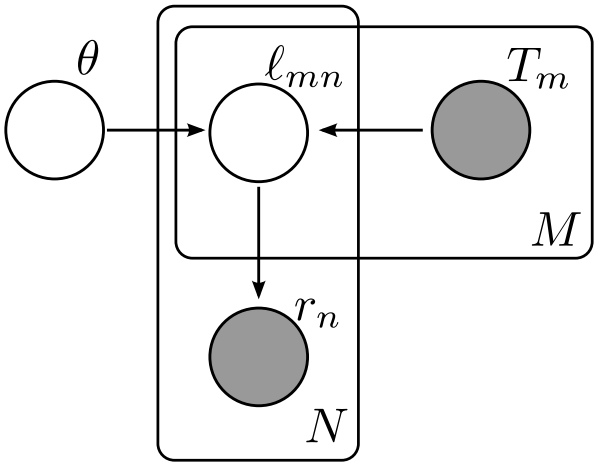

Figure 3: A graphical model representation of the BitSeq model. A d-separation test quickly shows that it is possible to marginalise θ given a variational parameterisation of $\mathbf{L}$.

Both θ and $\mathbf{L}$ are variables to be inferred, with $\boldsymbol{\theta}$ the main object of interest.

Writing the collection of all reads as $\mathbf{R} = \{\mathbf{r}_n\}_{n=1}^N$, the likelihood of a set of alignments $\mathbf{L}$ is

$$p(\mathbf{R}|\mathbf{T}, \mathbf{L}) = \prod_{n=1}^{N} p(\mathbf{r}_n|T_m)^{\ell_{nm}} \tag{19}$$

where T_m represents the mth transcript, $\mathbf{T}$ represents the transcriptome.

The values of $p(r_n|T_m)$ can be computed before performing inference in $\boldsymbol{\theta}$ since we are assuming a known transcriptome. We compute these values based on the quality of alignment of the read $\mathbf{r}_n$ to the transcript T_m, using a model which can correct for sequence specific or fragmentation biases. The method is described in detatil in Glaus et al. [2012].

We specify a conjugate Dirichlet prior over the vector $\boldsymbol{\theta}$.

$$p(\boldsymbol{\theta}) = \frac{\Gamma(\hat{\alpha}^o)}{\prod_{m=1}^{M} \Gamma(\alpha_m^o)} \prod_{m=1}^{M} \theta_m^{\alpha_m^o - 1} \tag{20}$$

with $\hat{\alpha}^o = \sum_{m=1}^{M} \alpha_m^o$. α_m^o represents our prior belief in the values of θ_m, and we use a relatively uninformative but proper prior $\alpha_m^o = 1 \, \forall m = 1 \dots M$. A priori, we assume that the concentrations are all equal, but with large uncertainty.

4.1 The collapsed bound

Figure 3 shows a graphical representation of the BitSeq model. It's clear that parameterisation of the latent variables will allow us to collapse θ, or vica-versa. Selecting again the latent variables for parameterisation, $\mathbf{X} = \{\mathbf{L}\}$, $\mathbf{Z} = \{\theta\}$, we first find the conditional bound as usual by:

$$\begin{aligned}
\ln p(\mathbf{R} \,|\, \mathbf{T}, \boldsymbol{\theta}) &= \ln \int p(\mathbf{R} \,|\, \mathbf{L}, \mathbf{T}) p(\mathbf{L} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\mathbf{L} \\
&\geq \mathbb{E}_{q(\mathbf{L})} \Big[\ln p(\mathbf{R} \,|\, \mathbf{L}, \mathbf{T}) + \ln p(\mathbf{L} \,|\, \boldsymbol{\theta}) - \ln q(\mathbf{L}) \Big] \\
&\geq \sum_{n=1}^{N} \sum_{m=1}^{M} \ell_{nm} \big(\ln p(\mathbf{r}_n \,|\, T_m) + \ln \theta_m - \ln \ell_{nm} \big) \\
&\geq \mathcal{L}_1
\end{aligned} \tag{21}$$

It's clear that this bound is conjugate to the prior for θ, so we can marginalise:

$$\ln p(\mathbf{R} \mid \mathbf{T}) \geq \mathcal{L}_{\text{KL}} = \sum_{n=1}^{N} \sum_{m=1}^{M} \ell_{nm} \big(\ln p(\mathbf{r}_n \mid T_m) - \ln \ell_{nm} \big) + \ln \Gamma(\hat{\alpha}^o) - \ln \Gamma(\hat{\alpha}^o + N) \\ - \sum_{m=1}^{M} \Big(\ln \Gamma(\alpha_m^o) - \ln \Gamma(\alpha_m^o + \hat{\ell}_m) \Big) \tag{22}$$

where $\hat{\ell}_m = \sum_{n=1}^{N} \ell_n$ and we also have that the approximate posterior distribution for $\boldsymbol{\theta}$ is a Dirichlet distribution with parameters $\alpha_m^o + \hat{\ell}_m$.

References

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.

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