1 Gradient of Kullback-Leibler divergence

Let $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$ be two sets of natural parameters of an exponential family, that is,

$$q(\boldsymbol{\beta};\boldsymbol{\lambda}) = h(\boldsymbol{\beta}) \exp\left(\boldsymbol{\lambda}^\top t(\boldsymbol{\beta}) - a(\boldsymbol{\lambda})\right). \tag{1}$$

The partial derivatives of their Kullback-Leibler divergence are given by

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\lambda}} D_{\mathrm{KL}}(\boldsymbol{\lambda},\boldsymbol{\lambda}') &= \frac{\partial}{\partial\boldsymbol{\lambda}} \mathbb{E}_{\boldsymbol{\lambda}}\left[\log\frac{q(\boldsymbol{\beta};\boldsymbol{\lambda})}{q(\boldsymbol{\beta};\boldsymbol{\lambda}')}\right] && (2)\\
&= \frac{\partial}{\partial\boldsymbol{\lambda}} \mathbb{E}_{\boldsymbol{\lambda}}\left[(\boldsymbol{\lambda}-\boldsymbol{\lambda}')^\top t(\boldsymbol{\beta}) - a(\boldsymbol{\lambda}) + a(\boldsymbol{\lambda}')\right] && (3)\\
&= (\boldsymbol{\lambda}-\boldsymbol{\lambda}')^\top \frac{\partial}{\partial\boldsymbol{\lambda}} \mathbb{E}_{\boldsymbol{\lambda}}\left[t(\boldsymbol{\beta})\right] + \mathbb{E}_{\boldsymbol{\lambda}}\left[t(\boldsymbol{\beta})\right] - \frac{\partial}{\partial\boldsymbol{\lambda}} a(\boldsymbol{\lambda}) && (4)\\
&= (\boldsymbol{\lambda}-\boldsymbol{\lambda}')^\top \frac{\partial^2}{\partial\boldsymbol{\lambda}^2} a(\boldsymbol{\lambda}) + \frac{\partial}{\partial\boldsymbol{\lambda}} a(\boldsymbol{\lambda}) - \frac{\partial}{\partial\boldsymbol{\lambda}} a(\boldsymbol{\lambda}) && (5)\\
&= (\boldsymbol{\lambda}-\boldsymbol{\lambda}')^\top I(\boldsymbol{\lambda}), && (6)
\end{aligned}$$

where we have used the exponential family identities

$$\frac{\partial}{\partial\boldsymbol{\lambda}} a(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\lambda}}[t(\boldsymbol{\beta})]^\top, \qquad \frac{\partial^2}{\partial\boldsymbol{\lambda}^2} a(\boldsymbol{\lambda}) = I(\boldsymbol{\lambda}). \tag{7}$$

2 Asymptotic behavior of trust-region method

Here we show that the trust-region update given below converges to a natural gradient step.

$$d\boldsymbol{\lambda} = \underset{d\boldsymbol{\lambda}}{\operatorname{argmax}} \left\{\mathcal{L}_n(\boldsymbol{\lambda}+d\boldsymbol{\lambda}) - \xi D_{\mathrm{KL}}(\boldsymbol{\lambda}+d\boldsymbol{\lambda},\boldsymbol{\lambda})\right\} \tag{8}$$

For large ξ, $d\boldsymbol{\lambda}$ will be close to zero so that we can focus on the target functions' first-order terms. For exponential families in canonical form, we have

$$\begin{aligned}
D_{\mathrm{KL}}(\boldsymbol{\lambda}+d\boldsymbol{\lambda},\boldsymbol{\lambda}) &= E_{\boldsymbol{\lambda}+d\boldsymbol{\lambda}}\left[\log\frac{q(\boldsymbol{\beta};\boldsymbol{\lambda}+d\boldsymbol{\lambda})}{q(\boldsymbol{\beta};\boldsymbol{\lambda})}\right] && (9)\\
&= E_{\boldsymbol{\lambda}+d\boldsymbol{\lambda}}\left[d\boldsymbol{\lambda}^\top t(\boldsymbol{\beta}) - a(\boldsymbol{\lambda}+d\boldsymbol{\lambda}) + a(\boldsymbol{\lambda})\right] && (10)\\
&= d\boldsymbol{\lambda}^\top \nabla a(\boldsymbol{\lambda}+d\boldsymbol{\lambda}) - a(\boldsymbol{\lambda}+d\boldsymbol{\lambda}) + a(\boldsymbol{\lambda}). && (11)
\end{aligned}$$

The gradient of the KL divergence in $d\boldsymbol{\lambda}$ is thus given by

$$\begin{aligned}
\nabla D_{\mathrm{KL}}(\boldsymbol{\lambda}+d\boldsymbol{\lambda},\boldsymbol{\lambda}) &= \nabla a(\boldsymbol{\lambda}+d\boldsymbol{\lambda}) + \nabla^2 a(\boldsymbol{\lambda}+d\boldsymbol{\lambda})d\boldsymbol{\lambda} - \nabla a(\boldsymbol{\lambda}+d\boldsymbol{\lambda}) && (12)\\
&= \nabla^2 a(\boldsymbol{\lambda}+d\boldsymbol{\lambda})d\boldsymbol{\lambda}. && (13)
\end{aligned}$$

Approximating the target function around $\boldsymbol{\lambda}$ yields

$$d\boldsymbol{\lambda}^\top \nabla\mathcal{L}_n(\boldsymbol{\lambda}) - \xi d\boldsymbol{\lambda}^\top \nabla^2 a(\boldsymbol{\lambda}) d\boldsymbol{\lambda}, \tag{14}$$

which when maximized gives an update proportional to the natural gradient direction,

$$d\boldsymbol{\lambda} = \frac{1}{2\xi}\left(\nabla^2 a(\boldsymbol{\lambda})\right)^{-1} \nabla\mathcal{L}_n(\boldsymbol{\lambda}). \tag{15}$$

3 Latent Dirichlet Allocation

In LDA we have global parameters $\boldsymbol{\beta}$ consisting of distributions over words $\boldsymbol{\beta}_k$ and local parameters $\boldsymbol{\theta}_n, \mathbf{z}_n$ with distributions

$$p(\boldsymbol{\beta}) = \prod_k \text{Dir}(\boldsymbol{\beta}_k; \boldsymbol{\eta}), \tag{16}$$

$$p(\boldsymbol{\theta}_n) = \text{Dir}(\boldsymbol{\theta}_n; \boldsymbol{\alpha}), \tag{17}$$

$$p(\boldsymbol{z}_n \mid \boldsymbol{\theta}_n) = \prod_m \theta_{n z_{nm}}, \tag{18}$$

$$p(\boldsymbol{x}_n \mid z_n, \boldsymbol{\beta}) = \prod_m \beta_{z_{nm} x_{nm}}. \tag{19}$$

We approximate the posterior distribution over $\boldsymbol{\beta}$ with

$$q(\boldsymbol{\beta}) = \prod_k \text{Dir}(\boldsymbol{\beta}_k; \boldsymbol{\lambda}_k). \tag{20}$$

Writing the likelihood in the form of Equation 2 of the main paper gives

$$p(\mathbf{x}_n, \mathbf{z}_n, \boldsymbol{\theta}_n \mid \boldsymbol{\beta}) = \text{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) \left(\prod_m \theta_{n z_{nm}} \right) \left(\prod_m \beta_{z_{nm} x_{nm}} \right) \tag{21}$$

$$= h(\boldsymbol{\theta}_n, \mathbf{z}_n) \exp \left(\sum_m \log \beta_{z_{nm} x_{nm}} \right) \tag{22}$$

$$= h(\boldsymbol{\theta}_n, \mathbf{z}_n) \exp \left(\langle t(\boldsymbol{\beta}), f(\mathbf{z}_n, \mathbf{x}_n) \rangle \right) \tag{23}$$

where h encompasses all terms which do not depend on $\boldsymbol{\beta}$ and

$$t(\boldsymbol{\beta}) = \log \boldsymbol{\beta}, \tag{24}$$

$$f(\mathbf{z}_n, \mathbf{x}_n) = \sum_m \mathbf{I}_{z_{nm} x_{nm}}, \tag{25}$$

where $\mathbf{I}_{ij}$ is a matrix with entry (i, j) set to 1 and all other entries set to 0. We here assume that $\boldsymbol{\beta}$ is a matrix whose rows are the topics $\boldsymbol{\beta}_k$ and

$$\langle \mathbf{A}, \mathbf{B} \rangle = \text{vec}(\mathbf{A})^\top \text{vec}(\mathbf{B}) = \text{tr}(\mathbf{AB}) \tag{26}$$

for matrices $\mathbf{A}$ and $\mathbf{B}$. Since the standard parametrization of the Dirichlet distribution is already in canonical form, we can immediately apply Equation 10 of the main paper to get the steps of the inner loop of the trust-region update,

$$\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda}_t + \rho_t \left(\boldsymbol{\eta} + N \mathbb{E}_{\boldsymbol{\phi}_n^*}[f(\mathbf{x}_n, \mathbf{z}_n)] \right). \tag{27}$$

The beliefs over $\mathbf{z}_n$ (i.e., $\boldsymbol{\phi}_n^*$) are computed in the usual manner [Blei et al., 2003].

4 Mixture models

Consider a mixture model where the local parameters are the cluster assignments k_n and the global parameters are the prior weights $\boldsymbol{\pi}$ and the components' parameters $\boldsymbol{\beta}_k$. We assume the following more specific form for the model,

$$\begin{aligned} p(\boldsymbol{\pi}) &= \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha}), && (28) \\ p(k_n \mid \boldsymbol{\pi}) &= \pi_{k_n}, && (29) \\ p(\boldsymbol{\beta}_k) &\propto h(\boldsymbol{\beta}_k) \exp\left(\boldsymbol{\eta}^\top t(\boldsymbol{\beta}_k)\right), && (30) \\ p(\mathbf{x}_n \mid k_n, \boldsymbol{\beta}) &= g(\mathbf{x}_n) \exp\left(t(\boldsymbol{\beta}_{k_n})^\top f(\mathbf{x}_n)\right) && (31) \end{aligned}$$

for suitable funtions t, f, g, h. The factors of a mean-field approximation to the posterior are given by

$$\begin{aligned} q(\boldsymbol{\pi}) &= \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\gamma}), && (32) \\ q(k_n) &= \phi_{nk_n}, && (33) \\ q(\boldsymbol{\beta}_k) &\propto h(\boldsymbol{\beta}_k) \exp\left(\boldsymbol{\lambda}_k^\top t(\boldsymbol{\beta}_k)\right). && (34) \end{aligned}$$

In each iteration, the trust-region method alternates between computing the optimal $\boldsymbol{\phi}_n$ and updating $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$. The approximate posterior over mixture components, slightly abusing notation, is given by

$$\begin{aligned} \boldsymbol{\phi}_n^* &= \operatorname*{argmax}_{\boldsymbol{\phi}_n} \mathcal{L}_n(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}_n) && (35) \\ &\propto \exp \mathbb{E}_q \left[\log p(\mathbf{x}_n \mid k_n, \boldsymbol{\beta}) p(k_n \mid \boldsymbol{\pi})\right] && (36) \\ &= \exp \mathbb{E}_q \left[\log p(\mathbf{x}_n \mid k_n, \boldsymbol{\beta})\right] \exp\left(\psi(\boldsymbol{\gamma}) - \psi\left(\sum_k \gamma_k\right)\right) && (37) \end{aligned}$$

where for the expected log-likelihood we have

$$\mathbb{E}_q \left[\log p(\mathbf{x}_n \mid k_n, \boldsymbol{\beta})\right] = \mathbb{E}_q \left[t(\boldsymbol{\beta}_{k_n})\right]^\top f(\mathbf{x}_n) + \log g(\mathbf{x}_n). \tag{38}$$

Once $\boldsymbol{\phi}_n^*$ is computed, we update $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ via

$$\begin{aligned} \boldsymbol{\gamma} &= (1 - \rho_t)\boldsymbol{\gamma}_t + \rho_t \left(\boldsymbol{\alpha} + N\boldsymbol{\phi}_n^*\right), && (39) \\ \boldsymbol{\lambda}_k &= (1 - \rho_t)\boldsymbol{\lambda}_k^t + \rho_t \left(\boldsymbol{\eta} + N\phi_{nk}^* f(\mathbf{x}_n)\right). && (40) \end{aligned}$$

For mini-batches of size B, these updates become

$$\begin{aligned} \boldsymbol{\gamma} &= (1 - \rho_t)\boldsymbol{\gamma}_t + \rho_t \left(\boldsymbol{\alpha} + \frac{N}{B}\sum_n \boldsymbol{\phi}_n^*\right), && (41) \\ \boldsymbol{\lambda}_k &= (1 - \rho_t)\boldsymbol{\lambda}_k^t + \rho_t \left(\boldsymbol{\eta} + \frac{N}{B}\sum_n \phi_{nk}^* f(\mathbf{x}_n)\right). && (42) \end{aligned}$$

To use these results with any concrete mixture model, we have to write down the prior over $\boldsymbol{\beta}$ in canonical form and implement the expected log-likelihood (Equation 38) and the update in Equation 42.

4.1 Mixture of multivariate Bernoullis

For the multivariate Bernoulli model,

$$p(x_n \mid k_n, \boldsymbol{\beta}) = \prod_i \beta_{ki}^{x_{ni}}(1-\beta_{ki})^{1-x_{ni}}, \tag{43}$$

we assume beta distributions for the prior and approximate posterior,

$$\begin{align} p(\boldsymbol{\beta}_k) &\propto \beta_{ki}^{a-1}(1-\beta_{ki})^{b-1}, \tag{44}\\ q(\boldsymbol{\beta}_k) &\propto \beta_{ki}^{a_{ki}-1}(1-\beta_{ki})^{b_{ki}-1}. \tag{45}\end{align}$$

$\boldsymbol{\lambda}_k = (\mathbf{a}_k, \mathbf{b}_k)$ are already the natural parameters of the beta distribution, so that the updates (Equation 42) become

$$\begin{align} \mathbf{a}_k &= (1-\rho_t)\mathbf{a}_k^t + \rho_t\left(a + \frac{N}{B}\sum_n \phi_{nk}^* x_{ni}\right), \tag{46}\\ \mathbf{b}_k &= (1-\rho_t)\mathbf{b}_k^t + \rho_t\left(b + \frac{N}{B}\sum_n \phi_{nk}^* (1-x_{ni})\right). \tag{47}\end{align}$$

The expected log-likelihood needed for the computation of $\boldsymbol{\phi}_n^*$ is given by

$$\mathbb{E}_q\left[\log p(\mathbf{x}_n \mid k_n, \boldsymbol{\beta})\right] = \mathbf{x}_n^\top \psi(\mathbf{a}_{k_n}) + (1-\mathbf{x}_n)^\top \psi(\mathbf{b}_{k_n}) - \mathbf{1}^\top \psi(\mathbf{a}_{k_n} + \mathbf{b}_{k_n}), \tag{48}$$

where the digamma function ψ is applied point-wise and $\mathbf{1}$ is a vector of ones.

4.2 Mixture of Gaussians

We assume a normal-inverse-Wishart distribution (NIW) for the parameters $\boldsymbol{\beta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ of each Gaussian distribution. The NIW for a singe component is given by

$$\begin{align} p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto \exp\left(-\frac{s}{2}(\boldsymbol{\mu}-\mathbf{m})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{m})\right)\exp\left(-\frac{1}{2}\text{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right)|\boldsymbol{\Sigma}|^{-\frac{\nu+D+1}{2}} \tag{49}\\ &= \exp\left(-\frac{s}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + s\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\mathbf{m} - \frac{1}{2}\text{tr}(s\mathbf{m}\mathbf{m}^\top\boldsymbol{\Sigma}^{-1}) - \frac{1}{2}\text{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}) - \frac{\nu}{2}\log|\boldsymbol{\Sigma}| - \frac{D+1}{2}\log|\boldsymbol{\Sigma}|\right) \tag{50}\\ &= \exp\left(\eta(\mathbf{m}, s, \boldsymbol{\Psi}, \nu)^\top t(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right) \tag{51}\end{align}$$

where

$$\begin{align} \eta(\mathbf{m}, s, \boldsymbol{\Psi}, \nu) &= \left(s, -2s\mathbf{m}, \text{vec}(s\mathbf{m}\mathbf{m}^\top + \boldsymbol{\Psi}), \nu\right) \tag{52}\\ &= (s, \mathbf{b}, \text{vec}\,\mathbf{C}, \nu) = \boldsymbol{\eta}, \tag{53}\\ t(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2}\left(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \text{vec}\,\boldsymbol{\Sigma}^{-1}, \log|\boldsymbol{\Sigma}|\right). \tag{54}\end{align}$$

are the natural parameters and sufficient statistics of the distribution, respectively. The likelihood for a single data point $\mathbf{x}$ is given by

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) / |\boldsymbol{\Sigma}|^{\frac{1}{2}} \tag{55}$$

$$= \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^\top\right) + \mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\log|\boldsymbol{\Sigma}|\right) \tag{56}$$

$$= \exp\left(t(\boldsymbol{\mu}, \boldsymbol{\Sigma})^\top f(\mathbf{x})\right), \tag{57}$$

where

$$f(\mathbf{x}) = \left(1, -2\mathbf{x}, \mathrm{vec}\left(\mathbf{x}\mathbf{x}^\top\right), 1\right). \tag{58}$$

Hence, assuming the natural parameters of all components are given by

$$\boldsymbol{\eta}_k = \eta(\mathbf{m}_k, s_k, \boldsymbol{\Psi}_k, \nu_k) = (s_k, \mathbf{b}_k, \mathrm{vec}\,\mathbf{C}_k, \nu_k), \tag{59}$$

an update of the inner loop of the trust-region method (Equation 42) is given by

$$s_k = (1-\rho_t)s_k^t + \rho_t\left(s + \frac{N}{B}\sum_n \phi_{nk}^*\right), \tag{60}$$

$$\mathbf{b}_k = (1-\rho_t)\mathbf{b}_k^t + \rho_t\left(\mathbf{b} - 2\frac{N}{B}\sum_n \phi_{nk}^*\mathbf{x}_n\right), \tag{61}$$

$$\mathbf{C}_k = (1-\rho_t)\mathbf{C}_k^t + \rho_t\left(\mathbf{C} + \frac{N}{B}\sum_n \phi_{nk}^*\mathbf{x}_n\mathbf{x}_n^\top\right), \tag{62}$$

$$\nu_k = (1-\rho_t)\nu_k^t + \rho_t\left(\nu + \frac{N}{B}\sum_n \phi_{nk}^*\right). \tag{63}$$

Or, in terms of the more traditional parametrization,

$$\mathbf{m}_k = (1-\rho_t)\frac{s_k^t}{s_k}\mathbf{m}_k^t + \frac{1}{s_k}\rho_t\left(s\mathbf{m} + \frac{N}{B}\sum_n \phi_{nk}^*\mathbf{x}_n\right), \tag{64}$$

$$\boldsymbol{\Psi}_k = (1-\rho_t)\left(\boldsymbol{\Psi}_k^t + s_k^t\mathbf{m}_k^t\mathbf{m}_k^{t\top}\right) + \rho_t\left(\boldsymbol{\Psi} + s\mathbf{m}\mathbf{m}^\top + \frac{N}{B}\sum_n \phi_{nk}^*\mathbf{x}_n\mathbf{x}_n^\top\right) - s_k\mathbf{m}_k\mathbf{m}_k^\top. \tag{65}$$

A proof that these updates leave $\boldsymbol{\Psi}_k$ positive definite is given below.

Positive definiteness of $\boldsymbol{\Psi}_k$

We show positive definiteness of $\boldsymbol{\Psi}_k$ in two steps. First, we show that the constraint is fulfilled for $\rho_t = 0$ and $\rho_t = 1$. Second, we show that the set of valid natural parameters induced by the constraint is convex, implying that the constraint must be fulfilled for any linear interpolation of two natural parameters.

For $\rho_t = 0$, we have $\mathbf{\Psi}_k = \mathbf{\Psi}_k^t$ since none of the natural parameters has changed. Thus, $\mathbf{\Psi}_k$ is positive definite if $\mathbf{\Psi}_k^t$ is positive definite. For $\rho_t = 1$, we have

$$\mathbf{\Psi}_k = \mathbf{\Psi} + s\mathbf{m}\mathbf{m}^\top + \frac{N}{B}\sum_n \phi_{nk}^* \mathbf{x}_n\mathbf{x}_n^\top - s_k\mathbf{m}_k\mathbf{m}_k^\top, \tag{66}$$

$$s_k\mathbf{m}_k\mathbf{m}_k = \frac{1}{s_k}\left(s\mathbf{m} + \frac{N}{B}\sum_n \phi_{nk}^* \mathbf{x}_n\right)\left(s\mathbf{m} + \frac{N}{B}\sum_n \phi_{nk}^* \mathbf{x}_n\right)^\top, \tag{67}$$

$$s_k = s + \frac{N}{B}\sum_n \phi_{nk}^*. \tag{68}$$

Note that $p_0 = s/s_k$ and $p_n = \frac{N}{B}\phi_{nk}/s_k$ are positive and sum to one and therefore can be considered probabilities. Let $\mathbf{X}$ be a random variable which takes on value $\mathbf{m}$ with probability p_0 and value $\mathbf{x}_n$ with probability p_n. Then we can rewrite Equation 66 as

$$\mathbf{\Psi}_k = \mathbf{\Psi} + s_k\mathbb{E}_p[\mathbf{X}\mathbf{X}^\top] - s_k\mathbb{E}_p[\mathbf{X}]\mathbb{E}_p[\mathbf{X}]^\top = \mathbf{\Psi} + s_k\mathbb{V}[\mathbf{X}]. \tag{69}$$

Since $\mathbf{\Psi}$ is positive definite and the covariance matrix $\mathbb{V}[\mathbf{X}]$ is at least positive semi-definite, $\mathbf{\Psi}_k$ must be positive definite.

We next show that the set of valid natural parameters,

$$\{(s, \mathbf{b}, \mathbf{C}, \nu) : s > 0, \nu > D - 1, \mathbf{C} - \frac{1}{4s}\mathbf{b}\mathbf{b}^\top \text{ is p. d. }\}, \tag{70}$$

is convex. Not that this set is convex iff

$$\{(s, \mathbf{b}, \mathbf{C}) : s > 0, \mathbf{C} - \frac{1}{s^2}\mathbf{b}\mathbf{b}^\top \text{ is p. d. }\}, \tag{71}$$

is convex. This set is convex iff

$$\{(s, \mathbf{b}, \mathbf{C}) : s > 0, \mathbf{C} - \mathbf{b}\mathbf{b}^\top \text{ is p. d. }\}, \tag{72}$$

is convex, since any *perspective function* preserves convexity and $P(s, \mathbf{b}, \mathbf{C}) = (s, \mathbf{b}/s, \mathbf{C})$ is a perspective function. Finally, this set is convex iff the following set is convex,

$$\Omega = \{(\mathbf{b}, \mathbf{C}) : \mathbf{C} - \mathbf{b}\mathbf{b}^\top \text{ is p. d. }\}. \tag{73}$$

Assume $(\mathbf{b}_1, \mathbf{C}_1), (\mathbf{b}_2, \mathbf{C}_2) \in \Omega$ and let $\rho \in [0, 1]$. Then

$$\rho\mathbf{C}_1 + (1-\rho)\mathbf{C}_2 - (\rho\mathbf{b}_1 + (1-\rho)\mathbf{b}_2)(\rho\mathbf{b}_1 + (1-\rho)\mathbf{b}_2)^\top \tag{74}$$

$$= \rho\mathbf{C}_1 + (1-\rho)\mathbf{C}_2 - (\rho(\mathbf{b}_1 - \mathbf{b}_2) + \mathbf{b}_2)(\mathbf{b}_1 + (1-\rho)(\mathbf{b}_2 - \mathbf{b}_1))^\top \tag{75}$$

$$= \rho(\mathbf{C}_1 - \mathbf{b}_1\mathbf{b}_1^\top) + (1-\rho)(\mathbf{C}_2 - \mathbf{b}_2\mathbf{b}_2^\top) + \rho(1-\rho)(\mathbf{b}_1 - \mathbf{b}_2)(\mathbf{b}_1 - \mathbf{b}_2)^\top, \tag{76}$$

which is a sum of positive definite and semi-definite matrices and therefore positive definite. Hence,

$$(\rho\mathbf{C}_1 + (1-\rho)\mathbf{C}_2, \rho\mathbf{b}_1 + (1-\rho)\mathbf{b}_2) \in \Omega \tag{77}$$

and Ω is convex.

Expected log-likelihood

To compute he expected log-likelihood (Equation 38), we need

$$
\begin{aligned}
\mathbb{E}_q\left[\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k\right] &= \mathbb{E}_q\left[\mathbb{E}_q\left[\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \mid \boldsymbol{\Sigma}_k\right]\right] && (78)\\
&= \mathbb{E}_q\left[\mathrm{tr}\left(s_k^{-1} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^{-1}\right) + \mathbf{m}_k^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{m}_k\right] && (79)\\
&= D s_k^{-1} + \nu_k \mathbf{m}_k^\top \boldsymbol{\Psi}_k^{-1} \mathbf{m}_k, && (80)\\
\mathbb{E}_q\left[\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k\right]^\top \mathbf{x} &= \nu_k \mathbf{m}_k^\top \boldsymbol{\Psi}_k^{-1} \mathbf{x}, && (81)\\
\mathbf{x}^\top \mathbb{E}_q\left[\boldsymbol{\Sigma}_k^{-1}\right] \mathbf{x} &= \nu_k \mathbf{x}^\top \boldsymbol{\Psi}_k^{-1} \mathbf{x}, && (82)\\
\mathbb{E}_q\left[\log |\boldsymbol{\Sigma}_k^{-1}|\right] &= \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \log 2 + \log |\boldsymbol{\Psi}_k^{-1}|. && (83)
\end{aligned}
$$

For Equation 79, see Petersen and Pedersen [2008]. For Equation 83, see Bishop [2006]. The expected log-likelihood is thus given by

$$
\begin{aligned}
\mathbb{E}_q\left[\log p(\mathbf{x} \mid k, \boldsymbol{\beta})\right] &= \mathbb{E}_q\left[-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x} + \mathbf{x} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k + \frac{1}{2} \log |\boldsymbol{\Sigma}_k^{-1}| - \frac{D}{2} \log(2\pi)\right] && (84)\\
&= -\frac{\nu_k}{2} \mathbf{x}^\top \boldsymbol{\Psi}_k^{-1} \mathbf{x} + \nu_k \mathbf{x}^\top \boldsymbol{\Psi}_k^{-1} \mathbf{m}_k - \frac{D}{2} s_k^{-1} - \frac{\nu_k}{2} \mathbf{m}_k^\top \boldsymbol{\Psi}_k^{-1} \mathbf{m}_k && (85)\\
&\quad + \frac{1}{2}\left(\sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \log 2 - \log |\boldsymbol{\Psi}_k|\right) - \frac{D}{2} \log(2\pi) && (86)\\
&= -\frac{\nu_k}{2} (\mathbf{x} - \mathbf{m}_k)^\top \boldsymbol{\Psi}_k^{-1} (\mathbf{x} - \mathbf{m}_k) - \frac{D}{2} s_k^{-1} + \frac{1}{2} \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + \frac{1}{2} \log |\boldsymbol{\Psi}_k^{-1}| - \frac{D}{2} \log \pi. && (87)
\end{aligned}
$$

References

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

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

K. Petersen and M. Pedersen. The Matrix Cookbook, 2008.