A collection of identities for variational inference with exponential-family models

Dominik Endres, Kathrin Pabst, Anna-Lena Eckert, Raphael Schween
dominik.endres@uni-marburg.de

September 5, 2022

Abstract

This is a collection of identities useful for variational inference with exponential family distributions/densities. All derviations were done by the authors, unless indicated otherwise. This does **not imply** that the results collected here have not appeared in the literature before. **DISCLAIMER**: this collection is a work in progress. It is certainly incomplete and probably buggy. Bug-reports and contributions are most welcome, please email dominik.endres@uni-marburg.de.

1 Exponential family distributions

A distribution is said to belong to the exponential family, if it can be written in the form [1]:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})) \tag{1}$$

where the random variates $\boldsymbol{x}$ may be discrete or continuous, the *sufficient statistics* $\boldsymbol{u}$ are functions of the $\boldsymbol{x}$, not necessarily of the same dimensionality. However, the $\boldsymbol{u}$ need to be linearly independent. The $\boldsymbol{\eta}$ are the *natural parameters*, one for each sufficient statistic. The function $g(\boldsymbol{\eta})$ is the normalization constant (replace the integral with a sum for discrete $\boldsymbol{x}$):

$$g(\boldsymbol{\eta})\int d\boldsymbol{x}\; h(\boldsymbol{x})\exp(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})) = 1 \tag{2}$$

1.1 Moments

Taking the gradient ∇ w.r.t. to $\boldsymbol{\eta}$ on both sides of equation 2, we find:

$$(\nabla g(\boldsymbol{\eta}))\underbrace{\int d\boldsymbol{x}\; h(\boldsymbol{x})\exp(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}))}_{=\frac{1}{g(\boldsymbol{\eta})}} + \underbrace{g(\boldsymbol{\eta})\int d\boldsymbol{x}\; h(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})\exp(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}))}_{\langle\boldsymbol{u}(\boldsymbol{x})\rangle} = \boldsymbol{0} \tag{3}$$

and thus the expectation $\langle\boldsymbol{u}(\boldsymbol{x})\rangle$ is:

$$\boxed{\langle\boldsymbol{u}(\boldsymbol{x})\rangle = -\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = -\nabla\log(g(\boldsymbol{\eta}))} \tag{4}$$

Computing the derivative of the i-th component of the l.h.s. of eqn. 3 w.r.t. $\boldsymbol{\eta}_j$ yields (noting that $\frac{\partial^2 \log(f(x,y))}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial x \partial y} / f(x,y) - \frac{\partial f(x,y)}{\partial x} \frac{\partial f(x,y)}{\partial y} / f^2(x,y)$):

$$\begin{aligned}
\frac{\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_i \partial \eta_v}}{g(\boldsymbol{\eta})} - \frac{\frac{\partial g(\boldsymbol{\eta})}{\partial \eta_i}\frac{\partial g(\boldsymbol{\eta})}{\partial \eta_j}}{g^2(\boldsymbol{\eta})} + \frac{\partial g(\boldsymbol{\eta})}{\partial \eta_j}\frac{\langle u_i(\boldsymbol{x})\rangle}{g(\boldsymbol{\eta})} &= -g(\boldsymbol{\eta}) \int d\boldsymbol{x}\; h(\boldsymbol{x}) u_i(\boldsymbol{x}) u_j(\boldsymbol{x}) \exp(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})) \\
\frac{\partial^2 \log(g(\boldsymbol{\eta}))}{\partial \eta_i \partial \eta_j} - \langle u_i(\boldsymbol{x})\rangle\langle u_j(\boldsymbol{x})\rangle &= -\langle u_i(\boldsymbol{x}) u_j(\boldsymbol{x})\rangle \\
\Rightarrow\; \langle u_i(\boldsymbol{x}) u_j(\boldsymbol{x})\rangle - \langle u_i(\boldsymbol{x})\rangle\langle u_j(\boldsymbol{x})\rangle &= -\frac{\partial^2 \log(g(\boldsymbol{\eta}))}{\partial \eta_i \partial \eta_j} \qquad (5)
\end{aligned}$$

Denoting the Hessian by $\nabla\nabla$, the covariance matrix of $\boldsymbol{u}(\boldsymbol{x})$ is thus given by

$$\boxed{\mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x})) = -\nabla\nabla \log(g(\boldsymbol{\eta}))} \qquad (6)$$

Higher order moments can be computed via higher order derviatives.

1.2 Maximum Likelihood

For maximum-likelihood approximations, we need the gradient of $\log(p(\boldsymbol{x}|\boldsymbol{\eta}))$ w.r.t. $\boldsymbol{\eta}$:

$$\nabla \log(p(\boldsymbol{x}|\boldsymbol{\eta})) = \nabla \log(g(\boldsymbol{\eta})) + \boldsymbol{u}(\boldsymbol{x}) = -\langle \boldsymbol{u}(\boldsymbol{x})\rangle + \boldsymbol{u}(\boldsymbol{x}). \qquad (7)$$

In other words, maximizing the likelihood (i.e. following the gradient towards higher likelihood values) amounts to making the expected value of the sufficient statistic more similiar to the actually observed sufficient statistic. For second-order optimization methods, the Hessian matrix may be needed, which is given by eqn. 6. For maximum likelihood parameter estimates, assume we had observed N i.i.d. datapoints $\boldsymbol{x}^n$. The parameter estimate is obtained by solving

$$\sum_{n=1}^{N} \nabla \log(p(\boldsymbol{x}^n|\boldsymbol{\eta})) = 0 \Rightarrow \langle \boldsymbol{u}(\boldsymbol{x})\rangle = \frac{\sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}^n)}{N} \qquad (8)$$

i.e. by setting the data mean of the sufficient statistics equal to the expectation.

1.3 Entropy

The (differential) entropy of $\boldsymbol{x}$ given $\boldsymbol{\eta}$ is defined as

$$H(\boldsymbol{x}|\boldsymbol{\eta}) = -\int d\boldsymbol{x}\; p(\boldsymbol{x}|\boldsymbol{\eta}) \log(p(\boldsymbol{x}|\boldsymbol{\eta})). \qquad (9)$$

note that this is not the conditional entropy of $\boldsymbol{x}$ given $\boldsymbol{\eta}$. Using the definition of the exponential family distribution (eqn. 1) , this can be written as

$$H(\boldsymbol{x}|\boldsymbol{\eta}) = -\int d\boldsymbol{x}\; p(\boldsymbol{x}|\boldsymbol{\eta}) \left(\log(g(\boldsymbol{\eta})) + \log(h(\boldsymbol{x})) + \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right)$$

and thus

$$\boxed{H(\boldsymbol{x}|\boldsymbol{\eta}) = -\log(g(\boldsymbol{\eta})) - \langle \log(h(\boldsymbol{x}))\rangle - \boldsymbol{\eta}^T \langle \boldsymbol{u}(\boldsymbol{x})\rangle} \qquad (10)$$

For the gradient of the entropy w.r.t. $\boldsymbol{\eta}$ we need

$$\begin{aligned}\frac{\partial}{\partial \eta_k}\,\boldsymbol{\eta}^T\langle \boldsymbol{u}(\boldsymbol{x})\rangle &= \langle u_k(\boldsymbol{x})\rangle + \sum_i \eta_i \frac{\partial \langle u_i(\boldsymbol{x})\rangle}{\partial \eta_k} \\ &= \langle u_k(\boldsymbol{x})\rangle - \sum_i \eta_i \frac{\partial^2 \log(g(\boldsymbol{\eta}))}{\partial \eta_i \partial \eta_k}. \end{aligned} \tag{11}$$

Using eqn. 6, we find

$$\nabla \boldsymbol{\eta}^T \langle \boldsymbol{u}(\boldsymbol{x})\rangle = \langle \boldsymbol{u}(\boldsymbol{x})\rangle + \mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x}))\cdot \boldsymbol{\eta} \tag{12}$$

We also need

$$\begin{aligned}\frac{\partial \langle \log(h(\boldsymbol{x}))\rangle}{\partial \eta_k} &= \frac{\partial}{\partial \eta_k}\left(g(\boldsymbol{\eta}) \int d\boldsymbol{x}\; h(\boldsymbol{x}) \log(h(\boldsymbol{x})) \exp(\eta^T \boldsymbol{u}(\boldsymbol{x}))\right) \\ &= \frac{\partial g(\boldsymbol{\eta})}{\partial \eta_k}\frac{\langle \log(h(\boldsymbol{x}))\rangle}{g(\boldsymbol{\eta})} + g(\boldsymbol{\eta}) \int d\boldsymbol{x}\; h(\boldsymbol{x}) \log(h(\boldsymbol{x})) u_k(\boldsymbol{x}) \exp(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})) \\ &= \frac{\partial \log(g(\boldsymbol{\eta}))}{\partial \eta_k}\langle \log(h(\boldsymbol{x}))\rangle + \langle \log(h(\boldsymbol{x})) u_k(\boldsymbol{x})\rangle \\ &= -\langle u_k(\boldsymbol{x})\rangle \langle \log(h(\boldsymbol{x}))\rangle + \langle \log(h(\boldsymbol{x})) u_k(\boldsymbol{x})\rangle \end{aligned} \tag{13}$$

and thus

$$\nabla \langle \log(h(\boldsymbol{x}))\rangle = -\langle \log(h(x))\rangle \langle \boldsymbol{u}(\boldsymbol{x})\rangle + \langle \log(h(\boldsymbol{x}))\boldsymbol{u}(\boldsymbol{x})\rangle \tag{14}$$

The gradient of the entropy w.r.t. the $\boldsymbol{\eta}$ is thus

$$\nabla H(\boldsymbol{x}|\boldsymbol{\eta}) = \underbrace{-\nabla \log(g(\boldsymbol{\eta}))}_{\langle \boldsymbol{u}(\boldsymbol{x})\rangle} - \nabla \langle \log(h(\boldsymbol{x}))\rangle - \nabla\; \boldsymbol{\eta}^T \langle \boldsymbol{u}(\boldsymbol{x})\rangle \tag{15}$$

$$\boxed{\nabla H(\boldsymbol{x}|\boldsymbol{\eta}) = \langle \log(h(\boldsymbol{x}))\rangle \langle \boldsymbol{u}(\boldsymbol{x})\rangle - \langle \log(h(\boldsymbol{x}))\boldsymbol{u}(\boldsymbol{x})\rangle - \mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x}))\cdot \boldsymbol{\eta}} \tag{16}$$

1.4 Kullback-Leibler divergence

The KL-divergence of a distribution with parameters $\tilde{\boldsymbol{\eta}}$ to a distribution with parameters $\boldsymbol{\eta}$ is:

$$\begin{aligned} D(p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}})||p(\boldsymbol{x}|\boldsymbol{\eta})) &= \int d\boldsymbol{x}\; p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}}) \log\left(\frac{p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}})}{p(\boldsymbol{x}|\boldsymbol{\eta})}\right) \\ &= \int d\boldsymbol{x}\; p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}}) \left[\log\left(p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}})\right) - \log\left(p(\boldsymbol{x}|\boldsymbol{\eta})\right)\right] \\ &= -H(\boldsymbol{x}|\tilde{\boldsymbol{\eta}}) - \int d\boldsymbol{x}\; p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}}) \log\left(p(\boldsymbol{x}|\boldsymbol{\eta})\right) \end{aligned} \tag{17}$$

For this, we compute the following expectation under $p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}})$:

$$\begin{aligned}\int d\boldsymbol{x}\; p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}}) \log\left(p(\boldsymbol{x}|\boldsymbol{\eta})\right) &= \int d\boldsymbol{x}\; p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}}) \left(\log(g(\boldsymbol{\eta})) + \log(h(\boldsymbol{x})) + \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right) \\ &= \log(g(\boldsymbol{\eta})) + \langle \log(h(\boldsymbol{x}))\rangle + \boldsymbol{\eta}^T \langle \boldsymbol{u}(\boldsymbol{x})\rangle \end{aligned} \tag{18}$$

and thus, using eqn. 10:

$$\boxed{D(p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}})||p(\boldsymbol{x}|\boldsymbol{\eta})) = \log(g(\tilde{\boldsymbol{\eta}})) - \log(g(\boldsymbol{\eta})) + (\tilde{\boldsymbol{\eta}}^T - \boldsymbol{\eta}^T)\langle \boldsymbol{u}(\boldsymbol{x})\rangle} \tag{19}$$

Its gradient w.r.t. $\tilde{\boldsymbol{\eta}}$ is then (using eqn. 4 and eqn. 6)

$$\begin{aligned} \nabla D(p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}})||p(\boldsymbol{x}|\boldsymbol{\eta})) &= \nabla \log(g(\tilde{\boldsymbol{\eta}})) + \nabla\, (\tilde{\boldsymbol{\eta}}^T - \boldsymbol{\eta}^T)\langle \boldsymbol{u}(\boldsymbol{x})\rangle \\ &= -\langle \boldsymbol{u}(\boldsymbol{x})\rangle + \langle \boldsymbol{u}(\boldsymbol{x})\rangle + \nabla\langle \boldsymbol{u}(\boldsymbol{x})\rangle(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) \end{aligned} \tag{20}$$

where $\nabla\langle \boldsymbol{u}(\boldsymbol{x})\rangle$ acts on $\boldsymbol{u}$ component-wise, yielding a matrix (ij) of derivatives $\frac{\partial u_i(\boldsymbol{x})}{\partial \tilde{\boldsymbol{\eta}}_j}$. Thus

$$\boxed{\nabla D(p(\boldsymbol{x}|\tilde{\boldsymbol{\eta}})||p(\boldsymbol{x}|\boldsymbol{\eta})) = -\nabla\nabla \log(g(\tilde{\boldsymbol{\eta}}))\cdot(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x}))(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta})} \tag{21}$$

2 Conjugate priors on exponential family distributions

For inference and learning in hierarchical models, conjugate priors on the parameters $\boldsymbol{\eta}$ are very useful, because inference/learning with i.i.d. observations translates into parameter updates (rather than complicated integrals). The conjugate prior on an exponential family distribution (eqn. 1) is given by

$$p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu) = f(\boldsymbol{\lambda}, \nu) m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^\nu \exp(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}) \tag{22}$$

where the $\boldsymbol{\lambda}, \nu$ are the parameters of the p(oste)rior, $g(\boldsymbol{\eta})$ is the same function as above and $m(\boldsymbol{\eta})$ is an arbitrary positive function (different from $g(\boldsymbol{\eta})$). To see that this is a conjugate prior on $p(\boldsymbol{x}|\boldsymbol{\eta})$, assume we had observed N datapoints $\boldsymbol{x}_n$. The posterior of $\boldsymbol{\eta}$ is then (using eqns. 1 and 22)

$$\begin{aligned} p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu, \boldsymbol{x}_{1:N}) &= \frac{p(\boldsymbol{x}_{1:N}, \boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)}{p(\boldsymbol{x}_{1:N}|\boldsymbol{\lambda}, \nu)} \\ &= \frac{\prod_{n=1}^N p(\boldsymbol{x}_n|\boldsymbol{\eta})\; p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)}{\int d\boldsymbol{\eta} \prod_{n=1}^N p(\boldsymbol{x}_n|\boldsymbol{\eta})\; p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)} \\ &= \frac{\prod_{n=1}^N g(\boldsymbol{\eta}) h(\boldsymbol{x}_n) \exp(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x}_n)) \cdot f(\boldsymbol{\lambda}, \nu) m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^\nu \exp(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda})}{\int d\boldsymbol{\eta} \prod_{n=1}^N g(\boldsymbol{\eta}) h(\boldsymbol{x}_n) \exp(\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x}_n)) \cdot f(\boldsymbol{\lambda}, \nu) m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^\nu \exp(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda})} \\ &= \frac{[\prod_n h(\boldsymbol{x}_n)]\, f(\boldsymbol{\lambda}, \nu) m(\boldsymbol{\eta})\; g(\boldsymbol{\eta})^{\nu+N} \exp\left(\boldsymbol{\eta}^T\, (\nu \boldsymbol{\lambda} + \sum_n \boldsymbol{u}(\boldsymbol{x}_n))\right)}{[\prod_n h(\boldsymbol{x}_n)]\, f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta}\; m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^{\nu+N} \exp\left(\boldsymbol{\eta}^T\, (\nu \boldsymbol{\lambda} + \sum_n \boldsymbol{u}(\boldsymbol{x}_n))\right)} \\ &= \frac{m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^{\nu+N} \exp\left(\boldsymbol{\eta}^T\, (\nu \boldsymbol{\lambda} + \sum_n \boldsymbol{u}(\boldsymbol{x}_n))\right)}{\int d\boldsymbol{\eta}\; m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^{\nu+N} \exp\left(\boldsymbol{\eta}^T\, (\nu \boldsymbol{\lambda} + \sum_n \boldsymbol{u}(\boldsymbol{x}_n))\right)} \end{aligned} \tag{23}$$

Note that this expression depends on the data $\boldsymbol{x}_{1:N}$ only through N and $\sum_n \boldsymbol{u}(\boldsymbol{x}_n)$. This is why the $\boldsymbol{u}(\boldsymbol{x})$ are called *sufficient statistics*: they contain all the information about $\boldsymbol{\eta}$ which we need from the data to determine the parameter posterior. A similar result holds for maximum-likelihood learning, see [1]. By introducting the

posterior parameters

$$\tilde{\nu} := \nu + N \tag{24}$$

$$\tilde{\boldsymbol{\lambda}} := \frac{\nu\boldsymbol{\lambda} + \sum_n \boldsymbol{u}(\boldsymbol{x}_n)}{\tilde{\nu}} \tag{25}$$

thus, $\nu\boldsymbol{\lambda} + \sum_N \boldsymbol{u}(\boldsymbol{x}_n) = \tilde{\nu}\tilde{\boldsymbol{\lambda}}$. We furthermore identify $f(\boldsymbol{\lambda}, \nu)$ as the normalization constant of the p(oste)rior, i.e.

$$f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu}) = \frac{1}{\int d\boldsymbol{\eta}\, m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\tilde{\nu}} \exp\left(\boldsymbol{\eta}^T \tilde{\nu}\tilde{\boldsymbol{\lambda}}\right)}$$

and finally, plugging these identities back into eqn. 23, we obtain:

$$\boxed{p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu, \boldsymbol{x}_{1:N}) = f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})\, m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\tilde{\nu}} \exp\left(\boldsymbol{\eta}^T \tilde{\nu}\tilde{\boldsymbol{\lambda}}\right) = p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}}, \tilde{\nu}).} \tag{26}$$

In other words, given an exponential family observation model, i.i.d. data and a conjugate prior, we obtain posterior just by replacing the prior parameters according to eqns. 24 and 25. Furthermore, note that

- According to eqn. 24, $\tilde{\nu}$ keeps track of the number of observed datapoints. Since it also contains prior information via ν, it is referred to as a *pseudocount*.
- For large enough N, the posterior is unimodal, and the log-posterior is convex. The width of the maximum is monotonically decreasing in $\tilde{\nu}$ (Can be shown by computing the Hessian of the log-posterior). Hence, $\tilde{\nu}$ is also called the *concentration parameter*.
- The posterior $\tilde{\boldsymbol{\lambda}}$ is just a weighted mean of the prior $\boldsymbol{\lambda}$ and and the observed data.
- These posterior updates can be iterated, i.e. the accumulation of the sufficient statistics can be restarted at any point. The extreme case of updating after every datapoint, i.e. *online learning*, boils down to keeping track of this weighted average datapoint per datapoint.

2.1 Maximum-a-posteriori (MAP) parameter estimates

Instead of working with the full posterior of the natural parameters $\boldsymbol{\eta}$ (eqn. 26), it is sometimes enough to use the parameter values which maximize the posterior, i.e. the numerator of eqn. 26 (the denominator does not depend on $\boldsymbol{\eta}$ after the integration). Setting the derivative of the log of the numerator to zero, we find

$$\begin{aligned} \nabla_{\boldsymbol{\eta}} \log(m(\boldsymbol{\eta})) + (\nu + N) \underbrace{\nabla_{\boldsymbol{\eta}} \log(g(\boldsymbol{\eta}))}_{-\langle \boldsymbol{u}(\boldsymbol{x}) \rangle, \text{eqn. 4}} + (\nu\boldsymbol{\lambda} + \sum_n \boldsymbol{u}(\boldsymbol{x}_n)) &\stackrel{!}{=} 0 \\ \Rightarrow \frac{\nu\boldsymbol{\lambda} + \sum_n \boldsymbol{u}(\boldsymbol{x}_n)}{N + \nu} + \frac{\nabla_{\boldsymbol{\eta}} \log(m(\boldsymbol{\eta}))}{N + \nu} &\stackrel{!}{=} \langle \boldsymbol{u}(\boldsymbol{x}) \rangle \end{aligned} \tag{27}$$

i.e. the posterior maximum is located at a point where the expected value of the natural parameters is equal to the quotient of the posterior parameters (eqns. 24,25) plus a term depending on $m(\boldsymbol{\eta})$. The latter is often zero, since $m(\boldsymbol{\eta}) = 1$ for many distributions (see tables in section 4).

2.2 Expectations

Computing parameter expectations (i.e. of $\boldsymbol{\eta}$ and functions thereof) of a conjugate p(oste)rior can be done similar to the expectations of an exponential family distribution. From the normalization equation

$$f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right) = 1 \tag{28}$$

follows

$$\begin{aligned}
\nabla_{\boldsymbol{\lambda}} \left[f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right) \right] &= 0 \\
\Rightarrow \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}, \nu) \underbrace{\int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right)}_{f(\boldsymbol{\lambda},\nu)^{-1}} &= -f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right) \nu \boldsymbol{\eta} \\
\Rightarrow \frac{\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}, \nu)}{f(\boldsymbol{\lambda}, \nu)} &= -\nu \langle \boldsymbol{\eta} \rangle
\end{aligned} \tag{29}$$

and thus

$$\boxed{\langle \boldsymbol{\eta} \rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda}, \nu))}{\nu}} \tag{30}$$

Likewise, from the derivative w.r.t. ν we find, noting that $\frac{\partial g(\boldsymbol{\eta})^{\nu}}{\partial \nu} = \log(g(\boldsymbol{\eta})) g(\boldsymbol{\eta})^{\nu}$:

$$\boxed{\langle \log(g(\boldsymbol{\eta})) \rangle + \boldsymbol{\lambda}^T \langle \boldsymbol{\eta} \rangle = -\frac{\partial \log(f(\boldsymbol{\lambda}, \nu))}{\partial \nu}} \tag{31}$$

For the second moments, take the derivatives of $\langle \boldsymbol{\eta} \rangle$ (see eqn. 29):

$$\begin{aligned}
\frac{\partial \langle \boldsymbol{\eta} \rangle}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j} \left[f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right) \boldsymbol{\eta} \right] \\
&= \frac{\partial f(\boldsymbol{\lambda}, \nu)}{\partial \lambda_j} \underbrace{\int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right) \boldsymbol{\eta}}_{= \frac{\langle \boldsymbol{\eta} \rangle}{f(\boldsymbol{\lambda}, \nu)}} \\
&\quad + f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right) \nu \boldsymbol{\eta} \eta_j \\
&= \frac{\partial \log(f(\lambda, \nu))}{\partial \lambda_j} \langle \boldsymbol{\eta} \rangle + \nu \langle \boldsymbol{\eta} \eta_j \rangle \\
&= -\nu \langle \eta_j \rangle \langle \boldsymbol{\eta} \rangle + \nu \langle \boldsymbol{\eta} \eta_j \rangle \qquad (32) \\
\Rightarrow \frac{1}{\nu} \frac{\partial \langle \eta_k \rangle}{\partial \lambda_j} &= \langle \eta_j \eta_k \rangle - \langle \eta_j \rangle \langle \eta_k \rangle \qquad (33)
\end{aligned}$$

and thus

$$\boxed{\mathrm{Cov}(\boldsymbol{\eta}) = -\frac{\nabla_{\boldsymbol{\lambda}}\nabla_{\boldsymbol{\lambda}}\log(f(\boldsymbol{\lambda},\nu))}{\nu^2}} \tag{34}$$

Likewise, computing the derivative of eqn. 31 on both sides yields

$$\frac{\partial\langle\log(g(\boldsymbol{\eta}))\rangle}{\partial\nu} = -\frac{\partial^2\log\left(f(\boldsymbol{\lambda},\nu)\right))}{\partial\nu^2} - \boldsymbol{\lambda}^T\frac{\partial\langle\boldsymbol{\eta}\rangle}{\partial\nu} \tag{35}$$

where

$$\begin{aligned}
\frac{\partial\langle\log(g(\boldsymbol{\eta}))\rangle}{\partial\nu} &= \frac{\partial\log(f(\boldsymbol{\lambda},\nu))}{\partial\nu}\langle\log(g(\boldsymbol{\eta}))\rangle + \left\langle\log(g(\boldsymbol{\eta}))^2\right\rangle + \boldsymbol{\lambda}^T\left\langle\log(g(\boldsymbol{\eta}))\boldsymbol{\eta}\right\rangle \\
&= \left\langle\log(g(\boldsymbol{\eta}))^2\right\rangle - \langle\log(g(\boldsymbol{\eta}))\rangle^2 + \boldsymbol{\lambda}^T\left\langle\log(g(\boldsymbol{\eta}))\boldsymbol{\eta}\right\rangle - \boldsymbol{\lambda}^T\left\langle\log(g(\boldsymbol{\eta}))\right\rangle\langle\boldsymbol{\eta}\rangle \\
&= \mathrm{Var}(\log(g(\boldsymbol{\eta}))) + \mathrm{Cov}(\log(g(\boldsymbol{\eta})), \boldsymbol{\lambda}^T\boldsymbol{\eta}) \\
\frac{\partial\langle\eta_i\rangle}{\partial\nu} &= \frac{\partial\log(f(\lambda,\nu))}{\partial\nu}\langle\eta_i\rangle + \langle\eta_i\log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle\boldsymbol{\eta}\eta_i\rangle \qquad (36) \\
&= \langle\eta_i\log(g(\boldsymbol{\eta}))\rangle - \langle\eta_i\rangle\langle\log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle\boldsymbol{\eta}\eta_i\rangle - \boldsymbol{\lambda}^T\langle\boldsymbol{\eta}\rangle\langle\eta_i\rangle \\
&= \mathrm{Cov}(\log(g(\boldsymbol{\eta})), \eta_i) + \mathrm{Cov}(\boldsymbol{\lambda}^T\boldsymbol{\eta}, \eta_i) \\
\Rightarrow \boldsymbol{\lambda}^T\frac{\partial\langle\boldsymbol{\eta}\rangle}{\partial\nu} &= \mathrm{Cov}(\log(g(\boldsymbol{\eta})), \boldsymbol{\lambda}^T\boldsymbol{\eta}) + \mathrm{Var}(\boldsymbol{\lambda}^T\boldsymbol{\eta}) \qquad (37)
\end{aligned}$$

and thus, noting that $\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y) + 2\,\mathrm{Cov}(x,y)$:

$$\boxed{\mathrm{Var}\left(\log(g(\boldsymbol{\eta})) + \boldsymbol{\lambda}^T\boldsymbol{\eta}\right) = -\frac{\partial^2\log\left(f(\boldsymbol{\lambda},\nu)\right))}{\partial\nu^2}} \tag{38}$$

Another expectation that can be computed from the normalization constant $f(\boldsymbol{\lambda},\nu)$ is

$$\langle g(\boldsymbol{\eta})^k\rangle = f(\boldsymbol{\lambda},\nu)\int d\boldsymbol{\eta}\; m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^{\nu+k}\exp\left(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda}\right). \tag{39}$$

To evaluate the integral, choose new parameters $\nu', \boldsymbol{\lambda}'$ such that

$$\begin{aligned}
\nu' &= \nu + k \qquad (40) \\
\nu'\boldsymbol{\lambda}' &= \nu\boldsymbol{\lambda} \Rightarrow \boldsymbol{\lambda}' = \frac{\nu}{\nu'}\boldsymbol{\lambda}. \qquad (41)
\end{aligned}$$

With these parameters, the integral is in exponential family normal form, and thus

$$\langle g(\boldsymbol{\eta})^k\rangle = \frac{f(\boldsymbol{\lambda},\nu)}{f(\boldsymbol{\lambda}',\nu')} \tag{42}$$

$$\boxed{\langle g(\boldsymbol{\eta})\rangle = \frac{f(\boldsymbol{\lambda},\nu)}{f(\frac{\nu}{\nu+1}\boldsymbol{\lambda},\nu+1)}} \tag{43}$$

$$\boxed{\mathrm{Var}(g(\boldsymbol{\eta})) = \frac{f(\boldsymbol{\lambda},\nu)}{f(\frac{\nu}{\nu+2}\boldsymbol{\lambda},\nu+2)} - \left[\frac{f(\boldsymbol{\lambda},\nu)}{f(\frac{\nu}{\nu+1}\boldsymbol{\lambda},\nu+1)}\right]^2} \tag{44}$$

2.3 Predictive distribution, entropy and log-likelihood

Let $\alpha > 0$. The predictive distribution and related quantities can be derived from the integral

$$\begin{aligned}
\int d\boldsymbol{\eta}\, p(\boldsymbol{x}|\boldsymbol{\eta})^\alpha p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) &= h(\boldsymbol{x})^\alpha f(\boldsymbol{\lambda},\nu) \int d\boldsymbol{\eta}\, g(\boldsymbol{\eta})^\alpha \exp\left(\alpha \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right) m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^\nu \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right) \\
&= h(\boldsymbol{x})^\alpha f(\boldsymbol{\lambda},\nu) \int d\boldsymbol{\eta}\, m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^{\nu+\alpha} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda} + \alpha \boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right) \\
&= h(\boldsymbol{x})^\alpha f(\boldsymbol{\lambda},\nu) \int d\boldsymbol{\eta}\, m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^{\nu+\alpha} \exp\left((\nu+\alpha) \boldsymbol{\eta}^T \frac{\nu\boldsymbol{\lambda} + \alpha \boldsymbol{u}(\boldsymbol{x})}{\nu+\alpha}\right) \\
\langle p(\boldsymbol{x}|\boldsymbol{\eta})^\alpha) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} &= h(\boldsymbol{x})^\alpha f(\boldsymbol{\lambda},\nu) \frac{1}{f\left(\frac{\nu\boldsymbol{\lambda}+\alpha\boldsymbol{u}(\boldsymbol{x})}{\nu+\alpha}, \nu+\alpha\right)} \qquad (45)
\end{aligned}$$

where the last line follows from the normalization equation 28. For $\alpha = 1$, the integral on the l.h.s. is the expectation of $p(\boldsymbol{x}|\boldsymbol{\eta})$ under the prior:

$$\boxed{p(\boldsymbol{x}|\boldsymbol{\lambda},\nu) = h(\boldsymbol{x}) \frac{f(\boldsymbol{\lambda},\nu)}{f\left(\frac{\nu\boldsymbol{\lambda}+\boldsymbol{u}(\boldsymbol{x})}{\nu+1}, \nu+1\right)}} \qquad (46)$$

Differentiating with respect to α yields:

$$\begin{aligned}
\frac{\partial}{\partial\alpha} \langle p(\boldsymbol{x}|\boldsymbol{\eta})^\alpha) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} &= \langle \log(p(\boldsymbol{x}|\boldsymbol{\eta})) p(\boldsymbol{x}|\boldsymbol{\eta})^\alpha \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} \\
&= \log(h(\boldsymbol{x}))\, h(\boldsymbol{x})^\alpha \frac{f(\boldsymbol{\lambda},\nu)}{f\left(\frac{\nu\boldsymbol{\lambda}+\alpha\boldsymbol{u}(\boldsymbol{x})}{\nu+\alpha}, \nu+\alpha\right)} \\
-\frac{h(x)^\alpha f(\boldsymbol{\lambda},\nu)}{f\left(\frac{\nu\boldsymbol{\lambda}+\alpha\boldsymbol{u}(\boldsymbol{x})}{\nu+\alpha}, \nu+\alpha\right)^2} &\cdot \left[\frac{\nu}{(\nu+\alpha)^2} \nabla_{\boldsymbol{\lambda}'} f(\boldsymbol{\lambda}',\nu')(\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{\lambda}) + \frac{\partial f(\boldsymbol{\lambda}',\nu')}{\partial \nu'}\right]
\end{aligned}$$

where the derivatives are evaluated at $\boldsymbol{\lambda}' = \frac{\nu\boldsymbol{\lambda}+\alpha\boldsymbol{u}(\boldsymbol{x})}{\nu+\alpha}$ and $\nu' = \nu+\alpha$. For $\alpha = 0$, we obtain the expected log-likelihood, using eqn. 30:

$$\boxed{\langle \log(p(\boldsymbol{x}|\boldsymbol{\eta})) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} = \log(h(x)) + \langle \boldsymbol{\eta}^T \rangle (\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{\lambda}) - \frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}} \qquad (47)$$

For $\alpha = 1$, we obtain expectations of the form $\langle \log(p(\boldsymbol{x}|\boldsymbol{\eta})) p(\boldsymbol{x}|\boldsymbol{\eta})) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)}$, which are the terms required for the computation of the expected entropy of $\boldsymbol{x}$:

$$\begin{aligned}
\boldsymbol{\lambda}' &= \frac{\nu\boldsymbol{\lambda} + \boldsymbol{u}(\boldsymbol{x})}{\nu+1} \qquad (48) \\
\nu' &= \nu + 1 \qquad (49) \\
\langle \log(p(\boldsymbol{x}|\boldsymbol{\eta})) p(\boldsymbol{x}|\boldsymbol{\eta}) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} &= p(\boldsymbol{x}|\boldsymbol{\lambda},\nu) \left[\log(h(\boldsymbol{x})) + \frac{\nu}{\nu'} \langle \boldsymbol{\eta}^T \rangle_{\boldsymbol{\lambda}',\nu'} (\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{\lambda}) \right. \\
&\quad \left. - \frac{\partial \log(f(\boldsymbol{\lambda}',\nu'))}{\partial \nu'}\right] \qquad (50)
\end{aligned}$$

2.4 Expected sufficient statistics

The expectations of the sufficient statistics $\boldsymbol{u}$ for fixed natural parameters $\boldsymbol{\eta}$ is given by eqn. 4:

$$\langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{\eta})} = -\nabla \log(g(\boldsymbol{\eta})) \tag{51}$$

If η is drawn from a conjugate prior, the expectation of $\boldsymbol{u}(\boldsymbol{x})$ under the prior is given by averaging the r.h.s. over $p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)$,

$$\langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{\eta})p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} = -\langle \nabla \log(g(\boldsymbol{\eta})) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} \tag{52}$$

To compute this expectation, one can use the Divergence Theorem from vector calculus (Ostrogradsky-Gauss). For a differentiable vector field $\mathbf{f}(\boldsymbol{x})$, this theorem states that

$$\int_V d\boldsymbol{x}\ \nabla \cdot \mathbf{f}(\boldsymbol{x}) = \oint_S d\mathbf{s}\ \mathbf{f}(\boldsymbol{x}) \tag{53}$$

where S is the surface enclosing the volume V. As a special case, consider the field $\mathbf{f}(\boldsymbol{x}) = \mathbf{c}z(\boldsymbol{x})$, with $\vec{c}$ a constant vector and $z(\boldsymbol{x})$ a smoothly differentiable scalar function. Then, using $\nabla \cdot \mathbf{c}z(\boldsymbol{x}) = \sum_i \frac{\partial c_i z(\boldsymbol{x})}{\partial x_i} = \mathbf{c}^T \nabla z(\boldsymbol{x})$ we find

$$\mathbf{c}^T \int_V d\boldsymbol{x}\ \nabla z(\boldsymbol{x}) = \mathbf{c}^T \oint_S d\mathbf{s}\ \mathbf{z}(\boldsymbol{x}) \tag{54}$$

and since this holds for any $\mathbf{c}$, it follows that

$$\int_V d\boldsymbol{x}\ \nabla \vec{z}(\boldsymbol{x}) = \oint_S d\mathbf{s}\ \mathbf{z}(\boldsymbol{x}) \tag{55}$$

This identity can be used to compute the expectation on the r.h.s. of eqn. 52 by integrating the gradient of $p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)$:

$$\begin{aligned}
\int d\boldsymbol{\eta} \nabla_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu) &= f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta} \nabla_{\boldsymbol{\eta}} \left(m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^{\nu} \exp\left(\nu \boldsymbol{\eta}^T \boldsymbol{\lambda}\right)\right) \\
&= f(\boldsymbol{\lambda}, \nu) \int d\boldsymbol{\eta}\ \left(\frac{\nabla_{\boldsymbol{\eta}} m(\boldsymbol{\eta})}{m(\boldsymbol{\eta})} + \nu \frac{\nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} + \nu\lambda \right) p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu) \\
&= \langle \nabla_{\boldsymbol{\eta}} \log(m(\boldsymbol{\eta})) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} + \nu \langle \nabla_{\boldsymbol{\eta}} \log(g(\boldsymbol{\eta})) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} + \nu\boldsymbol{\lambda} \\
&= \oint_S d\mathbf{s}\ p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)
\end{aligned} \tag{56}$$

where the surface S encloses the range of $\boldsymbol{\eta}$. Hence, the expectation is:

$$\boxed{\langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{\eta})p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} = \boldsymbol{\lambda} + \frac{\langle \nabla_{\boldsymbol{\eta}} \log(m(\boldsymbol{\eta})) \rangle_{p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} - \oint d\mathbf{s}\ p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)}{\nu}} \tag{57}$$

If $\boldsymbol{\eta} \in \mathbb{R}^D$ with no further constraints, then $p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu) \to 0$ on the surface of the range of $\boldsymbol{\eta}$, since $p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu)$ has to be normalizable. Hence, the surface integral must be zero. This is e.g. the case for the Multinomial distribution and the Poisson distribution. Furthermore, if $m(\boldsymbol{\eta}) = \text{const.}$, then the gradient vanishes (Dirichlet,Gamma, Stick-breaking for $c_i = 0$, see tables in appendix). In those cases, the above expression simplifies to

$$\langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{\eta})p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} = \boldsymbol{\lambda} \tag{58}$$

The expectation for the multivariate Gaussian is also computable, see appendix.

2.5 Maximum likelihood

For maximum-likelihood approximations, we need the gradient of $\log(p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu))$ w.r.t. $\boldsymbol{\lambda}$ and ν:

$$\begin{aligned}
\nabla_{\boldsymbol{\lambda}} \log(p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) &= \nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu)) + \nu\boldsymbol{\eta} \\
&= -\nu\langle\boldsymbol{\eta}\rangle + \nu\boldsymbol{\eta} = \nu(\boldsymbol{\eta} - \langle\boldsymbol{\eta}\rangle) \qquad (59) \\
\frac{\partial}{\partial\nu} \log(p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) &= \frac{\partial f(\boldsymbol{\lambda},\nu)}{\partial\nu} + \log(g(\boldsymbol{\eta})) + \boldsymbol{\eta}^T\boldsymbol{\lambda} \\
&= (\log(g(\boldsymbol{\eta})) - \langle\log(g(\boldsymbol{\eta}))\rangle) + \boldsymbol{\lambda}^T(\boldsymbol{\eta} - \langle\boldsymbol{\eta}\rangle) \qquad (60)
\end{aligned}$$

Similar to the gradient of the exponential family distributions, this gradient points towards the actual value of $\boldsymbol{\eta}$ and away from the expectation.

2.6 Entropy

The differential entropy (not the conditional entropy) of $\boldsymbol{\eta}$ given $\boldsymbol{\lambda}$ and ν is

$$\begin{aligned}
H(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) &= -f(\boldsymbol{\lambda},\nu) \int d\boldsymbol{\eta}\, m(\boldsymbol{\eta})\, g(\boldsymbol{\eta})^\nu \exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda}) \left[\log(f(\boldsymbol{\lambda},\nu)) + \log(m(\boldsymbol{\eta})) + \nu\log(g(\boldsymbol{\eta})) + \nu\boldsymbol{\eta}^T\boldsymbol{\lambda}\right] \\
&= -\log(f(\boldsymbol{\lambda},\nu)) - \langle\log(m(\boldsymbol{\eta}))\rangle - \nu\left[\langle\log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle\boldsymbol{\eta}\rangle\right] \qquad (61)
\end{aligned}$$

where the expectations are w.r.t. the p(oste)rior eqn. 22. Using eqn. 31, this can be rewritten as

$$\boxed{H(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = -\log(f(\boldsymbol{\lambda},\nu)) - \langle\log(m(\boldsymbol{\eta}))\rangle + \nu\frac{\partial\log(f(\boldsymbol{\lambda},\nu))}{\partial\nu}} \tag{62}$$

The derivates of this entropy are therefore:

$$\boxed{\frac{\partial H(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)}{\partial\nu} = \nu\frac{\partial^2\log\left(f(\boldsymbol{\lambda},\nu)\right)}{\partial\nu^2} - \frac{\partial\langle\log(m(\boldsymbol{\eta}))\rangle}{\partial\nu}} \tag{63}$$

and (using eqn. 36 and 30):

$$\begin{aligned}
\nabla_{\boldsymbol{\lambda}} H(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) &= -\nabla_{\boldsymbol{\lambda}}\log(f(\boldsymbol{\lambda},\nu)) + \nu\nabla_{\boldsymbol{\lambda}}\frac{\partial\log(f(\boldsymbol{\lambda},\nu))}{\partial\nu} - \nabla_{\boldsymbol{\lambda}}\langle\log(m(\boldsymbol{\eta}))\rangle \\
&= \nu\langle\boldsymbol{\eta}\rangle + \nu\frac{\partial}{\partial\nu}\nabla_\lambda\log(f(\boldsymbol{\lambda},\nu)) - \nabla_{\boldsymbol{\lambda}}\langle\log(m(\boldsymbol{\eta}))\rangle \\
&= \nu\langle\boldsymbol{\eta}\rangle - \nu\frac{\partial}{\partial\nu}\nu\langle\boldsymbol{\eta}\rangle - \nabla_{\boldsymbol{\lambda}}\langle\log(m(\boldsymbol{\eta}))\rangle \\
&= \nu\langle\boldsymbol{\eta}\rangle - \nu\langle\boldsymbol{\eta}\rangle - \nu^2\frac{\partial\langle\boldsymbol{\eta}\rangle}{\partial\nu} - \nabla_{\boldsymbol{\lambda}}\langle\log(m(\boldsymbol{\eta}))\rangle \\
&= -\nu^2\left[\mathrm{Cov}(\log(g(\eta)),\boldsymbol{\eta}) + \mathrm{Cov}(\lambda^T\boldsymbol{\eta},\boldsymbol{\eta})\right] - \nabla_{\boldsymbol{\lambda}}\langle\log(m(\boldsymbol{\eta}))\rangle
\end{aligned}$$

$$\boxed{\nabla_{\boldsymbol{\lambda}} H(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = -\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu)) + \nu \nabla_{\boldsymbol{\lambda}} \frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu} - \nabla_{\boldsymbol{\lambda}} \langle \log(m(\boldsymbol{\eta})) \rangle} \tag{64}$$

$$\boxed{\nabla_{\boldsymbol{\lambda}} H(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = -\nu^2 \frac{\partial \langle \boldsymbol{\eta} \rangle}{\partial \nu} - \nabla_{\boldsymbol{\lambda}} \langle \log(m(\boldsymbol{\eta})) \rangle} \tag{65}$$

2.7 Kullback-Leibler divergence

The KL-divergence of a distribution with parameters $\tilde{\boldsymbol{\lambda}}, \tilde{\nu}$ to a distribution with parameters $\boldsymbol{\lambda}, \nu$ is given by

$$\begin{aligned} D(p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})||p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) &= \int d\boldsymbol{\eta}\, p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu}) \left[\log(p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})) - \log(p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) \right] \\ &= -H(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu}) - \int d\boldsymbol{\eta}\, p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu}) \log(p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) \end{aligned} \tag{66}$$

The second term on the r.h.s. is given by (expectations w.r.t $p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})$)

$$\int d\boldsymbol{\eta}\, p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu}) \log(p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) = \log(f(\boldsymbol{\lambda},\nu)) + \langle \log(m(\boldsymbol{\eta})) \rangle + \nu \langle \log(g(\boldsymbol{\eta})) \rangle + \nu \boldsymbol{\lambda} \langle \boldsymbol{\eta} \rangle \tag{67}$$

and thus, using eqn. 61 and 31 we find

$$\boxed{D(p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})||p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu}-\nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu \langle \boldsymbol{\eta} \rangle} \tag{68}$$

The derivatives are:

$$\begin{aligned} \nabla_{\tilde{\boldsymbol{\lambda}}} D(p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})||p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) &= \nabla_{\tilde{\boldsymbol{\lambda}}} \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})) - (\tilde{\nu}-\nu)\frac{\partial}{\partial \tilde{\nu}} \nabla_{\tilde{\boldsymbol{\lambda}}} \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})) \\ &\quad + \nu \langle \boldsymbol{\eta} \rangle - \frac{\nu}{\tilde{\nu}} (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T) \nabla_{\tilde{\boldsymbol{\lambda}}} \nabla_{\tilde{\boldsymbol{\lambda}}} \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})) \\ &= -\tilde{\nu} \langle \boldsymbol{\eta} \rangle + (\tilde{\nu}-\nu)\frac{\partial}{\partial \tilde{\nu}}\, \tilde{\nu} \langle \boldsymbol{\eta} \rangle \\ &\quad + \nu \langle \boldsymbol{\eta} \rangle - \frac{\nu}{\tilde{\nu}} (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T) \nabla^2_{\tilde{\boldsymbol{\lambda}}} \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})) \\ &= \tilde{\nu}(\tilde{\nu}-\nu)\frac{\partial \langle \boldsymbol{\eta} \rangle}{\partial \tilde{\nu}} - \frac{\nu}{\tilde{\nu}} (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T) \nabla^2_{\tilde{\boldsymbol{\lambda}}} \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})) \\ &= \tilde{\nu}(\tilde{\nu}-\nu)\frac{\partial \langle \boldsymbol{\eta} \rangle}{\partial \tilde{\nu}} + \nu (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T) \nabla_{\tilde{\boldsymbol{\lambda}}} \langle \boldsymbol{\eta} \rangle \end{aligned} \tag{69}$$

$$\boxed{\nabla_{\tilde{\boldsymbol{\lambda}}} D(p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})||p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) = \tilde{\nu}(\tilde{\nu}-\nu)\frac{\partial \langle \boldsymbol{\eta} \rangle}{\partial \tilde{\nu}} + \nu (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T) \nabla_{\tilde{\boldsymbol{\lambda}}} \langle \boldsymbol{\eta} \rangle} \tag{70}$$

where $\nabla_{\tilde{\boldsymbol{\lambda}}} \langle \boldsymbol{\eta} \rangle$ is a matrix with entries $\left(\nabla_{\tilde{\boldsymbol{\lambda}}} \langle \boldsymbol{\eta} \rangle\right)_{i,j} = \frac{\partial \langle \boldsymbol{\eta} \rangle_i}{\partial \tilde{\lambda}_j}$. Likewise,

$$\boxed{\frac{\partial}{\partial \tilde{\nu}} D(p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})||p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) = -(\tilde{\nu}-\nu)\frac{\partial^2 f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{\partial \tilde{\nu}^2} - \nu (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\frac{\partial \langle \tilde{\boldsymbol{\eta}} \rangle}{\partial \tilde{\nu}}} \tag{71}$$

3 Variational approximations with exponential family distributions

3.1 Variational inference with conjugate p(oste)riors

In variational inference, we replace an intractable distribution (or density) $q(\boldsymbol{X}|\boldsymbol{d})$ (i.e. one where marginals and conditionals are hard to compute) with a tractable, factorized approximation $q(\boldsymbol{X})$. $\boldsymbol{d}$ is the observed data. Strictly speaking, $q(\boldsymbol{X}) = q(\boldsymbol{X}|\boldsymbol{d})$ but it is customary to omit writing this conditioning, since it is only approximate. The approximation is linked to the correct distribution via the variational bound also 'evidence lower bound (ELBO)':

$$\begin{aligned}
\log(q(\boldsymbol{d})) &= \log\left(\sum_{\boldsymbol{x}} q(\boldsymbol{d},\boldsymbol{x})\right) = \log\left(\sum_{\boldsymbol{x}} q(\boldsymbol{x}) p(\boldsymbol{d}|\boldsymbol{x})\frac{p(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \\
&\geq \int_{\boldsymbol{x}} q(\boldsymbol{x}) \log\left(p(\boldsymbol{d}|\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) = \langle \log p(\boldsymbol{d}|\boldsymbol{x})\rangle_{q(\boldsymbol{X})} - D(q(\boldsymbol{X})||p(\boldsymbol{X})) \\
\Rightarrow L(q,\boldsymbol{d}) &= \langle \log p(\boldsymbol{d}|\boldsymbol{x})\rangle_{q(\boldsymbol{X})} - D(q(\boldsymbol{X})||p(\boldsymbol{X})) \leq \log(p(\boldsymbol{d})) \qquad (72)
\end{aligned}$$

where the second line follows from Jensen's inequality for convex functions and the definition of the Kullback-Leibler divergence. $\mathcal{L}(q,\boldsymbol{d})$ is a lower bound on the log-marginal-likelihood, which we try to maximize w.r.t. $q(\boldsymbol{x})$. The resulting $q(\boldsymbol{x})$ is an approximate version of the correct posterior $p(\boldsymbol{x}|\boldsymbol{d})$ which will be exact iff $p(\boldsymbol{x}|\boldsymbol{d})$ is contained in the class of distributions which can be modeled by $q(\boldsymbol{x})$. In that case (and only in that case), the bound will be tight.

In the following, we will derive the posterior update rules for the case $p(\boldsymbol{X})$ is conjugate to $q(\boldsymbol{d}|\boldsymbol{X})$ and both are in the exponential family. We also assume that $q(\boldsymbol{X})$ is conjugate to the likelihood, such that posterior updated reduce to parameter updates, like in section 2. Lastly, assume that the data are comprised of N i.i.d. observations, i.e. $p(\boldsymbol{d}|\boldsymbol{X}) = \prod_{i=1}^{N} p(\boldsymbol{d}_i|\boldsymbol{X})$. We use a generalized version of the ELBO, which has an inverse temperature parameter $\beta \geq 0$:

$$L(q,\boldsymbol{d}) = \langle \log p(\boldsymbol{d}|\boldsymbol{x})\rangle_{q(\boldsymbol{X})} - \beta D(q(\boldsymbol{X})||p(\boldsymbol{X})) \leq \log(P(\boldsymbol{d})) \qquad (73)$$

$\beta \neq 1$ can be used to model deviations from optimal inference, or for stochastic updating in minibatches etc.. Denote the prior parameters with $\nu, \boldsymbol{\lambda}$ and the posterior parameters with $\tilde{\nu}, \tilde{\boldsymbol{\lambda}}$.

The expected log-likelihood under the posterior $\langle \log p(\boldsymbol{d}|\boldsymbol{x})\rangle_{q(\boldsymbol{X})}$ for N datapoints can then be computed from eqn. (47):

$$\langle \log(p(\boldsymbol{d}|\boldsymbol{\eta}))\rangle_{q(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})} = \sum_{i=1}^{N} \log(h(\boldsymbol{d}_i)) + \langle \boldsymbol{\eta}^T\rangle (\sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{d}_i) - N\,\tilde{\boldsymbol{\lambda}}) - N\,\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} \qquad (74)$$

where $\langle \boldsymbol{\eta}^T\rangle$ is given by eqn. (30) and the KL-divergence $D(q(\boldsymbol{X})||p(\boldsymbol{X}))$ by eqn. (68). To maximize eqn. (73) w.r.t. the posterior parameters $\tilde{\boldsymbol{\lambda}}, \tilde{\nu}$, we will rewrite the elbo as difference between one part that does not depend on the posterior parameters, and a KL-divergence between the posterior, and a distribution in the same exponential family as the posterior that depends on parameters $\boldsymbol{\lambda}', \nu'$ Because the KL-divergence is zero exactly if the two distributions that enter it

are pointwise equal, we can then compute the maximal ELBO by setting $\tilde{\boldsymbol{\lambda}} = \lambda'$ and $\nu' = \tilde{\nu}$.

$$
\begin{aligned}
L(q, \boldsymbol{d}) &= \langle \log p(\boldsymbol{d}|\boldsymbol{x}) \rangle_{q(\boldsymbol{X})} - \beta D(q(\boldsymbol{X})||p(\boldsymbol{X})) \\
&= \sum_{i=1}^{N} \log(h(\boldsymbol{d}_i)) + \langle \boldsymbol{\eta}^T \rangle \left(\sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{d}_i) - N\, \tilde{\boldsymbol{\lambda}} \right) - N\, \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu}))}{\partial \tilde{\nu}} \\
&- \beta \left[\log \left(\frac{f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})}{f(\boldsymbol{\lambda}, \nu)} \right) - (\tilde{\nu} - \nu) \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T) \nu \langle \boldsymbol{\eta} \rangle \right] \\
&= \sum_{i=1}^{N} \log(h(\boldsymbol{d}_i)) - \beta\, \log \left(\frac{f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})}{f(\boldsymbol{\lambda}, \nu)} \right) \\
&+ \langle \boldsymbol{\eta}^T \rangle \left(\sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{d}_i) + \beta \nu \boldsymbol{\lambda} - (N + \beta \nu) \tilde{\boldsymbol{\lambda}} \right) \\
&- (N + \beta \nu - \beta \tilde{\nu}) \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu}))}{\partial \tilde{\nu}} \qquad (75)
\end{aligned}
$$

Define

Posterior parameters for β-variational update

$$\nu' = \frac{N}{\beta} + \nu \qquad (76)$$

$$\boldsymbol{\lambda}' = \frac{\sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{d}_i) + \beta \nu \tilde{\boldsymbol{\lambda}}}{N + \beta \nu} = \frac{\nu \boldsymbol{\lambda} + \sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{d}_i)/\beta}{\nu'} \qquad (77)$$

and note the similarity of these definitions with the exact posterior updates eqn. (26) – all data-related quantities have been divided by β. Collecting terms in the ELBO eqn. (75) and plugging in these definitions, we find

$$
\begin{aligned}
L(q, \boldsymbol{d}) &= \sum_{i=1}^{N} \log(h(\boldsymbol{d}_i)) - \beta\, \log \left(\frac{f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})}{f(\boldsymbol{\lambda}, \nu)} \right) \\
&+ \beta \nu' \langle \boldsymbol{\eta}^T \rangle \left(\boldsymbol{\lambda}' - \tilde{\boldsymbol{\lambda}} \right) \\
&- \beta (\tilde{\nu} - \nu') \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu}))}{\partial \tilde{\nu}} \qquad (78)
\end{aligned}
$$

Comparing the last two lines to the expression for the KL divergence eqn. (68), we find that up to a $\log \left(\frac{f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})}{f(\boldsymbol{\lambda}', \nu')} \right)$ term and a factor β, these lines are a KL-divergence. Inserting and subtracting this term, we find

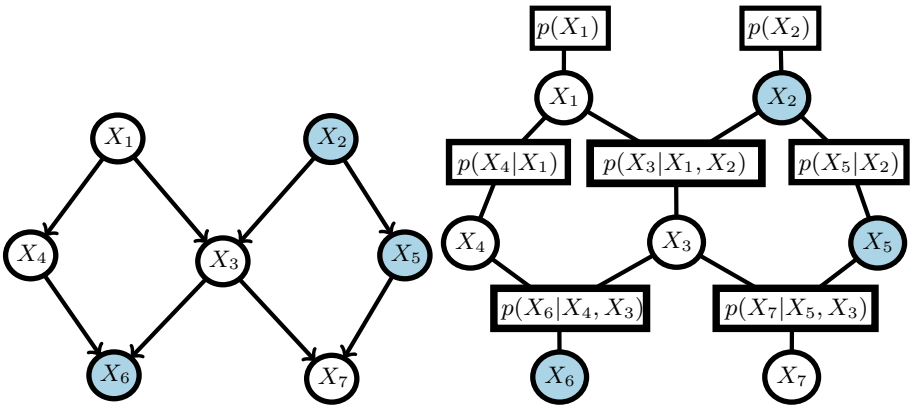

Figure 1: **Left:** Bayesian network with (undirected) loops. Exact sum-product can not be run on this graph. **Right**: corresponding factor graph. For variational message passing, a node (e.g. X_3) needs to receive messages from all members of its Markov blanket, which are the variables connected to neighbouring factors (for X_3: all other nodes). through the connecting factors.

$$
\begin{aligned}
L(q, \boldsymbol{d}) &= \sum_{i=1}^{N} \log(h(\boldsymbol{d}_i)) - \beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})}{f(\boldsymbol{\lambda}, \nu)}\right) + \beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})}{f(\boldsymbol{\lambda}', \nu')}\right) \\
&- \beta D(q(\boldsymbol{X}|\tilde{\boldsymbol{\lambda}}, \tilde{\nu})||p(\boldsymbol{X}|\boldsymbol{\lambda}', \nu')) \\
&= \sum_{i=1}^{N} \log(h(\boldsymbol{d}_i)) - \beta \log\left(\frac{f(\boldsymbol{\lambda}', \nu')}{f(\lambda, \nu)}\right) \qquad (79) \\
&- \beta D(q(\boldsymbol{X}|\tilde{\boldsymbol{\lambda}}, \tilde{\nu})||p(\boldsymbol{X}|\boldsymbol{\lambda}', \nu')) \qquad (80)
\end{aligned}
$$

which is maximal if $D(q(\boldsymbol{X}|\tilde{\boldsymbol{\lambda}}, \tilde{\nu})||p(\boldsymbol{X}|\boldsymbol{\lambda}', \nu')) = 0$, which happens if and only if $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}'$ and $\nu' = \tilde{\nu}$. Thus, the maximal ELBO after the posterior update is given by eqn. (79), which is achieved for the parameters given in eqns. (76).

Now we derive an expression for the expected log-likelihood eqn. (74) that depends only on the p(oste)rior parameters. This is possible because these parameters are computed from the sufficient statistics (which are by definition sufficient to determine the likelihood). Rewrite the posterior parameters as $N = \beta(\nu' - \nu)$ and $\beta(\boldsymbol{\lambda}'\nu' - \boldsymbol{\lambda}\nu)$ and substitute these expressions into the log-likelihood, then

$$
\langle \log(p(\boldsymbol{d}|\boldsymbol{\eta})) \rangle_{q(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}}, \tilde{\nu})} = \sum_{i=1}^{N} \log(h(\boldsymbol{d}_i)) + \beta\nu \langle \boldsymbol{\eta}^T \rangle (\boldsymbol{\lambda}' - \boldsymbol{\lambda}) - \beta(\nu' - \nu) \frac{\partial \log(f(\boldsymbol{\lambda}', \nu'))}{\partial \nu'} \qquad (81)
$$

3.2 Variational message passing

Denote the set of indexes of latent variables by $\mathbf{L}$, the set of indexes of observed variables with $\mathbf{O}$, and the set of all indexes by $\mathbf{N}$ such that $\mathbf{L} \cap \mathbf{O} = \emptyset$ and

$\mathbf{L} \cup \mathbf{O} = \mathbf{N}$. We consider a fully factorized approximation, i.e. one where the density of the latent variables

$$Q(\boldsymbol{x}) = \prod_{i \in \mathbf{L}} Q_i(x_i) = \prod_{i \in \mathbf{L}} Q(x_i)$$

is a product over distributions of individual variables. Strictly speaking, the notation in the middle is correct because there is one density per variable. We omit the extra index and assume the reader knows what is meant. Let the correct density be expressed as a Bayes net,

$$P(\boldsymbol{d}, \boldsymbol{x}) = \prod_{j \in \mathbf{O}} P(d_j | \mathrm{pa}_{X_j}) \prod_{i \in \mathbf{L}} P(x_i | \mathrm{pa}_{X_i})$$

and d_j is the observed data at node X_j. Furthermore, to simplify notation, we introduce the "sum-product" symbol:

$$\sum_{i \in \mathbf{K}} \prod Q(x_i) := \sum_{x_i : i \in \mathbf{K}} \prod_{i \in \mathbf{K}} Q(x_i) \tag{82}$$

The bound then is:

$$\begin{aligned}
L(Q, \boldsymbol{d}) &= \sum_{\boldsymbol{x}} \prod_{i \in \mathbf{L}} Q(x_i) \left[\sum_{j \in \mathbf{O}} \log(P(d_j | \mathrm{pa}_{X_j})) + \sum_{k \in \mathbf{L}} \log(P(x_k | \mathrm{pa}_{X_k})) - \sum_{k \in \mathbf{L}} \log(Q(x_k)) \right] \\
&= \sum_{\boldsymbol{x}} \prod_{i \in \mathbf{L}} Q(x_i) \left[\sum_{j \in \mathbf{O}} \log(P(d_j | \mathrm{pa}_{X_j})) + \sum_{k \in \mathbf{L}} \log(P(x_k | \mathrm{pa}_{X_k})) \right] \\
&\quad - \sum_{i \in \mathbf{L}} \sum_{x_i} Q(x_i) \log(Q(x_i)) \\
&= \sum_{j \in \mathbf{O}} \sum_{i \in \mathbf{L} \cap \mathrm{pa}_{X_j}} \prod Q(x_i) \log(p(d_j | \mathrm{pa}_{X_j})) \\
&\quad + \sum_{k \in \mathbf{L}} \sum_{i \in \mathbf{L} \cap \mathrm{pa}_{X_k} \cup \{k\}} \prod Q(x_i) \log(P(x_k | \mathrm{pa}_{X_k})) \\
&\quad - \sum_{i \in \mathbf{L}} \sum_{x_i} Q(x_i) \log(Q(x_i))
\end{aligned} \tag{83}$$

To find the $Q(\boldsymbol{x})$ that maximizes the bound, we take the derivative w.r.t. $q(\boldsymbol{x})$ and set it to zero. This is a necessary condition for a maximum, it can be shown that it is sufficient, too. We furthermore impose the constraint that all $q(x_i)$ have to be distributions, i.e. $q(x_i) \geq 0$ and $\sum_{x_i} q(x_i) = 1$. It will turn out that we do not have to impose the first constraint, we do however need to make sure the second one is fulfilled. This can be achieved by a Lagrange multiplier for each distribution. The Lagrangian functional therefore is

$$\mathcal{L}(Q, \boldsymbol{d}) = L(Q, \boldsymbol{d}) + \sum_{i \in \mathbf{L}} \tau_i \left(\sum_{x_i} Q(x_i) - 1 \right) \tag{84}$$

A necessary condition for an extemum of $L(Q, \boldsymbol{d})$ is a stationary point of $\mathcal{L}(Q, \boldsymbol{d})$, i.e. the derivatives w.r.t. the components of $Q(\boldsymbol{x})$ have to vanish:

$$\begin{aligned}\frac{\delta\mathcal{L}(Q,\boldsymbol{d})}{\delta Q(X_m)} &= \sum_{j\in\mathbf{O}\cap\mathrm{ch}_{X_m}}\sum_{i\in\mathbf{L}\cap\mathrm{pa}_{X_j}\setminus\{m\}}\prod Q(x_i)\log(P(d_j|\mathrm{pa}_{X_j}))\\ &+\sum_{i\in\mathbf{L}\cap\mathrm{pa}_{X_m}}\prod Q(x_i)\log(P(x_m|\mathrm{pa}_{X_m}))\\ &+\sum_{k\in\mathbf{L}\cap\mathrm{ch}_{X_m}}\sum_{i\in\mathbf{L}\cap\mathrm{pa}_{X_k}\cup\{k\}\setminus\{m\}}\prod Q(x_i)\log(P(x_k|\mathrm{pa}_{X_k}))\\ &-\log(Q(x_m))-1+\tau_m \qquad (85)\end{aligned}$$

which we set to 0 and solve for $Q(x_m)$ to find the optimal approximate posterior marginals. To do so, we interpret the terms in eqn. 85 as *messages* sent to node X_m on the factor graph corresponding to the Bayesian network (see fig. 1 for an example).

Define the messages sent a variable node X_m to a neighbouring factor node $f(X_m,\ldots)$ as just the variational posterior marginals for unobserved nodes, and a 1-or-0 message for observed nodes:

$$\begin{aligned}\forall m\in\mathbf{L}:\mu_{X_m\to f(\ldots,X_m,\ldots)}(X_m) &= Q(X_m) \qquad (86)\\ \forall j\in\mathbf{O}:\mu_{X_j\to f(\ldots,X_j,\ldots)}(X_j) &= \delta_{x_j,d_j} \qquad (87)\end{aligned}$$

and the messages sent from a factor $f(\ldots,X_m,\ldots)$ depending on variables with indices $j\in\mathbf{F},m\in\mathbf{F}$ to a neighbouring variable node as:

$$\begin{aligned}\mu_{f(\ldots,X_m,\ldots)\to X_m}(X_m) &= \sum_{j\in\mathbf{F}\setminus\{m\}}\prod Q(x_j)f(\ldots,x_m,\ldots) \qquad (88)\\ &= \sum_{j\in\mathbf{F}\setminus\{m\}}\prod f(\ldots,x_m,\ldots)\mu_{X_j\to f(\ldots,X_j,\ldots)}(x_j) \qquad (89)\end{aligned}$$

i.e. the average over the factor with respect to the posterior of all variables that connect to it, except for the variable where the message is being sent to. With these message definitions, eqn. 85 becomes

$$\begin{aligned}\frac{\delta L(Q,\boldsymbol{d})}{\delta Q(X_m)} &= \sum_{j\in\mathbf{O}\cap\mathrm{ch}_{X_m}}\mu_{\log(P(X_j|\mathrm{pa}_{X_j}))\to X_m}(X_m)\\ &\quad+\mu_{\log(P(X_m|\mathrm{pa}_{X_m}))\to X_m}(X_m)\\ &\quad+\sum_{k\in\mathbf{L}\cap\mathrm{ch}_{X_m}}\mu_{\log(P(X_k|\mathrm{pa}_{X_k}))\to X_m}(X_m)\\ &\quad-\log(Q(x_m))-1+\tau_m \qquad (90)\\ &= \sum_{j\in\mathrm{ch}_{X_m}}\mu_{\log(P(X_j|\mathrm{pa}_{X_j}))\to X_m}(X_m)\\ &\quad+\mu_{\log(P(X_m|\mathrm{pa}_{X_m}))\to X_m}(X_m)\\ &\quad-\log(Q(X_m))-1+\tau_m \qquad (91)\end{aligned}$$

Defining $\exp(\tau_m - 1) = \frac{1}{Z_m}$, we can solve for $Q(X_m)$ now:

$$Q(X_m) = \frac{1}{Z_m} \exp\left(\sum_{j \in \mathrm{ch}_{X_m}} \mu_{\log(P(X_j|\mathrm{pa}_{X_j}))\to X_m}(X_m) + \mu_{\log(P(X_m|\mathrm{pa}_{X_m}))\to X_m}(X_m) \right) \tag{92}$$

In other words, the variational posterior at variable node X_m is computed by adding up all incoming messages, exponentiating, and normalizing (since Z_i is computed from the Lagrange multiplier which enforces normalization). This message-passing scheme has to be iterated until convergence, which is guaranteed since the bound $L(Q, D)$ is a Lyapunov function of the iteration dynamics.

Another way of deriving this algorithm without computing derivatives is via the KL divergence $D(Q(X)||\tilde{Q}(X))$. Recall that the KL divergence is positive, and zero if and only if the distributions are equal everywhere. Assume again we wanted to maximize eqn. 83 w.r.t. $Q(X_k)$. To carry out this maximization, we only need to consider terms which depend on $Q(X_k)$, which in turn depend on the members of X_k's Markov blanket:

$$\begin{aligned} \operatorname*{argmax}_{Q(X_k)} [L(Q, \boldsymbol{d})] \quad = \quad & \operatorname*{argmax}_{Q(X_k)} \Bigg[\sum_{j \in \mathbf{O} \cap \mathrm{ch}_{X_k}} \sum\!\!\!\!\!\!\prod_{i \in \mathbf{L} \cap \mathrm{pa}_{X_j}} Q(x_i) \log(P(d_j|\mathrm{pa}_{X_j})) \\ & + \sum_{j \in \mathbf{L} \cap \mathrm{ch}_{X_k} \cup \{k\}} \sum\!\!\!\!\!\!\prod_{i \in \mathbf{L} \cap \mathrm{pa}_{X_j} \cup \{j\}} \log(P(x_j|\mathrm{pa}_{X_j})) \\ & - \sum_{x_k} Q(x_k) \log(Q(x_k)) \Bigg] \end{aligned} \tag{93}$$

Note that all terms on the r.h.s. include a factor $Q(x_k)$ and a summation over x_k. We can therefore pull it out:

$$\begin{aligned} \operatorname*{argmax}_{Q(X_k)} [L(Q, \boldsymbol{d})] \quad = \quad & \operatorname*{argmax}_{Q(X_k)} \Bigg[\sum_{x_k} Q(x_k) \Bigg(\sum_{j \in \mathbf{O} \cap \mathrm{ch}_{X_k}} \sum\!\!\!\!\!\!\prod_{i \in \mathbf{L} \cap \mathrm{pa}_{X_j} \setminus \{k\}} Q(x_i) \log(P(d_j|\mathrm{pa}_{X_j})) \\ & + \sum\!\!\!\!\!\!\prod_{i \in \mathbf{L} \cap \mathrm{pa}_{X_k}} \log(P(x_k|\mathrm{pa}_{X_k})) \\ & + \sum_{j \in \mathbf{L} \cap \mathrm{ch}_{X_k}} \sum\!\!\!\!\!\!\prod_{i \in \mathbf{L} \cap \mathrm{pa}_{X_j} \cup \{j\} \setminus \{k\}} \log(P(x_j|\mathrm{pa}_{X_j})) - \log(Q(x_k)) \Bigg) \Bigg] \end{aligned} \tag{94}$$

With the message definitions (eqns. 86 – 88) the r.h.s. can be written as

$$\begin{aligned}\underset{Q(X_k)}{\operatorname{argmax}}\left[L(Q,\boldsymbol{d})\right] &= \underset{Q(X_k)}{\operatorname{argmax}}\left[\sum_{x_k} Q(x_k)\left(\sum_{j\in\mathbf{O}\cap\mathrm{ch}_{X_k}} \mu_{\log(P(d_j|\mathrm{pa}_{X_j}))\to X_k}(x_k)\right.\right.\\ &\quad +\mu_{\log(P(X_k|\mathrm{pa}_{X_k}))\to X_k}(x_k)\\ &\quad \left.\left.+\sum_{j\in\mathbf{L}\cap\mathrm{ch}_{X_k}} \mu_{\log(P(X_j|\mathrm{pa}_{X_j})\to X_k}(x_k)-\log(Q(x_k))\right)\right]\end{aligned} \tag{95}$$

i.e. we need the incoming messages from all neighbouring factor nodes to compute this expression. Note that the unions of the index sets of the sums in the first and the last line are simply the indexes of all children of X_k, whereas the message on the second line is the incoming message from the parents. Thus, define

$$\log(U(x_k)) = \sum_{j\in\mathrm{ch}_{X_k}} \mu_{\log(P(d_j|\mathrm{pa}_{X_j}))\to X_k}(x_k) + \mu_{\log(P(X_k|\mathrm{pa}_{X_k}\to X_k}(x_k) \tag{96}$$

and let

$$U(x_k) = Z_k\tilde{Q}(x_k) \tag{97}$$

with $\sum_{x_k}\tilde{Q}(x_k)=1$ and $Z_k>0$, i.e. $\tilde{Q}(X_k)$ is a probability distribution. With these definition we obtain

$$\begin{aligned}\underset{Q(X_k)}{\operatorname{argmax}}\left[L(Q,\boldsymbol{d})\right] &= \underset{Q(X_k)}{\operatorname{argmax}}\left[\sum_{x_k} Q(x_k)\log(U(x_k))-\log(Q(x_k))\right]\\ &= \underset{Q(X_k)}{\operatorname{argmax}}\left[\sum_{x_k} Q(x_k)\log(Z_k)+\sum_{x_k} Q(x_k)\log\left(\frac{\tilde{Q}(x_k)}{Q(x_k)}\right)\right]\\ &= \log(Z_k)-\underset{Q(X_k)}{\operatorname{argmax}}\left[D(Q(X_k)||\tilde{Q}(X_k))\right]\end{aligned} \tag{98}$$

Since the KL-divergence is ≥ 0, it follows that the variational bound $L(Q,\boldsymbol{d})$ is maximized if $Q(X_k)=\tilde{Q}(X_k)$. In other words, to compute the optimal distribution at a given variable node given the distributions of the variables in its Markov blanket, do the following:

- sum all incoming messages from neighbouring factor nodes,
- exponentiate,
- normalize.

The factor nodes collect messages from their neighbouring variable nodes, and compute messages by summing their log-factor over all variables except the one where the message is being sent to, similar to sum-product message passing.

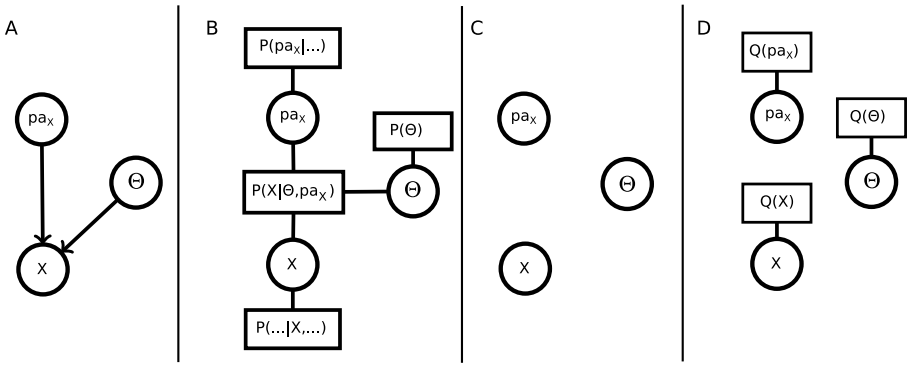

Figure 2: Variational approximation of parameter (Θ) learning in a Bayesian network with a fully factorized approximation. **A**: fragment of a Bayesian network. $\mathrm{pa}_{\boldsymbol{X}}$ are the parents of $\boldsymbol{X}$, which may have parents and other children. For learning the parameters Θ, these other children/parents are not relevant. **B**: corresponding factor graph fragment. The parameters Θ appear only in the factor connecting $\boldsymbol{X}$ to its parents. **C**: fully factorized approximation and **D** corresponding factor graph. When computing the variational bound, $\log(P(\boldsymbol{X}|\Theta, \mathrm{pa}_{\boldsymbol{X}}))$ has to be averaged over all variables that appear in it, which are the neighbours of the factor node $P(\boldsymbol{X}|\Theta, \mathrm{pa}_{\boldsymbol{X}})$ in **B**.

3.3 Learning parameters with exponential family distributions

To apply variational message passing, it is necessary to know the factors, which are the conditional probability distributions in case of a Bayesian network. If we want to learn these factors from data, then it is useful to have a compact parametrization of them, which can be done with exponential family distributions and their conjugate p(oste)riors. Consider the network (fragment) in fig. 2. Assume we wanted to learn the conditional distribution of $\boldsymbol{X}$ given its parents, and parametrize this distribution by Θ. We lump these parents together in one supernode. $\boldsymbol{X}$ may be continuous or discrete, but we assume that the parents of $\boldsymbol{X}$, $\mathrm{pa}_{\boldsymbol{X}}$ are discrete (in some special cases, continuous models are tractable). Also assume that the distribution of $\boldsymbol{X}$ given $\mathrm{pa}_{\boldsymbol{X}}$ is from the exponential family, i.e.

$$p(\boldsymbol{X}|\mathrm{pa}_{\boldsymbol{X}}, \boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}}) = h(\boldsymbol{x})g(\boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}}) \exp(\boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}}^T \boldsymbol{u}(\boldsymbol{x})) \tag{99}$$

i.e. there is one parameter vector $\boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}}$ for each value of $\mathrm{pa}_{\boldsymbol{X}}$, and Θ is the concatentation of these parameter vectors. The conjugate prior on each $\boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}}$ is then

$$p(\boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}}|\boldsymbol{\lambda}_{\mathrm{pa}_{\boldsymbol{X}}}, \nu_{\mathrm{pa}_{\boldsymbol{X}}}) = f(\boldsymbol{\lambda}_{\mathrm{pa}_{\boldsymbol{X}}}, \nu_{\mathrm{pa}_{\boldsymbol{X}}}) g(\boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}})^{\nu} \exp(\nu_{\mathrm{pa}_{\boldsymbol{X}}} \boldsymbol{\eta}_{\mathrm{pa}_{\boldsymbol{X}}}^T \boldsymbol{\lambda}_{\mathrm{pa}_{\boldsymbol{X}}}) \tag{100}$$

Assume now we had observed $n = 1, \ldots, N$ datapoints $\boldsymbol{d}_{1:N}$ and computed the corresponding latent variable distributions $Q(X^n)$, $Q(\mathrm{pa}_{\boldsymbol{X}}^n)$. If any of the nodes are observed, replace the corresponding distribution with a distribution concentrated at the observed value. Note that in a fully factorized approximation, each $Q(\mathrm{pa}_{\boldsymbol{X}^n})$ is a actually a product over the parents, $Q(\mathrm{pa}_{\boldsymbol{X}^n}) =$

$\prod_{\boldsymbol{Y}^n \in \mathrm{pa}_{\boldsymbol{X}^n}} Q(Y^n)$. Looking at fig. 2, we note that there is exactly one factor connecting the parameter node to $\boldsymbol{X}$ and $\mathrm{pa}_{\boldsymbol{X}}$, and the prior factor for Θ. Using an approximating posterior for $Q(\Theta)$ which has the same form as the prior (eqn. 100), the summands in the variational bound which depend on the posterior distribution of Θ are (where $\boldsymbol{y}^n, \boldsymbol{y} \in \mathrm{range}(\mathrm{pa}_{\boldsymbol{X}^n})$):

$$\begin{aligned}
\mathcal{L}_\Theta &= \sum_{n=1}^{N} \sum_{\boldsymbol{x}^n} \sum_{\boldsymbol{y}^n} Q(\boldsymbol{x}^n) Q(\boldsymbol{y}^n) \left\langle \log(P(\boldsymbol{x}^n|\Theta, \boldsymbol{y}^n)) \right\rangle_{Q(\Theta)} - \beta D(Q(\Theta)||P(\Theta)) \\
&= \sum_{n=1}^{N} \sum_{\boldsymbol{x}^n} \sum_{\boldsymbol{y}^n} Q(\boldsymbol{x}^n) Q(\boldsymbol{y}^n) \left\langle \log(P(\boldsymbol{x}^n|\boldsymbol{\eta}_{\boldsymbol{y}^n})) \right\rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}^n}|\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}^n}, \tilde{\nu}_{\boldsymbol{y}^n})} \\
&\quad - \sum_{\boldsymbol{y}} \beta D(Q(\boldsymbol{\eta}_{\boldsymbol{y}}|\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})||P(\boldsymbol{\eta}_{\boldsymbol{y}}|\boldsymbol{\lambda}_{\boldsymbol{y}}, \nu_{\boldsymbol{y}})) \qquad (101)
\end{aligned}$$

We rewrite the first term as a sum over the possible values of $\boldsymbol{x}$ and pa_X:

$$\begin{aligned}
\mathcal{L}_\Theta &= \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} \left\langle \log(P(\boldsymbol{x}|\boldsymbol{\eta}_{\boldsymbol{y}})) \right\rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}}|\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})} \sum_{n=1}^{N} Q(\boldsymbol{X}^n = \boldsymbol{x}) Q(\mathrm{pa}_{X^n} = \boldsymbol{y}) \\
&\quad - \sum_{\boldsymbol{y}} \beta D(Q(\boldsymbol{\eta}_{\boldsymbol{y}}|\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})||P(\boldsymbol{\eta}_{\boldsymbol{y}}|\boldsymbol{\lambda}_{\boldsymbol{y}}, \nu_{\boldsymbol{y}})) \qquad (102)
\end{aligned}$$

and define the "responsibilities" (because they measure how much a given setting of the latent variables contributes to explaining a datapoint)

$$r_{\boldsymbol{x},\boldsymbol{y}} = \sum_{n=1}^{N} Q(\boldsymbol{X}^n = \boldsymbol{x}) Q(\mathrm{pa}_{X^n} = \boldsymbol{y}) \qquad (103)$$

Using the definition of the exponential familiy distribution (eqn. 1), the Kullback-Leibler divergence between conjugate p(oste)riors (eqn. 68) and denoting $Q(\boldsymbol{\eta}_{\boldsymbol{y}}) = Q(\boldsymbol{\eta}_{\boldsymbol{y}}|\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})$, we find

$$\begin{aligned}
\mathcal{L}_\Theta &= \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}} \left[\log(h(\boldsymbol{x})) + \langle \log(g(\boldsymbol{\eta}_{\boldsymbol{y}})) \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})} + \boldsymbol{u}(\boldsymbol{x})^T \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})} \right] r_{\boldsymbol{x},\boldsymbol{y}} \\
&\quad - \sum_{\boldsymbol{y}} \left[\beta \log \left(\frac{f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{f(\boldsymbol{\lambda}_{\boldsymbol{y}}, \nu_{\boldsymbol{y}})} \right) - \beta(\tilde{\nu}_{\boldsymbol{y}} - \nu_{\boldsymbol{y}}) \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}}))}{\partial \tilde{\nu}_{\boldsymbol{y}}} + \beta(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T - \boldsymbol{\lambda}_{\boldsymbol{y}}^T) \nu_{\boldsymbol{y}} \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})} \right] \qquad (104)
\end{aligned}$$

The expectations $\langle \log(g(\boldsymbol{\eta}_{\boldsymbol{y}})) \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})}$ can be expressed using eqn. 31:

$$\langle \log(g(\boldsymbol{\eta}_{\boldsymbol{y}})) \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})} = -\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{\partial \tilde{\nu}_{\boldsymbol{y}}} - \tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})} \qquad (105)$$

Inserting this expression into eqn. 104:

$$\begin{aligned}\mathcal{L}_\Theta &= \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}}\left[\log(h(\boldsymbol{x})) - \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{\partial \tilde{\nu}_{\boldsymbol{y}}} - \tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})} + \boldsymbol{u}(\boldsymbol{x})^T \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})}\right] r_{\boldsymbol{x},\boldsymbol{y}} \\ &\quad - \sum_{\boldsymbol{y}}\left[\beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{f(\boldsymbol{\lambda}_{\boldsymbol{y}}, \nu_{\boldsymbol{y}})}\right) - \beta(\tilde{\nu}_{\boldsymbol{y}} - \nu_{\boldsymbol{y}}) \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}}))}{\partial \tilde{\nu}_{\boldsymbol{y}}} + \beta(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T - \boldsymbol{\lambda}_{\boldsymbol{y}}^T)\nu_{\boldsymbol{y}} \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})}\right]\end{aligned} \tag{106}$$

Collecting terms, we find:

$$\begin{aligned}\mathcal{L}_\Theta &= \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}} \log(h(\boldsymbol{x})) r_{\boldsymbol{x},\boldsymbol{y}} - \sum_{\boldsymbol{y}} \beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{f(\boldsymbol{\lambda}_{\boldsymbol{y}}, \nu_{\boldsymbol{y}})}\right) \\ &\quad + \sum_{\boldsymbol{y}} \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{\partial \tilde{\nu}_{\boldsymbol{y}}} \left[\beta\left(\tilde{\nu}_{\boldsymbol{y}} - \nu_{\boldsymbol{y}}\right) - \sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}\right] \\ &\quad - \sum_{\boldsymbol{y}} \nu_{\boldsymbol{y}} \left[\beta\left(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T - \boldsymbol{\lambda}_{\boldsymbol{y}}^T\right) + \tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\nu_{\boldsymbol{y}}} - \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}} \boldsymbol{u}(\boldsymbol{x})^T}{\nu_{\boldsymbol{y}}}\right] \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})}\end{aligned} \tag{107}$$

With the expressions

$$\begin{aligned}\hat{\nu}_{\boldsymbol{y}} &= \nu_{\boldsymbol{y}} + \frac{\sum_{\boldsymbol{x},\boldsymbol{y}} r_{\boldsymbol{x},\boldsymbol{y}}}{\beta} = \nu_{\boldsymbol{y}}\left(1 + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\nu_{\boldsymbol{y}}\beta}\right) \\ \hat{\boldsymbol{\lambda}}_{\boldsymbol{y}} &= \frac{\nu_{\boldsymbol{y}}\boldsymbol{\lambda}_{\boldsymbol{y}} + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\beta}\boldsymbol{u}(\boldsymbol{x})}{\nu_{\boldsymbol{y}} + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\beta}} = \frac{\nu_{\boldsymbol{y}}}{\hat{\nu}_{\boldsymbol{y}}}\left(\boldsymbol{\lambda}_{\boldsymbol{y}} + \frac{\frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\beta}\boldsymbol{u}(\boldsymbol{x})}{\nu_{\boldsymbol{y}}}\right)\end{aligned} \tag{108}$$

and noting that

$$-\beta\hat{\nu}_{\boldsymbol{y}} = -\beta\nu_{\boldsymbol{y}} - \sum_{\boldsymbol{x},\boldsymbol{y}} r_{\boldsymbol{x},\boldsymbol{y}} \tag{109}$$

$$\begin{aligned}\nu_{\boldsymbol{y}}\left[\beta\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T + \tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\nu_{\boldsymbol{y}}}\right] &= \beta\nu_{\boldsymbol{y}}\left[\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T + \tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\nu_{\boldsymbol{y}}\beta}\right] \\ &= \beta\nu_{\boldsymbol{y}}\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T\left[1 + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\nu_{\boldsymbol{y}}\beta}\right] = \beta\hat{\nu}_{\boldsymbol{y}}\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T\end{aligned} \tag{110}$$

$$\begin{aligned}\nu_{\boldsymbol{y}}\left[-\beta\boldsymbol{\lambda}_{\boldsymbol{y}}^T - \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}\boldsymbol{u}(\boldsymbol{x})^T}{\nu_{\boldsymbol{y}}}\right] &= -\nu_{\boldsymbol{y}}\beta\left[\boldsymbol{\lambda}_{\boldsymbol{y}}^T + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}\boldsymbol{u}(\boldsymbol{x})^T}{\nu_{\boldsymbol{y}}\beta}\right] \\ &= -\beta\hat{\nu}_{\boldsymbol{y}}\hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T\end{aligned} \tag{111}$$

we insert a zero by adding and subtracting the term $\sum_{\boldsymbol{y}} \beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{f(\hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \hat{\nu}_{\boldsymbol{y}})}\right)$

and write (using eqn. 68)

$$\begin{aligned}
\mathcal{L}_\Theta &= \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}} \log(h(\boldsymbol{x}))r_{\boldsymbol{x},\boldsymbol{y}} - \sum_{\boldsymbol{y}} \beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{f(\boldsymbol{\lambda}_{\boldsymbol{y}}, \nu_{\boldsymbol{y}})}\right) \\
&\quad + \sum_{\boldsymbol{y}} \beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{f(\hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \hat{\nu}_{\boldsymbol{y}})}\right) - \sum_{\boldsymbol{y}} \beta \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{f(\hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \hat{\nu}_{\boldsymbol{y}})}\right) \\
&\quad + \sum_{\boldsymbol{y}} \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}})}{\partial \tilde{\nu}_{\boldsymbol{y}}} \beta(\tilde{\nu}_{\boldsymbol{y}} - \hat{\nu}_{\boldsymbol{y}}) \\
&\quad - \sum_{\boldsymbol{y}} \beta \hat{\nu}_{\boldsymbol{y}} (\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T - \hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}^T) \langle \boldsymbol{\eta}_{\boldsymbol{y}} \rangle_{Q(\boldsymbol{\eta}_{\boldsymbol{y}})} \qquad (112) \\
&= \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}} \log(h(\boldsymbol{x}))r_{\boldsymbol{x},\boldsymbol{y}} - \sum_{\boldsymbol{y}} \beta \log\left(\frac{f(\hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \hat{\nu}_{\boldsymbol{y}})}{f(\boldsymbol{\lambda}_{\boldsymbol{y}}, \nu_{\boldsymbol{y}})}\right) \\
&\quad - \sum_{\boldsymbol{y}} \beta D(Q(\boldsymbol{\eta}_{\boldsymbol{y}}|\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \tilde{\nu}_{\boldsymbol{y}}) || Q(\boldsymbol{\eta}_{\boldsymbol{y}}|\hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}, \hat{\nu}_{\boldsymbol{y}})) \qquad (113)
\end{aligned}$$

The first part of this expression is constant w.r.t. to the $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}}$, $\tilde{\nu}_{\boldsymbol{y}}$, and the second part is a sum of KL-divergences. To maximize $\mathcal{L}_\Theta$, we therefore choose the posterior parameters so that each KL-divergence is zero, i.e. $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}} = \hat{\boldsymbol{\lambda}}_{\boldsymbol{y}}$ and $\tilde{\nu}_{\boldsymbol{y}} = \hat{\nu}_{\boldsymbol{y}}$:

variational posterior parameters

$$\begin{aligned}
r_{\boldsymbol{x},\boldsymbol{y}} &:= \sum_{n=1}^{N} Q(\boldsymbol{X}^n = \boldsymbol{x}) Q(\mathrm{pa}_{X^n} = \boldsymbol{y}) \qquad (114) \\
\tilde{\nu}_{\boldsymbol{y}} &:= \nu_{\boldsymbol{y}} + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\beta} \qquad (115) \\
\tilde{\boldsymbol{\lambda}}_{\boldsymbol{y}} &:= \frac{\nu_{\boldsymbol{y}}\boldsymbol{\lambda}_{\boldsymbol{y}} + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\beta}\boldsymbol{u}(\boldsymbol{x})}{\nu_{\boldsymbol{y}} + \frac{\sum_{\boldsymbol{x}} r_{\boldsymbol{x},\boldsymbol{y}}}{\beta}} \qquad (116)
\end{aligned}$$

Comparing these learning rules to the conjugate update rules (eqns. 24,25), we see that in the variational framework, a datapoint can be "shared" between different values of the latent variables in the model. This sharing comes about because the responsibilities are (approximate) probability, rather than deterministic 0s or 1s. Otherwise the rules are identical. Note in particular that if $r_{\boldsymbol{x},\boldsymbol{y}} \in \{0, 1\}$, i.e. the variables are known with certainty, then the variational rules reduce to the exact update rules. Also, the β-variational update rules can be obtained by dividing all responsibilities by β. Thus, if $\beta > 1$, the prior becomes 'stiffer' and tends to ignore the data, whereas for $\beta \to 0$, we get maximum likelihood updates.

4 Frequently used special cases

This section contains frequently used conjugate pairs and relevant quanitities computed thereof.

4.1 Bernoulli-Beta

For a discrete random variable $x \in \{0; 1\}$, where 1 is alternatively called "success" (e.g. when betting on coin tosses), is given by

$$P(x|q) = q^x(1-q)^{(1-x)}. \tag{117}$$

Its canonical conjugate prior is a Beta distribution in q with density

$$p(q|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} q^{\alpha-1}(1-q)^{\beta-1}. \tag{118}$$

To transform these expressions into the exponential family normal form (eqns. 1 and 22), introduce the *logit*

$$\eta = \log\left(\frac{q}{1-q}\right) \tag{119}$$

whence $q = \frac{1}{1+\exp(-\eta)}$, $1-q = \frac{1}{1+\exp(\eta)}$ and $\frac{dq}{d\eta} = -\frac{\exp(-\eta)}{(1+\exp(-\eta))^2} = -q(1-q)$. Substitute η in eqn. (117):

$$\begin{aligned} P(x|\eta) &= \exp\left(x\log(q) + (1-x)\log(1-q)\right) \\ &= \exp\left(x\log\left(\frac{q}{1-q}\right) + \log(1-q)\right) \\ &= (1-q)\exp(\eta x) = \frac{1}{1+\exp(\eta)}\exp(\eta x) \end{aligned} \tag{120}$$

Hence, $h(x) = 1$, $g(\eta) = \frac{1}{1+\exp(\eta)}$ (cf. eqn. (1)) and

$$P(x|\eta) = h(x)g(\eta)\exp(\eta x). \tag{121}$$

To transform the Beta density into exponential family normal form, note that densities transform like $p(\eta) = p(q(\eta))\left|\frac{dq}{d\eta}\right|$:

$$\begin{aligned} p(\eta|\alpha, \beta) &= \frac{1}{B(\alpha, \beta)} q^{\alpha-1}(1-q)^{\beta-1}\left|\frac{dq}{d\eta}\right| \\ &= \frac{1}{B(\alpha, \beta)} q^{-1}(1-q)^{-1}\exp\left(\alpha\log(q) + \beta\log(1-q)\right) q(1-q) \\ &= \frac{1}{B(\alpha, \beta)}\exp\left(\alpha\log\left(\frac{q}{1-q}\right) + (\alpha+\beta)\log(1-q)\right) \\ &= \frac{1}{B(\alpha, \beta)}\left[\frac{1}{1+\exp(\eta)}\right]^{\alpha+\beta}\exp\left(\alpha\eta\right) \end{aligned} \tag{122}$$

Letting $\nu := \alpha + \beta$, $\lambda := \frac{\alpha}{\nu}$, $f(\lambda, \nu) = \frac{1}{B(\nu\lambda, \nu(1-\lambda))}$, $m(\eta) = 1$ we find

$$\begin{aligned} p(\eta|\lambda,\nu) &= \frac{1}{B(\nu\lambda,\nu(1-\lambda))}\left[\frac{1}{1+\exp(\eta)}\right]^{\nu}\exp(\nu\lambda\eta) \\ &= f(\lambda,\nu)\, m(\eta)\, g(\eta)^{\nu}\exp(\nu\lambda\eta). \end{aligned} \tag{123}$$

Bernoulli distribution

standard form	$q^x(1-q)^{1-x}$
constraints x	$x \in \{0,1\}$
constraints q	$q \in [0,1]$
$u(x)$	x
η	$\log\left(\frac{q}{1-q}\right)$
q	$\frac{1}{1+\exp(-\eta)}$
constraints η	$\eta \in \mathbb{R}$
$g(\eta)$	$\frac{1}{1+\exp(\eta)} = 1-q$
$h(x)$	1
$\langle u(x)\rangle$	$\frac{1}{1+\exp(-\eta)} = q$
$\mathrm{Var}(u(x))$	$\frac{\exp(-\eta)}{(1+\exp(-\eta))^2} = q(1-q)$

Beta distribution

standard form	$\frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)}$
constraints α, β	$\alpha, \beta \in \mathbb{R}^+$
λ	$\frac{\alpha}{\alpha+\beta}$
ν	$\alpha + \beta$
α	$\nu\lambda$
β	$\nu(1-\lambda)$
constraints ν	$\nu \in \mathbb{R}^+$
constraints λ	$\lambda \in [0,1]$
$f(\lambda,\nu)$	$\frac{1}{B(\nu\lambda,\nu(1-\lambda))}$
$m(\eta)$	1
$\langle\eta\rangle$	$\psi(\nu\lambda) - \psi(\nu(1-\lambda))$
$\mathrm{Var}(\eta)$	$\frac{\psi(\nu\lambda)-\psi(\nu(1-\lambda))}{\nu^2} + \psi'(\nu\lambda) + \psi'(\nu(1-\lambda))$
$\frac{\partial \log(f(\lambda,\nu))}{\partial\nu}$	$\psi(\nu) - \lambda\psi(\nu\lambda) - (1-\lambda)\psi(\nu(1-\lambda))$
$\frac{\partial^2 \log(f(\lambda,\nu))}{\partial\nu^2}$	$\psi'(\nu) - \lambda^2\psi'(\nu\lambda) - (1-\lambda)^2\psi'(\nu(1-\lambda))$
$\langle g(\eta)\rangle$	$1-\lambda = \frac{\beta}{\alpha+\beta} = 1 - \langle q\rangle$
$\mathrm{Var}(g(\eta))$	$\frac{(1-\lambda)\lambda}{\nu+1} = \frac{\frac{\beta}{\alpha+\beta}\frac{\alpha}{\alpha+\beta}}{\alpha+\beta+1}$
$p(x\|\lambda,\nu)$	$\lambda x + (1-\lambda)(1-x) = x\langle q\rangle + (1-x)(1-\langle q\rangle)$
$\langle u(x)\rangle_{p(x\|\lambda,\nu)}$	λ

Table 1: Bernoulli distribution and conjugate Beta prior

4.2 Multinomial-Dirichlet

The multinomial distribution is the generalization of the Bernoulli distribution to K possible outcomes. It is convenient to represent multinomial random

variates $\boldsymbol{x}$ by vectors with K components, such that $x_k \in \{0;1\}$ and $\sum_{k=1}^{K} x_k = 1$, whence $x_K = 1 - \sum_{k=1}^{K-1} x_k$. This is called *1-of-K* representation, because exactly one component of $\boldsymbol{x}$ is 1 and all others are 0. Let $\boldsymbol{q} = (q_1, \ldots, q_K)$ be the probabilities of the K possible $\boldsymbol{x}$, such that $q_K = 1 - \sum_{k=1}^{K-1} q_k$. Then, the multinomial distribution can be written as

$$P(\boldsymbol{x}|\boldsymbol{q}) = \prod_{k=1}^{K-1} q_k^{x_k} \left(1 - \sum_{k=1}^{K-1} q_k\right)^{x_K} \tag{124}$$

This expression can be transformed into exponential family form via

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{q}) &= \exp\left(\sum_{k=1}^{K-1} x_k \log(q_k) + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log\left(1 - \sum_{k=1}^{K-1} q_k\right)\right) \\ &= \left(1 - \sum_{k=1}^{K-1} q_k\right) \exp\left(\sum_{k=1}^{K-1} x_k \log\left(\frac{q_k}{1 - \sum_{i=1}^{K-1} q_i}\right)\right) \end{aligned} \tag{125}$$

and by introducing the generalized logit $\eta_k = \log\left(\frac{q_k}{1-\sum_{i=1}^{K-1} q_i}\right)$ we find that

$$\forall k = 1, \ldots, K-1 \quad : \quad q_k = \frac{\exp(\eta_k)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} \tag{126}$$

$$q_K = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} \tag{127}$$

(alternatively, we could fix $\eta_K = 0$, and let $\boldsymbol{q} = \text{softmax}(\boldsymbol{\eta})$). Hence, $h(\boldsymbol{x}) = 1$ (after $x_k \in \{0;1\}$ has been enforced) and

$$g(\eta) = \left(1 - \sum_{k=1}^{K-1} q_k\right) = q_K = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}$$

.

The standard conjugate prior on the multinomial is the Dirichlet distribution. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k \geq 0$ and

$$M = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \tag{128}$$

then the density of the Dirichlet distribution is

$$p(\boldsymbol{q}|\boldsymbol{\alpha}) = M \prod_{k=1}^{K-1} q_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K-1} q_k\right)^{\alpha_K - 1} \tag{129}$$

$$\begin{aligned} &= M \exp\left(\sum_{k=1}^{K-1} (\alpha_k - 1) \log(q_k) + (\alpha_K - 1) \log\left(1 - \sum_{k=1}^{K-1} q_k\right)\right) \\ &= M \prod_{k=1}^{K-1} q_k^{-1} (1 - \sum_{k=1}^{K-1} q_k)^{-1} \exp\left(\sum_{k=1}^{K-1} \alpha_k \log\left(\frac{q_k}{q_K}\right) + \sum_{k=1}^{K} \alpha_K \log(q_K)\right) \\ &= M\, q_K^{\sum_{k=1}^{K} \alpha_K} \prod_{k=1}^{K-1} q_k^{-1} (1 - \sum_{k=1}^{K-1} q_k)^{-1} \exp\left(\boldsymbol{\alpha}^T \boldsymbol{\eta}\right) \end{aligned} \tag{130}$$

To substitute $\boldsymbol{q}$ by $\boldsymbol{\eta}$, we need to compute the determinant of the Jacobian $\frac{d\boldsymbol{q}}{d\boldsymbol{\eta}}$. Let $Z = 1 + \sum_{k=1}^{K-1} \exp(\eta_k)$, i.e. $q_k = \frac{\exp(\eta_k)}{Z}$. Then

$$\begin{aligned}
\left|\frac{d\boldsymbol{q}}{d\boldsymbol{\eta}}\right| &= \left| \begin{pmatrix} \frac{\exp(\eta_1)}{Z} - \frac{\exp(2\eta_1)}{Z^2} & \cdots & -\frac{\exp(\eta_1)\exp(\eta_K)}{Z^2} \\ & \ddots & \\ -\frac{\exp(\eta_K)\exp(\eta_1)}{Z^2} & \cdots & \frac{\exp(\eta_K)}{Z} - \frac{\exp(2\eta_K)}{Z^2} \end{pmatrix} \right| \\
&= \left| \begin{pmatrix} q_1 & \ldots & 0 \\ 0 & \ddots & 0 \\ 0 & \ldots & q_K \end{pmatrix} - \begin{pmatrix} q_1 \\ \vdots \\ q_K \end{pmatrix} (q_1 \ldots q_K) \right| \\
&= \left| \begin{pmatrix} q_1 & \ldots & 0 \\ 0 & \ddots & 0 \\ 0 & \ldots & q_K \end{pmatrix} \left[\mathbf{1}_{K\times K} - \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (q_1 \ldots q_K) \right] \right| \\
&= \prod_{k=1}^{K-1} q_k \left(1 - \sum_{k=1}^{K-1} q_k\right) \qquad (131)
\end{aligned}$$

And thus

$$\begin{aligned}
p(\boldsymbol{\eta}|\boldsymbol{\alpha}) &= p(\boldsymbol{q}(\boldsymbol{\eta})|\boldsymbol{\alpha}) \left|\frac{d\boldsymbol{q}}{d\boldsymbol{\eta}}\right| \\
&= M \left(\frac{1}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}\right)^{\sum_{k=1}^{K} \alpha_K} \exp\left(\boldsymbol{\alpha}^T \boldsymbol{\eta}\right) \qquad (132)
\end{aligned}$$

With $\nu = \sum_{k=1}^{K} \alpha_k$, $\lambda_k = \frac{\alpha_k}{\nu}$, $f(\boldsymbol{\lambda}, \nu) = M$, $m(\boldsymbol{\eta}) = 1$ and $g(\boldsymbol{\eta})$ as above, we find

$$p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \nu) = f(\boldsymbol{\lambda}, \nu) m(\boldsymbol{\eta}) g(\boldsymbol{\eta})^{\nu} \exp(\nu \boldsymbol{\eta} \boldsymbol{\lambda}) \qquad (133)$$

multinomial distribution

standard form	$\prod_{k=1}^{K-1} q_k^{x_k} \left(1 - \sum_{k=1}^{K-1} q_k\right)^{x_K}$
constraints $\boldsymbol{x}$	$x_k \in \{0,1\}$, $\sum_{k=1}^{K} x_k = 1$
constraints $\boldsymbol{q}$	$q_k \in [0,1]$, $\sum_{k=1}^{K-1} q_k \leq 1$
$\boldsymbol{u}(\boldsymbol{x})$	$\boldsymbol{x}$
$\boldsymbol{\eta}$	$\log\left(\frac{\boldsymbol{q}}{1-\sum_{i=1}^{K-1} q_i}\right)$
$\boldsymbol{q}$	$\frac{\exp(\boldsymbol{\eta})}{1+\sum_{i=1}^{K-1}\exp(\eta_i)}$
constraints $\boldsymbol{\eta}$	$\eta_k \in \mathbb{R}$
$g(\eta)$	$\frac{1}{1+\sum_{i=1}^{K-1}\exp(\eta_i)} = 1 - \sum_{i=1}^{K-1} q_i = q_K$
$h(x)$	1
$\langle \boldsymbol{u}(\boldsymbol{x}) \rangle$	$\frac{\exp(\boldsymbol{\eta})}{1+\sum_{i=1}^{K-1}\exp(\eta_i)} = \boldsymbol{q}$
$\mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x}))$	$C_{kl} = \frac{\delta_{kl}\exp(\eta_k)}{1+\sum_{i=1}^{K-1}\exp(\eta_i)} - \frac{\exp(\eta_k+\eta_l)}{\left(1+\sum_{i=1}^{K-1}\exp(\eta_i)\right)^2} = \delta_{kl}q_l - q_l q_k$

Dirichlet distribution

standard form	$\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K-1} q_k^{\alpha_k - 1} \left(1 - \sum_{k=1}^{K-1} q_k\right)^{\alpha_K - 1}$
constraints $\boldsymbol{\alpha}$	$\alpha_k \in \mathbb{R}^+$
λ	$\forall k = 1, \ldots, K-1 : \lambda_k = \frac{\alpha_k}{\sum_{i=1}^{K}\alpha_i}$
ν	$\sum_{i=1}^{K}\alpha_i$
α	$\nu\boldsymbol{\lambda}$
constraints ν	$\nu \in \mathbb{R}^+$
constraints $\boldsymbol{\lambda}$	$\lambda_i \in [0,1]$, $\sum_{i=1}^{K-1}\lambda_i \leq 1$
$f(\boldsymbol{\lambda}, \nu)$	$\frac{\Gamma(\nu)}{\Gamma\left(\nu(1-\sum_{k=1}^{K-1}\lambda_k)\right)\prod_{i=1}^{K-1}\Gamma(\nu\lambda_i)}$
$m(\boldsymbol{\eta})$	1
$\langle \boldsymbol{\eta} \rangle$	$\psi(\nu\boldsymbol{\lambda}) - \psi\left(\nu(1-\sum_{i=1}^{K-1}\lambda_k)\right)$
$\mathrm{Cov}(\eta_i, \eta_j)$	$\delta_{i,j}\psi'(\nu\lambda_i) + \psi'\left(\nu(1-\sum_{i=1}^{K-1}\lambda_k)\right)$
$\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$	$\psi(\nu) - (1-\sum_{i=1}^{K-1}\lambda_i)\psi\left(\nu(1-\sum_{i=1}^{K-1}\lambda_i)\right) - \sum_{i=1}^{K-1}\lambda_i\psi(\nu\lambda_i)$
$\frac{\partial^2 \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu^2}$	$\psi'(\nu) - (1-\sum_{i=1}^{K-1}\lambda_i)^2\psi'\left(\nu(1-\sum_{i=1}^{K-1}\lambda_i)\right) - \sum_{i=1}^{K-1}\lambda_i^2\psi'(\nu\lambda_i)$
$\langle g(\boldsymbol{\eta}) \rangle$	$1 - \sum_{i=1}^{K-1}\lambda_i = \frac{\alpha_K}{\sum_{i=1}^{K}\alpha_i} = \langle q_K \rangle$
$\mathrm{Var}(g(\boldsymbol{\eta}))$	$\frac{(1-\sum_{i=1}^{K-1}\lambda_i)\sum_{i=1}^{K-1}\lambda_i}{\nu+1} = \frac{\langle q_K\rangle\langle 1-q_K\rangle}{\nu+1}$
$p(\boldsymbol{x}\|\boldsymbol{\lambda}, \nu)$	$\sum_{i=1}^{K-1} x_i\lambda_i + x_K(1-\sum_{i=1}^{K-1}\lambda_i) = \sum_{i=1}^{K} x_i\langle q_i \rangle$
$\langle \boldsymbol{u}(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}\|\boldsymbol{\lambda},\nu)}$	$\boldsymbol{\lambda}$

Table 2: Multiomial distribution and conjugate Dirichlet prior

4.3 Multinomial-StickBreaking

The stick-breaking construction is another way of parameterizing multinomial distributions. It has attracted a lot of attention in Machine Learning around 2005, because it is a convenient way of representing *infinite* multinomials with a Dirichlet-process prior. To re-parameterize the distribution of a K component multinomial variable $\boldsymbol{x}$ (in 1-of-K representation) with probabilities $\boldsymbol{q}$, introduce the variables $\boldsymbol{v} = (v_1, \ldots, v_{K-1})$ such that

$$\begin{aligned} q_1 &= v_1 && (134)\\ q_2 &= (1-v_1)v_2 && (135)\\ q_3 &= (1-v_1)(1-v_2)v_3 && (136)\\ &\vdots \\ q_{K-1} &= (1-v_1)(1-v_2)\ldots v_{K-1} && (137)\\ q_K &= (1-v_1)(1-v_2)\ldots(1-v_{K-1}). && (138) \end{aligned}$$

The distribution of $\boldsymbol{x}$ can then be written as

$$p(\boldsymbol{x}|\boldsymbol{v}) = \prod_{i=1}^{K-1} v_i^{x_i}(1-v_i)^{\sum_{j=i+1}^{K} x_j} \tag{139}$$

We now introduce the sufficient statistics $\boldsymbol{u}(\boldsymbol{x}) = (u_0, \ldots, u_{K-1})$

$$u_k = \sum_{i=k+1}^{K} x_i \tag{140}$$

hence $u_0 = 1$. $u_k = 1$ implies that $x_i = 1$ with $i > k$. Eqn. 139 becomes

$$p(\boldsymbol{x}|\boldsymbol{v}) = \prod_{i=1}^{K-1} v_i^{u_{i-1}-u_i}(1-v_i)^{u_i} = \exp\left(\sum_{i=1}^{K-1}(u_{i-1}-u_i)\log(v_i) + u_i\log(1-v_i)\right) \tag{141}$$

To transform this expression into exponential family normal form, rewrite the exponent on the r.h.s. as

$$\log(v_1) + \sum_{i=1}^{K-2} u_i \log\left(\frac{1-v_i}{v_i}v_{i+1}\right) + u_{K-1}\log\left(\frac{1-v_{K-1}}{v_{K-1}}\right) \tag{142}$$

and introduce the natural parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{K-1})$:

$$\begin{aligned} \eta_{K-1} &= \log\left(\frac{1-v_{K-1}}{v_{K-1}}\right) && (143)\\ \forall k = K-2, \ldots, 1 : \eta_k &= \log\left(\frac{1-v_k}{v_k}v_{k+1}\right) && (144) \end{aligned}$$

Solving for $\boldsymbol{v}$ yields

$$v_{K-1} = \frac{1}{1+\exp(\eta_{K-1})} \tag{145}$$

$$v_{K-2} = \frac{1}{1+\exp(\eta_{K-2})(1+\exp(\eta_{K-1}))} \tag{146}$$

$$\vdots$$

$$v_1 = \frac{1}{1+\exp(\eta_1)(1+\exp(\eta_2)(1+\ldots(1+exp(\eta_{K-1}))\ldots)} \tag{147}$$

Thus

$$P(\boldsymbol{x}|\boldsymbol{\eta}) = v_1(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})\right) \tag{148}$$

which is in exponential family normal form, with $h(\boldsymbol{x}) = 1$ and $g(\boldsymbol{\eta}) = v_1(\boldsymbol{\eta})$.

The conjugate p(oste)rior on $\boldsymbol{v}$ is given by a product of Beta distributions. Let $\boldsymbol{\alpha}, \boldsymbol{\beta}$ be the parameters of these Beta distributions, then

$$p(\boldsymbol{v}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{i=1}^{K-1}\frac{1}{B(\alpha_i,\beta_i)}v_i^{\alpha_i-1}(1-v_i)^{\beta_i-1} \tag{149}$$

To derive the corresponding density in $\boldsymbol{\eta}$ we need the determinant of the Jacobian $\left|\frac{d\boldsymbol{v}}{d\boldsymbol{\eta}}\right|$. Note that as a consequence of eqns. 145-147, the Jacobian is an upper triangular matrix, hence the determinant is given by the product of the diagonal entries:

$$\frac{dv_i}{d\eta_i} = -\frac{\exp(\eta_i)(1+\exp(\eta_{i+1})...)}{(1+\exp(\eta_i)(1+\exp(\eta_{i+1})(....)))^2} = -v_i(1-v_i) \tag{150}$$

$$\Rightarrow \left|\frac{d\boldsymbol{v}}{d\boldsymbol{\eta}}\right| = \prod_{i=1}^{K-1} v_i(1-v_i). \tag{151}$$

Now we can rewrite eqn. 149 in exponential form

$$p(\boldsymbol{v}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{i=1}^{K-1}\frac{v_i^{-1}(1-v_i)^{-1}}{B(\alpha_i,\beta_i)}\exp\left(\sum_{k=1}^{K-1}\alpha_i\log(v_i)+\beta_i\log(1-v_i)\right) \tag{152}$$

and rearrange the exponent as:

$$\begin{aligned}
&\sum_{k=1}^{K-1}\alpha_i\log(v_i)+\beta_i\log(1-v_i) \\
&= \sum_{k=1}^{K-1}(\alpha_i+\beta_i)\log(v_i)+\beta_i\log\left(\frac{1-v_i}{v_i}\right) && (153)\\
&= \beta_{K-1}\log\left(\frac{1-v_{K-1}}{v_{K-1}}\right)+\sum_{i=1}^{K-2}\beta_i\log\left(\frac{1-v_i}{v_i}v_{i+1}\right) \\
&\quad +(\alpha_1+\beta_1)\log(v_1)+\sum_{i=2}^{K-1}\underbrace{(\alpha_i+\beta_i-\beta_{i-1})}_{=:c_i}\log(v_i) && (154)\\
&= \beta_{K-1}\eta_{K-1}+\sum_{i=1}^{K-2}\beta_i\eta_i+(\alpha_1+\beta_1)\log(g(\boldsymbol{\eta}))+\sum_{i=2}^{K-1}c_i\log(v_i) && (155)
\end{aligned}$$

Hence, letting $N(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{K-1} B(\alpha_i, \beta_i)^{-1}$ we obtain for the density of $\boldsymbol{\eta}$:

$$p(\boldsymbol{\eta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{v}(\boldsymbol{\eta})|\alpha, \beta) \left| \frac{d\boldsymbol{v}}{d\boldsymbol{\eta}} \right| \tag{156}$$

$$= N(\boldsymbol{\alpha}, \boldsymbol{\beta}) \prod_{i=2}^{K-1} v_i(\boldsymbol{\eta})^{c_i} g(\boldsymbol{\eta})^{\alpha_1+\beta_1} \exp\left(\boldsymbol{\eta}^T \boldsymbol{\beta}\right) \tag{157}$$

which is almost in exponential family normal form, except for:

- $\prod_{i=2}^{K-1} v_i(\boldsymbol{\eta})^{c_i}$ should be $m(\boldsymbol{\eta})$, i.e. it must not depend on the data, hence the c_i need to be constant. Since $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will be updated on observation, this requires that $\forall i = 2, \ldots, K-1 : \ \alpha_i = c_i - \beta_i + \beta_{i-1}$. This implies that $\boldsymbol{\alpha}$ is determined by $\boldsymbol{\beta}$ up to the c_i, except for α_1.
- $\nu := \alpha_1 + \beta_1$ and $\lambda_i := \frac{\beta_i}{\nu}$, hence $\alpha_1 = \nu(1 - \lambda_1)$.
- $f(\boldsymbol{\lambda}, \nu) = N(\boldsymbol{\alpha}(\boldsymbol{\lambda}, \nu), \nu\boldsymbol{\lambda})$

In the standard parametrization of the prior on the stick-breaking construction (eqn. 149), the α_i are the pseudocounts on the number of instances where $x_i = 1$. To see that our definition

$$\alpha_i := c_i - \beta_i + \beta_{i-1} \tag{158}$$

has the same meaning, assume that we had observed N datapoints $\boldsymbol{x}^{1:N}$ and corresponding sufficient statistics $\boldsymbol{u}^{1:N}$. Using the exponential family update rules (eqns. 24 and 25), we find for the posterior parameters

$$\alpha_1^N = \nu - \beta_1^N = \alpha_1 + \beta_1 + N - \beta_1 - \sum_{n=1}^{N} u_1^n$$

$$= \alpha_1 + N - \sum_{n=1}^{N} u_1^n$$

$$= \alpha_1 + \sum_{n=1}^{N} x_1 \tag{159}$$

$$\forall i = 2, \ldots, K-1 : \beta_i^N = \beta_i + \sum_{n=1}^{N} u_i^n \tag{160}$$

$$\alpha_i^N = c_i - \beta_i^N + \beta_{i-1}^N$$

$$= c_i - \beta_i + \beta_{i-1} + \sum_{n=1}^{N} (u_{i-1}^n - u_i^n)$$

$$= c_i + \beta_{i-1} - \beta_i + \sum_{n=1}^{N} x_i^n$$

$$= \alpha_i + \sum_{n=1}^{N} x_i^n \tag{161}$$

Hence, the meaning of $\boldsymbol{\alpha}$ is preserved.

In this table, let

$$\gamma_k = 1 + \exp(\eta_k)(1 + \exp(\eta_{k+1})(1 + \ldots (1 + \exp(\eta_{K-1}))\ldots))$$

multinomial stick-breaking distribution

standard form	$\prod_{k=1}^{K-1} v_k^{x_k} (1 - v_k)^{\sum_{i=k+1}^{K} x_i}$
constraints $\boldsymbol{x}$	$x_k \in \{0, 1\}$, $\sum_{k=1}^{K} x_k = 1$
constraints $\boldsymbol{v}$	$v_k \in [0, 1]$
$\boldsymbol{u}(\boldsymbol{x})$	$\forall k = 1, \ldots, K-1 : u_k = \sum_{i=k+1}^{K} x_i$
$\boldsymbol{\eta}$	$\eta_{K-1} = \log\left(\frac{1-v_{K-1}}{v_{K-1}}\right)$, $\forall k < K-1 : \eta_k = \log\left(\frac{1-v_k}{v_k} v_{k+1}\right)$
$\boldsymbol{v}$	$v_{K-1} = \frac{1}{1+\exp(\eta_{K-1})}$, $\forall k < K-1 : v_i = \frac{1}{1+\frac{\exp(\eta_i)}{v_{i+1}}}$
constraints $\boldsymbol{\eta}$	$\eta_k \in \mathbb{R}$
$g(\boldsymbol{\eta})$	$\frac{1}{\gamma_1} = v_1$
$h(x)$	1
$\langle \boldsymbol{u}(\boldsymbol{x}) \rangle$	$\langle u_k \rangle = \frac{\exp(\sum_{i=1}^{k} \eta_i)\gamma_{k+1}}{\gamma_1} = \prod_{i=1}^{k}(1 - v_i)$
$\mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x}))$	$C_{kl} = \frac{\exp\left(\sum_{i=1}^{max(k,l)} \eta_i\right)\gamma_{max(k,l)+1}}{\gamma_1} - \frac{\exp\left(\sum_{i=1}^{k} \eta_i\right)\gamma_{k+1} \exp\left(\sum_{i=1}^{k} \eta_i\right)\gamma_{l+1}}{\gamma_1^2}$ $= \prod_{i=1}^{max(k,l)}(1 - v_i) - \prod_{i=1}^{k}(1 - v_i) \prod_{j=1}^{l}(1 - v_j)$

Table 3: The stick-breaking distribution for multinomial variables

stick-breaking prior

standard form	$\prod_{i=1}^{K-1} \frac{1}{B(\alpha_i,\beta_i)} v_i^{\alpha_i-1}(1-v_i)^{\beta_i-1}$
constraints $\boldsymbol{\alpha},\boldsymbol{\beta}$	$\alpha_k,\beta_k \in \mathbb{R}^+$
$\boldsymbol{\lambda}$	$\frac{\boldsymbol{\beta}}{\nu}$
ν	$\alpha_1+\beta_1$
constraints $\boldsymbol{c}$	$\forall i=2,\ldots,K-1:\ c_i > \beta_i - \beta_{i-1}$
$\boldsymbol{\alpha}$	$\alpha_1=\nu(1-\lambda_1),\ \forall i=2,\ldots,K-1:\ \alpha_i = c_i-\beta_i+\beta_{i-1}$
constraints ν	$\nu \in \mathbb{R}^+$
constraints $\boldsymbol{\lambda}$	$\lambda_i \in \mathbb{R}^+$
$f(\boldsymbol{\lambda},\nu)$	$\frac{\Gamma(\nu)}{\Gamma(\nu\lambda_1)\Gamma(\nu(1-\lambda_1))}\prod_{i=2}^{K-1}\frac{\Gamma(c_i+\nu\lambda_{i-1})}{\Gamma(\nu\lambda_i)\Gamma(c_i-\nu\lambda_i+\nu\lambda_{i-1})}$
$m(\boldsymbol{\eta})$	$\prod_{i=2}^{K-1} v_i(\boldsymbol{\eta})^{c_i}$
$\forall i<K-1:\langle\eta_i\rangle$	$\psi(\nu\lambda_i)-\psi(c_i-\nu\lambda_i+\nu\lambda_{i-1})+\psi(c_{i+1}-\nu\lambda_{i+1}+\nu\lambda_i)-\psi(c_{i+1}+\nu\lambda_i)$ $=\psi(\beta_i)-\psi(\alpha_i)+\psi(\alpha_{i+1})-\psi(\alpha_{i+1}+\beta_{i+1})$
$\langle\eta_{K-1}\rangle$	$\psi(\nu\lambda_i)-\psi(c_i-\nu\lambda_i+\nu\lambda_{i-1})=\psi(\beta_i)-\psi(\alpha_i)$
$\mathrm{Var}(\eta_1)$	$\psi'(\nu\lambda_1)+\psi'(\nu(1-\lambda_1))-\psi'(c_2+\nu\lambda_1)+\psi'(c_2-\nu\lambda_2+\nu\lambda_1)$ $=\psi'(\beta_1)+\psi'(\alpha_1)-\psi'(\alpha_2+\beta_2)+\psi'(\alpha_2)$
$\forall 1<i<K-1:\mathrm{Var}(\eta_i)$	$\psi'(\nu\lambda_i)+\psi'(c_i-\nu\lambda_i+\nu\lambda_{i-1})-\psi'(c_{i+1}+\nu\lambda_i)+\psi'(c_{i+1}-\nu\lambda_{i+1}+\nu\lambda_i)$ $=\psi'(\beta_i)+\psi'(\alpha_i)-\psi'(\alpha_{i+1}+\beta_{i+1})+\psi'(\alpha_{i+1})$
$\mathrm{Cov}(\eta_i,\eta_{i+1})$	$\psi'(c_{i+1}-\nu\lambda_{i+1}+\nu\lambda_i)=\psi'(\alpha_{i+1})$
$\mathrm{Var}(\eta_{K-1})$	$\psi'(\nu\lambda_{K-1})+\psi'(c_{K-1}-\nu\lambda_{K-1}+\nu\lambda_{K-2})=\psi'(\beta_{K-1})+\psi'(\alpha_{K-1})$
$\frac{\partial\log(f(\boldsymbol{\lambda},\nu))}{\partial\nu}$	$\psi(\nu)-\lambda_1\psi(\nu\lambda_1)-(1-\lambda_1)\psi(\nu(1-\lambda_1))$ $+\sum_{i=2}^{K-1}[\lambda_{i-1}\psi(c_i+\nu\lambda_{i-1})-\lambda_i\psi(\nu\lambda_i)-(\lambda_{i+1}-\lambda_i)\psi(c_i-\nu\lambda_i+\nu\lambda_{i+1})]$
$\frac{\partial^2\log(f(\boldsymbol{\lambda},\nu))}{\partial\nu^2}$	$\psi'(\nu)-\lambda_1^2\psi'(\nu\lambda_1)-(1-\lambda_1)^2\psi'(\nu(1-\lambda_1))$ $+\sum_{i=2}^{K-1}\left[\lambda_{i-1}^2\psi'(c_i+\nu\lambda_{i-1})-\lambda_i^2\psi'(\nu\lambda_i)-(\lambda_{i+1}-\lambda_i)^2\psi'(c_i-\nu\lambda_i+\nu\lambda_{i+1})\right]$
$\langle g(\boldsymbol{\eta})\rangle$	$1-\lambda_1=\frac{\alpha_1}{\nu}$
$\mathrm{Var}(g(\eta))$	$\frac{\lambda_1(1-\lambda_1)}{nu+1}=\frac{\alpha_1\beta_1}{\nu^2(1-\nu)}$
$p(\boldsymbol{x}\|\boldsymbol{\lambda},\nu)$	$x_1=1 \quad : \quad 1-\lambda_1=\frac{\alpha_1}{\alpha_1+\beta_1}$ $x_{i\in\{2;\ldots;K-2\}}=1 \quad : \quad \lambda_1\prod_{j=2}^{i-1}\frac{\nu\lambda_j}{c_j+\nu\lambda_{j-1}}\frac{c_i-\nu\lambda_i+\nu\lambda_{i-1}}{c_i+\nu\lambda_{i-1}}=\sum_{j=1}^{i-1}\frac{\beta_j}{\alpha_j+\beta_j}\frac{\alpha_i}{\alpha_j+\beta_j}$ $x_K=1 \quad : \quad \lambda_1\prod_{j=2}^{K-1}\frac{\nu\lambda_j}{c_j+\nu\lambda_{j-1}}=\prod_{j=1}^{K-1}\frac{\beta_j}{\alpha_j+\beta_j}$

Table 4: The conjugate prior on the stick-breaking distribution for multinomial variables

4.4 Poisson-Gamma

The Poisson distribution is a distribution over a univariate integer-valued random variable x, e.g. spike count or radioactive decay events. In standard form, it is given by

$$p(x|r) = \frac{r^x \exp(-r)}{x!} \tag{162}$$

where $r \in \mathbb{R}_0^+$ is the rate. Its sufficient statistic and natural parameter are

$$u(x) = x \tag{163}$$

$$\eta = \log(r) \tag{164}$$

and hence

$$p(x|\eta) = \underbrace{\frac{1}{\Gamma(x+1)}}_{h(x)} \underbrace{\exp(-\exp(\eta))}_{g(\eta)} \exp(\eta\, u(x)). \tag{165}$$

Note that for $x \in \mathbb{N}$, $\Gamma(x+1) = x!$. The conjugate prior on r is a Gamma distribution with density

$$p(r|\alpha, S) = \frac{1}{\Gamma(\alpha) S^\alpha} r^{\alpha-1} \exp\left(-\frac{r}{S}\right) \tag{166}$$

where $\alpha \in \mathbb{R}_0^+$ is the shape parameter, and $S \in \mathbb{R}^+$ is the scale. To tranform this into exponential family form, let

$$\nu = \frac{1}{S}\lambda = \alpha S \tag{167}$$

and note that $\left|\frac{dr(\eta)}{d\eta}\right| = \exp(\eta) = r(\eta)$. Thus, we find

$$\begin{aligned} p(\eta|\nu,\lambda) &= \frac{\nu^{\nu\lambda}}{\Gamma(\nu\lambda)} r(\eta)^{\nu\lambda-1} \exp\left(-\nu r(\eta)\right) r(\eta) \\ &= \frac{\nu^{\nu\lambda}}{\Gamma(\nu\lambda)} \exp\left(\nu\lambda \log(r(\eta))\right) \exp\left(-\nu r(\eta)\right) \\ &= \underbrace{\frac{\nu^{\nu\lambda}}{\Gamma(\nu\lambda)}}_{f(\lambda,\nu)} \underbrace{\exp(-\exp(\eta)))^\nu}_{g(\eta)^\nu} \exp(\nu\lambda\eta) \end{aligned} \tag{168}$$

4.5 Multivariate Gaussian with Gauss-Wishart prior

The multivariate Gaussian is widely used, e.g. all finite-sized marginals of a Gaussian process are Gaussian. But it is also a standard ingredient in parametric models for regression, e.g. linear or other basis function. In the standard form, a multivariate Gaussian density of a vector-valued random variable with variates $\boldsymbol{x}$, $dim(\boldsymbol{x}) = D$ is parameterized by a mean vector $\boldsymbol{\mu}$ and a symmetric, positive definite covariance matrix $\boldsymbol{\Sigma}$:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi}^D \sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-0.5 \cdot (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{169}$$

Poisson distribution

standard form	$\frac{r^x \exp(-r)}{x!}$
constraints	$x \in \mathbb{N}_0$, $r \in \mathbb{R}_0^+$
$u(x)$	x
η	$\log(r)$
constraints η	$\eta \in \mathbb{R}$
$g(\eta)$	$\exp(-\exp(\eta))$
$h(x)$	$\frac{1}{\Gamma(x+1)}$
$\langle u(x) \rangle$	$\exp(\eta) = r$
$\mathrm{Var}(u(x))$	$\exp(\eta) = r$

Gamma prior for Poisson-distributed RV

standard form	$\frac{1}{\Gamma(\alpha) S^\alpha} r^{\alpha-1} \exp\left(-\frac{r}{S}\right)$
constraints	$\alpha \in \mathbb{R}_0^+$, $S \in \mathbb{R}^+$
λ	αS
ν	$\frac{1}{S}$
constraints	$\nu \in \mathbb{R}^+$, $\lambda \in \mathbb{R}_0^+$
$f(\lambda, \nu)$	$\frac{\nu^{\nu\lambda}}{\Gamma(\nu\lambda)}$
$m(\eta)$	1
$\langle \boldsymbol{\eta} \rangle$	$\psi(\nu\lambda) - \log(\nu)$
$\mathrm{Var}(\eta)$	$\psi'(\nu\lambda)$
$\frac{\partial \log(f(\boldsymbol{\lambda}, \nu))}{\partial \nu}$	$\lambda\left(\log(\nu) + 1 - \psi(\nu\lambda)\right)$
$\frac{\partial^2 \log(f(\boldsymbol{\lambda}, \nu))}{\partial \nu^2}$	$\lambda\left(\frac{1}{\nu} - \lambda\psi'(\nu\lambda)\right)$
$\langle g(\boldsymbol{\eta}) \rangle$	$\left(1 + \frac{1}{\nu}\right)^{-\lambda\nu}$
$\mathrm{Var}(g(\boldsymbol{\eta}))$	$\left(1 + \frac{2}{\nu}\right)^{-\lambda\nu} - \left(1 + \frac{1}{\nu}\right)^{-2\lambda\nu}$
$p(\boldsymbol{x}\|\boldsymbol{\lambda}, \nu)$	$\frac{\Gamma(\nu\lambda + x)}{\Gamma(\nu\lambda)\Gamma(x+1)} \frac{(\nu+1)^{-(\nu\lambda+x)}}{\nu^{-\lambda\nu}}$
$\langle u(x) \rangle_{p(x\|\lambda,\nu)}$	λ

Table 5: Poisson distribution and conjugate Gamma prior

It is often convenient to use the inverse of $\boldsymbol{\Sigma}$, called *precision matrix* $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1}$, which is also symmetric ($\boldsymbol{P} = \boldsymbol{P}^T$) and positive definite:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{P}) = \frac{\sqrt{|\boldsymbol{P}|}}{\sqrt{2\pi}^D} \exp\left(-\frac{1}{2} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{P} (\boldsymbol{x} - \boldsymbol{\mu})\right) \tag{170}$$

To transform the Gaussian into the exponential family normal form, we rewrite the exponent as

$$-\frac{1}{2} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{P} (\boldsymbol{x} - \boldsymbol{\mu}) = -\frac{1}{2} \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} + \boldsymbol{\mu}^T \boldsymbol{P} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{P} \boldsymbol{\mu} \tag{171}$$

Note that

$$\begin{aligned} \boldsymbol{x}^T \boldsymbol{P} \boldsymbol{x} &= \sum_i \sum_j \boldsymbol{x}_i \boldsymbol{P}_{i,j} \boldsymbol{x}_j = \sum_i \boldsymbol{P}_{i,i} \boldsymbol{x}_i^2 + 2 \sum_i \sum_{j<i} \boldsymbol{P}_{i,j} \boldsymbol{x}_i \boldsymbol{x}_j \\ \boldsymbol{\mu}^T \boldsymbol{P} \boldsymbol{x} &= \sum_i (\boldsymbol{P}\boldsymbol{\mu})_i \boldsymbol{x}_i \end{aligned} \tag{172}$$

We therefore introduce sufficient statistics and natural parameters comprised of three parts: first, from the diagonal elements in eqn. 172, the row vectors

$$\begin{aligned} \boldsymbol{u}_d &= (\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \ldots, \boldsymbol{x}_D^2)^T && (173) \\ \boldsymbol{\eta}_d &= -\frac{1}{2}(\boldsymbol{P}_{1,1}, \boldsymbol{P}_{2,2}, \ldots, \boldsymbol{P}_{d,d})^T. && (174) \end{aligned}$$

Second, we order the off-diagonal elements (with $i < j$) in some arbitrary fashion (e.g. lectically) and construct the vectors

$$\begin{aligned} \boldsymbol{u}_c &= (\boldsymbol{x}_2\boldsymbol{x}_1, \boldsymbol{x}_3\boldsymbol{x}_1, \boldsymbol{x}_3\boldsymbol{x}_2, \ldots, \boldsymbol{x}_D\boldsymbol{x}_{D-1})^T = \mathrm{lt}(\boldsymbol{x}\boldsymbol{x}^T) && (175) \\ \boldsymbol{\eta}_c &= -(\boldsymbol{P}_{2,1}, \boldsymbol{P}_{3,1}, \boldsymbol{P}_{3,2}, \ldots, \boldsymbol{P}_{D,D-1})^T = -\mathrm{lt}(\boldsymbol{P}) && (176) \end{aligned}$$

i.e. $\boldsymbol{\eta}_c$ contains the lower triangle of $\boldsymbol{P}$, and the lt() operator extracts the lower triangle of a matrix, excluding the diagonal. Third, from eqn. 172:

$$\begin{aligned} \boldsymbol{u}_\mu &= \boldsymbol{x} && (177) \\ \boldsymbol{\eta}_\mu &= \boldsymbol{P}\boldsymbol{\mu}. && (178) \end{aligned}$$

We stack these vectors into the total sufficient statistics and natural parameters

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{x}) &= \begin{pmatrix} \boldsymbol{u}_d \\ \boldsymbol{u}_c \\ \boldsymbol{u}_\mu \end{pmatrix} && (179) \\ \boldsymbol{\eta} &= \begin{pmatrix} \boldsymbol{\eta}_d \\ \boldsymbol{\eta}_c \\ \boldsymbol{\eta}_\mu \end{pmatrix}. && (180) \end{aligned}$$

Noting that $\boldsymbol{\mu}^T\boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\mu}^T\boldsymbol{P}^T\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu$ and $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{\eta})$, eqn. 171 can be written as

$$\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}) - \frac{1}{2}\boldsymbol{\mu}^t\boldsymbol{P}\boldsymbol{\mu} = \underbrace{\boldsymbol{\eta}_d^T\boldsymbol{u}_d + \boldsymbol{\eta}_c^T\boldsymbol{u}_c}_{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{P}\boldsymbol{x}} + \underbrace{\boldsymbol{\eta}_\mu^T\boldsymbol{u}_\mu}_{\mu^T\boldsymbol{P}\boldsymbol{x}} - \frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}(\boldsymbol{\eta})^{-1}\boldsymbol{\eta}_\mu \qquad (181)$$

With these substitutions, the exponential family normal form of the multivariate Gaussian is therefore

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \underbrace{\frac{\sqrt{|\boldsymbol{P}(\boldsymbol{\eta})|}}{\sqrt{2\pi}^D}\exp\left(-\frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}(\boldsymbol{\eta})^{-1}\boldsymbol{\eta}_\mu\right)}_{=:g(\boldsymbol{\eta})}\exp\left(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})\right) \qquad (182)$$

and thus

$$\log(g(\boldsymbol{\eta})) = -\frac{D}{2}\log(2\pi) + \frac{1}{2}\log\left(|\boldsymbol{P}(\boldsymbol{\eta})|\right) - \frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}(\boldsymbol{\eta})^{-1}\boldsymbol{\eta}_\mu. \qquad (183)$$

To compute moments, we need the gradient of this expression w.r.t. $\boldsymbol{\eta}$, whose components can be computed via the chain rule. Since

$$\begin{aligned} \frac{\partial \boldsymbol{P}_{i,j}}{\partial(\boldsymbol{\eta}_d)_k} &= -2\,\delta_{i,k}\delta_{j,k} && (184) \\ \frac{\partial \boldsymbol{P}_{i,j}}{\partial(\boldsymbol{\eta}_c)_{(k,l)}} &= -\delta_{i,k}\delta_{j,l} - \delta_{i,l}\delta_{j,k} && (185) \end{aligned}$$

and $\boldsymbol{P}$ does not depend on $\boldsymbol{\eta}_\mu$, we find

$$\frac{\partial \log(g(\boldsymbol{\eta}))}{\partial(\boldsymbol{\eta}_\mu)_k} = -\frac{1}{2} \cdot 2(\boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu)_k = -\boldsymbol{\mu}_k \tag{186}$$

$$\frac{\partial \log(g(\boldsymbol{\eta}))}{\partial(\boldsymbol{\eta}_d)_k} = \sum_{i,j} \frac{1}{2}\left[\boldsymbol{P}^{-1}_{i,j} + \boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu\boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\right] \cdot (-2)\delta_{i,k}\delta_{j,k} = -\boldsymbol{P}^{-1}_{k,k} - \boldsymbol{\mu}_k^2 \tag{187}$$

$$\frac{\partial \log(g(\boldsymbol{\eta}))}{\partial(\boldsymbol{\eta}_c)_{(k,l)}} = \sum_{i,j} \frac{1}{2}\left[\boldsymbol{P}^{-1}_{i,j} + \boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu\boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\right] \cdot (-\delta_{i,k}\delta_{j,l} - \delta_{i,l}\delta_{j,k}) = -\boldsymbol{P}^{-1}_{k,l} - \boldsymbol{\mu}_k\boldsymbol{\mu}_l \tag{188}$$

which implies, by virtue of eqn. 4:

$$\langle \boldsymbol{u}_\mu(\boldsymbol{x})\rangle = \boldsymbol{\mu} \tag{189}$$

$$\langle \boldsymbol{u}_d(\boldsymbol{x})\rangle = -\mathrm{diag}(\boldsymbol{P}^{-1}) - \boldsymbol{\mu}^T\boldsymbol{\mu} \tag{190}$$

$$\langle \boldsymbol{u}_c(\boldsymbol{x})_{(kl)}\rangle = -\boldsymbol{P}^{-1}_{k,l} - \boldsymbol{\mu}_k\boldsymbol{\mu}_l \tag{191}$$

i.e. the well-known expressions for expectations of Gaussians. The second derivatives (necessary for the evaluation of the gradient of the KL divergence) are omitted here, they can be computed by automatic differentiation from the above expressions, e.g. by `Theano`.

The prior on the parameters of the Gaussian is given by a (multivariate) Gauss-Wishart distribution [1]. In standard form, it is given by

$$p(\boldsymbol{\mu}, \boldsymbol{P}|\beta, \boldsymbol{\mu}_0, \gamma, \boldsymbol{V}) = p(\boldsymbol{\mu}|\beta, \boldsymbol{P})p(\boldsymbol{P}|\gamma, \boldsymbol{V}) \tag{192}$$

$$p(\boldsymbol{\mu}|\beta, \boldsymbol{P}) = \frac{\beta^{\frac{D}{2}}|\boldsymbol{P}|^{\frac{1}{2}}}{\sqrt{2\pi}^D} \exp\left(-\frac{1}{2}\beta(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{P}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right) \tag{193}$$

$$p(\boldsymbol{P}|\gamma, \boldsymbol{V}) = \frac{|\boldsymbol{P}|^{\frac{\gamma-D-1}{2}}}{2^{\frac{\gamma D}{2}}|\boldsymbol{V}|^{\frac{\gamma}{2}}\Gamma_D\left(\frac{\gamma}{2}\right)} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{P})\right) \tag{194}$$

where $\Gamma_D(x)$ is the multivariate gamma function [3]:

$$\Gamma_D(x) = \pi^{\frac{D(D-1)}{4}} \prod_{j=1}^{D} \Gamma\left(x + \frac{1-j}{2}\right) \tag{195}$$

We reparameterize as follows:

$$\beta = \nu \tag{196}$$

$$\gamma = \nu + \alpha,\ \alpha > 0 \text{ and const.} \tag{197}$$

$$\boldsymbol{\lambda}_\mu = \boldsymbol{\mu}_0 \tag{198}$$

$$\boldsymbol{B} = \boldsymbol{V}^{-1} \tag{199}$$

$$(\boldsymbol{\lambda}_d)_i = \left(\frac{B_{i,i}}{\nu} + (\boldsymbol{\mu}_0)_i^2\right) \tag{200}$$

$$(\boldsymbol{\lambda}_c)_{(k,l)} = \frac{B_{k,l}}{\nu} + (\boldsymbol{\mu}_0)_k(\boldsymbol{\mu}_0)_l = \frac{B_{l,k}}{\nu} + (\boldsymbol{\mu}_0)_k(\boldsymbol{\mu}_0)_l \tag{201}$$

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_d, \boldsymbol{\lambda}_c)^T \tag{202}$$

Using these substitutions and $\boldsymbol{\mu}^T\boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu$, the arguments of the exponentials in eqn. 192 can be written as

$$\begin{aligned} & -\nu\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{P}\boldsymbol{\mu} + \nu\boldsymbol{\mu}_0^T\boldsymbol{P}\boldsymbol{\mu} - \nu\frac{1}{2}\boldsymbol{\mu}_0^T\boldsymbol{P}\boldsymbol{\mu}_0 - \frac{1}{2}\mathrm{tr}(\boldsymbol{B}\boldsymbol{P}) \\ = \;& \nu\left(-\frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu + \boldsymbol{\eta}_\mu^T\boldsymbol{\mu}_0 - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\left(\frac{\boldsymbol{B}}{\nu} + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^T\right)\right)\right) \\ = \;& \nu\left(-\frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu + \boldsymbol{\eta}_\mu^T\boldsymbol{\lambda}_\mu + \boldsymbol{\eta}_d^T\boldsymbol{\lambda}_d s + \boldsymbol{\eta}_c^T\boldsymbol{\lambda}_c\right) \end{aligned} \tag{203}$$

Thus eqn. 192 becomes

$$\begin{aligned} p(\boldsymbol{\mu},\boldsymbol{P}|\nu,\boldsymbol{\mu}_0,\boldsymbol{B}) \;=\; & \sqrt{2\pi}^{(\nu-1)D}\,\frac{\nu^{-\frac{D}{2}}|B(\boldsymbol{\lambda})|^{\frac{\nu+\alpha}{2}}}{2^{\frac{(\nu+\alpha)D}{2}}\Gamma_D\left(\frac{\nu+\alpha}{2}\right)}|P|^{\frac{\alpha-D}{2}} \\ \times\; & \left(\frac{1}{\sqrt{2\pi}^D}\right)^\nu |P|^{\frac{\nu}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu\right)^\nu \\ \times\; & \exp\left(\nu\boldsymbol{\eta}_\mu^T\boldsymbol{\lambda}_\mu + \nu\boldsymbol{\eta}_d^T\boldsymbol{\lambda}_d + \nu\boldsymbol{\eta}_c^T\boldsymbol{\lambda}_c\right). \end{aligned} \tag{204}$$

To transform this into the desired exponential family form, we need to change $\boldsymbol{\mu}, \boldsymbol{P}$ into the natural parameters of the multivariate Gaussian (see eqns. 174,176,178). This can be accomplished by multiplying eqn.192 with the determinant of the Jacobian $\mathbf{J}$ of that transformation, which can be constructed as a block matrix as follows: stack the diagonal elements of $\boldsymbol{P}$ into the vector $\boldsymbol{P}_d$ of dimensionality D, the off-diagonal lower triangle into the vector $\boldsymbol{P}_c$ of dimensionality $Q = \frac{D(D-1)}{2}$. The Jacobian then has the following structure, which follows from the definitions in eqns. 174,176,178 (rows and columns labelled with variable names):

$$\mathbf{J} = \begin{array}{c|c|c|c|} & \boldsymbol{\mu}_1 \;\ldots\; \boldsymbol{\mu}_D & (\boldsymbol{P}_d)_1 \;\ldots\; (\boldsymbol{P}_d)_D & (\boldsymbol{P}_c)_1 \;\ldots\; (\boldsymbol{P}_c)_Q \\ \hline (\boldsymbol{\eta}_\mu)_1 \\ \vdots \\ (\boldsymbol{\eta}_\mu)_D & \boldsymbol{P}^{-1} & \mathbf{0} & \mathbf{0} \\ \hline (\boldsymbol{\eta}_d)_1 \\ \vdots \\ (\boldsymbol{\eta}_d)_D & \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\eta}_d} & -2\cdot\mathbf{1}_D & \mathbf{0} \\ (\boldsymbol{\eta}_c)_1 \\ \vdots \\ (\boldsymbol{\eta}_c)_Q & \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\eta}_c} & \mathbf{0} & -\mathbf{1}_Q \\ \hline \end{array} \tag{205}$$

By virtue of eqns. 397 and 398 in [2], the absolute value of the determinant of $\mathbf{J}$ is therefore

$$|J| = 2^D|P^{-1}(\boldsymbol{\eta})| \tag{206}$$

Reparameterizing eqn. 204 in terms of $\boldsymbol{\eta}$ and multiplying with this expression

yields (cf. eqn. 182 for the definition of $g(\boldsymbol{\eta})$):

$$\begin{aligned} p(\boldsymbol{\eta}|\nu,\boldsymbol{\mu}_0,\boldsymbol{B}) &= \underbrace{2^D\sqrt{2\pi}^{(\nu-1)D}\frac{\nu^{\frac{D}{2}}|\boldsymbol{B}(\boldsymbol{\lambda})|^{\frac{\nu+\alpha}{2}}}{2^{\frac{(\nu+\alpha)D}{2}}\Gamma_D\left(\frac{\nu+\alpha}{2}\right)}}_{=:f(\nu,\boldsymbol{\lambda})}\underbrace{|\boldsymbol{P}|^{\frac{\alpha-D}{2}-1}}_{=:m(\boldsymbol{\eta})} \\ &\times \underbrace{\left(\frac{1}{\sqrt{2\pi}^D}\right)^{\nu}|P|^{\frac{\nu}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}^{-1}\boldsymbol{\eta}_\mu\right)^{\nu}}_{=g(\boldsymbol{\eta})^\nu} \\ &\times \exp\left(\nu\boldsymbol{\eta}_\mu^T\boldsymbol{\lambda}_\mu+\nu\boldsymbol{\eta}_d^T\boldsymbol{\lambda}_d+\nu\boldsymbol{\eta}_c^T\boldsymbol{\lambda}_c\right) \\ &= f(\nu,\boldsymbol{\lambda})m(\boldsymbol{\eta})g(\boldsymbol{\eta})^\nu\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda}) \end{aligned} \tag{207}$$

which is the exponential family normal form of the Gauss-Wishart prior on the parameters of the multivariate Gaussian.

To calculate expectations, we need the derivatives of $\log(f(\nu,\lambda))$:

$$\log(f(\nu,\lambda)) = C+\frac{D\nu}{2}\log(2\pi)+\frac{D}{2}\log(\nu)-\log\left(\Gamma_D\left(\frac{\nu+\alpha}{2}\right)\right)+\frac{\nu+\alpha}{2}\log(|\boldsymbol{B}(\boldsymbol{\lambda})|) \tag{208}$$

with $C = D\log(2) - \frac{D}{2}\log(2\pi) - \frac{\alpha D}{2}\log(2)$. Thus

$$\frac{\partial f(\nu,\lambda)}{\partial\nu} = \frac{D}{2}\log(2\pi)+\frac{D}{2\nu}-\frac{1}{2}\Psi_D\left(\frac{\nu+\alpha}{2}\right)+\frac{1}{2}\log(|\boldsymbol{B}(\boldsymbol{\lambda})|) \tag{209}$$

where $\Psi_D(x) = \frac{\partial\log(\Gamma_D(x))}{\partial x}$ is the multivariate digamma function. For the derivatives w.r.t. $\boldsymbol{\lambda}$, note that eqns. 196-202 imply $B_{i,i} = \nu(2(\lambda_d)_i - (\lambda_m u)_i^2$ and $\forall i > j : B_i j = \nu((\lambda_c)_{i,j} - (\lambda_\mu)_i(\lambda_\mu)_j$, $\forall j > i : B_i j = \nu((\lambda_c)_{j,i} - (\lambda_\mu)_i(\lambda_\mu)_j$, hence

$$\frac{\partial B_{i,j}}{\partial(\boldsymbol{\lambda}_\mu)_k} = -\nu((\boldsymbol{\lambda}_\mu)_i\delta_{j,k}+(\boldsymbol{\lambda}_\mu)_j\delta_{i,k}) \tag{210}$$

$$\frac{\partial B_{i,j}}{\partial(\boldsymbol{\lambda}_d)_k} = \nu\delta_{i,k}\delta_{j,k} \tag{211}$$

$$\frac{\partial B_{i,j}}{\partial(\boldsymbol{\lambda}_c)_{k,l}} = \nu(\delta_{i,k}\delta_{j,l}+\delta_{i,l}\delta_{j,k}) \tag{212}$$

With formula 57 from [2], we therefore find

$$\begin{aligned} \frac{\partial\log(f(\nu,\boldsymbol{\lambda}))}{\partial(\boldsymbol{\lambda}_d)_k} &= \sum_{m,n}\frac{\nu+\alpha}{2}\frac{\partial\log(|\boldsymbol{B}|)}{\partial B_{m,n}}\frac{\partial B_{m,n}}{\partial(\boldsymbol{\lambda}_d)_k} \\ &= \frac{\nu(\nu+\alpha)}{2}\sum_{m,n}\boldsymbol{B}^{-1}_{m,n}\delta_{m,k}\delta_{n,k} = -\nu(\nu+\alpha)\boldsymbol{B}^{-1}_{k,k} \end{aligned} \tag{213}$$

$$\frac{\partial\log(f(\nu,\boldsymbol{\lambda}))}{\partial(\boldsymbol{\lambda}_c)_{k,l}} = \nu(\nu+\alpha)\boldsymbol{B}^{-1}_{k,l} \tag{214}$$

$$\begin{aligned} \frac{\partial\log(f(\nu,\boldsymbol{\lambda}))}{\partial(\boldsymbol{\lambda}_\mu)_k} &= -\nu\frac{(\nu+\alpha)}{2}\sum_{m,n}\boldsymbol{B}^{-1}_{m,n}((\lambda_\mu)_m\delta_{n,k}+(\lambda_\mu)_n\delta_{m,k}) \\ &= -\nu(\nu+\alpha)\sum_n\boldsymbol{B}^{-1}_{k,n}(\boldsymbol{\lambda}_\mu)_n \end{aligned} \tag{215}$$

The expectations of the sufficient statistics of the multivariate Gaussian are thus (lt($\boldsymbol{X}$): vectorized form of lower triangle of $\boldsymbol{X}$)

$$\begin{aligned}
\langle \boldsymbol{\eta}_\mu \rangle &= \langle \boldsymbol{P}\boldsymbol{\mu} \rangle = (\nu+\alpha)\boldsymbol{B}^{-1}(\boldsymbol{\lambda})\lambda_\mu = \gamma \boldsymbol{V}\boldsymbol{\mu}_0 && (216)\\
\langle \boldsymbol{\eta}_d \rangle &= -\frac{1}{2}\langle \mathrm{diag}(\boldsymbol{P}) \rangle = -\frac{(\nu+\alpha)}{2}\mathrm{diag}(\boldsymbol{B}^{-1}) && (217)\\
\langle \boldsymbol{\eta}_c \rangle &= -\langle \mathrm{lt}(\boldsymbol{P}) \rangle = -(\nu+\alpha)\mathrm{lt}(\boldsymbol{B}^{-1}) && (218)\\
&\Rightarrow \quad \langle \boldsymbol{P} \rangle = \gamma \boldsymbol{V} && (219)
\end{aligned}$$

Second derivatives can be obtained similarly, or more easily by automatic differentiation.

The expectation of the sufficient statistics $\boldsymbol{u}(\boldsymbol{x})$ can be computed from eqn. 57. We already computed $\boldsymbol{\lambda}$ (eqn. 202). The surface integral vanishes, because $p(\boldsymbol{x}|\boldsymbol{\lambda},\nu) \to 0$ as $\boldsymbol{\lambda}_\mu \to \infty$, since the density must be normalized. Furthermore, in the subspace where $|\boldsymbol{P}| = 0$, the density is also zero if $\gamma > D-1$. Since $m(\boldsymbol{\eta})$ does not depend on $\boldsymbol{\mu}_0$, we see that

$$\langle \boldsymbol{u}_\mu(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{\lambda},\boldsymbol{\mu})} = \boldsymbol{\lambda}_\mu = \boldsymbol{\mu}_0 \tag{220}$$

For $\boldsymbol{u}_d,\ \boldsymbol{u}_c$, it remains to evaluate the expectation

$$\begin{aligned}
\langle \nabla_{\boldsymbol{\eta}} \log(m(\boldsymbol{\eta})) \rangle_{p(\boldsymbol{x}|\boldsymbol{\lambda},\boldsymbol{\mu})} &= -2\left\langle \left(\frac{\alpha-D}{2}-1\right)\boldsymbol{P}^{-1} \right\rangle_{p(\boldsymbol{x}|\boldsymbol{\lambda},\boldsymbol{\mu})}\\
&= -(\alpha-D-2)\frac{\boldsymbol{V}^{-1}}{\nu+\alpha-D-1} && (221)
\end{aligned}$$

where the factor -2 results from the differentation of the elements of $\boldsymbol{P}$ w.r.t. $\boldsymbol{\eta}$, see also the Jacobian above. The entries of $\boldsymbol{V}^{-1}$ have to be suitably rearranged to match the entries of $\boldsymbol{\eta}$. Thus we find

$$\begin{aligned}
\langle \boldsymbol{u}_d(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{\lambda},\boldsymbol{\mu})} &= \mathrm{diag}(\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T) + \frac{\mathrm{diag}(\boldsymbol{V}^{-1})}{\nu} - \frac{\alpha-D-2}{\nu}\frac{\mathrm{diag}(\boldsymbol{V}^{-1})}{\nu+\alpha-D-1}\\
&= \mathrm{diag}(\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T) + \frac{\nu+1}{\nu}\frac{\mathrm{diag}(\boldsymbol{V}^{-1})}{\nu+\alpha-D-1} && (222)
\end{aligned}$$

and likewise for $\boldsymbol{u}_c$.

4.5.1 Univariate Gauss-Gauss-Gamma

trivial, see above.

5 Random identities that might be useful

Decomposition of Kullback-Leibler divergence for multivariate Gaussians: let

$$p(\boldsymbol{x}_1,\boldsymbol{x}_2|\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{K}_{11},\boldsymbol{K}_{21},\boldsymbol{K}_{22}) = \mathcal{N}\left((\boldsymbol{x}_1,\boldsymbol{x}_2)|(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2),\begin{pmatrix}\boldsymbol{K}_{11} & \boldsymbol{K}_{21}^T\\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22}\end{pmatrix}\right) \tag{223}$$

Multivariate Gaussian distribution

standard form	$\frac{1}{\sqrt{2\pi}^D\sqrt{\lvert\boldsymbol{\Sigma}\rvert}}\exp\left(-0.5\cdot(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$
constraints	$\boldsymbol{x},\boldsymbol{\mu}\in\mathbb{R}^D$, $\boldsymbol{\Sigma}$ positive semidefinite and symmetric
$\boldsymbol{u}(\boldsymbol{x})$	$\boldsymbol{u}_d=\mathrm{diag}(\boldsymbol{x}\boldsymbol{x}^T)$,$\boldsymbol{u}_c=\mathrm{lt}(\boldsymbol{x}\boldsymbol{x}^T)$,$\boldsymbol{u}_\mu=\boldsymbol{x}$
$\boldsymbol{\eta}$	$\boldsymbol{\eta}_d=-\frac{1}{2}\mathrm{diag}(\boldsymbol{P})$, $\boldsymbol{\eta}_c=-\mathrm{lt}(\boldsymbol{P})$, $\boldsymbol{\eta}_\mu=\boldsymbol{P}\boldsymbol{\mu}$
constraints η	$\boldsymbol{\eta}_\mu\in\mathbb{R}^D$, $\boldsymbol{\eta}_d\leq 0$, $\boldsymbol{\eta}_c$ s.t. $\boldsymbol{P}$ pos.semidef.
$g(\boldsymbol{\eta})$	$\frac{\sqrt{\lvert\boldsymbol{P}(\boldsymbol{\eta})\rvert}}{\sqrt{2\pi}^D}\exp\left(-\frac{1}{2}\boldsymbol{\eta}_\mu^T\boldsymbol{P}(\boldsymbol{\eta})^{-1}\boldsymbol{\eta}_\mu\right)$
$h(x)$	1
$\langle\boldsymbol{u}(\boldsymbol{x})\rangle$	$\langle\boldsymbol{u}_\mu\rangle=\boldsymbol{\mu}$, $\langle\boldsymbol{u}_d\rangle=-\mathrm{diag}(\boldsymbol{P}^{-1}+\boldsymbol{\mu}\boldsymbol{\mu}^T)$, $\langle\boldsymbol{u}_c\rangle=-\mathrm{lt}(\boldsymbol{P}^{-1}+\boldsymbol{\mu}\boldsymbol{\mu}^T)$

Gauss-Wishart prior for multivariate Gaussian RV

standard form	$\frac{\beta^{\frac{D}{2}}\lvert P\rvert^{\frac{1}{2}}}{\sqrt{2\pi}^D}\exp\left(-\frac{1}{2}\beta(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{P}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right)\frac{\lvert\boldsymbol{P}\rvert^{\frac{\gamma-D-1}{2}}}{2^{\frac{\gamma D}{2}}\lvert\boldsymbol{V}\rvert^{\frac{\gamma}{2}}\Gamma_D\left(\frac{\gamma}{2}\right)}\exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{P})\right)$
constraints	$\boldsymbol{\mu},\boldsymbol{\mu}_0\in\mathbb{R}^D$, $\boldsymbol{P},\boldsymbol{V}$ pos.semidef., $\beta>0$,$\gamma>D-1$
$\boldsymbol{\lambda}$	$\boldsymbol{B}=\boldsymbol{V}^{-1}$, $\boldsymbol{\lambda}_d=\mathrm{diag}\left(\frac{\boldsymbol{B}}{\nu}+\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T\right)$, $\boldsymbol{\lambda}_c=\mathrm{lt}\left(\frac{\boldsymbol{B}}{\nu}+\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T\right)$, $\boldsymbol{\lambda}_\mu=\boldsymbol{\mu}_0$
ν	$\nu=\beta$, $\gamma=\nu+\alpha$
constraints	$\nu\in\mathbb{R}^+$, α const. and s.t. $\nu+\alpha>D-1$
$f(\nu,\boldsymbol{\lambda})$	$2^{\frac{D(1-\alpha)}{2}}\sqrt{\pi}^{(\nu-1)D}\frac{\nu^{-\frac{D}{2}}\lvert\boldsymbol{B}(\boldsymbol{\lambda})\rvert^{\frac{\nu+\alpha}{2}}}{\Gamma_D\left(\frac{\nu+\alpha}{2}\right)}$
$m(\boldsymbol{\eta})$	$\lvert\boldsymbol{P}(\boldsymbol{\eta})\rvert^{\frac{\alpha-D}{2}-1}$
$\langle\boldsymbol{\eta}\rangle$	$\langle\boldsymbol{\eta}_\mu\rangle=(\nu+\alpha)\boldsymbol{B}^{-1}(\boldsymbol{\lambda})\boldsymbol{\lambda}_\mu$, $\langle\boldsymbol{\eta}_d\rangle=-(\nu+\alpha)\mathrm{diag}(\boldsymbol{B}^{-1})$, $\langle\boldsymbol{\eta}_c\rangle=-(\nu+\alpha)\mathrm{lt}(\boldsymbol{B}^{-1})$
$\frac{\partial\log(f(\boldsymbol{\lambda},\nu))}{\partial\nu}$	$\frac{D}{2}\log(\pi)-\frac{D}{2\nu}-\frac{1}{2}\Psi_D\left(\frac{\nu+\alpha}{2}\right)+\frac{1}{2}\log(\lvert\boldsymbol{B}(\boldsymbol{\lambda})\rvert)$
$p(\boldsymbol{x}\lvert\boldsymbol{\lambda},\nu)$	$\frac{\Gamma_D\left(\frac{\nu+\alpha+1}{2}\right)}{\sqrt{\pi}\Gamma_D\left(\frac{\nu+\alpha}{2}\right)}\left(\frac{\nu}{\nu+1}\right)^{\frac{D}{2}}\frac{\lvert\boldsymbol{B}(\boldsymbol{\lambda})\rvert^{\frac{\nu+\alpha}{2}}}{\lvert\boldsymbol{B}\left(\frac{\nu\boldsymbol{\lambda}+\boldsymbol{u}(\boldsymbol{x})}{\nu+1}\right)\rvert^{\frac{\nu+\alpha+1}{2}}}$
$\langle\boldsymbol{u}_\mu(\boldsymbol{x})\rangle_{p(\boldsymbol{x}\lvert\boldsymbol{\lambda},\mu)}$	$\boldsymbol{\lambda}_u=\boldsymbol{\mu}_0$
$\langle\boldsymbol{u}_d(\boldsymbol{x})\rangle_{p(\boldsymbol{x}\lvert\boldsymbol{\lambda},\mu)}$	$\boldsymbol{\lambda}_d-\frac{\alpha-D-2}{\nu}\frac{\mathrm{diag}(\boldsymbol{V}^{-1})}{\nu+\alpha-D-1}=\mathrm{diag}(\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T)+\frac{\nu+1}{\nu}\mathrm{diag}(\boldsymbol{V}^{-1})$
$\langle\boldsymbol{u}_c(\boldsymbol{x})\rangle_{p(\boldsymbol{x}\lvert\boldsymbol{\lambda},\mu)}$	$\boldsymbol{\lambda}_c-\frac{\alpha-D-2}{\nu}\frac{\mathrm{lt}(\boldsymbol{V}^{-1})}{\nu+\alpha-D-1}=\mathrm{lt}(\boldsymbol{\mu}_0\boldsymbol{\mu}_0^T)+\frac{\nu+1}{\nu}\mathrm{lt}(\boldsymbol{V}^{-1})$

Table 6: Multivariate Gaussian distribution and conjugate Gauss-Wishart prior

be a joint multivariate Gaussian on $\boldsymbol{x}_1,\boldsymbol{x}_2$. The conditional distribution of $\boldsymbol{x}_2$ given $\boldsymbol{x}_1$ is then also multivariate Gaussian (see [2]):

$$\begin{aligned}\boldsymbol{M}_{21} &= \boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1} &(224)\\ \boldsymbol{K}_{2|1} &= \boldsymbol{K}_{22}-\boldsymbol{M}_{21}\boldsymbol{K}_{11}^{-1}\boldsymbol{M}_{21}^T &(225)\\ p(\boldsymbol{x}_1|\boldsymbol{x}_2) &= \mathcal{N}\left(\boldsymbol{x}_2|\boldsymbol{\mu}_2+\boldsymbol{M}_{21}(\boldsymbol{x}_1-\boldsymbol{\mu}_1),\boldsymbol{K}_{2|1}\right) &(226)\end{aligned}$$

Assume now we had a variational posterior for $\boldsymbol{x}_1,\boldsymbol{x}_2$, which we decompose in the following way (tilde indicates variational parameters):

$$q(\boldsymbol{x}_1,\boldsymbol{x}_2)=q(\boldsymbol{x}_2|\boldsymbol{x}_1)q(\boldsymbol{x}_1)=\mathcal{N}\left(\boldsymbol{x}_2|\tilde{\boldsymbol{\mu}}_2+\tilde{\boldsymbol{M}}_{21}\boldsymbol{x}_1,\tilde{\boldsymbol{K}}_{2|1}\right)\mathcal{N}\left(\boldsymbol{x}_1|\tilde{\boldsymbol{\mu}}_1,\tilde{\boldsymbol{K}}_1\right) \quad (227)$$

which is a generalized form of the conditional Gaussian above, because $\tilde{\boldsymbol{\mu}}_{2|1},\tilde{\boldsymbol{M}}_{21},\tilde{\boldsymbol{K}}_{2|1}$ are now free variational parameters. Note the following decomposition property

of the KL-divergence, which follows directly from its definition:

$$D\left(q(\boldsymbol{x}_1,\boldsymbol{x}_2)|p(\boldsymbol{x}_1,\boldsymbol{x}_2)\right) = \left\langle D\left(q(\boldsymbol{x}_2|\boldsymbol{x}_1)|p(\boldsymbol{x}_2|\boldsymbol{x}_1)\right)\right\rangle_{q(\boldsymbol{x}1)} + D\left(q(\boldsymbol{x}_1)|p(\boldsymbol{x}_1)\right) \tag{228}$$

Using the above distributions, the second term on the right hand side is given by the usual expression for the KL-divergence between multivariate Gaussians:

$$\begin{aligned} D\left(q(\boldsymbol{x}_1)|p(\boldsymbol{x}_1)\right) &= \frac{1}{2}\Big(\mathrm{tr}\left[\boldsymbol{K}_1^{-1}\tilde{\boldsymbol{K}}_1\right] + (\tilde{\boldsymbol{\mu}}_1-\boldsymbol{\mu}1)^T\boldsymbol{K}_1^{-1}(\tilde{\boldsymbol{\mu}}_1-\boldsymbol{\mu}1) \\ &\quad -\dim[\boldsymbol{x}_1] + \log(|\boldsymbol{K}_1|) - \log(|\tilde{\boldsymbol{K}}_1|)\Big) \end{aligned} \tag{229}$$

and the first term is

$$\begin{aligned} &\left\langle D\left(q(\boldsymbol{x}_2|\boldsymbol{x}_1)|p(\boldsymbol{x}_2|\boldsymbol{x}_1)\right)\right\rangle_{q(\boldsymbol{x}1)} \\ = \;& \frac{1}{2}\Big(\mathrm{tr}\left[\boldsymbol{K}_{2|1}^{-1}\tilde{\boldsymbol{K}}_{2|1}\right] + \\ &+\Big(\tilde{\boldsymbol{\mu}}_2+\tilde{\boldsymbol{M}}_{2|1}\tilde{\boldsymbol{\mu}}_1-(\boldsymbol{\mu}_2+\boldsymbol{M}_{2|1}(\tilde{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1))\Big)^T\boldsymbol{K}_1^{-1}\Big(\tilde{\boldsymbol{\mu}}_2+\tilde{\boldsymbol{M}}_{2|1}\tilde{\boldsymbol{\mu}}_1-(\boldsymbol{\mu}_2+\boldsymbol{M}_{2|1}(\tilde{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1))\Big) \\ &-\dim[\boldsymbol{x}_2] + \log(|\boldsymbol{K}_{2|1}|) - \log(|\tilde{\boldsymbol{K}}_{2|1}|) \\ &+\mathrm{tr}\left[(\tilde{\boldsymbol{M}}_{2|1}-\boldsymbol{M}_{2|1})^T\boldsymbol{K}_{2|1}^{-1}(\tilde{\boldsymbol{M}}_{2|1}-\boldsymbol{M}_{2|1})\tilde{\boldsymbol{K}}_{2|1}\right]\Big) \end{aligned} \tag{230}$$

This expression is zero iff: $\tilde{\boldsymbol{M}}_{2|1}=\boldsymbol{M}_{2|1}, \boldsymbol{K}_{2|1}=\tilde{\boldsymbol{K}}_{2|1}, \tilde{\boldsymbol{\mu}}_1=\boldsymbol{\mu}_1, \tilde{\boldsymbol{\mu}}_2=\boldsymbol{\mu}_2-\boldsymbol{M}_{2|1}\boldsymbol{\mu}_1$. Note that a zero expectation can not be achieved if we had made the Ansatz $q(\boldsymbol{x}_2|\boldsymbol{x}_1)=\mathcal{N}\left(\boldsymbol{x}_2|\tilde{\boldsymbol{\mu}}_{2|1},\tilde{\boldsymbol{K}}_{2|1}\right)$, because the KL-divergence is positive, thus for its expectation to be zero, it has to be zero for all values of $\boldsymbol{x}_1$ which requires an explicit representation of the mean projection matrix $\boldsymbol{M}$.

References

[1] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2007.

[2] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Version 20121115, http://www.math.uwaterloo.ca/ hwolkowi/matrixcookbook.pdf

[3] A.T. James. *Distributions of Matrix Variates and Latent Roots Derived from Normal Samples*. Ann. Math. Statist. 35 (1964), no. 2, 475–501.