Supplementary material for Non-conjugate Variational Message Passing for Multinomial and Binary Regression

October 29, 2011

1 Alternative derivation

We will focus on a particular factor f_a and variable x_i, with the aim of calculating an exponential family message $m_{a\to i}(x_i;\phi)$, parameterised by the natural parameter ϕ. Consider the local exclusive KL with exponential family $q_i(x_i;\theta)$, with natural parameter θ. Here $q^{\setminus a}(\mathbf{x})$ is the cavity distribution and $q^{\setminus i}(\mathbf{x}^{\setminus i}) = \prod_{j\neq i} q_j(x_j)$ is the current variational distribution over variables other than x_i.

$$\mathrm{KL}(q_i(x_i;\theta)q^{\setminus i}(\mathbf{x}^{\setminus i})||f_a(\mathbf{x})q^{\setminus a}(\mathbf{x})) = \int q_i(x_i;\theta)\log q_i(x_i;\theta)dx_i \tag{1}$$

$$- \int q_i(x_i;\theta)q^{\setminus i}(\mathbf{x}^{\setminus i})\log f_a(\mathbf{x})d\mathbf{x} \tag{2}$$

$$- \int q_i(x_i;\theta)q^{\setminus i}(\mathbf{x}^{\setminus i})\log q^{\setminus a}(\mathbf{x})d\mathbf{x} + \text{const.} \tag{3}$$

The cavity distribution itself factorises as $q^{\setminus a}(\mathbf{x}) = q_i(x_i;\theta^{\setminus a})q^{\setminus a,i}(\mathbf{x}^{\setminus i})$, where $q_i(x_i;\theta^{\setminus a})$ is the product of all the other incoming messages to x_i.

$$\mathrm{KL}(\theta) = - H[q_i(x_i;\theta)] \tag{4}$$

$$- \int q_i(x_i;\theta)\langle\log f_a(\mathbf{x})\rangle_{\sim q_i(x_i)}dx_i \tag{5}$$

$$- \int q_i(x_i;\theta)\log q_i(x_i;\theta^{\setminus a})dx_i + \text{const.} \tag{6}$$

$$= \theta^T\frac{\partial\kappa}{\partial\theta} - \kappa(\theta) - S(\theta) - \theta^{\setminus a}\frac{\partial\kappa}{\partial\theta} + \kappa(\theta^{\setminus a}) + \text{const.} \tag{7}$$

where we have used the fact that the expectation of the sufficient statistics of an exponential family are given by the derivatives of κ. The variational posterior $q_i(x_i;\theta)$ will be updated to $m_{a\to i}(x_i;\phi)q_i(x_i;\theta^{\setminus a})$, so we have the relationship

$\theta = \theta^{\setminus a} + \phi$. We assume that $\theta^{\setminus a}$ is fixed (which is at least true once the algorithm has converged), so differentiating wrt to θ and ϕ is equivalent:

$$\frac{\partial \mathrm{KL}}{\partial \phi} = \frac{\partial \mathrm{KL}}{\partial \theta} = H(\theta)\theta + \frac{\partial \kappa}{\partial \theta} - \frac{\partial \kappa}{\partial \theta} - \frac{\partial S(\theta)}{\partial \theta} - H(\theta)\theta^{\setminus a} \tag{8}$$

$$= H(\theta)\phi - \frac{\partial S(\theta)}{\partial \theta} \tag{9}$$

where $H(\theta)$ is the Hessian of $\kappa(\theta)$. Setting this derivative to zero corresponds to a fixed point scheme for ϕ, and recovers the update for ϕ, the gradient matching scheme for an exponential family message.

2 NCVMP as moment matching

Gradient matching can be seen as analogous to moment matching in EP. The gradient of the true S is

$$\begin{aligned}
\frac{\partial S(\theta)}{\partial \theta} &= \int \frac{\partial q_i(x_i;\theta)}{\partial \theta} \langle \log f_a(\mathbf{x}) \rangle_{\neg q_i(x_i)} dx_i \\
&= \int \frac{\partial \log q_i(x_i;\theta)}{\partial \theta} q_i(x_i;\theta) \langle \log f_a(\mathbf{x}) \rangle_{\neg q_i(x_i)} dx_i \\
&= \int (\mathbf{u}(x_i) - \langle \mathbf{u}(x_i) \rangle_{q_i(x_i;\theta)}) q_i(x_i;\theta) \langle \log f_a(\mathbf{x}) \rangle_{\neg q_i(x_i)} dx_i.
\end{aligned}$$

Whereas the gradient of the approximate $\tilde{S}$ is

$$\begin{aligned}
\frac{\partial \tilde{S}(\theta,\phi)}{\partial \theta} &= \int \frac{\partial q_i(x_i;\theta)}{\partial \theta} \log m_{a\to i}(x_i;\phi) dx_i \\
&= \int \frac{\partial \log q_i(x_i;\theta)}{\partial \theta} q_i(x_i;\theta) \langle \log m_{a\to i}(x_i;\phi) dx_i \\
&= \int (\mathbf{u}(x_i) - \langle \mathbf{u}(x_i) \rangle_{q_i(x_i;\theta)}) q_i(x_i;\theta) \log m_{a\to i}(x_i;\phi) dx_i.
\end{aligned}$$

We see that matching gradients is equivalent to matching moments of the true and approximate log factors, given the current variational posterior.

3 NCVMP is parameterisation invariant

NCVMP is based on matching gradients at the current estimate

$$\left.\frac{\partial \tilde{S}(\theta;\phi)}{\partial \theta}\right|_{\theta=\theta^{(t)}} = \left.\frac{\partial S(\theta)}{\partial \theta}\right|_{\theta=\theta^{(t)}}$$

Now if we reparameterise in terms of ψ with a bijective mapping $\theta = g(\psi)$ then we would work in terms of $S_\psi(\psi) = S(g(\psi))$ and $\tilde{S}_\psi(\psi;\phi) = \tilde{S}(g(\psi);\phi)$:

$$\begin{aligned}
&\frac{\partial \tilde{S}_\psi(\psi;\phi)}{\partial \psi} = \frac{\partial S_\psi(\psi)}{\partial \psi} \\
\Leftrightarrow\ &\frac{\partial \tilde{S}(g(\psi);\phi)}{\partial \psi} = \frac{\partial S(g(\psi))}{\partial \psi} \\
\Leftrightarrow\ &\frac{\partial \tilde{S}(\theta;\phi)}{\partial \theta}\frac{\partial \theta}{\partial \psi} = \frac{\partial S(\theta)}{\partial \theta}\frac{\partial \theta}{\partial \psi}
\end{aligned}$$

The Jacobian matrix $\frac{\partial \theta}{\partial \psi}$ is full rank since g is bijective, so the original gradient matching scheme is recovered.

4 Optimising the variational parameter for the quadratic softmax bound

The quadratic softmax bound is

$$\log \sum_{k=1}^{K} e^{x_k} \leq a + \sum_{k=1}^{K} \frac{x_k - a - t_k}{2} + \lambda(t_k)[(x_k - a)^2 - t_k^2] - \log \sigma(-t_k) \tag{10}$$

where t are new variational parameters and $\lambda(t) = \frac{1}{2t}\left[\frac{1}{1+e^{-t}} - \frac{1}{2}\right]$. Taking the expectation wrt to x we have

$$\langle \log \sum_{k=1}^{K} e^{x_k} \rangle \leq F(a) = a + \sum_{k=1}^{K} \frac{m_k - a - t_k}{2} + \lambda(t_k)[(m_k - a)^2 + v_k - t_k^2] - \log \sigma(-t_k) \tag{11}$$

Setting the derivatives of wrt a and t equal to zero gives the following fixed point updates for a and t to make the bound as tight as possible:

$$a \leftarrow \frac{2\sum_{k=1}^{K} m_k \lambda(t_k) + K/2 - 1}{2\sum_{k=1}^{K} \lambda(t_k)} \tag{12}$$

$$t_k^2 \leftarrow m_k^2 + v_k - 2m_k a + a^2 \qquad \forall k \tag{13}$$

For small dimensionality and counts these fixed point iterations converge very fast. However, for large counts and dimensionality K we found that the coupling between t and α was very strong and co-ordinate-wise optimization was highly inefficient. In this regime an effective solution is to substitute the expression for t_k in Equation 13 into the objective function to give a univariate optimization problem in α, which can solved efficiently using Newton's method. See the supplementary material for details. The overall bound for the factor is

$$\log f(d|x) \geq \sum_{k=1}^{K} d_k x_k - a - \sum_{k=1}^{K} \left[\frac{x_k - a - t_k}{2} + \lambda(t_k)[(x_k - a)^2 - t_k^2] - \log \sigma(-t_k)\right] \tag{14}$$

Calculating the messages to x and p, and the evidence, are now conjugate operations. The message to x_k will have precision $2S\lambda(t_k)$ and mean times precision $d_k - 1 - (\frac{1}{2} - 2a\lambda(t_k))$. At the minimum we have

$$t_k(a) = \sqrt{(m_k - a)^2 + v_k} \tag{15}$$

Using this expression we can simplify Equation 11 to get

$$F(a) = a + \sum_{k=1}^{K} \frac{m_k - a - t_k}{2} - \log \sigma(-t_k) \tag{16}$$

The derivatives of t_k wrt a are

$$t_k'(a) = -(m_k - a)/t_k(a) \tag{17}$$

$$t_k''(a) = 1/t_k(a) - (m_k - a)^2/t_k(a)^3 \tag{18}$$

Using the chain rule we now find:

$$F'(a) = 1 + \sum_k -(1 + t_k'(a))/2 + t_k'(a)\sigma(t_k(a)) \tag{19}$$

$$F''(a) = \sum_k t_k''(a)(\sigma(t_k(a)) - .5) + t_k'(a)^2\sigma(t_k(a))\sigma(-t_k(a)) \tag{20}$$

We can then use a Newton algorithm with LM line search to cope with small $F''(a)$.

5 Derivation of tilted bound

The tilted bound can be derived as follows, analogously to the univariate bound

$$\langle \log \sum_i e^{x_i} = \langle \log e^{\sum_j a_j x_j} e^{-\sum_j a_j x_j} \sum_j e^{x_i} \rangle \tag{21}$$

$$\leq \sum_i a_j m_i + \log \sum_i \langle e^{x_i - \sum_j a_j x_j} \rangle \tag{22}$$

$$\leq \frac{1}{2} \sum_j a_j^2 v_j + \log \sum_i e^{m_i + (1-2a_i)v_i/2} =: \mathcal{T}(m, v, a) \tag{23}$$

Taking derivatives wrt a_k gives

$$\nabla_{a_k} \mathcal{T}(m, v, a) = a_k v_k - v_k \sigma_k \left[\mathbf{m} + \frac{1}{2}(\mathbf{1} - 2\mathbf{a}) \cdot \mathbf{v} \right] \tag{24}$$

Setting this expression equal to zero results in the fixed point update

$$\mathbf{a} := \sigma \left[\mathbf{m} + \frac{1}{2}(\mathbf{1} - 2\mathbf{a}) \cdot \mathbf{v} \right] \tag{25}$$

6 Taylor series expansion for log-sum-exp

We can use a Taylor series expansion about the mean of x. This will not give a bound, but may be more accurate and is cheap to compute.

$$\log \sum_i e^{x_i} \approx \log \sum_i e^{m_i} + \sum_i (x_i - m_i)\sigma_i(\mathbf{m}) + \frac{1}{2}\sum_i (x_i - m_i)^2 \sigma_i(\mathbf{m})[1 - \sigma_i(\mathbf{m})]$$

We ignore the cross terms of the Hessian because we are using a fully factorised variational posterior for x. Taking expectations we find

$$\langle \log \sum_i e^{x_i} \rangle_q \approx \log \sum_i e^{m_i} + \frac{1}{2}\sum_i v_i \sigma_i(\mathbf{m})[1 - \sigma_i(\mathbf{m})] \tag{26}$$

This approximation is similar in spirit to Laplace's approximation, expect that we calculate the curvature around an approximation mean (calculated using VMP) rather than the MAP. Using the notation in the paper the messages to x_k will be given by:

$$\frac{1}{v_{kf}} = (d_\cdot - K)\sigma_k(\mathbf{m})(1 - \sigma_k(\mathbf{m})) \tag{27}$$

$$\frac{m_{kf}}{v_{kf}} = d_k - 1 + \frac{m_k}{v_{kf}} - (d_\cdot - K)\sigma_k(\mathbf{m}) \tag{28}$$

This message will always be proper (have positive variance) but there is no guarantee of global convergence since this approximation is not a bound.

7 Bohning's bound

Bohning's bound has the same form as the Taylor series expansion, only with a different approximation to the Hessian matrix H of $\log \sum \exp$, specifically using the bound

$$H \geq \frac{1}{2}(I - \mathbf{1}\mathbf{1}^T/K) =: H_B \tag{29}$$

the following bound on $\log \sum \exp$:

$$\log \sum_i e^{x_i} \leq \log \sum_i e^{m_i} + \sum_i (x_i - m_i)\sigma_i(\mathbf{m}) + \frac{1}{4}\sum_{ij} (x_i - m_i)(x_j - m_j)(\delta_{ij} - \frac{1}{K})$$

In the case of a fully factorised distribution on x taking expectations we have:

$$\langle \log \sum_i e^{x_i} \rangle_q \leq \log \sum_i e^{m_i} + \frac{1}{4}\sum_i \left(1 - \frac{1}{K}\right) v_i \tag{30}$$

Analogously to the Taylor series expansion, we have the following message to x_k:

$$\frac{1}{v_{kf}} = \frac{1}{2}(d_\cdot - K)\left(1 - \frac{1}{K}\right) \tag{31}$$

$$\frac{m_{kf}}{v_{kf}} = d_k - 1 + \frac{m_k}{v_{kf}} - (d_\cdot - K)\sigma_k(\mathbf{m}) \tag{32}$$

Note here that the variance is constant and depends only on $d_\cdot$ and K, and is always less than or equal to the variance of the message calculated using the Taylor series expansion.