Supplemental Material

Jason L. Pacheco
Department of Computer Science
Brown University
Providence, RI
pachecoj@cs.brown.edu

Erik B. Sudderth
Department of Computer Science
Brown University
Providence, RI
sudderth@cs.brown.edu

1 Gradient Calculations

1.1 Discrete

The gradients can be expressed more compactly by first defining the discrete BP fixed points given by [1],

$$\tau_s^{BP}(x_s;\lambda) = \varphi_s(x_s)\exp\left\{\frac{1}{n_s-1}\sum_{t\in N(s)}\lambda_{ts}(x_s)\right\} \tag{1}$$

$$\tau_{st}^{BP}(x_s,x_t;\lambda) = \phi_{st}(x_s,x_t)\exp\Big\{\lambda_{ts}(x_s)+\lambda_{st}(x_t)\Big\}. \tag{2}$$

The gradients take an intuitive then take the intuitive form,

$$\frac{\partial\mathcal{L}_c}{\partial\tau_s(x_s)} = (n_s-1)\Big[\log\tau_s^{BP}(x_s)-\log\tau_s(x_s)-1\Big]-\xi_s+c\left[C_{ts}(x_s;\tau)-C_s(\tau)\right] \tag{3}$$

$$\frac{\partial\mathcal{L}_c}{\partial\tau_{st}(x_s)} = \log\tau_{st}(x_s,x_t)+1-\log\tau_{st}^{BP}(x_s,x_t)-c\Big[C_{ts}(x_s;\tau)+C_{st}(x_t;\tau)\Big]. \tag{4}$$

It is then obvious that any zero-gradient must not only satisfy the constraints, but also be of the form defined by BP fixed-point equations.

1.2 Gaussian

The derivative w.r.t. the node variance is given by,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial V_s} = \frac{n_s-1}{2}&\left[V_s^{-1}-A_s-\frac{1}{n_s-1}\sum_{t\in N(s)}\lambda_{st}\right]\\ &+c\sum_{t\in N(s)}[V_s-V_{ts}]+\kappa\sum_{t\in N(s)}[\log V_s-\log V_{ts}]V_s^{-1},\end{aligned} \tag{5}$$

and for the diagonal and off-diagonal elements of the pairwise variance as,

$$\frac{\partial\mathcal{L}}{\partial V_{ts}} = \frac{1}{2}\left[A_s+\lambda_{st}-|\Sigma_{st}|^{-1}V_{st}\right]+c[V_{ts}-V_s]+\kappa[\log V_{ts}-\log V_s] \tag{6}$$

$$\frac{\partial\mathcal{L}}{\partial\Sigma_{st}^{st}} = J_{st}+|\Sigma_{st}|^{-1}\Sigma_{st}^{st}. \tag{7}$$

1.3 Conditional Gaussian

The full joint distribution of the model is,

$$\begin{aligned} p(x,z) &= \varphi_0(x)\prod_{i=1}^n \psi_0(z_i)\varphi_i(x,z_i;y_i) \\ &= N(x \mid \mu_0, P_0)\prod_{i=1}^n (1-\beta_0)^{1-z_i}\beta_0^{z_i} N(y_i \mid 0, \sigma_0^2)^{1-z_i} N(y_i \mid x, \sigma_1^2)^{z_i}. \end{aligned} \tag{8}$$

Using the chain rule for entropy $H(X,Z) = H(Z) + H(X \mid Z)$ we compute the (negative) Bethe entropy as,

$$\begin{aligned} -H(X,Z) &= -\sum_{i=1}^n (H(Z_i) + H(X \mid Z_i)) \\ &= \sum_i ((1-\beta_i)\log(1-\beta_i) + \beta_i \log \beta_i) - \sum_i ((1-\beta_i)\frac{1}{2}\log 2\pi e V_{i0} + \beta_i \frac{1}{2}\log 2\pi e V_{i1}) \end{aligned} \tag{9}$$

The Bethe free energy for the conditional Gaussian model is,

$$\mathcal{F}_{CGB}(m,V,\beta) = \sum_{i=1}^n \mathbb{E}_i[\log q_i(x,z_i) - \log \phi_i(x,z_i)] - (n-1)\mathbb{E}_i[\log q_0(x) - \log \varphi_0(x)],$$

where $\phi_i(x,z_i) = \varphi_0(x)\psi_0(z_i)\varphi_i(x,z;y_i)$. Expanding terms we have,

$$\mathcal{F}_{CGB}(m,V,\beta) = (N-1)\frac{1}{2}\log V_0 - (N-1)\frac{1}{2}(V_0 + m_0^2)P_0^{-1} + (N-1)m_0 P_0^{-1}\mu_0 \tag{10}$$

$$\sum_i (1-\beta_i)\left\{\log(1-\beta_i) - \frac{1}{2}\log V_{i0} - \gamma_{i0} + \frac{1}{2}(V_{i0}+m_{i0}^2)P_0^{-1} - m_{i0}P_0^{-1}\mu_0 - \log(1-\beta_0)\right\} +$$

$$\sum_i \beta_i \left\{\log \beta_i - \frac{1}{2}\log V_{i1} - \gamma_{i1} + \frac{1}{2}(V_{i1}+m_{i1}^2)(P_0^{-1}+\sigma_1^{-2}) - m_{i1}(P_0^{-1}\mu_0 + \sigma_1^{-2}y_i) - \log \beta_0\right\}$$

with the shorthand notation $\gamma_{ij} = \log N(y_i \mid 0, \sigma_j^2)$. Note that while the free energy is bounded on the set of expectation constraints [2] the entorpy term $\log V_0$ means that the free energy is unbounded below off of the constraint set as $V_0 \to \infty$ at an exponential rate. Such an objective can be problematic for MoM optimization and so we add an additional penalty,

$$\mathcal{F}_{CGB}(m,V,\beta) + \frac{\kappa}{2}\sum_i |\log V_0 - \log \bar{V}_i|^2,$$

for some fixed $\kappa \geq 1$ where the Gaussian mixture variance is denoted,

$$\begin{aligned} \bar{V}_i &= (1-\beta_i)V_{i0} + \beta_i V_{i1} + (1-\beta_i)(m_{i0} - \bar{m}_i)^2 + \beta_i(m_{i1} - \bar{m}_i)^2 \\ \bar{m}_i &= (1-\beta_i)m_{i0} + \beta_i m_{i1}. \end{aligned}$$

This added term is quadratic in $\log V_0$, thus bounding the objective off of the constraint set. The augmented Lagrangian is,

$$\begin{aligned} \mathcal{L}_c(m,V,\beta) = \mathcal{F}(m,V,\beta) &+ \frac{\kappa}{2}\sum_i [\log V_0 - \log \bar{V}_i]^2 + \sum_i \eta_i[m_0 - \bar{m}_i] + \sum_i \lambda_i[V_0 - \bar{V}_i] \\ &+ \frac{c}{2}\sum_i [m_0 - \bar{m}_i]^2 + \frac{c}{2}\sum_i [V_0 - \bar{V}_i]^2 \end{aligned}$$

Gradients of the Gaussian marginal moments are,

$$\frac{\partial \mathcal{L}_c}{\partial V_0} = (N-1)\frac{1}{2}V_0^{-1} - (N-1)\frac{1}{2}P_0^{-1} + \sum_i \lambda_i + c\sum_i [V_0 - \bar{V}_i] + \kappa V_0^{-1}\sum_i [\log V_0 - \log \bar{V}_i]$$

$$\frac{\partial \mathcal{L}_c}{\partial m_0} = -(N-1)m_o P_0^{-1} + (N-1)P_0^{-1}\mu_0 + \sum_i \eta_i + c\sum_i (m_0 - \bar{m}_i).$$

Gradients of the mixture variances,

$$\frac{\partial \mathcal{L}_c}{\partial V_{i0}} = (1-\beta_i)\left\{\frac{1}{2}P_0^{-1} - \frac{1}{2}V_{i0}^{-1} - \lambda_i - c(V_0 - \bar{V}_i) - \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1}\right\}$$
$$\frac{\partial \mathcal{L}_c}{\partial V_{i1}} = \beta_i\left\{\frac{1}{2}(P_0^{-1} + \sigma_1^{-2}) - \frac{1}{2}V_{i1}^{-1} - \lambda_i - c(V_0 - \bar{V}_i) - \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1}\right\}.$$

Gradients of the mixture means,

$$\begin{aligned}\frac{\partial \mathcal{L}_c}{\partial m_{i0}} &= (1-\beta_i)\Big\{m_{i0}P_0^{-1} - P_0^{-1}\mu_0 - \eta_i - c(m_0 - \bar{m}_i) \\ &\qquad + 2\beta_i(m_{i1} - m_{i0})\left[\lambda_i + c(V_0 - \bar{V}_i) + \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1}\right]\Big\} \\ \frac{\partial \mathcal{L}_c}{\partial m_{i1}} &= \beta_i\Big\{m_{i1}(P_0^{-1} + \sigma_1^{-2}) - P_0^{-1}\mu_0 - \sigma_1^{-2}y_i - \eta_i - c(m_0 - \bar{m}_i) \\ &\qquad 2(1-\beta_i)(m_{i0} - m_{i1})\left[\lambda_i + c(V_0 - \bar{V}_i) + \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1}\right]\Big\}\end{aligned}$$

For the mixture weights we first introduce some shorthand notation,

$$\xi_{i0}(m, V, \beta) = \log(1-\beta_i) - \frac{1}{2}\log V_{i0} - \gamma_{i0} + \frac{1}{2}(V_{i0} + m_{i0}^2)P_0^{-1} - m_{i0}P_0^{-1}\mu_0$$
$$\xi_{i1}(m, V, \beta) = \log \beta_i - \frac{1}{2}\log V_{i1} - \gamma_{i1} + \frac{1}{2}(V_{i1} + m_{i1}^2)(P_0^{-1} + \sigma_1^{-2}) - m_{i1}(P_0^{-1}\mu_0 + \sigma_1^{-1}y_i),$$

we similarly define shorthand for partials of the mean and variance constraints,

$$\begin{aligned}m' = \frac{\partial C_i^{mean}}{\partial \beta_i} &= m_{i0} - m_{i1} \\ V' = \frac{\partial C_i^{var}}{\partial \beta_i} &= V_{i0} - V_{i1} + (m_{i0} - \bar{m}_i)^2 - (m_{i1} - \bar{m}_i)^2 \\ &\qquad - 2 * (m_{i0} - m_{i1})((1-\beta_i)(m_{i0} - \bar{m}) + \beta_i(m_{i1} - \bar{m}_i))\end{aligned}$$

and the derivative w.r.t. the mixture weights is given by,

$$\begin{aligned}\frac{\partial \mathcal{L}_c}{\partial \beta_i} &= -\xi_{i0}(m, V, \beta) + \xi_{i1}(m, V, \beta) \\ &\qquad + m'(\eta_i + c(m_0 - \bar{m}_i)) + V'(\lambda_i + c(V_0 - \bar{V}_i) + c\bar{V}_i^{-1}(\log V_0 - \log \bar{V}_i))\end{aligned}$$

References

[1] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

[2] T. P. Minka. Expectation propagation for approximate bayesian inference. *Uncertainty in Artificial Intelligence*, 17:362–369, 2001.