1 Gradient Calculations

1.1 Discrete

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

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

\[ \tau_{BP}^{st}(x_s, x_t; \lambda) = \varphi_{st}(x_s, x_t) \exp \left\{ \lambda_{ts}(x_s) + \lambda_{st}(x_t) \right\}. \]

The gradients take an intuitive then take the intuitive form,

\[ \frac{\partial L_c}{\partial \tau_s(x_s)} = \left( n_s - 1 \right) \left[ \log \tau_{BP}^s(x_s) - \log \tau_s(x_s) - 1 \right] - \xi_s + c \left[ C_{ts}(x_s; \tau) - C_s(\tau) \right] \]

\[ \frac{\partial L_c}{\partial \tau_{st}(x_s)} = \log \tau_{st}(x_s, x_t) + 1 - \log \tau_{BP}^{st}(x_s, x_t) - c \left[ C_{ts}(x_s, \tau) + C_{st}(x_t; \tau) \right]. \]

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,

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

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

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

\[ \frac{\partial L}{\partial \Sigma_{st}} = J_{st} + |\Sigma_{st}|^{-1} \Sigma_{st}. \]
1.3 Conditional Gaussian

The full joint distribution of the model is,

\[ p(x, z) = \varphi_0(x) \prod_{i=1}^{n} \psi_0(z_i) \varphi_i(x, z_i; y_i) \]

\[ = N(x \mid \mu_0, P_0) \prod_{i=1}^{n} (1 - \beta_0)^{1-z_i} \beta_0^{z_i} N(y_i \mid 0, \sigma_0^2) \prod_{i=1}^{n} (1 - z_i) N(y_i \mid x, \sigma_1^2 z_i). \tag{8} \]

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

\[ -H(X, Z) = - \sum_{i=1}^{n} (H(Z_i) + H(X \mid Z_i)) \]

\[ = \sum_{i} ((1 - \beta_i) \log(1 - \beta_i) + \beta_i \log \beta_i) - \sum_{i} ((1 - \beta_i) \frac{1}{2} \log 2\pi e V_{i0} + \beta_i \frac{1}{2} \log 2\pi e V_{i1}) \tag{9} \]

The Bethe free energy for the conditional Gaussian model is,

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

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

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

\[ \sum_{i} (1 - \beta_i) \left\{ \log(1 - \beta_i) - \frac{1}{2} \log V_{i0} - \gamma_{i0} + \frac{1}{2} (V_{i0} + m_{i0}^2) P_0^{-1} - m_{i0} P_0^{-1} \mu_0 - \log(1 - \beta_0) \right\} + \]

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

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 entropy 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,

\[ \bar{V}_i = (1 - \beta_i) V_{i0} + \beta_i V_{i1} + (1 - \beta_i) (m_{i0} - \bar{m}_i)^2 + \beta_i (m_{i1} - \bar{m}_i)^2 \]

\[ \bar{m}_i = (1 - \beta_i) m_{i0} + \beta_i m_{i1}. \]

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

\[ \mathcal{L}_c(m, V, \beta) = \mathcal{F}(m, V, \beta) + \frac{\kappa}{2} \sum_{i} |\log V_0 - \log \bar{V}_i|^2 + \sum_{i} \eta_i [m_{i0} - \bar{m}_i] + \sum_{i} \lambda_i [V_0 - \bar{V}_i] \]

\[ + \frac{\epsilon}{2} \sum_{i} [m_{i0} - \bar{m}_i]^2 + \frac{\epsilon}{2} \sum_{i} [V_0 - \bar{V}_i]^2 \]

Gradients of the Gaussian marginal moments are,

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

\[ \frac{\partial \mathcal{L}_c}{\partial m_{i0}} = -(N - 1) m_{i0} P_0^{-1} + (N - 1) P_0^{-1} \mu_0 + \sum_{i} \eta_i + \epsilon \sum_{i} (m_{i0} - \bar{m}_i). \]
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,
\[ \frac{\partial \mathcal{L}_c}{\partial m_{i0}} = (1 - \beta_i) \left\{ m_{i0} P_{0}^{-1} - P_{0}^{-1} \mu_0 - \eta_i - c(m_0 - \bar{m}_i) \right\} \]
\[ + 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] \}
\[ \frac{\partial \mathcal{L}_c}{\partial m_{i1}} = \beta_i \left\{ m_{i1} (P_{0}^{-1} + \sigma_1^{-2}) - P_{0}^{-1} \mu_0 - \sigma_1^{-2} y_i - \eta_i - c(m_0 - \bar{m}_i) \right\} \]
\[ 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] \}

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,
\[ m' = \frac{\partial \mathcal{C}_c^{mean}}{\partial \beta_i} = m_{i0} - m_{i1} \]
\[ \nu' = \frac{\partial \mathcal{C}_c^{var}}{\partial \beta_i} = V_{i0} - V_{i1} + (m_{i0} - \bar{m}_i)^2 - (m_{i1} - \bar{m}_i)^2 \]
\[ - 2 * (m_{i0} - m_{i1}) ((1 - \beta_i) (m_{i0} - \bar{m}) + \beta_i (m_{i1} - \bar{m}_i)) \]
and the derivative w.r.t. the mixture weights is given by,
\[ \frac{\partial \mathcal{L}_c}{\partial \beta_i} = -\xi_{i0}(m, V, \beta) + \xi_{i1}(m, V, \beta) \]
\[ + m'(\eta_i + c(m_0 - \bar{m}_i)) + \nu'(\bar{\lambda}_i + c(V_0 - \bar{V}_i) + c\bar{V}_i^{-1}(\log V_0 - \log \bar{V}_i)) \]

References
