

# Machine Learning: Data to Models

## Assignment 3a

Qun Gao, JHED ID qgao6:  
Li-Yi Lin, JHED ID: llin34

### 2.1 Empirical Questions [40 points]

**a. [10 points]** The objective function is

$$f(q) = \sum_{c \in C} \sum_{\mathbf{x}_c} \theta_c(\mathbf{x}_c) \prod_{i \in c} q_i(x_i) - \sum_{i \in V} \sum_{x_i} q_i(x_i) \log q_i(x_i)$$

The constraint is

$$C_i(q) = \sum_{x_i} q_i(x_i) - 1 = 0, \forall i \in V$$

Let

$$\begin{aligned} L(q, \lambda) &= f(q) - \sum_{i \in V} \lambda_i C_i(q) \\ &= \sum_{c \in C} \sum_{\mathbf{x}_c} \theta_c(\mathbf{x}_c) \prod_{i \in c} q_i(x_i) - \sum_{i \in V} \sum_{x_i} q_i(x_i) \log q_i(x_i) - \sum_{i \in V} \lambda_i (\sum_{x_i} q_i(x_i) - 1) \\ &= \sum_{c \in C} \sum_{\mathbf{x}_c} \theta_c(\mathbf{x}_c) \prod_{i \in c} q_i(x_i) - \sum_{i \in V} \sum_{x_i} (q_i(x_i) \log q_i(x_i) + \lambda_i q_i(x_i)) + \sum_{i \in V} \lambda_i \\ \frac{\partial L}{\partial q_i(x_i)} &= \sum_{c: i \in c} \sum_{\mathbf{x}_{c-i}} \theta_c(\mathbf{x}_c) \prod_{j \in c-i} q_j(x_j) - \log q_i(x_i) - 1 - \lambda_i \end{aligned}$$

Let  $\frac{\partial L}{\partial q_i(x_i)} = 0$ , then

$$q_i(x_i) = \exp(\sum_{c: i \in c} \sum_{\mathbf{x}_{c-i}} \theta_c(\mathbf{x}_c) \prod_{j \in c-i} q_j(x_j) - 1 - \lambda_i) \propto \exp(\sum_{c: i \in c} \sum_{\mathbf{x}_{c-i}} \theta_c(\mathbf{x}_c) \prod_{j \in c-i} q_j(x_j))$$

From the Bayesian network and  $\theta_c(\mathbf{x}_c) = \log \phi_c(\mathbf{x}_c)$ , we have

$$\theta_A(A) = \log P(A), \theta_{A,B}(A, B) = \log P(B|A), \theta_{A,B,C}(A, B, C) = \log P(C|A, B)$$

$$\theta_{B,D}(B, D) = \log P(D|B), \theta_{C,D,E}(C, D, E) = \log P(E|C, D), \theta_{D,F}(D, F) = \log P(F|D)$$

and  $i = 1, 2, \forall i \in \{A, B, C, D, E, F\}$ .

So the update equations for each of the marginal variational distributions are

$$Q(A) = q_A(a) \propto \exp(\log P(a) + \sum_b q_B(b) \log P(b|a) + \sum_{b,c} q_B(b) q_C(c) \log P(c|a, b))$$

$$Q(B) = q_B(b) \propto \exp(\sum_a q_A(a) \log P(b|a) + \sum_{a,c} q_A(a) q_C(c) \log P(c|a, b) + \sum_d q_D(d) \log P(d|b))$$

$$Q(C) = q_C(c) \propto \exp(\sum_{a,b} q_A(a) q_B(b) \log P(c|a, b) + \sum_{d,e} q_D(d) q_E(e) \log P(e|c, d))$$

$$Q(D) = q_D(d) \propto \exp(\sum_b q_B(b) \log P(d|b) + \sum_{c,e} q_C(c) q_E(e) \log P(e|c, d) + \sum_f q_F(f) \log P(f|d))$$

$$Q(E) = q_E(e) \propto \exp(\sum_{c,d} q_C(c) q_D(d) \log P(e|c, d))$$

$$Q(F) = q_F(f) \propto \exp(\sum_d q_D(d) \log P(f|d))$$

where  $a, b, c, d, e, f = \{1, 2\}$ .

**b. [10 points]** Using mean field, the KL divergence is 0.8290152. This seems a little large so the naive mean field approximation is not close to the original network and doesn't seem reasonable in this problem.

**c. [10 points]** The derivation is the same as the previous problem but here  $q(x) = q_{A,B,C}(A, B, C) q_{D,E,F}(D, E, F)$ . Since

$$q_i(x_i) \propto \exp\left(\sum_{c:i \in c} \sum_{\mathbf{x}_{c-i}} \theta_c(\mathbf{x}_c) \prod_{j \in c-i} q_j(x_j)\right)$$

So the update equations for each of the structured mean-field variational distributions are

$$Q(A, B, C) = q_{A,B,C}(a, b, c) \propto \exp(\log P(a) + \log P(b|a) + \log P(c|a, b) + \sum_{d,e,f} q_{D,E,F}(d, e, f) \log P(d|b) + \sum_{d,e,f} q_{D,E,F}(d, e, f) \log P(e|c, d))$$

$$Q(D, E, F) = q_{D,E,F}(d, e, f) \propto \exp(\sum_{a,b,c} q_{A,B,C}(a, b, c) \log P(d|b) + \sum_{a,b,c} q_{A,B,C}(a, b, c) \log P(e|c, d) + \log P(f|d))$$

**d. [10 points]** The KL divergence between the structured mean-field and the true one is 0.001898526, which is much smaller than the naive mean-field. This means the structured mean-field approximation in this case perform better than naive mean-field. This is because in this network the variables A,B and C have

strong dependencies within each other, and the same as D,E and F. However, the correlations between cluster A,B,C and D,E,F is weak. So this structured mean field approximation is close to the original network but simpler for inference. And it's better than naive mean-field, which breaks much of the original network.