

10-708 PGM (Spring 2020): Homework 1

Andrew ID: [your Andrew ID]
 Name: [your first and last name]
 Collaborators: [Andrew IDs of all collaborators, if any]

1 Bayesian Networks [20 Points] (Ben)

State True or False, and briefly justify your answer within 3 lines. The statements are either direct consequences of theorems in Koller and Friedman (2009, Ch. 3), or have a short proof. In the follows, P is a distribution and \mathcal{G} is a BN structure.

1. [2 points] If $A \perp B \mid C$ and $A \perp C \mid B$, then $A \perp B$ and $A \perp C$. (Suppose the joint distribution of A, B, C is positive.) (This is a general probability question not related to BNs.)

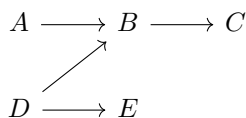


Figure 1: A Bayesian network.

2. [2 points] In Figure 1, $E \perp C \mid B$.
3. [2 points] In Figure 1, $A \perp E \mid C$.

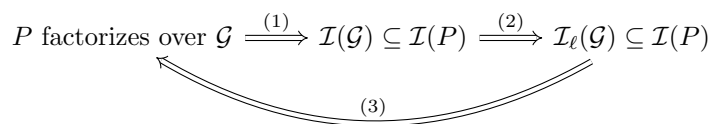


Figure 2: Some relations in Bayesian networks.

Recall the definitions of local and global independences of \mathcal{G} and independences of P .

$$\mathcal{I}_\ell(\mathcal{G}) = \{(X \perp \text{NonDescendants}_{\mathcal{G}}(X) \mid \text{Parents}_{\mathcal{G}}(X))\} \quad (1)$$

$$\mathcal{I}(\mathcal{G}) = \{(X \perp Y \mid Z) : \text{d-separated}_{\mathcal{G}}(X, Y \mid Z)\} \quad (2)$$

$$\mathcal{I}(P) = \{(X \perp Y \mid Z) : P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)\} \quad (3)$$

4. [2 points] In Figure 2, relation (1) is true.
5. [2 points] In Figure 2, relation (2) is true.
6. [2 points] In Figure 2, relation (3) is true.
7. [2 points] If \mathcal{G} is an I-map for P , then P may have extra conditional independencies than \mathcal{G} .

8. **[2 points]** Two BN structures \mathcal{G}_1 and \mathcal{G}_2 are I-equivalent iff they have the same skeleton and the same set of v-structures.
9. **[2 points]** If \mathcal{G}_1 is an I-map of distribution P , and \mathcal{G}_1 has fewer edges than \mathcal{G}_2 , then \mathcal{G}_2 is not a minimal I-map of P .
10. **[2 points]** The P-map of a distribution, if it exists, is unique.

2 Markov Networks [30 points] (Xun)

Let $\mathbf{X} = (X_1, \dots, X_d)$ be a random vector (not necessarily Gaussian) with mean $\boldsymbol{\mu}$ and covariance matrix Σ . The partial correlation matrix R of \mathbf{X} is a $d \times d$ matrix where each entry $R_{ij} = \rho(X_i, X_j | \mathbf{X}_{-ij})$ is the partial correlation between X_i and X_j given the $d - 2$ remaining variables \mathbf{X}_{-ij} . Let $\Theta = \Sigma^{-1}$ be the inverse covariance matrix of \mathbf{X} .

We will prove the relation between R and Θ , and furthermore how Θ characterizes conditional independence in Gaussian graphical models.

1. [10 points] Show that

$$\begin{pmatrix} \Theta_{ii} & \Theta_{ij} \\ \Theta_{ji} & \Theta_{jj} \end{pmatrix} = \begin{pmatrix} \text{Var}[e_i] & \text{Cov}[e_i, e_j] \\ \text{Cov}[e_i, e_j] & \text{Var}[e_j] \end{pmatrix}^{-1} \quad (4)$$

for any $i, j \in [d]$, $i \neq j$. Here e_i is the residual resulting from the linear regression of \mathbf{X}_{-ij} to X_i , and similarly e_j is the residual resulting from the linear regression of \mathbf{X}_{-ij} to X_j .

2. [10 points] Show that

$$R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}}\sqrt{\Theta_{jj}}} \quad (5)$$

3. [10 points] From the above result and the relation between independence and correlation, we know $\Theta_{ij} = 0 \iff R_{ij} = 0 \iff X_i \perp X_j \mid \mathbf{X}_{-ij}$. Note the last implication only holds in one direction.

Now suppose $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ is jointly Gaussian. Show that $R_{ij} = 0 \implies X_i \perp X_j \mid \mathbf{X}_{-ij}$.

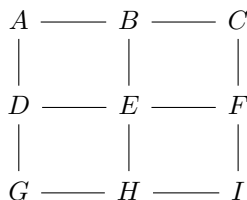
3 Exact Inference [20 points] (Yiwen)

Reference materials for this problem:

- Jordan textbook Ch. 3, available at <https://people.eecs.berkeley.edu/~jordan/prelims/chapter3.pdf>
- Koller and Friedman (2009, Ch. 9 and Ch. 10)

3.1 Variable elimination on a grid [10 points]

Consider the following Markov network:

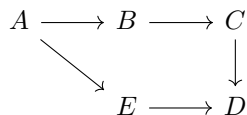


We are going to see how *tree-width*, a property of the graph, is related to the intrinsic complexity of variable elimination of a distribution.

1. [2 points] Write down largest clique(s) for the elimination order $E, D, H, F, B, A, G, I, C$.
2. [2 points] Write down largest clique(s) for the elimination order $A, G, I, C, D, H, F, B, E$.
3. [2 points] Which of the above ordering is preferable? Explain briefly.
4. [4 points] Using this intuition, give a reasonable ($\ll n^2$) upper bound on the tree-width of the $n \times n$ grid.

3.2 Junction tree (a.k.a Clique Tree) [10 points]

Consider the following Bayesian network \mathcal{G} :



We are going to construct a junction tree \mathcal{T} from \mathcal{G} . Please sketch the generated objects in each step.

1. [1 points] Moralize \mathcal{G} to construct an undirected graph \mathcal{H} .
2. [3 points] Triangulate \mathcal{H} to construct a chordal graph \mathcal{H}^* .
(Although there are many ways to triangulate a graph, for the ease of grading, please try adding fewest additional edges possible.)
3. [3 points] Construct a cluster graph \mathcal{U} where each node is a maximal clique C_i from \mathcal{H}^* and each edge is the sepset $S_{i,j} = C_i \cap C_j$ between adjacent cliques C_i and C_j .
4. [3 points] The junction tree \mathcal{T} is the maximum spanning tree of \mathcal{U} .
(The cluster graph is small enough to calculate maximum spanning tree in one's head.)

4 Parameter Estimation [30 points] (Xun)

Consider an HMM with T time steps, M discrete states, and K -dimensional observations as in Figure 3, where $\mathbf{z}_t \in \{0, 1\}^M$, $\sum_s z_{ts} = 1$, $\mathbf{x}_t \in \mathbb{R}^K$ for $t \in [T]$.

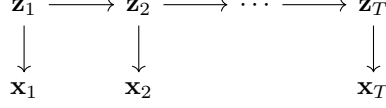


Figure 3: A hidden Markov model.

The joint distribution factorizes over the graph:

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = p(\mathbf{z}_1) \prod_{t=2}^T p(\mathbf{z}_t | \mathbf{z}_{t-1}) \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{z}_t). \quad (6)$$

Now consider the parameterization of CPDs. Let $\boldsymbol{\pi} \in \mathbb{R}^M$ be the initial state distribution and $A \in \mathbb{R}^{M \times M}$ be the transition matrix. The emission density $f(\cdot)$ is parameterized by ϕ_i at state i . In other words,

$$p(z_{1i} = 1) = \pi_i, \quad p(\mathbf{z}_1) = \prod_{i=1}^M \pi_i^{z_{1i}}, \quad (7)$$

$$p(z_{tj} = 1 | z_{t-1,i} = 1) = a_{ij}, \quad p(\mathbf{z}_t | \mathbf{z}_{t-1}) = \prod_{i=1}^M \prod_{j=1}^M a_{ij}^{z_{t-1,i} z_{tj}}, \quad t = 2, \dots, T \quad (8)$$

$$p(\mathbf{x}_t | z_{ti} = 1) = f(\mathbf{x}_t; \phi_i), \quad p(\mathbf{x}_t | \mathbf{z}_t) = \prod_{i=1}^M f(\mathbf{x}_t; \phi_i)^{z_{ti}}, \quad t = 1, \dots, T. \quad (9)$$

Let $\theta = (\boldsymbol{\pi}, A, \{\phi_i\}_{i=1}^M)$ be the set of parameters of the HMM. Given the empirical distribution \hat{p} of $\mathbf{x}_{1:T}$, we would like to find MLE of θ by solving the following problem:

$$\max_{\theta} \mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} [\log p_{\theta}(\mathbf{x}_{1:T})]. \quad (10)$$

However the marginal likelihood is intractable due to summation over M^T terms:

$$p_{\theta}(\mathbf{x}_{1:T}) = \sum_{\mathbf{z}_{1:T}} p_{\theta}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}). \quad (11)$$

An alternative is to use the EM algorithm as we saw in the class.

1. [10 points] Show that the EM updates can take the following form:

$$\theta^* \leftarrow \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} [F(\mathbf{x}_{1:T}; \theta)] \quad (12)$$

where

$$F(\mathbf{x}_{1:T}; \theta) := \sum_{i=1}^M \gamma(z_{1i}) \log \pi_i + \sum_{t=2}^T \sum_{i=1}^M \sum_{j=1}^M \xi(z_{t-1,i}, z_{tj}) \log a_{ij} + \sum_{t=1}^T \sum_{i=1}^M \gamma(z_{ti}) \log f(\mathbf{x}_t; \phi_i) \quad (13)$$

and γ and ξ are the posterior expectations over current parameters $\hat{\theta}$:

$$\gamma(z_{ti}) := \mathbb{E}_{\mathbf{z}_{1:T} \sim p_{\hat{\theta}}(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} [z_{ti}] = p_{\hat{\theta}}(z_{ti} = 1 | \mathbf{x}_{1:T}), \quad t = 1, \dots, T \quad (14)$$

$$\xi(z_{t-1,i}, z_{tj}) := \mathbb{E}_{\mathbf{z}_{1:T} \sim p_{\hat{\theta}}(\mathbf{z}_{1:T} | \mathbf{x}_{1:T})} [z_{t-1,i} z_{tj}] = p_{\hat{\theta}}(z_{t-1,i} z_{tj} = 1 | \mathbf{x}_{1:T}), \quad t = 2, \dots, T \quad (15)$$

2. **[0 points]** (No need to answer.) Suppose γ and ξ are given, and we use isotropic Gaussian $\mathbf{x}_t|z_{ti} = 1 \sim N(\boldsymbol{\mu}_i, \sigma_i^2 I)$ as the emission distribution. Then the parameter updates have the following closed form:

$$\pi_i^* \propto \mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} [\gamma(z_{1i})] \quad (16)$$

$$a_{ij}^* \propto \mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} \left[\sum_{t=2}^T \xi(z_{t-1,i}, z_{tj}) \right] \quad (17)$$

$$\mu_{ik}^* = \frac{\mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} \left[\sum_{t=1}^T \gamma(z_{ti}) \mathbf{x}_t \right]}{\mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} \left[\sum_{t=1}^T \gamma(z_{ti}) \right]} \quad (18)$$

$$\sigma_i^{2*} = \frac{\mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} \left[\sum_{t=1}^T \gamma(z_{ti}) \|\mathbf{x}_t - \boldsymbol{\mu}_i\|_2^2 \right]}{\mathbb{E}_{\mathbf{x}_{1:T} \sim \hat{p}} \left[\sum_{t=1}^T \gamma(z_{ti}) K \right]} \quad (19)$$

3. **[10 points]** We will use the belief propagation algorithm (Koller and Friedman, 2009, Alg. 10.2) to perform inference for *all* marginal queries:

$$\gamma(\mathbf{z}_t) = p_{\hat{\theta}}(\mathbf{z}_t | \mathbf{x}_{1:T}), \quad t = 1, \dots, T \quad (20)$$

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = p_{\hat{\theta}}(\mathbf{z}_{t-1}, \mathbf{z}_t | \mathbf{x}_{1:T}), \quad t = 2, \dots, T \quad (21)$$

For convenience, the notation $\hat{\theta}$ will be omitted from now on.

Derive the following BP updates:

$$\gamma(\mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot s(\mathbf{z}_t) \quad (22)$$

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot c(\mathbf{z}_{t-1}, \mathbf{z}_t) \quad (23)$$

$$(24)$$

where

$$s(\mathbf{z}_t) = \alpha(\mathbf{z}_t) \beta(\mathbf{z}_t), \quad t = 1, \dots, T \quad (25)$$

$$c(\mathbf{z}_{t-1}, \mathbf{z}_t) = p(\mathbf{z}_t | \mathbf{z}_{t-1}) p(\mathbf{x}_t | \mathbf{z}_t) \alpha(\mathbf{z}_{t-1}) \beta(\mathbf{z}_t), \quad t = 2, \dots, T \quad (26)$$

$$Z(\mathbf{x}_{1:T}) = \sum_{\mathbf{z}_t} s(\mathbf{z}_t) \quad (27)$$

and

$$\alpha(\mathbf{z}_1) = p(\mathbf{z}_1) p(\mathbf{x}_1 | \mathbf{z}_1) \quad (28)$$

$$\alpha(\mathbf{z}_t) = p(\mathbf{x}_t | \mathbf{z}_t) \sum_{\mathbf{z}_{t-1}} p(\mathbf{z}_t | \mathbf{z}_{t-1}) \alpha(\mathbf{z}_{t-1}), \quad t = 2, \dots, T \quad (29)$$

$$\beta(\mathbf{z}_{t-1}) = \sum_{\mathbf{z}_t} p(\mathbf{z}_t | \mathbf{z}_{t-1}) p(\mathbf{x}_t | \mathbf{z}_t) \beta(\mathbf{z}_t), \quad t = 2, \dots, T \quad (30)$$

$$\beta(\mathbf{z}_T) = 1 \quad (31)$$

4. **[0 points]** (No need to answer.) Implemented as above, the (α, β) -recursion is likely to encounter numerical instability due to repeated multiplication of small values. One way to mitigate the numerical issue is to scale (α, β) messages at each step t , so that the scaled values are always in some appropriate range, while not affecting the inference result for (γ, ξ) .

Recall that the forward message is in fact a joint distribution

$$\alpha(\mathbf{z}_t) = p(\mathbf{x}_{1:t}, \mathbf{z}_t). \quad (32)$$

Define scaled messages by re-normalizing α w.r.t. \mathbf{z}_t :

$$\hat{\alpha}(\mathbf{z}_t) := \frac{1}{Z(\mathbf{x}_{1:t})} \cdot \alpha(\mathbf{z}_t), \quad (33)$$

$$Z(\mathbf{x}_{1:t}) = \sum_{\mathbf{z}_t} \alpha(\mathbf{z}_t). \quad (34)$$

Furthermore, define

$$r_1 := Z(\mathbf{x}_1), \quad (35)$$

$$r_t := \frac{Z(\mathbf{x}_{1:t})}{Z(\mathbf{x}_{1:t-1})}, \quad t = 2, \dots, T \quad (36)$$

Notice that $Z(\mathbf{x}_{1:t}) = r_1 \cdots r_t$, hence

$$\hat{\alpha}(\mathbf{z}_t) = \frac{1}{r_1 \cdots r_t} \cdot \alpha(\mathbf{z}_t). \quad (37)$$

Plugging $\hat{\alpha}$ into forward messages, the new $\hat{\alpha}$ -recursion is

$$\hat{\alpha}(\mathbf{z}_1) = \frac{1}{r_1} \cdot \underbrace{p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1)}_{\tilde{\alpha}(\mathbf{z}_1)} \quad (38)$$

$$\hat{\alpha}(\mathbf{z}_t) = \frac{1}{r_t} \cdot \underbrace{p(\mathbf{x}_t|\mathbf{z}_t) \sum_{\mathbf{z}_{t-1}} p(\mathbf{z}_t|\mathbf{z}_{t-1}) \hat{\alpha}(\mathbf{z}_{t-1})}_{\tilde{\alpha}(\mathbf{z}_t)}, \quad t = 2, \dots, T \quad (39)$$

Since $\hat{\alpha}$ is normalized, each r_t serves as the normalizing constant:

$$r_t = \sum_{\mathbf{z}_t} \tilde{\alpha}(\mathbf{z}_t). \quad (40)$$

Now switch focus to β . In order to make the inference for (γ, ξ) invariant of scaling, β has to be scaled in a way that counteracts the scaling on α . Plugging $\hat{\alpha}$ into the marginal queries,

$$\gamma(\mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot r_1 \cdots r_t \cdot \hat{\alpha}(\mathbf{z}_t) \beta(\mathbf{z}_t), \quad (41)$$

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot p(\mathbf{z}_t|\mathbf{z}_{t-1}) p(\mathbf{x}_t|\mathbf{z}_t) \cdot r_1 \cdots r_{t-1} \cdot \hat{\alpha}(\mathbf{z}_{t-1}) \beta(\mathbf{z}_t). \quad (42)$$

Since $Z(\mathbf{x}_{1:T}) = r_1 \cdots r_T$, a natural scaling scheme for β is

$$\hat{\beta}(\mathbf{z}_{t-1}) := \frac{1}{r_t \cdots r_T} \cdot \beta(\mathbf{z}_{t-1}), \quad t = 2, \dots, T \quad (43)$$

$$\hat{\beta}(\mathbf{z}_T) := \beta(\mathbf{z}_T), \quad (44)$$

which simplifies the expression for marginals (γ, ξ) to

$$\gamma(\mathbf{z}_t) = \hat{\alpha}(\mathbf{z}_t) \hat{\beta}(\mathbf{z}_t), \quad (45)$$

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{r_t} \cdot p(\mathbf{z}_t|\mathbf{z}_{t-1}) p(\mathbf{x}_t|\mathbf{z}_t) \hat{\alpha}(\mathbf{z}_{t-1}) \hat{\beta}(\mathbf{z}_t). \quad (46)$$

The new $\hat{\beta}$ -recursion can be obtained by plugging $\hat{\beta}$ into backward messages:

$$\hat{\beta}(\mathbf{z}_{t-1}) = \frac{1}{r_t} \cdot \sum_{\mathbf{z}_t} p(\mathbf{z}_t | \mathbf{z}_{t-1}) p(\mathbf{x}_t | \mathbf{z}_t) \hat{\beta}(\mathbf{z}_t), \quad t = 2, \dots, T \quad (47)$$

$$\hat{\beta}(\mathbf{z}_T) = 1. \quad (48)$$

In other words, $\hat{\beta}(\mathbf{z}_{t-1})$ is scaled by $1/r_t$, the normalizer of $\hat{\alpha}(\mathbf{z}_t)$.

The full algorithm is summarized below.

Algorithm 1 Exact inference for (γ, ξ)

(a) Scaled forward message for $t = 1$:

$$\tilde{\alpha}(\mathbf{z}_1) = p(\mathbf{z}_1) p(\mathbf{x}_1 | \mathbf{z}_1) \quad (49)$$

$$r_1 = \sum_{\mathbf{z}_1} \tilde{\alpha}(\mathbf{z}_1) \quad (50)$$

$$\hat{\alpha}(\mathbf{z}_1) = \frac{1}{r_1} \cdot \tilde{\alpha}(\mathbf{z}_1) \quad (51)$$

(b) Scaled forward message for $t = 2, \dots, T$:

$$\tilde{\alpha}(\mathbf{z}_t) = p(\mathbf{x}_t | \mathbf{z}_t) \sum_{\mathbf{z}_{t-1}} p(\mathbf{z}_t | \mathbf{z}_{t-1}) \hat{\alpha}(\mathbf{z}_{t-1}) \quad (52)$$

$$r_t = \sum_{\mathbf{z}_t} \tilde{\alpha}(\mathbf{z}_t) \quad (53)$$

$$\hat{\alpha}(\mathbf{z}_t) = \frac{1}{r_t} \cdot \tilde{\alpha}(\mathbf{z}_t) \quad (54)$$

(c) Scaled backward message for $t = T + 1$:

$$\hat{\beta}(\mathbf{z}_T) = 1 \quad (55)$$

(d) Scaled backward message for $t = T, \dots, 2$:

$$\hat{\beta}(\mathbf{z}_{t-1}) = \frac{1}{r_t} \cdot \sum_{\mathbf{z}_t} p(\mathbf{z}_t | \mathbf{z}_{t-1}) p(\mathbf{x}_t | \mathbf{z}_t) \hat{\beta}(\mathbf{z}_t) \quad (56)$$

(e) Singleton marginal for $t = 1, \dots, T$:

$$\gamma(\mathbf{z}_t) = \hat{\alpha}(\mathbf{z}_t) \hat{\beta}(\mathbf{z}_t) \quad (57)$$

(f) Pairwise marginal for $t = 2, \dots, T$:

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{r_t} \cdot p(\mathbf{z}_t | \mathbf{z}_{t-1}) p(\mathbf{x}_t | \mathbf{z}_t) \hat{\alpha}(\mathbf{z}_{t-1}) \hat{\beta}(\mathbf{z}_t) \quad (58)$$

5. [10 points] We will implement the EM algorithm (also known as Baum-Welch algorithm), where E-step performs exact inference and M-step updates parameter estimates. Please complete the TODO blocks in the provided template `baum_welch.py` and submit it to Gradescope. The template contains a toy problem to play with. The submitted code will be tested against randomly generated problem instances.