10-708 PGM (Spring 2020): Homework 1

Andrew ID: changshi
Name: Chang Shi
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.)

Solution

Since $A \perp B \mid C$ and $A \perp C \mid B$, we have $P(A \mid C) = P(A \mid B, C) = P(A \mid B)$. Then, using the property we have

$$\begin{aligned} P(A,C)P(B) &= P(B)P(A \mid C)P(C) \\ &= P(B)P(A \mid B)P(C) \\ &= P(A,B)P(C) \\ \sum_b LHS &= \sum_b RHS \\ P(A,C) &= P(A)P(C) \end{aligned}$$

Thus, $A \perp C$. In a similar way, from sum over c about both side of equation $P(A,B)P(C) = P(A,C)P(B)$, we can get $P(A,B) = P(A)P(B)$, thus $A \perp B$.

$$\begin{tikzcd} A \arrow[r] & B \arrow[r] & C \\ D \arrow[ur] \arrow[r] & E & \end{tikzcd}$$

Figure 1: A Bayesian network.

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

Solution

True. The local independencies state that each node X_i is conditionally independent of its nondescendants given its parents. Thus, given its parents B, the node C is conditionally independent of its nondescendant E.

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

Solution

False. Because with v-structure $A \to B \leftarrow D$, B's descendant C is given, while no other node along the trail $A \rightleftharpoons B \rightleftharpoons D \rightleftharpoons E$ is given, the trail $A \rightleftharpoons B \rightleftharpoons D \rightleftharpoons E$ is active given C. Thus, with an active trail between A and E, $A \not\perp E | C$.

$$\begin{tikzcd} P \text{ factorizes over } \mathcal{G} \arrow[r, Rightarrow, "(1)"] & \mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P) \arrow[r, Rightarrow, "(2)"] & \mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(P) \arrow[ll, Rightarrow, bend left, "(3)"] \end{tikzcd}$$

Figure 2: Some relations in Bayesian networks.

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

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

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

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

4. **[2 points]** In Figure 2, relation (1) is true.

Solution

True. According to Theorem 3.2 that "If P factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is an I-map for P", and the definition of I-map ($\mathcal{G}$ is an I-map for P means $\mathcal{G}$ is an I-map for $\mathcal{I}(P)$), we can get $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.

5. **[2 points]** In Figure 2, relation (2) is true.

Solution

True. Treating the $\text{Parents}_{\mathcal{G}}(X)$ as the Z in the definition of $\mathcal{I}(\mathcal{G})$, we can know that all the local independencies satisfy the requirements of global Markov independencies, thus $\mathcal{I}_\ell(\mathcal{G})$ is a subset of $\mathcal{I}(\mathcal{G})$, leading to $\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.

6. **[2 points]** In Figure 2, relation (3) is true.

Solution

True. According to the proof of Theorem 3.1 in Koller and Friedman (2009, Ch. 3), we assume a topological ordering $X_1, \dots, X_n$ of variables. We first use the chain rule for probabilities.

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \dots, X_{i-1})$$

then using $\{X_1, \dots, X_{i-1}\} = Pa_{X_i} \bigcup \boldsymbol{Z}$ where $\boldsymbol{Z} \subseteq NonDescendants_{X_i}$ and $((X_i \perp \boldsymbol{Z} | Pa_{X_i})$, We have that $P(X_1, \dots, X_n) = P(X_i | Pa_{X_i})$. Applying this transformation to all of the factors in the chain rule decomposition, the result follows.

7. [**2 points**] If $\mathcal{G}$ is an I-map for P, then P may have extra conditional independencies than $\mathcal{G}$.

Solution

True. According to the explanation under the Definition 3.3 of I-map,for $\mathcal{G}$ to be an I-map of P, it is necessary that $\mathcal{G}$ does not mislead us regarding independencies in P: any independence that $\mathcal{G}$ asserts must also hold in P. Conversely, P may have additional independencies that are not reflected in $\mathcal{G}$.

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.

Solution

False. If $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same skeleton and the same set of v-structure then they are I-equivalent, however if $\mathcal{G}_1$ and $\mathcal{G}_2$ are I-equivalent, we cannot conclude that they have the same set of v-structures. One Counterexample is that any two complete graphs are I-equivalent, although they have the same skeleton, they invariably have different 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.

Solution

False. Different topological orderings will give different minimal I-maps. I-map $\mathcal{G}_2$ with more edges than I-map $\mathcal{G}_1$ does not mean $\mathcal{G}_2$ is not a minimal I-map of P, this may just due to the choice of topological ordering, as long as the removal of even a single edge from $\mathcal{G}_2$ renders it not an I-map, $\mathcal{G}_2$ is called a minimal I-map.

10. [**2 points**] The P-map of a distribution, if it exists, is unique.

Solution

False. P-map is not unique. For example, $x_1 \rightarrow x_2$ and $x_1 \leftarrow x_2$ can have precisely the same independence assumptions bad same distribution, while the P-maps are different.

2 Markov Networks [30 points] (Xun)

Let $\mathbf{X} = (X_1, \ldots, X_d)$ 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

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

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.

Solution

Without losing generality, we discuss about the situation when $i = 1, j = 2$. According to definition of partial correlation matrix

$$R_{12} = \rho(X_1, X_2|\mathbf{X}_r) = \frac{\mathrm{Cov}[e_1, e_2]}{\sqrt{\mathrm{Var}[e_1]}\sqrt{\mathrm{Var}[e_2]}},\ r = (3, 4, \ldots, d)$$

where

$$e_1 = X_1 - (\mathbf{X}_r^T \beta_1 + \beta_1^{(0)})$$
$$e_2 = X_2 - (\mathbf{X}_r^T \beta_2 + \beta_2^{(0)})$$

while all the βs should be the parameters of best linear predictor of X_r to X_1 or X_2. Thus, taking X_1 as an example, the population least square problem would be

$$\begin{aligned}
\min_{\beta, \beta^{(0)}} L &= \mathbb{E}[(X_1 - (X_r^T \beta + \beta^{(0)}))^2] \\
&= \mathbb{E}[X_1^2 - 2X_1(X_r^T \beta + \beta^{(0)}) + (X_r^T \beta + \beta^{(0)})^2] \\
&= \mathbb{E}[X_1^2] - 2\beta^T \mathbb{E}[X_1 X_r] - 2\beta^{(0)} \mathbb{E}[X_1] + \beta^T \mathbb{E}[X_r X_r^T]\beta + 2\beta^{(0)}\beta^T \mathbb{E}[X_r] + {\beta^{(0)}}^2
\end{aligned}$$

Taking partial derivative of L w.r.t. β and $\beta^{(0)}$, we have

$$\begin{cases} \dfrac{\partial L}{\partial \beta} = -2\,\mathbb{E}[X_1 X_r] + 2\,\mathbb{E}[X_r X_r^T]\beta + 2\beta^{(0)}\,\mathbb{E}[X_r] = 0 \\ \dfrac{\partial L}{\partial \beta^{(0)}} = -2\,\mathbb{E}[X_1] + 2\beta^T\,\mathbb{E}[X_r] + 2\beta^{(0)} = 0 \end{cases}$$

Solving the equation set, we can get the best prediction parameter β would be

$$\begin{aligned}
(\mathbb{E}[X_r X_r^T] - (\mathbb{E}[X_2]\,\mathbb{E}[X_r]^T))\beta &= \mathbb{E}[X_1 X_r] - \mathbb{E}[X_1]\,\mathbb{E}[X_r] \\
\mathrm{Var}[X_r]\beta &= \mathrm{Cov}[X_1, X_r] \\
\beta &= (\mathrm{Var}[X_r])^{-1}\mathrm{Cov}[X_1, X_r]
\end{aligned}$$

Thus, in e_1 and e_2 we will have

$$\beta_1 = (\mathrm{Var}[X_r])^{-1}\mathrm{Cov}[X_1, X_r]$$
$$\beta_2 = (\mathrm{Var}[X_r])^{-1}\mathrm{Cov}[X_2, X_r]$$

While $\beta_1^{(0)}$ and $\beta_2^{(0)}$ have no impact during the calculation of variance and covariance value, we omit them here.
Next, we represent the covariance matrix of X with subblocks

$$\Sigma = \left(\begin{array}{cc|ccc} \Sigma_{11} & \Sigma_{12} & — & \Sigma_{1r} & — \\ \Sigma_{21} & \Sigma_{22} & — & \Sigma_{2r} & — \\ \hline | & | & & & \\ \Sigma_{r1} & \Sigma_{r2} & & \Sigma_{rr} & \\ | & | & & & \end{array}\right)$$

Then with the property of inverse matrix of block matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & \dots \\ \dots & \dots \end{pmatrix}$$

$$\begin{aligned}
\begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{pmatrix} &= \left[\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} - \begin{pmatrix} — & \Sigma_{1r} & — \\ — & \Sigma_{2r} & — \end{pmatrix} \Sigma_{rr}^{-1} \begin{pmatrix} | & | \\ \Sigma_{r1} & \Sigma_{r2} \\ | & | \end{pmatrix} \right]^{-1} \\
&= \left[\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{1r}\Sigma_{rr}^{-1}\Sigma_{r1} & \Sigma_{1r}\Sigma_{rr}^{-1}\Sigma_{r2} \\ \Sigma_{2r}\Sigma_{rr}^{-1}\Sigma_{r1} & \Sigma_{2r}\Sigma_{rr}^{-1}\Sigma_{r2} \end{pmatrix} \right]^{-1} \\
&= \begin{pmatrix} \Sigma_{11} - \Sigma_{1r}\Sigma_{rr}^{-1}\Sigma_{r1} & \Sigma_{12} - \Sigma_{1r}\Sigma_{rr}^{-1}\Sigma_{r2} \\ \Sigma_{21} - \Sigma_{2r}\Sigma_{rr}^{-1}\Sigma_{r1} & \Sigma_{22} - \Sigma_{2r}\Sigma_{rr}^{-1}\Sigma_{r2} \end{pmatrix}^{-1}
\end{aligned}$$

While

$$\begin{aligned}
\mathrm{Cov}[e_1, e_2] &= \mathrm{Cov}[X_1 - (\mathbf{X}_r^T\beta_1 + \beta_1^{(0)}),\ X_2 - (\mathbf{X}_r^T\beta_2 + \beta_2^{(0)})] \\
&= \mathrm{Cov}[X_1, X_2] - \mathrm{Cov}[X_1, X_r]^T\beta_2 - \mathrm{Cov}[X_2, X_r]^T\beta_1 + \beta_1^T\mathrm{Cov}[X_r, X_r]\beta_2 \\
&= \Sigma_{12} - \Sigma_{1r}^T\Sigma_{rr}^{-1}\Sigma_{2r} - \Sigma_{2r}^T\Sigma_{rr}^{-1}\Sigma_{1r} + \Sigma_{rr}^{-1}\Sigma_{1r}\Sigma_{rr}\Sigma_{rr}^{-1}\Sigma_{2r} \\
&= \Sigma_{12} - \Sigma_{2r}^T\Sigma_{rr}^{-1}\Sigma_{1r} \\
\mathrm{Var}[e_1] &= \mathrm{Var}[X_1 - (\mathbf{X}_r^T\beta_1 + \beta_1^{(0)})] \\
&= \mathrm{Var}[X_1] + \mathrm{Var}[\mathbf{X}_r^T\beta_1] - 2\mathrm{Cov}[X_1, \mathbf{X}_r^T\beta_1] \\
&= \mathrm{Var}[X_1] + \beta_1^T\mathrm{Var}[\mathbf{X}_r^T]\beta_1 - 2\beta_1^T\mathrm{Cov}[X_1, \mathbf{X}_r^T] \\
&= \Sigma_{11} + \Sigma_{1r}^T\Sigma_{rr}^{-1}\Sigma_{rr}\Sigma_{rr}^{-1}\Sigma_{1r} - 2\Sigma_{1r}^T\Sigma_{rr}^{-1}\Sigma_{1r} \\
&= \Sigma_{11} - \Sigma_{1r}^T\Sigma_{rr}^{-1}\Sigma_{1r}
\end{aligned}$$

Thus, we can generalize from X_1 and X_2 to X_i and X_j, that

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

2. **[10 points]** Show that

$$R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}}\sqrt{\Theta_{jj}}} \tag{5}$$

Solution

Since the inverse matrix of the 2x2 matrix obeys

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

We get

$$\begin{aligned}\begin{pmatrix} \Theta_{ii} & \Theta_{ij} \\ \Theta_{ji} & \Theta_{jj} \end{pmatrix} &= \begin{pmatrix} \mathrm{Var}[e_i] & \mathrm{Cov}[e_i, e_j] \\ \mathrm{Cov}[e_i, e_j] & \mathrm{Var}[e_j] \end{pmatrix}^{-1} \\ &= \frac{1}{\mathrm{Var}[e_i]\mathrm{Var}[e_j] - \mathrm{Cov}[e_i, e_j]\mathrm{Cov}[e_i, e_j]}\begin{pmatrix} \mathrm{Var}[e_j] & -\mathrm{Cov}[e_i, e_j] \\ -\mathrm{Cov}[e_i, e_j] & \mathrm{Var}[e_i] \end{pmatrix}\end{aligned}$$

Thus,

$$\Theta_{ii} = \frac{\mathrm{Var}[e_j]}{\mathrm{Var}[e_i]\mathrm{Var}[e_j] - \mathrm{Cov}[e_i, e_j]\mathrm{Cov}[e_i, e_j]}$$

$$\Theta_{jj} = \frac{\mathrm{Var}[e_i]}{\mathrm{Var}[e_i]\mathrm{Var}[e_j] - \mathrm{Cov}[e_i, e_j]\mathrm{Cov}[e_i, e_j]}$$

$$\Theta_{ij} = \frac{-\mathrm{Cov}[e_i, e_j]}{\mathrm{Var}[e_i]\mathrm{Var}[e_j] - \mathrm{Cov}[e_i, e_j]\mathrm{Cov}[e_i, e_j]}$$

$$-\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}}\sqrt{\Theta_{jj}}} = \frac{\mathrm{Cov}[e_i, e_j]}{\sqrt{\mathrm{Var}[e_j]}\sqrt{\mathrm{Var}[e_i]}} = R_{ij}$$

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

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

Solution

According to (5) of question 2, from $R_{ij} = 0$, we know that $\Theta_{ij} = \Theta_{ji} = 0$. Without loosing generality, we first prove the case when $i = 1, j = 2$. Since $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, we can get $X_1, X_2|\mathbf{X}_{-ij}$ is also Gaussian. Thus, from the property of Gaussian distribution, we can get the covariance matrix of conditional random variables

$$\begin{aligned}\begin{pmatrix} \mathrm{Var}(X_1|X_r) & \mathrm{Cov}(X_1, X_2|X_r) \\ \mathrm{Cov}(X_1, X_2|X_r) & \mathrm{Var}(X_2|X_r) \end{pmatrix} &= \begin{pmatrix} \Theta_{ii} & \Theta_{ij} \\ \Theta_{ji} & \Theta_{jj} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \Theta_{ii} & 0 \\ 0 & \Theta_{jj} \end{pmatrix}^{-1} \\ &= \frac{1}{\Theta_{ii}\Theta_{jj}}\begin{pmatrix} \Theta_{jj} & 0 \\ 0 & \Theta_{ii} \end{pmatrix}\end{aligned}$$

Thus, $\mathrm{Cov}(X_1, X_2|X_r) = 0$, leads to $X_1 X_2 \mid \mathbf{X}_{-12}$. Generalize to X_i and X_j, we have $X_i X_j \mid \mathbf{X}_{-ij}$.

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]

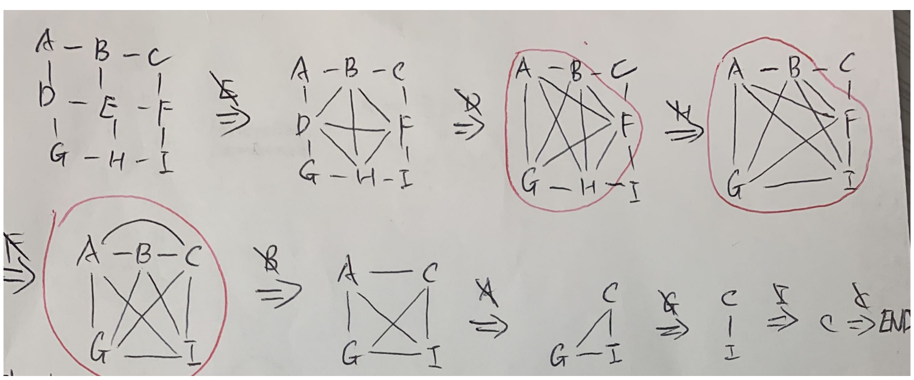
Consider the following Markov network:

$$\begin{tikzcd}
A \arrow[r, dash] \arrow[d, dash] & B \arrow[r, dash] \arrow[d, dash] & C \arrow[d, dash] \\
D \arrow[r, dash] \arrow[d, dash] & E \arrow[r, dash] \arrow[d, dash] & F \arrow[d, dash] \\
G \arrow[r, dash] & H \arrow[r, dash] & I
\end{tikzcd}$$

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 $E, D, H, F, B, A, G, I, C$.

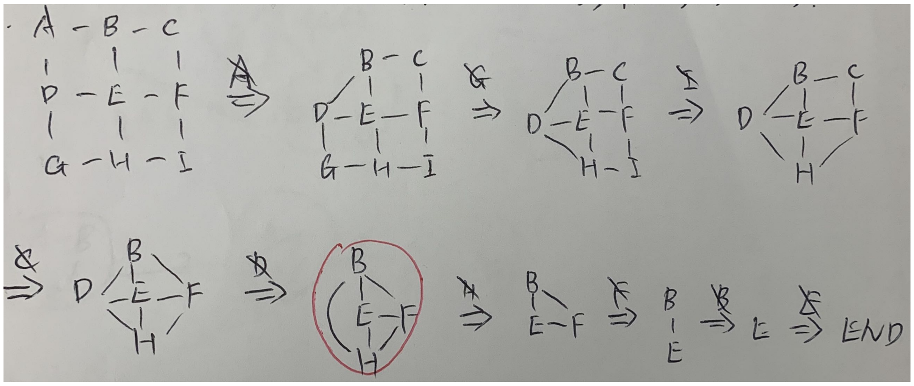
Solution



The largest cliques are {A,B,F,G,H},{A,B,F,G,I},{A,B,C,G,I}.

2. [**2 points**] Write down largest clique(s) for the elimination order $A, G, I, C, D, H, F, B, E$.

Solution



The largest clique is {B,E,F,H}.

3. **[2 points]** Which of the above ordering is preferable? Explain briefly.

Solution

The second order $A, G, I, C, D, H, F, B, E$ is preferable, because the overall complexity is determined by the number of the largest elimination clique, the second elimination ordering lead to smaller clique and hence reduce complexity.

4. **[4 points]** Using this intuition, give a reasonable ($\ll n^2$) upper bound on the tree-width of the $n \times n$ grid.

Solution

A reasonable upper bound of the tree-width of the $n \times n$ grid would be n. Since a particular elimination order of eliminating nodes in the grid row by row, from left to right, up to down will gives maximum clique with n number of nodes, the tree-width of $n \times n$ grid would be definitely smaller than n.

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

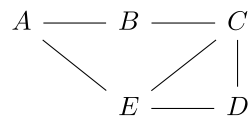
Consider the following Bayesian network $\mathcal{G}$:

$$\begin{tikzcd}
A \arrow[r] \arrow[dr] & B \arrow[r] & C \arrow[d] \\
 & E \arrow[r] & D
\end{tikzcd}$$

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 $\mathcal{H}$.

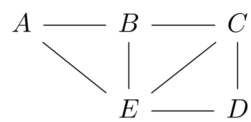
Solution



2. [**3 points**] Triangulate $\mathcal{H}$ to construct a chordal graph $\mathcal{H}^*$.

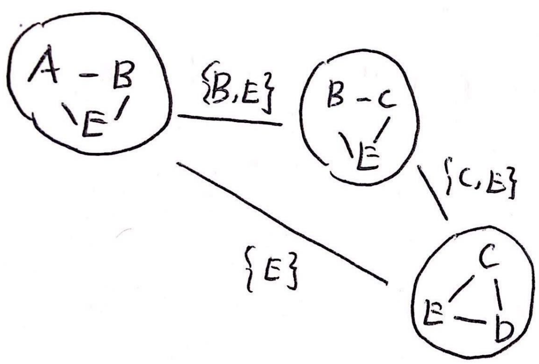
(Although there are many ways to triangulate a graph, for the ease of grading, please try adding fewest additional edges possible.)

Solution



3. [**3 points**] Construct a cluster graph $\mathcal{U}$ where each node is a maximal clique $\boldsymbol{C}_i$ from $\mathcal{H}^*$ and each edge is the sepset $\boldsymbol{S}_{i,j} = \boldsymbol{C}_i \cap \boldsymbol{C}_j$ between adjacent cliques $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$.

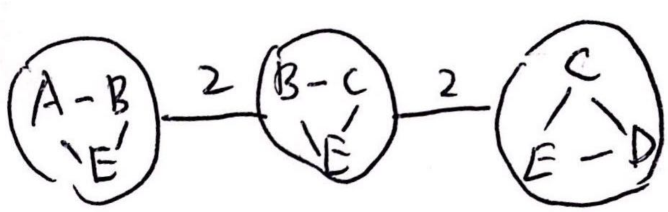
Solution



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.)

Solution



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 $t \in [T]$.

$$
\begin{tikzcd}
\mathbf{z}_1 \arrow[r] \arrow[d] & \mathbf{z}_2 \arrow[r] \arrow[d] & \cdots \arrow[r] & \mathbf{z}_T \arrow[d] \\
\mathbf{x}_1 & \mathbf{x}_2 & & \mathbf{x}_T
\end{tikzcd}
$$

Figure 3: A hidden Markov model.

The joint distribution factorizes over the graph:

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

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 $\boldsymbol{\phi}_i$ at state i. In other words,

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

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

$$p(\mathbf{x}_t|z_{ti} = 1) = f(\mathbf{x}_t; \boldsymbol{\phi}_i), \qquad p(\mathbf{x}_t|\mathbf{z}_t) = \prod_{i=1}^{M} f(\mathbf{x}_t; \boldsymbol{\phi}_i)^{z_{ti}}, \qquad t = 1, \ldots, T. \tag{9}$$

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

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

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

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

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:

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

where

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

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

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

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

Solution

Since the marginal likelihood $\log p_\theta(\mathbf{x}_{1:T})$ is intractable, we can give it a lower bound by applying Jensen's inequality.

$$\begin{aligned}
\log p_\theta(\mathbf{x}_{1:T}) &= \log \sum_{\mathbf{z}_{1:T}} p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) \\
&= \log \sum_{\mathbf{z}_{1:T}} q(\mathbf{z}_{1:T}) \frac{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{q(\mathbf{z}_{1:T})} \\
&\geq \sum_{\mathbf{z}_{1:T}} q(\mathbf{z}_{1:T}) \log \frac{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{q(\mathbf{z}_{1:T})} \qquad (Jensen's\ inequality) \\
&= \mathbb{E}_{q(\mathbf{z}_{1:T})} \log p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) + H[q(\mathbf{z}_{1:T})]
\end{aligned}$$

where the second term $H[q(\mathbf{z}_{1:T})] = -\mathbb{E}_{q(\mathbf{z}_{1:T})} \log q(\mathbf{z}_{1:T})$ is the Shannon Entropy not related to θ. Thus, maximizing $\log p_\theta(\mathbf{x}_{1:T})$ is the same as maximizing the first term $\mathbb{E}_{q(\mathbf{z}_{1:T})} \log p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})$.

$$\begin{aligned}
\log p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) &= \log \left[p(\mathbf{z}_1) \prod_{t=2}^{T} p(\mathbf{z}_t|\mathbf{z}_{t-1}) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{z}_t) \right] \\
&= \log p(\mathbf{z}_1) + \sum_{t=2}^{T} \log p(\mathbf{z}_t|\mathbf{z}_{t-1}) + \sum_{t=1}^{T} \log p(\mathbf{x}_t|\mathbf{z}_t) \\
&= \log \prod_{i=1}^{M} \pi_i^{z_{1i}} + \sum_{t=2}^{T} \log \prod_{i=1}^{M} \prod_{j=1}^{M} a_{ij}^{z_{t-1,i} z_{tj}} + \sum_{t=1}^{T} \log \prod_{i=1}^{M} f(\mathbf{x}_t; \boldsymbol{\phi}_i)^{z_{ti}} \\
&= \sum_{i=1}^{M} p_{\hat{\theta}}(z_{1i} = 1|\mathbf{x}_{1:T}) \log \pi_i + \sum_{t=2}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} p_{\hat{\theta}}(z_{t-1,i} z_{tj} = 1|\mathbf{x}_{1:T}) \log a_{ij} \\
&\qquad + \sum_{t=1}^{T} \sum_{i=1}^{M} p_{\hat{\theta}}(z_{ti} = 1|\mathbf{x}_{1:T}) \log f(\mathbf{x}_t; \boldsymbol{\phi}_i) \\
&= \sum_{i=1}^{M} \gamma(z_{1i}) \log \pi_i + \sum_{t=2}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} \xi(z_{t-1,i}, z_{tj}) \log a_{ij} + \sum_{t=1}^{T} \sum_{i=1}^{M} \gamma(z_{ti}) \log f(\mathbf{x}_t; \boldsymbol{\phi}_i) \\
&= F(\mathbf{x}_{1:T}; \theta)
\end{aligned}$$

So solving $\max_\theta\ \mathbb{E}_{\mathbf{x}_{1:T} \sim \widehat{p}} [\log p_\theta(\mathbf{x}_{1:T})]$ is equivalent as doing EM updates taking the following form

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

.

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:

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

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

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

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

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

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

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

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

Derive the following BP updates:

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

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

$$\tag{24}$$

where

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

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

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

and

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

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

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

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

Solution

$$\begin{tikzcd} \mathbf{z}_1 \arrow[r] \arrow[d] & \mathbf{z}_2 \arrow[r] \arrow[d] & \cdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \end{tikzcd}$$

From (28), $\alpha(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) = p(\mathbf{x}_1, \mathbf{z}_1)$. From (29), we derive

$$\begin{aligned}
\alpha(\mathbf{z}_2) &= p(\mathbf{x}_2|\mathbf{z}_2)\sum_{\mathbf{z}_1} p(\mathbf{z}_2|\mathbf{z}_1)\alpha(\mathbf{z}_1) \\
&= p(\mathbf{x}_2|\mathbf{z}_2)\sum_{\mathbf{z}_1} p(\mathbf{z}_2|\mathbf{z}_1)p(\mathbf{x}_1, \mathbf{z}_1) \\
&= p(\mathbf{x}_2|\mathbf{z}_2)\sum_{\mathbf{z}_1} p(\mathbf{z}_2|\mathbf{x}_1, \mathbf{z}_1)p(\mathbf{x}_1, \mathbf{z}_1) \\
&= p(\mathbf{x}_2|\mathbf{z}_2)\sum_{\mathbf{z}_1} p(\mathbf{z}_2, \mathbf{z}_1, \mathbf{x}_1) \\
&= p(\mathbf{x}_2|\mathbf{z}_2)p(\mathbf{z}_2, \mathbf{x}_1) \\
&= p(\mathbf{x}_2|\mathbf{z}_2, \mathbf{x}_1)p(\mathbf{z}_2, \mathbf{x}_1) \\
&= p(\mathbf{z}_2, \mathbf{x}_1, \mathbf{x}_2)
\end{aligned}$$

By recursively substituting back into (29), we can get

$$\alpha(\mathbf{z}_t) = p(\mathbf{z}_t, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)$$

$$\begin{tikzcd} \cdots \arrow[r] & \mathbf{z}_{T-1} \arrow[r] \arrow[d] & \mathbf{z}_T \arrow[d] \\ & \mathbf{x}_{T-1} & \mathbf{x}_T \end{tikzcd}$$

Similarly, since $\beta(\mathbf{z}_T) = 1$, from (30),we derive

$$\begin{aligned}
\beta(\mathbf{z}_{T-1}) &= \sum_{\mathbf{z}_T} p(\mathbf{z}_T|\mathbf{z}_{T-1})p(\mathbf{x}_T|\mathbf{z}_T)\beta(\mathbf{z}_T) \\
&= \sum_{\mathbf{z}_T} p(\mathbf{z}_T|\mathbf{z}_{T-1})p(\mathbf{x}_T|\mathbf{z}_T, \mathbf{z}_{T-1}) \\
&= \sum_{\mathbf{z}_T} \frac{p(\mathbf{z}_{T-1})p(\mathbf{z}_T|\mathbf{z}_{T-1})p(\mathbf{x}_T|\mathbf{z}_T, \mathbf{z}_{T-1})}{p(\mathbf{z}_{T-1})} \\
&= \sum_{\mathbf{z}_T} \frac{p(\mathbf{x}_T, \mathbf{z}_T, \mathbf{z}_{T-1})}{p(\mathbf{z}_{T-1})} \\
&= \sum_{\mathbf{z}_T} p(\mathbf{x}_T, \mathbf{z}_T|\mathbf{z}_{T-1}) \\
&= p(\mathbf{x}_T|\mathbf{z}_{T-1})
\end{aligned}$$

By recursively substituting back into (30), we can get

$$\beta(\mathbf{z}_t) = p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{z}_t)$$

Thus, from (25),

$$\begin{aligned}
s(\mathbf{z}_t) &= \alpha(\mathbf{z}_t)\beta(\mathbf{z}_t) \\
&= p(\mathbf{z}_t, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{z}_t) \\
&= p(\mathbf{z}_t, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{z}_t, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t) \\
&= p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T, \mathbf{z}_t) \\
&= p(\mathbf{x}_{1:T}, \mathbf{z}_t)
\end{aligned}$$

According to (27), $Z(\mathbf{x}_{1:T}) = \sum_{\mathbf{z}_t} s(\mathbf{z}_t) = p(\mathbf{x}_{1:T})$, then from (20)

$$\gamma(\mathbf{z}_t) = p_{\hat{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:T}) = \frac{p_{\hat{\theta}}(\mathbf{z}_t, \mathbf{x}_{1:T})}{p_{\hat{\theta}}(\mathbf{x}_{1:T})} = \frac{s(\mathbf{z}_t)}{Z(\mathbf{x}_{1:T})}$$

We successfully derive the BP update formula (22).

According to (26)

$$\begin{aligned}
c(\mathbf{z}_{t-1}, \mathbf{z}_t) &= p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t)\alpha(\mathbf{z}_{t-1})\beta(\mathbf{z}_t) \\
&= p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{z}_{t-1})p(\mathbf{z}_{t-1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1})p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{z}_t) \\
&= p(\mathbf{x}_t, \mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{z}_{t-1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1})p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{z}_t) \\
&= p(\mathbf{x}_t, \mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1})p(\mathbf{z}_{t-1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1})p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{z}_t) \\
&= p(\mathbf{x}_{1:t}, \mathbf{z}_{t-1}, \mathbf{z}_t)p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{z}_t) \\
&= p(\mathbf{x}_{1:t}, \mathbf{z}_{t-1}, \mathbf{z}_t)p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T|\mathbf{x}_{1:t}, \mathbf{z}_t, \mathbf{z}_{t-1}) \\
&= p(\mathbf{x}_{1:T}, \mathbf{z}_{t-1}, \mathbf{z}_t)
\end{aligned}$$

From (21),

$$\begin{aligned}
\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) &= p_{\hat{\theta}}(\mathbf{z}_{t-1}, \mathbf{z}_t|\mathbf{x}_{1:T}) \\
&= \frac{p_{\hat{\theta}}(\mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{x}_{1:T})}{p_{\hat{\theta}}(\mathbf{x}_{1:T})} \\
&= \frac{c(\mathbf{z}_{t-1}, \mathbf{z}_t)}{Z(\mathbf{x}_{1:T})}
\end{aligned}$$

We successfully derive the BP update formula (23).

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

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

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

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

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

Furthermore, define

$$r_1 \coloneqq Z(\mathbf{x}_1), \tag{35}$$

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

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

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

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

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

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

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

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

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,

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

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

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

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

$$\hat{\beta}(\mathbf{z}_T) \coloneqq \beta(\mathbf{z}_T), \tag{44}$$

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

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

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

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

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

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

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.

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.

Algorithm 1 Exact inference for (γ, ξ)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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