Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

QUIZ 2 ANSWER BOOKLET

Thursday, December 3, 2020
9:00 am – 11:00 am / 7:00 pm – 9:00 pm

NAME: Andy Markov

- Don't forget to put your name on all sheets.
- Remember that only this answer booklet will be considered in the grading of your exam.
- Be sure to **show all relevant work and reasoning.**
- Please be neat! You may want to first work things through on scratch paper and then neatly transfer to this answer booklet the work you would like us to look at.
- **Please remember that the problems add up to** 70 **points but the maximum score that you can get is 60 points. If you get** x **points, then the score that contributes to your final grades is** $\min\{x, 60\}$. For example: if you receive 45/70, your score will be 45; if you receive 65/70, your score will be 60.

Problem	Your score
1	20
2	10
3	15
4	25
Total	$\min\{70, 60\} = 60$

Problem 1

(a) $\log p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y};\theta) =$

$$\sum_{i=1}^{M} \underbrace{x_0^i}_{\triangleq \zeta_i} \log \tau_i + \sum_{i=1}^{M}\sum_{j=1}^{M} \underbrace{x_0^i x_1^j}_{\triangleq q_{ij}} \log b_{ij} + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} \underbrace{\left[\sum_{t=0}^{T-2} x_t^i x_{t+1}^j x_{t+2}^k\right]}_{\triangleq q_{ijk}} \log b_{ijk} + \sum_{i=1}^{M}\sum_{j=1}^{M} \underbrace{\left[\sum_{t=0}^{T} x_t^i y_t^j\right]}_{\triangleq n_{ij}} \log m_{ij}$$

$\widehat{\tau}_i = \hat{p}(x_0 = i) = \zeta_i$

$\widehat{b}_{ijk} = \hat{p}(x_{t+2} = k | x_{t+1} = j, x_t = i) = \frac{q_{ijk}}{\sum_{k'=1}^{M} q_{ijk'}}$

$\widehat{m}_{ij} = \hat{p}(y_t = j | x_t = i) = \frac{n_{ij}}{\sum_{k'=1}^{M} n_{ik'}}$

Reasoning/Work to be looked at for Problem 1(a):

The log-likelihood of the data is

$$\begin{aligned}
&\log p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y};\theta) \\
&= \log\left\{\tau_{x_0} b_{x_0,x_1} \prod_{t=0}^{T-2} b_{x_t,x_{t+1},x_{t+2}} \prod_{t=0}^{T} \beta_{x_t,y_t}\right\} \\
&= \log\left\{\left(\prod_{i=1}^{M} \tau_i^{x_0^i}\right)\left(\prod_{i=1}^{M}\prod_{j=1}^{M} b_{ij}^{x_0^i x_1^j}\right)\left(\prod_{t=0}^{T-2}\prod_{i=1}^{M}\prod_{j=1}^{M}\prod_{k=1}^{M} b_{ijk}^{x_t^i x_{t+1}^j x_{t+2}^k}\right)\left(\prod_{t=0}^{T}\prod_{i=1}^{M}\prod_{j=1}^{M} m_{ij}^{x_t^i y_t^j}\right)\right\} \\
&= \sum_{i=1}^{M} x_0^i \log \tau_i + \sum_{i=1}^{M}\sum_{j=1}^{M} x_0^i x_1^j \log b_{ij} + \sum_{t=0}^{T-2}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} x_t^i x_{t+1}^j x_{t+2}^k \log b_{ijk} + \sum_{t=0}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M} x_t^i y_t^j \log m_{ij} \\
&= \sum_{i=1}^{M} \underbrace{x_0^i}_{\triangleq \zeta_i} \log \tau_i + \sum_{i=1}^{M}\sum_{j=1}^{M} \underbrace{x_0^i x_1^j}_{\triangleq q_{ij}} \log b_{ij} + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} \underbrace{\left[\sum_{t=0}^{T-2} x_t^i x_{t+1}^j x_{t+2}^k\right]}_{\triangleq q_{ijk}} \log b_{ijk} + \sum_{i=1}^{M}\sum_{j=1}^{M} \underbrace{\left[\sum_{t=0}^{T} x_t^i y_t^j\right]}_{\triangleq n_{ij}} \log m_{ij}.
\end{aligned}$$

The ML estimates will simply be the empirical distributions of each parameter:

$$\begin{aligned}
\hat{\tau}_i &= \hat{p}(x_0 = i) = \zeta_i \\
\hat{b}_{ijk} &= \hat{p}(x_{t+2} = k | x_{t+1} = j, x_t = i) = \frac{q_{ijk}}{\sum_{k'=1}^{M} q_{ijk'}} \\
\hat{m}_{ij} &= \hat{p}(y_t = j | x_t = i) = \frac{n_{ij}}{\sum_{k'=1}^{M} n_{ik'}}
\end{aligned}$$

Note: A common mistake was for students to normalize the messages incorrectly. b_{ijk} is the probability of transitioning to k given the previous two states were i, j, and thus we must normalize so $\sum_{k'=1}^{M} b_{ijk'} = 1$.

(b)(i) $\mathbb{P}(x_t = i, x_{t+1} = j, x_{t+2} = k|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)}) \cong$
$m_{(t-2)\to t}(x_t) m_{t-1\to t+1}(x_{t+1}) m_{t+3\to t+1}(x_{t+1}) m_{(t+4)\to t+2}(x_{t+2}) m_{(t-1)\to t}(x_t) m_{t+3\to t+2}(x_{t+2})$
$\phi_t(x_t)\phi(x_{t+1})\phi(x_{t+2})\psi_{t,t+1}(x_t, x_{t+1})\psi_{t,t+2}(x_t, x_{t+2})\psi_{t+1,t+2}(x_{t+1}, x_{t+2})$

(b)(ii) $\widehat{\tau}_i = \mathbb{P}(x_0 = i|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)})$

$\widehat{b}_{ijk} = \frac{\sum_{t=0}^{T-2} \mathbb{P}(x_t=i, x_{t+1}=j, x_{t+2}=k|\mathbf{y}=\mathbf{y}; \theta_M^{(\ell)})}{\sum_{t=0}^{T-1} \mathbb{P}(x_t=i, x_{t+1}=j|\mathbf{y}=\mathbf{y}; \theta_M^{(\ell)}}$

$\widehat{m}_{ij} = \frac{\sum_{t=0}^{T} \mathbb{P}(x_t=i|\mathbf{y}=\mathbf{y}; \theta_M^{(\ell)}) \mathbf{1}(y_t=j)}{\sum_{t=0}^{T} \mathbb{P}(x_t=i|\mathbf{y}=\mathbf{y}; \theta_M^{(\ell)})}$

Reasoning/Work to be looked at for Problem 1(b):

b(i): To produce the marginal belief over the three nodes by eliminating all other nodes. This involves multiplying the incoming messages, node potentials of the three variables, and their edge potentials:

$$\begin{aligned}
&\mathbb{P}(x_t = i, x_{t+1} = j, x_{t+2} = k|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)}) \\
\cong &m_{(t-2)\to t}(x_t) m_{t-1\to t+1}(x_{t+1}) m_{t+3\to t+1}(x_{t+1}) m_{(t+4)\to t+2}(x_{t+2}) m_{(t-1)\to t}(x_t) m_{t+3\to t+2}(x_{t+2}) \\
&\phi_t(x_t)\phi(x_{t+1})\phi(x_{t+2})\psi_{t,t+1}(x_t, x_{t+1})\psi_{t,t+2}(x_t, x_{t+2})
\end{aligned}$$

A common minor mistake was to leave out the incoming messages to x_{t+1}.

b(ii): We find the ML estimate in the M-step by maximizing $\mathbb{E}_{p(x|y;\theta^{(\ell)})}[\log \mathbb{P}_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}; \theta)]$, which boils down to calculating $\mathbb{E}[x_0^i], \mathbb{E}[x_t^i x_{t+1}^j x_{t+2}^k], \mathbb{E}[x_t^i y_{t+1}^j]$ over $\mathbb{P}_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}; \theta)$. We can reuse our ML estimates from part (a), but with our new estimates of $\zeta_i, q_{ijk}, \eta_{ij}$, to get:

$$\begin{aligned}
\widehat{\tau}_i &= \mathbb{P}(x_0 = i|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)}) \\
\widehat{b}_{ijk} &= \frac{\sum_{t=0}^{T-2} \mathbb{P}(x_t = i, x_{t+1} = j, x_{t+2} = k|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)})}{\sum_{t=0}^{T-1} \mathbb{P}(x_t = i, x_{t+1} = j|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)})} \\
\widehat{m}_{ij} &= \frac{\sum_{t=0}^{T} \mathbb{P}(x_t = i|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)}) \mathbf{1}(y_t = j)}{\sum_{t=0}^{T} \mathbb{P}(x_t = i|\mathbf{y} = \mathbf{y}; \theta_M^{(\ell)})}
\end{aligned}$$

(c)(i) $\mathbf{s} = \{s_0, \ldots, s_T\}$ where $s_t = \{x_{t-1}, x_t\}$ for $t = 1, \ldots, T$ and $s_0 = x_0$

$\mathbf{v} = \mathbf{y}$

(c)(ii) $\widehat{\tau}_i^{(\ell+1)} = \mathbb{P}(s_0 = i | \mathbf{v} = \mathbf{v}; \theta^{(\ell)})$

$\widehat{b}_{ijk}^{(\ell+1)} = \frac{\sum_{t=1}^{T-1} \mathbb{P}(s_t = \{i,j\}, s_{t+1} = \{j,k\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)})}{\sum_{t=1}^{T-1} \mathbb{P}(s_t = \{i,j\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)})}$

$\widehat{m}_{ij}^{(\ell+1)} = \frac{\sum_{t=1}^{T} \sum_{k=1}^{M} \mathbb{P}(s_t = \{k,i\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)}) \mathbf{1}(v_t = j)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \mathbb{P}(s_t = \{k,i\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)})}$

Reasoning/Work to be looked at for Problem 1(c):

c(i): We define s_t to be a supernode containing the set of variables $\{x_{t-1}, x_t\}$ and keep $v_t = y_t$. The initial s_0 simply contains x_0. Notice that v_t is independent of all other nodes conditioned on s_t, and s_{t+1} only depends on s_t. Thus we have our original HMM where y_t are the observed states and s_t are the hidden states at each timestep. The parameters would be:

$$\begin{aligned} \pi_i &= \mathbb{P}(s_0 = i) = \tau_i \\ a_{\{ij\}\{j'k\}} &= \mathbb{P}(s_{t+1} = \{j', k\} | s_t = \{i, j\}) = \mathbf{1}(j = j') b_{ijk} \\ \eta_{\{ij\}k} &= \mathbb{P}(y_t = k | s_t = \{i, j\}) = m_{jk} \end{aligned}$$

Students gave multiple other alternative solutions which were all acceptable, including $s_t = \{x_t, x_{t+1}\}, v_t = \{y_t, y_{t+1}\}$ and $s_t = \{x_{2t}, x_{2t+1}\}, v_t = \{y_{2t}, y_{2t+1}\}$. Some solutions which were overly complex and which would not lend itself to the Baum Welch algorithm easily were penalized slightly.

c(ii): We first write the modified HMM parameters in terms of the original HMM parameters, and then plug in the solutions to the original HMM parameters in terms of the edge and node potentials.

$$\begin{aligned} \widehat{\tau}_i^{(\ell+1)} &= \widehat{\pi}_i^{(\ell+1)} = \mathbb{P}(s_0 = i | \mathbf{v} = \mathbf{v}; \theta^{(\ell)}) \\ \widehat{b}_{ijk}^{(\ell+1)} &= \widehat{a}_{\{ij\}\{jk\}} \\ &= \frac{\sum_{t=1}^{T-1} \mathbb{P}(s_t = \{i,j\}, s_{t+1} = \{j,k\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)})}{\sum_{t=1}^{T-1} \mathbb{P}(s_t = \{i,j\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)})} \\ \widehat{m}_{ij}^{(\ell+1)} &= \frac{\sum_{t=1}^{T} \sum_{k=1}^{M} \mathbb{P}(s_t = \{k,i\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)}) \mathbf{1}(v_t = j)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \mathbb{P}(s_t = \{k,i\} | \mathbf{v} = \mathbf{v}; \theta^{(\ell)})} \end{aligned}$$

Depending on the answer to c(i), the solution to c(ii) varied, but the solution was correct as long as it was analogous to the one provided above.

Problem 2

(a) $P(Y = 1|do(Z = 0)) = 0.162$

Reasoning/Work to be looked at for Problem 2(a):

We can write $P(Y = 1|do(Z = 0))$ as

$$
\begin{aligned}
& P(Y = 1|do(Z = 0)) \\
=& \sum_{u=0}^{1} P_u(u) \sum_{x=0}^{1} P_{x|z,u}(x|0, u) P_{Y|X,Z}(1|x, 0) \\
=& P(u = 0)[P(x = 0|z = 0, u = 0)P(y = 1|x = 0, z = 0)+ \\
& P(x = 1|z = 0, u = 0)P(y = 1|x = 0, z = 0)]+ \\
& P(u = 1)[P(x = 0|z = 0, u = 1)P(y = 1|x = 0, z = 0)+ \\
& P(x = 1|z = 0, u = 1)P(y = 1|x = 1, z = 0)] \\
=& 0.6(0.5 * 0.2 + 0.5 * 0.1) + 0.4(0.8 * 0.2 + 0.2 * 0.1) \\
=& 0.162
\end{aligned}
$$

Note: Students received partial credit for writing the correct interventional distribution, but incorrectly computing the final answer.

(b)(i) **Reasoning/Work to be looked at for Problem 2(b)(i):**

We can write the do-operation as

$$\begin{aligned}\mathbb{P}(y|do(z=z')) &= \sum_{u=0}^{1}\sum_{x=0}^{1} P(u)P(x|u,z')P(y|x,z') \\ &= \sum_{x=0}^{1} P(y|x,z')\sum_{u=0}^{1} P(u)P(x|u,z').\end{aligned}$$

It's not possible to remove the dependence on u, so it's not possible to determine the causal effect without observing u.

(b)(ii) **Reasoning/Work to be looked at for Problem 2(b)(ii):**

$$\begin{aligned}\mathbb{P}(y|do(x=x')) &= \sum_{z=0}^{1}\sum_{u=0}^{1} P(y|x',z)P(z|u)P(u) \\ &= \sum_{z=0}^{1} P(y|x',z)\sum_{u=0}^{1} P(z|u)P(u) \\ &= \sum_{z=0}^{1} P(y|x',z)P(z)\end{aligned}$$

This causal effect is possible to calculate using only the observed data, because there is no dependency on u.

Note: An alternate valid explanation is to state that $\{z\}$ is a backdoor adjustment. However, stating that $\{z\}$ satisfies the parent adjustment is not correct, since $\pi_x = \{u, z\}$.

Problem 3

(a) Which of the following is true? (Please provide a proof for your answer.)

(i) $\mathcal{G}_1 = \mathcal{G}_\mathcal{T}$.

✓ $\mathcal{G}_1 = \mathcal{G}_c$.

(iii) None of the above

Reasoning/Work to be looked at for Problem 3(a):

(ii) $\mathcal{G}_1 = \mathcal{G}_c$ is True.

Since $\mathbf{J}$ is positive definite, $\mathbf{\Lambda}$ is also positive definite (Inverse of a positive definite matrix is positive definite). Further, $\mathbf{\Lambda}_{ii} \neq 0$ for all $i \in \{1, \cdots, p\}$ because all the diagonal entries of a positive definite matrix are positive. Therefore, we have

$$\rho_{i,j} = \frac{\mathbf{\Lambda}_{ij}}{\sqrt{\mathbf{\Lambda}_{ii}\mathbf{\Lambda}_{jj}}} \tag{1}$$

Using hint (A), $\mathbf{\Lambda}_{ij} \neq 0$ for all $(i,j) \in \mathcal{E}_\mathcal{T}$. Therefore, $\rho_{i,j} \neq 0$ for all $(i,j) \in \mathcal{E}_\mathcal{T}$. Consider any $m \neq n \in \{1, \cdots, p\}$. There is a unique path connecting nodes m and n in the tree $\mathcal{G}_\mathcal{T}$. For every edge (s,t) along this path, $\rho_{s,t} \neq 0$. Using hint (B), we see that $\rho_{m,n} \neq 0$. Therefore, $\mathcal{E}_1 = \mathcal{E}_c$ implying $\mathcal{G}_1 = \mathcal{G}_c$.

(b) Which of the following is true? (Please provide a proof for your answer.)

✓ $\mathcal{G}_2 = \mathcal{G}_{\mathcal{T}}$.

(ii) $\mathcal{G}_2 = \mathcal{G}_c$.

(iii) None of the above

Reasoning/Work to be looked at for Problem 3(b):

(i) $\mathcal{G}_2 = \mathcal{G}_{\mathcal{T}}$ is True.

Let $\mathbf{A}$ denote the meta-variable (x_i, x_j) and let $\mathbf{B}$ denote the meta-variable $(x_{\mathcal{V}\setminus\{i,j\}})$. We can write down the covariance matrix $\mathbf{\Lambda}$ as follows:

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda_{AA}} & \mathbf{\Lambda_{AB}} \\ \mathbf{\Lambda_{BA}} & \mathbf{\Lambda_{BB}} \end{bmatrix} \tag{2}$$

Similarly, we can write down the information matrix $\mathbf{J}$ as follows:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J_{AA}} & \mathbf{J_{AB}} \\ \mathbf{J_{BA}} & \mathbf{J_{BB}} \end{bmatrix} \tag{3}$$

Using the Schur's complement, we have

$$[\mathbf{J_{AA}}]^{-1} = \mathbf{\Lambda_{AA}} - \mathbf{\Lambda_{AB}}[\mathbf{\Lambda_{BB}}]^{-1}\mathbf{\Lambda_{BA}} \tag{4}$$

We know that the covariance matrix of $x_i, x_j | x_{\mathcal{V}\setminus\{i,j\}}$ is $\mathbf{\Lambda_{AA}} - \mathbf{\Lambda_{AB}}[\mathbf{\Lambda_{BB}}]^{-1}\mathbf{\Lambda_{BA}}$. Putting it all together, we have

$$r_{i,j} \triangleq \frac{\mathrm{cov}(x_i, x_j | x_{\mathcal{V}\setminus\{i,j\}})}{\sqrt{\mathrm{var}(x_i | x_{\mathcal{V}\setminus\{i,j\}})}\sqrt{\mathrm{var}(x_j | x_{\mathcal{V}\setminus\{i,j\}})}} = \frac{[\mathbf{J_{AA}}]^{-1}_{12}}{\sqrt{[\mathbf{J_{AA}}]^{-1}_{11}[\mathbf{J_{AA}}]^{-1}_{22}}} \tag{5}$$

Now, let us compute $[\mathbf{J_{AA}}]^{-1}$. We have

$$\mathbf{J_{AA}} = \begin{bmatrix} \mathbf{J}_{ii} & \mathbf{J}_{ij} \\ \mathbf{J}_{ji} & \mathbf{J}_{jj} \end{bmatrix} \tag{6}$$

Computing the inverse of the above 2×2 matrix, we have

$$[\mathbf{J_{AA}}]^{-1} = \frac{1}{\mathbf{J}_{ii}\mathbf{J}_{jj} - \mathbf{J}_{jj}^2} \begin{bmatrix} \mathbf{J}_{jj} & -\mathbf{J}_{ij} \\ -\mathbf{J}_{ji} & \mathbf{J}_{ii} \end{bmatrix} \tag{7}$$

Combining (5) and (7), we have

$$r_{i,j} = \frac{-\mathbf{J}_{ij}}{\sqrt{\mathbf{J}_{ii}\mathbf{J}_{jj}}} \tag{8}$$

Therefore, $r_{i,j} = 0$ iff $\mathbf{J}_{ij} = 0$. This implies that $r_{i,j} = 0$ iff $(i,j) \in \mathcal{E}_{\mathcal{T}}$ i.e., $r_{i,j} \neq 0$ for all $(i,j) \notin \mathcal{E}_{\mathcal{T}}$. Therefore, $\mathcal{E}_2 = \mathcal{E}_{\mathcal{T}}$ implying $\mathcal{G}_2 = \mathcal{G}_{\mathcal{T}}$.

Common mistakes. Many students argued the equivalence

$$\mathrm{cov}(x_i, x_j \mid x_{\mathcal{V}\setminus\{i,j\}}) = 0 \iff x_i \perp\!\!\!\perp x_j \mid x_{\mathcal{V}\setminus\{i,j\}}.$$

The "only if" ($\Longrightarrow$) part is not true in general, although the "if" ($\Longleftarrow$) part is true for any distributions. The "only if" part is true for *Gaussian* distributions, but this has to be proved (e.g., by the argument above) or explicitly pointed out in the solution.

(c) **Reasoning/Work to be looked at for Problem 3(c):**

Following the hint, let (u, v) be that particular edge of $\mathcal{G}_c$ which is not present in $\hat{\mathcal{G}}$. Let $\mathbf{A} = \{u, v\}$ and $\mathbf{B} = \mathcal{V} \setminus \mathbf{A}$. We can re-write $\hat{\mathbf{J}}^b$ and $\hat{\mathbf{J}}^a$ as follows:

$$\hat{\mathbf{J}}^b = \begin{bmatrix} \hat{\mathbf{J}}^b_{\mathbf{AA}} & \hat{\mathbf{J}}^b_{\mathbf{AB}} \\ \hat{\mathbf{J}}^b_{\mathbf{BA}} & \hat{\mathbf{J}}^b_{\mathbf{BB}} \end{bmatrix} \qquad \hat{\mathbf{J}}^a = \begin{bmatrix} \hat{\mathbf{J}}^a_{\mathbf{AA}} & \hat{\mathbf{J}}^b_{\mathbf{AB}} \\ \hat{\mathbf{J}}^b_{\mathbf{BA}} & \hat{\mathbf{J}}^b_{\mathbf{BB}} \end{bmatrix} \tag{9}$$

where

$$\hat{\mathbf{J}}^b_{\mathbf{AA}} = \begin{bmatrix} 1 & -\hat{r}_{u,v} \\ -\hat{r}_{u,v} & 1 \end{bmatrix} \qquad \hat{\mathbf{J}}^a_{\mathbf{AA}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{10}$$

The Schur complement of $\hat{\mathbf{J}}^b_{\mathbf{BB}}$ is as follows:

$$\hat{\mathbf{J}}^b / \hat{\mathbf{J}}^b_{\mathbf{BB}} = \hat{\mathbf{J}}^b_{\mathbf{AA}} - \hat{\mathbf{J}}^b_{\mathbf{AB}} [\hat{\mathbf{J}}^b_{\mathbf{BB}}]^{-1} \hat{\mathbf{J}}^b_{\mathbf{BA}} \tag{11}$$

The Schur complement of $\hat{\mathbf{J}}^a_{\mathbf{BB}}$ is as follows:

$$\hat{\mathbf{J}}^a / \hat{\mathbf{J}}^b_{\mathbf{BB}} = \hat{\mathbf{J}}^a_{\mathbf{AA}} - \hat{\mathbf{J}}^b_{\mathbf{AB}} [\hat{\mathbf{J}}^b_{\mathbf{BB}}]^{-1} \hat{\mathbf{J}}^b_{\mathbf{BA}} \tag{12}$$

Positive definiteness of $\hat{\mathbf{J}}^b$ implies positive definiteness of $\hat{\mathbf{J}}^b_{\mathbf{BB}}$ and $\hat{\mathbf{J}}^b / \hat{\mathbf{J}}^b_{\mathbf{BB}}$. Because $\hat{\mathbf{J}}^b_{\mathbf{BB}}$ is positive definite, $\hat{\mathbf{J}}^a$ is positive definite if and only if $\hat{\mathbf{J}}^a / \hat{\mathbf{J}}^b_{\mathbf{BB}}$ is positive definite. In order to show positive definiteness of $\hat{\mathbf{J}}^a / \hat{\mathbf{J}}^b_{\mathbf{BB}}$, we will exploit the positive definiteness of $\hat{\mathbf{J}}^b / \hat{\mathbf{J}}^b_{\mathbf{BB}}$.

Using the fact that $\hat{\mathbf{J}}^b$ is MTP$_2$, we have $\hat{\mathbf{J}}^b_{\mathbf{BB}}$ is MTP$_2$ (any conditional distribution of a MTP$_2$ distribution is a MTP$_2$). This implies that $[\hat{\mathbf{J}}^b_{\mathbf{BB}}]^{-1}_{ij} \geq 0$ for all i, j. Therefore, all entries in $\hat{\mathbf{J}}^b_{\mathbf{AB}} [\hat{\mathbf{J}}^b_{\mathbf{BB}}]^{-1} \hat{\mathbf{J}}^b_{\mathbf{BA}}$ are non-negative. Therefore we can write $\hat{\mathbf{J}}^b / \hat{\mathbf{J}}^b_{\mathbf{BB}}$ and $\hat{\mathbf{J}}^a / \hat{\mathbf{J}}^b_{\mathbf{BB}}$ as:

$$\hat{\mathbf{J}}^b / \hat{\mathbf{J}}^b_{\mathbf{BB}} = \begin{bmatrix} 1-\alpha & -(\hat{r}_{u,v} + \beta) \\ -(\hat{r}_{u,v} + \beta) & 1-\delta \end{bmatrix} \qquad \hat{\mathbf{J}}^b / \hat{\mathbf{J}}^b_{\mathbf{BB}} = \begin{bmatrix} 1-\alpha & \beta \\ \beta & 1-\delta \end{bmatrix} \tag{13}$$

where $\alpha, \delta \in [0, 1)$ and $\beta \geq 0$. $\alpha, \beta, \delta \geq 0$ follows from non-negativity of $\hat{\mathbf{J}}^b_{\mathbf{AB}} [\hat{\mathbf{J}}^b_{\mathbf{BB}}]^{-1} \hat{\mathbf{J}}^b_{\mathbf{BA}}$ and $\alpha, \delta < 1$ follows because the diagonal entries of a positive definite matrix are positive. Note: $\hat{r}_{u,v} \geq 0$ because $\hat{\mathbf{J}}^b_{u,v} \leq 0$.

Since $\hat{\mathbf{J}}^b$ is positive definite we have

$$(\hat{r}_{u,v} + \beta)^2 < (1-\alpha)(1-\delta) \tag{14}$$

and hence

$$\beta^2 < (1-\alpha)(1-\delta) - \hat{r}^2_{u,v} - 2\hat{r}_{u,v}\beta \leq (1-\alpha)(1-\delta) \tag{15}$$

This implies positive definiteness of $\hat{\mathbf{J}}^a / \hat{\mathbf{J}}^b_{\mathbf{BB}}$. Therefore, $\hat{\mathbf{J}}^a$ is positive definite.

The implication of this result is that $\hat{\mathbf{J}}^a$ is always positive if all the empirical partial correlations are non-negative.

Alternative solution. Several students solved this problem with a different approach, without relying on Schur complements. Since $\hat{\mathbf{J}}^b$ is an M-matrix, we have $\hat{r}_{i,j} \geq 0$ and the off-diagonal entries of $\hat{\mathbf{J}}^b$ are nonpositive. Assume without loss of generality that there is an edge $(1,2) \in \mathcal{G}_c$ deleted from $\mathcal{G}_c$ when we threshold to obtain $\hat{\mathcal{G}}$ and $\hat{\mathbf{J}}^a$.

In order to prove that $\hat{\mathbf{J}}^a$ is a positive definite matrix, we have to show that for any nonzero $\mathbf{z} \in \mathbb{R}^p$, we have $\mathbf{z}^{\mathrm{T}}\hat{\mathbf{J}}^a\mathbf{z} > 0$. Note that, for any nonzero $\mathbf{z} \in \mathbb{R}^p$,

$$\begin{aligned}\mathbf{z}^{\mathrm{T}}\hat{\mathbf{J}}^a\mathbf{z} &= \sum_{i=1}^{p}\mathbf{z}_i^2 - \sum_{\substack{i\neq j\\(i,j)\neq(1,2)\\(i,j)\neq(2,1)}}\hat{r}_{i,j}\mathbf{z}_i\mathbf{z}_j \geq \sum_{i=1}^{p}|\mathbf{z}_i|^2 - \sum_{\substack{i\neq j\\(i,j)\neq(1,2)\\(i,j)\neq(2,1)}}\hat{r}_{i,j}|\mathbf{z}_i||\mathbf{z}_j| \\ &\geq \sum_{i=1}^{p}|\mathbf{z}_i|^2 - \sum_{i\neq j}\hat{r}_{i,j}|\mathbf{z}_i||\mathbf{z}_j| = |\mathbf{z}|^T\hat{\mathbf{J}}^b|\mathbf{z}|,\end{aligned}$$

where the two inequalities used that $\hat{r}_{i,j} \geq 0$. By positive definiteness of $\hat{\mathbf{J}}^b$, we have $|\mathbf{z}|^T\hat{\mathbf{J}}^b|\mathbf{z}| > 0$, therefore proving $\mathbf{z}^{\mathrm{T}}\hat{\mathbf{J}}^a\mathbf{z} > 0$.

Common mistakes. Many students argued

i. A Gaussian distribution $\mathtt{N}^{-1}(\mathbf{0}, \mathbf{J})$ is MTP_2 if and only if $\mathbf{J}_{ij} \leq 0$ for all $i \neq j$.

ii. Since $\hat{\mathbf{J}}^a_{ij} \leq 0$ for $i \neq j$ after thresholding $\hat{\mathbf{J}}^b_{ij}$, $\hat{\mathbf{J}}^a$ is the information matrix of an MTP_2 Gaussian distribution and hence positive definite.

While the first statement is true, one has to note that "$\mathtt{N}^{-1}(\mathbf{0}, \mathbf{J})$ is a Gaussian" already implicitly requires that $\mathbf{J}$ is a positive (semi)definite matrix. Therefore, $\mathtt{N}^{-1}(\mathbf{0}, \mathbf{J})$ being MTP_2 is equivalent to $\mathbf{J}_{ij} \leq 0$ for all $i \neq j$ *and* $\mathbf{J}$ being positive definite. As a result, showing $\hat{\mathbf{J}}^a_{ij} \leq 0$ for $i \neq j$ does not necessarily prove that $\hat{\mathbf{J}}^a$ is a positive definite matrix.

Problem 4

(a) **Reasoning/Work to be looked at for Problem 4(a):**

$$\mathbf{\Lambda}_{ij} = [\mathbf{J}^{-1}]_{ij} = [(\mathbf{I} - \mathbf{R})^{-1}]_{ij} = \sum_l [\mathbf{R}^l]_{ij} = \sum_l \phi(i \xrightarrow{l} j) = \phi(i \to j) \tag{16}$$

Common mistakes. A few students missed the step: $\sum_l \phi(i \xrightarrow{l} j) = \phi(i \to j)$.

(b) **Reasoning/Work to be looked at for Problem 4(b):**

Observe that the decomposition of $\mathcal{W}^k(i \xrightarrow{\setminus i} i)$ into the product of k single-revisit self-return walks is a valid decomposition. From Property 3, we have

$$\phi(\mathcal{W}^k(i \xrightarrow{\setminus i} i)) = \phi^k(i \xrightarrow{\setminus i} i) = \alpha_i^k \tag{17}$$

Observe that for any $k_1, k_2 \in \{0, 1, \cdots\}$, the subsets $\mathcal{W}^{k_1}(i \xrightarrow{\setminus i} i)$ and $\mathcal{W}^{k_2}(i \xrightarrow{\setminus i} i)$ are disjoint. Then, by $\mathbf{\Lambda}_{ii} = \phi(i \to i), \mathcal{W}(i \to i) = \cup_{k \geq 0} \mathcal{W}^k(i \xrightarrow{\setminus i} i)$, Property 4 and (17):

$$\mathbf{\Lambda}_{ii} = \phi(i \to i) = \sum_k \phi^k(i \xrightarrow{\setminus i} i) = \sum_k \alpha_i^k = \frac{1}{1 - \alpha_i} \tag{18}$$

Path-weightability of the model implies convergence of the geometric series (i.e., $|\alpha_i| < 1$).

Common mistakes. A few students did not show that the subsets $\mathcal{W}^k$ are disjoints. This is required to use Property 4. Also, a few students got confused with the limits of k: it should go from 0 to ∞.

(c)
- The recursive relationship for $\alpha_{i\to j}$ in terms of $\alpha_{k\to i}$ where $k \in \mathcal{N}(i)\backslash\{j\}$ is:

$$\alpha_{i\to j} = r_{i,j}^2 \frac{1}{1 - \sum_{k\in\mathcal{N}(i)\backslash j} \alpha_{k\to i}} \tag{19}$$

- $\alpha_{i\to j} = -\mathbf{J}_{i\to j}$
- $\gamma_{i\backslash j} = (1 + \sum_{k\in\mathcal{N}(i)\backslash\{j\}} \mathbf{J}_{k\to i})^{-1}$

Reasoning/Work to be looked at for Problem 4(c):

To calculate the path-weight for multiple-revisit self-return paths in $T_{i\backslash j}$, we can use the single-revisit counterpart as done in the sub-problem (b):

$$\gamma_{i\backslash j} = \phi(i \to i|T_{i\backslash j}) = \frac{1}{1 - \phi(i \xrightarrow{\backslash i} i|T_{i\backslash j})} \tag{20}$$

Now, we decompose the single-revisit paths in the subtree $T_{i\backslash j}$ in terms of the possible first step of the path (i, k), where $k \in \mathcal{N}(i) \setminus j$. Hence,

$$\phi(i \xrightarrow{\backslash i} i|T_{i\backslash j}) = \sum_{k\in\mathcal{N}(i)\backslash j} \phi(i \xrightarrow{\backslash i} i|T_{k\to i}) \tag{21}$$

Using $\alpha_{i\to j} = \phi(j \xrightarrow{\backslash j} j|T_{i\to j}) = r_{i,j}^2\phi(i \to i|T_{i\backslash j})$, (20), and (21) we are able to represent the path-weight $\phi(j \xrightarrow{\backslash j} j|T_{i\to j})$ in terms of the path-weights $\phi(i \xrightarrow{\backslash i} i|T_{k\to i})$ on smaller subtrees $T_{k\to i}$. This is the basis of the following recursive calculation:

$$\alpha_{i\to j} = r_{i,j}^2 \frac{1}{1 - \sum_{k\in\mathcal{N}(i)\backslash j} \alpha_{k\to i}} \tag{22}$$

These equations look strikingly similar to the belief propagation updates:

$$\mathbf{J}_{i\to j} = -\mathbf{J}_{ij}^2 \Big(1 + \sum_{k\in\mathcal{N}(i)\backslash j} \mathbf{J}_{k\to i}\Big)^{-1} \tag{23}$$

It is evident that the recursive path-weight equations can be mapped exactly to belief propagation updates. Observing that $r_{i,j}^2 = \mathbf{J}_{ij}^2$, we have the message update $\alpha_{i\to j} = -\mathbf{J}_{i\to j}$ and the variance estimate in the subtree $T_{i\backslash j}$ is $\gamma_{i\backslash j} = (1 + \sum_{k\in\mathcal{N}(i)\backslash\{j\}} \mathbf{J}_{k\to i})^{-1}$.

Common mistakes. A few students got the correct recursion but incorrect mapping.

(d) **Reasoning/Work to be looked at for Problem 4(d):**
The true variance $\mathbf{\Lambda}_{ii}$ is a path-weight over all self-return paths that start and end at i in $\mathcal{G}$. However, paths in $\mathcal{G}$ that start and end at i may map to paths that start at the root node of T_i but end at a replica of the root node instead of the root. These paths are not captured by the loopy BP variance estimate. The paths for the variance estimate $\hat{\mathbf{\Lambda}}_{ii}$ are self-return paths $\mathcal{W}(0 \to 0|T_i)$ that start and end at the root node in the computation tree.

For example, in $\mathcal{G}_1$ (Figure 4(b)), the path $(1, 2, 3, 1)$ is a self-return path in the original graph $\mathcal{G}_1$ but is not a self-return path in the computation tree of node 1 in $\mathcal{G}_1$ (Figure 6: (right)). Loopy BP variances capture only those self-return paths of the original graph $\mathcal{G}$ that are also self-return paths in the computation tree. For example, the path $(1, 3, 2, 3, 4, 3, 1)$ is a self-return path in both Figure 4(b) and Figure 6: (right). These paths are the backtracking paths.

Thus, the loopy BP variance estimate at each node is a sum over the backtracking self-return paths in $\mathcal{G}$, a subset of all self-return paths needed to calculate the correct variance.

Note: Providing a valid counterexample (a non-backtracking path) is a complete answer.

Common mistakes. A few students argued incorrectly that the variance estimate using the computation trees only captures finite paths and the true variance captures inifite paths.

(e) **Reasoning/Work to be looked at for Problem 4(e):**
The back-tracking paths for the variances have positive costs, since each edge in the path is traversed an even number of times. With each loopy BP iteration the computation tree grows and new back-tracking paths are included, hence variance estimates grow monotonically.

Note: Assuming $r_{i,j} \geq 0$ is incorrect.

Common mistakes. A few students did not argue why the costs of paths are positive / non-negative.