Problem 1 – I-maps
1.1 For each of the DAG’s below, draw an I-map in the form of an undirected graph. Keep the number of edges as small as possible.

1.2 For each of the undirected graphs below, draw the corresponding I-map in the form of a DAG. Keep the number of edges as small as possible.

Problem 2 – Triangulation, Junction Tree, Variable Elimination
2.1 Moralize and triangulate this graphical model. Add as few edges as possible during triangulation. List the variables eliminated, the cliques that appear, and the edges added.
2.2 Draw a junction tree for this graphical model; label all cliques and separators.

2.3 We want to find $P_{E|G=1,H=0}$ by variable elimination. You are to choose the elimination ordering so as to make the computation as efficient as possible. Write the initial potentials (i.e. the conditional probability tables), then, for each step of the elimination, write what potentials are eliminated, and what new potential is created.

In order to obtain the most efficient elimination, you must identify the potentials that are equal to 1 and use that to your advantage.

Original potentials:

$$P_V = P_GP_D|GP_E|GP_F|GP_B|DP_C|EP_A|BP_HC$$

<table>
<thead>
<tr>
<th>Variable eliminated</th>
<th>Potential created and potentials eliminated</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H) (Reduced) $\phi_C = P_{H=0</td>
<td>C}$</td>
</tr>
<tr>
<td>(G) (Reduced) $P_G =$ (constant), $\phi_D = P_{D</td>
<td>G=1}$, $\phi_E = P_{E</td>
</tr>
<tr>
<td>A  $1 = \sum_a P_A</td>
<td>BC$</td>
</tr>
<tr>
<td>B  $1 = \sum_b P_B</td>
<td>D$</td>
</tr>
<tr>
<td>D  $1 = \sum_d P_D</td>
<td>G=1 = \sum_d \phi_D$</td>
</tr>
<tr>
<td>C  $\phi_{EF} = \sum_c \phi_C P_C</td>
<td>EF$</td>
</tr>
<tr>
<td>F  $\phi_E = \sum_f \phi_{EF} \phi_F$</td>
<td></td>
</tr>
</tbody>
</table>

$$P_{E|H=0,G=1} \propto \phi_E \phi_E^t$$

Problem 3 – ML estimation in Bayes Net

In the graphical model below all variables have domain \{0, 1\}. We observe data containing a total of $N = 10$ samples.
3.1 Write the distribution $P_{ABCD}$ in factored form according to the DAG.

$$P_{ABCD} = P_A P_{B|A} P_C P_{D|BC}$$

3.2 What are the sufficient statistics for the parameters $\theta_C$ and $\theta_{D|BC}$ (numerical answer only)?

$$N_C = \begin{array}{c|c} 0 & 1 \\ \hline 7 & 3 \end{array}$$

$$N_{D|BC} = \begin{array}{c|c|c} BC = 00 & 1 & 5 \\ \hline 1 & 0 & 1 \\ 10 & 0 & 1 \\ 11 & 0 & 1 \end{array}$$

3.3 Compute the Maximum likelihood estimates of the parameters $\theta_C$ and $\theta_{D|BC}$ (numerical values only).

$$\theta_C = \begin{array}{c|c} 0 & 1 \\ \hline 0.7 & 0.3 \end{array}$$

$$\theta_{D|BC} = \begin{array}{c|c} BC = 00 & 1/6 & 5/6 \\ \hline 1 & 0 \\ 10 & 1 \\ 11 & 0 & 1 \end{array}$$

Problem 4 – Belief propagation

Consider the following Junction Tree and denote by $\phi_{AB}, \phi_{BC}, \phi_{CD}, \phi_{DE}$, and $\phi_B, \phi_C, \phi_D$ respectively the clique and separator potential tables.

Assume the clique and separator potentials contain the normalized marginals of the respective variable sets under a joint distribution $P_V$.

We perform

1. $\text{MaxAbsorb}(AB \rightarrow BC)$

2. $\text{MaxAbsorb}(BC \rightarrow CD)$

3. $\text{MaxAbsorb}(CD \rightarrow DE)$

in this order. The following questions refer to the potential tables after this sequence of operations.

4.1 What is $\max_{d,e} \phi_{DE}$?

$$\max_{x \in V} P_V(x_V)$$

4.2 What is $\phi_{DE}(d, e)$?
We start again with the clique and separator potentials containing the normalized marginals of the respective variable sets under the joint distribution $P_V$. We perform

1. EnterEvidence($A = a^*$), where $a^*$ is some fixed value in $\Omega_A$

2. Normalize($\phi_{AB}$)

3. Absorb($AB \rightarrow BC$)

4. Absorb($BC \rightarrow CD$)

5. MaxAbsorb($CD \rightarrow DE$)

in this order. This and the following questions refer to the potential tables after this sequence of operations. Show your work.

What is $\phi_{CD}$?

Below are shown, step by step, the potentials that change, and to what.

after 1. $\phi_{AB} = P_{AB}(a^*, b), b \in \Omega_B$

after 2. $\phi_{AB} = P_{B|A=a^*}$

after 3. $\phi_B = P_{B|A=a^*}$

$\phi_{BC} = P_{BC|A=a^*}$

after 4. $\phi_C = P_{C|A=a^*}$

$\phi_{CD} = P_{CD|A=a^*} P_C = P_{CD|A=a^*}$

4.4 What is $\phi_{DE}(d, e)$? Show your work.

after 5:

$$\phi_D = \max_c \phi_{CD} = \max_c P_{CD|A=a^*}$$

$$\phi_{DE} = \max_c \phi_{CD} \frac{\phi_{DE}}{\phi_D} = \max_c (P_{CD|A=a^*} \phi_{E|D}) = \max_c (P_{CDE|A=a^*})$$

4.5 What is $\max_{d,e} \phi_{DE}(d, e)$? Show your work.

$$\max_{d,e} \phi_{DE} = \max_{c,d,e} P_{CDE|A=a^*}$$

[This problem is optional, for extra credit. You can solve it instead of any other problem(s) on the exam.]

Problem 5 – the likelihood gradient in the Ising model

An Ising model over $V = \{1, 2, \ldots n\}$ is defined by

$$P_V(x_V) = \frac{1}{Z} e^{-\sum_{i \in V} a_i x_i - \sum_{ij \in E} b_{ij} x_i x_j}$$

In the above, $E$ is a given set of edges (i.e pairs of variables), $x_i \in \{0, 1\}, i \in V$ and $a_i, i \in V, b_{ij}, ij \in E$ are real parameters.

5.1. Write the expression of the normalization constant $Z$ of this model as a function of $a_i, i \in$
5.2. Derive the expression of the log-likelihood $l = \frac{1}{N} \ln P_V(D)$ for a data set of size $N$. What are the sufficient statistics?

$$l = \frac{1}{N} \sum_{x \in D} \ln P_V(x)$$

In (2) the terms containing $a_i$ are counted only when the respective $x_i$ is 1, and the terms containing $b_{ij}$ are counted only when the respective $x_i, x_j$ are both equal to 1. In (??) we denote by $N_i$ and $N_{ij}$ respectively the number of times variable $x_i$ is 1, and variables $x_i, x_j$ are simultaneously equal to 1, and in (??) we denote by $\hat{p}_i$ and $\hat{p}_{ij}$ respectively the probabilities of variable $x_i$ to be 1, and of variables $x_i, x_j$ to be simultaneously equal to 1, as relative frequencies in the data set.

5.3. Derive the formula for the partial derivatives $\frac{\partial Z}{\partial a_i}, \frac{\partial Z}{\partial b_{ij}}$.

Then show that the partial derivatives of $\ln Z$, $\frac{\partial \ln Z}{\partial a_i}, \frac{\partial \ln Z}{\partial b_{ij}}$ have a simple expression in terms of marginal probabilities.

[ OK to solve the problem in the special case $i = 1$. ]

For the special case $i = 1$ we have

$$Z = \sum_{x_1 x_2 \ldots x_n} e^{-\sum_{i \geq 2} a_i x_i - \sum_{i,j \neq 1} b_{ij} x_i x_j} (1 + e^{-a_1 - \sum_{j \neq 1} b_{1j} x_j})$$

$$\frac{\partial Z}{\partial a_1} = \sum_{x_1 x_2 \ldots x_n} e^{-\sum_{i \geq 2} a_i x_i - \sum_{i,j \neq 1} b_{ij} x_i x_j} \frac{\partial}{\partial a_1} (1 + e^{-a_1 - \sum_{j \neq 1} b_{1j} x_j})$$

$$= \sum_{x_2 x_3 \ldots x_n} e^{-\sum_{i \geq 2} a_i x_i - \sum_{i,j \neq 1} b_{ij} x_i x_j} (-e^{-a_1 - \sum_{j \neq 1} b_{1j} x_j}) = -ZP(x_1 = 1)$$

$$\frac{\partial \ln Z}{\partial a_1} = \frac{1}{Z} \frac{\partial Z}{\partial a_1} = -P(x_1 = 1)$$

$$\frac{\partial Z}{\partial b_{1j}} = \sum_{x_2 x_3 \ldots x_n} e^{-\sum_{i \geq 2} a_i x_i - \sum_{i,j \neq 1} b_{ij} x_i x_j} \frac{\partial}{\partial b_{1j}} (1 + e^{-a_1 - \sum_{j \neq 1} b_{1j} x_j})$$

$$= \sum_{x_2 x_3 \ldots x_n} e^{-\sum_{i \geq 2} a_i x_i - \sum_{i,j \neq 1} b_{ij} x_i x_j} (-x_j e^{-a_1 - \sum_{j \neq 1} b_{1j} x_j}) = -ZP(x_1 = 1, x_j = 1)$$

$$\frac{\partial \ln Z}{\partial b_{1j}} = \frac{1}{Z} \frac{\partial Z}{\partial b_{1j}} = -P(x_1 = 1, x_j = 1)$$

In general,

$$\frac{\partial \ln Z}{\partial a_i} = -P(x_i = 1) \quad \frac{\partial \ln Z}{\partial b_{ij}} = -P(x_i = 1, x_j = 1)$$
where \( P(x_i = 1), P(x_i = 1, x_j = 1) \) are the model marginals for the respective events.

5.4. Now use your result in 5.3 to derive the formula for the gradient of the log-likelihood:

\[
\frac{\partial l}{\partial a_i} = -\hat{p}_i + P(x_i = 1)
\]

\[
\frac{\partial l}{\partial b_{ij}} = -\hat{p}_{ij} + P(x_i = 1, x_j = 1)
\]

You can readily see analogies with the Gradient Ascent algorithm for a general MRF? The gradients are differences between empirical marginals (calculated from the data) and model marginals. For parameter \( a_i \), the corresponding gradient component depends on the marginals of \( x_i \); similarly, for \( b_{ij} \), the gradient depends on the marginal of \( x_i, x_j = 1 \).

The result above is an instance of a general property of exponential family model (coming up in 538): that the partial derivatives of \( \ln Z \) w.r.t the model parameters are moments.