# Probabilistic Graphical Models Final Exam

---

**Please fill your name and I.D.:**

Name: .....
I.D.: .....

Duration: 3 hours.

**Guidelines:**

1. The test is composed of five questions. The credit for each question is 25 points. You should accumulate as many points as you can from the possible **125** points. Notice that you will gain credit for each section you solve, **regardless** of whether you answer the complete question or not.
   Your grade in the test will be the number of points you obtained (where grades above 100 will be rounded to 100).
2. Concise answers are preferred. But, make sure that your arguments are clear and well supported.
3. **Notations and definitions:**
   (a) Val(X) set of possible values of RV (Random Variable) X.
   (b) Upper case letters denote RVs (e.g., X, Y, Z).
   (c) Upper case bold letters denote set of RVs (e.g., $\mathbf{X}$, $\mathbf{Y}$).
   (d) Lower case letters denote RV values (e.g., x, y, z).
   (e) Lower case bold letters denote RV set values (e.g., $\mathbf{x}$).
   (f) Values for categorical RVs with $|Val(X)| = k : x^1, x^2, \ldots, x^k$.
   (g) Marginal distribution over $\mathbf{X}$: P($\mathbf{X}$).
   (h) $Pa_x$ - Parents of RV X.
   (i) $P \models (A \perp B)$ - $P$ satisfies : $A$ is independent of $B$
   (j) $P \models (A \perp_c B)$ - under context $c$, $P$ satisfies that $A$ is independent of $B$
   (k) A positive distribution: a distribution $P$ where $\forall_x P(\mathbf{X} = \mathbf{x}) > 0$
   (l) Entropy of a discrete variable $Z$: $H(Z) = \sum_z -p(z) \log p(z)$.
   (m) Entropy of a discrete variable $Z$ given evidence $\mathbf{e}$ : $H(Z|\mathbf{e}) = \sum_{z:z\models\mathbf{e}} -p(z|\mathbf{e}) \log p(z|\mathbf{e})$.
   (n) The Kullback Leibler distance, KL, between two distribution defined on the RV $\mathbf{S}$: $KL(Pr(\mathbf{S}), Pr'(\mathbf{S})) = \sum_{\mathbf{s}} Pr(\mathbf{s}) \log \frac{Pr(\mathbf{s})}{Pr'(\mathbf{s})}$.
   (o) The Kullback Leibler distance, KL, between two distribution defined on the RV $\mathbf{S}$, under evidence $\mathbf{e}$: $KL(Pr(\mathbf{S}|e), Pr'(\mathbf{S}|e)) = \sum_{\mathbf{s}:\mathbf{s}\models\mathbf{e}} Pr(\mathbf{s}|\mathbf{e}) \log \frac{Pr(\mathbf{s}|\mathbf{e})}{Pr'(\mathbf{s}|\mathbf{e})}$.

**Good luck!**

**Questions:**

1. **Learning Naive Bayes network.**
   In this question we will deal with learning a *Naive Bayes* classification network. The network, illustrated in Figure 1, is composed of a discrete variable $C$ with domain $\{c^1 \cdots c^K\}$, and $N$ boolean variables: $\mathbf{X} = X_1 \cdots X_N$.

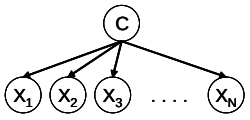


Figure 1: Naive Bayes network

   (a) **[5 points]** Assume you have $M$ I.I.D instances of complete data: $D = d_1 \cdots d_M$, where $d_i = \{c[i], \mathbf{x}[i]\}$ is the i-th instance of the data in which $C = c[i]$ and $\mathbf{X} = \mathbf{x}[i]$. We define also the notation of $x_j[i]$ to be the value of $X_j$ in data instance $i$.
   Write the formulas of the maximum likelihood estimate (MLE) for the network parameters i.e. the parameters $\Theta$ that maximize the likelihood of the data $L(D : \Theta)$. More specifically write the formulas of $\Theta_{X_i|C}$ and $\Theta_C$.
   *Remark: write the final solution without the entire derivation (e.g., no derivatives and no integrals).*

   (b) **[10 points]** Now, we would like to compute the Bayesian estimators for the parametrs incorporating a Dirichlet prior over the different variables.
   Recall that the Dirichlet prior is defined by:

$$P(\Theta) = \frac{1}{Z}\prod_k \Theta_k^{\alpha_k - 1} \text{ where } Z = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)} \text{ and } \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

   Write the following computations using Dirichlet prior over $c$ and $\mathbf{X}$:

   i. $P(X_j[M+1] = x^1 | D, C[M+1] = c^1)$
   ii. $P(C[M+1] = c^1 | D)$

   *Remark: write the final solution without the entire derivation (e.g., no derivatives and no integrals).*

   (c) **[10 points]** In real applications the value of the classification variable $C$ is often hidden. In this setting one option for parameter estimation employs the EM algorithm learned in class.

   i. **[2 points]** Briefly describe what is the type of computation that is done in the E-step and the M-step of an EM algorithm.
   ii. **[8 points]** Write the explicit formulas for the computation of the E-step and the M-step when learning the above network where $C$ is hidden.

2. **Context Specific Independence (CSI)**

(a) **[15 points]** Consider a BN $B$ with a variable $X$ that has a tree-CPD. Let $c$ be some context - a specific assignment of values for a subset $\mathbf{C}$ of $X's$ parents ($\mathbf{Pa_X}$). i.e., $\mathbf{c} \in Val(\mathbf{C})$ for $\mathbf{C} \subseteq \mathbf{Pa_X}$. Let $Z \notin \mathbf{C}$ be another variable in $B$ s.t. $Z \in \mathbf{Pa_X}$.
Propose a (simple) assumption on the distributions at the leaves of $X$'s tree-CPD under which it holds that if $P_B \models (X \perp_c Z|\mathbf{Pa_X} - Z, \mathbf{c})$, then $T_c$ (the tree-CPD of $X$ restricted to the branches consistent with $\mathbf{c}$) does not test $Z$. You may assume that all of the variables in $B$ are discrete binary variables.

(b) **[10 points]** Prove the following statement, or disprove it by finding a counterexample: CSI-sep (CSI - context specific independence) statements are monotonic in the context; i.e., let $\mathbf{c}$ be an assignment to some set of variables $\mathbf{C} \neq \emptyset$ in a BN, and let $\mathbf{C} \subset \mathbf{C}'$ , $\mathbf{c}'$ - an assignment to $\mathbf{C}'$ that is consistent with $\mathbf{c}$, then if $X$ and $Y$ are some variables in the BN that are CSI-separated given $\mathbf{c}$, they are also CSI-separated given $\mathbf{c}'$.
**Reminder - CSI-sep:** Let $B$ be a BN, $\mathbf{c}$ be a context, and let $\mathbf{X}, \mathbf{Y}$ be sets of variables. if $CSI - sep(B, \mathbf{c}, \mathbf{X}, \mathbf{Y}) = true$ then $P_B \models (\mathbf{X} \perp_c \mathbf{Y}|\mathbf{c})$.

3. **I-equivalence**
   In this question you will prove the following theorem about I-equivalent networks:

   If $\mathcal{G}$ and $\mathcal{G}'$ are two I-equivalent networks, then there is a sequence of networks $\mathcal{G} = \mathcal{G}_1 ... \mathcal{G}_k = \mathcal{G}'$ that are all I-equivalent, such that the difference between any network $\mathcal{G}_i$ to $\mathcal{G}i + 1$ is a single edge reversal.

   (a) **[2 points]**
      i. Define I-equivalence between two networks.
      ii. State the two structural properties that must hold in any two I-equivalent Bayesian networks.

   (b) **[3 points]** Consider the network structures shown in Figure 2. Which of them belong to the same I-equivalence class and which do not?

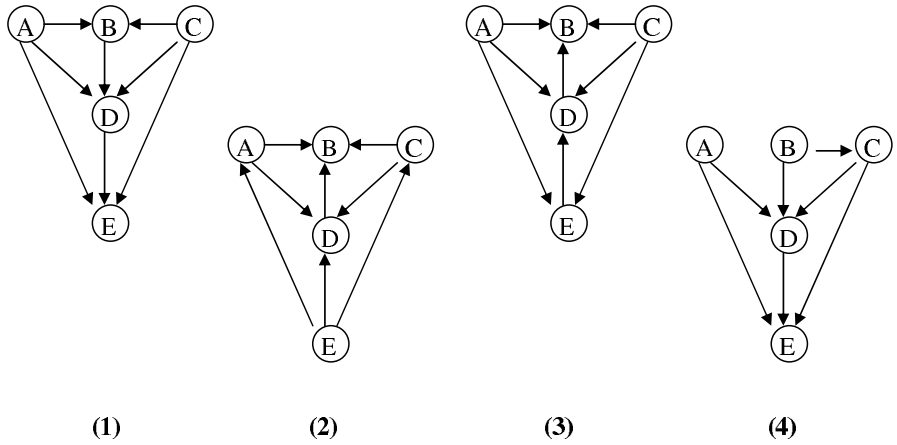


Figure 2: 4 different network structures

   (c) **[5 points]** Choose two I-equivalent graphs from Figure 2, and mark them as $\mathcal{G}$ and $\mathcal{G}'$.
      i. Show a series of networks $\mathcal{G} = \mathcal{G}_1 ... \mathcal{G}_k = \mathcal{G}'$ all I-equivalent such that the difference between any $\mathcal{G}_i$ to $\mathcal{G}i + 1$ is a single edge reversal.
      ii. Show a series of networks $\mathcal{G} = \mathcal{G}_1 ... \mathcal{G}_k = \mathcal{G}'$ such that the difference between any $\mathcal{G}_i$ to $\mathcal{G}i + 1$ is a single edge reversal, but the graphs are not all I-equivalent.

   (d) **[5 points]** Consider a network $\mathcal{G}$ and a single edge reversal in this network that results in a new network $\mathcal{G}'$, such that $X \rightarrow Y \in \mathcal{G}$ and $X \leftarrow Y \in \mathcal{G}'$.
   State a general rule such that if it holds for an edge $X \rightarrow Y \in \mathcal{G}$ then reversing this edge will result in an I-equivalent network $\mathcal{G}'$ (Hint: Consider the parents of nodes $X$ and $Y$). Prove your claim.

   (e) **[5 points]** Consider two I-equivalent networks $\mathcal{G}$ and $\mathcal{G}'$. Let $R$ be the set of all edges $X \rightarrow Y \in \mathcal{G}$ that have a different direction in $\mathcal{G}'$ (i.e., $X \rightarrow Y \in \mathcal{G}$ and $X \leftarrow Y \in \mathcal{G}'$). Show that there must exist at least one edge in $R$ that the rule you defined in the previous section holds for.

   (f) **[5 points]** Use the above section to show that if $\mathcal{G}$ and $\mathcal{G}'$ are two I-equivalent networks, then you can find a sequence of networks $\mathcal{G} = \mathcal{G}_1 ... \mathcal{G}_k = \mathcal{G}'$ such that the difference between any $\mathcal{G}_i$ to $\mathcal{G}i + 1$ is a single edge reversal.

4. **Approximate Inference by Edge Deletion**
**[25 points]** In this question we consider the problem of deleting edges from a Bayesian network for the purpose of simplifying models in probabilistic inference.
Let $B$ be a Bayesian network over variables $\chi$ with node $X$ having parents $Y$ and $\mathbf{U}$. The network $B'$ which results from deleting edge $Y \rightarrow X$ from $B$ given evidence $\mathbf{e}$ is defined as follows:

   - $B'$ has the same structure as $B$ except that edge $Y \rightarrow X$ is removed.
   - The CPT (conditional probability table) for variable $X$ in $B'$ is given by:

$$\theta' x|\mathbf{u} \doteq \sum_{y} \theta_{x|y,\mathbf{u}} Pr(y|\mathbf{e}).$$

   - The CPTs for variables other than $X$ in $B'$ are the same as those in $N$.

   (a) **[8 points]** Can the transformation described above be performed in polynomial time in general? be precise.

   For the following parts we will use the definitions of the entropy and the KL distance from the notations and definitions section.

   Assume that all distributions are positive and that the RVs which are instantiated in $\mathbf{e}$ do not contain $X$,$Y$ or $\mathbf{U}$.

   (b) **[8 points]** Let $B$ and $B'$ be two Bayesian networks as given in the definition. Prove that,

$$KL(Pr(\chi|\mathbf{e}), Pr'(\chi|\mathbf{e})) = log\frac{Pr'(\mathbf{e})}{Pr(\mathbf{e})} - \sum_{y,\mathbf{u}} Pr(y,\mathbf{u}|\mathbf{e}) \sum_{x} Pr(x|y,\mathbf{u},\mathbf{e}) log\Big(\frac{\theta' x|\mathbf{u}}{\theta x|y,\mathbf{u}}\Big)$$

   (c) **[9 points]** Let $B$ and $B'$ be two Bayesian networks as given in the definition. Use the previous result to prove that,

$$KL(Pr(\chi|\mathbf{e}), Pr'(\chi|\mathbf{e})) \leq \log \frac{Pr'(\mathbf{e})}{Pr(\mathbf{e})} + H(Y|\mathbf{e})$$

5. **Generalized Belief Propagation (GBP)**
**[25 points]** In class we saw that we can improve on the basic *loopy belief propagation* algorithm by propagating messages on a general cluster graph $\mathcal{K}$, resulting in the *generalized belief propagation* (GBP) algorithm. Recall that such a graph is specified both by the scope of its factors (the variables over which the factor are defined), and by the scope of the edges between these factors. In this question we are interested in applying GBP to directed models. To do so we need to transform a general Bayesian network $\mathcal{B}$ into an undirected general cluster graph.

**Reminder:**

Recall the following definitions of *sepset* and *running intersection* in cluster graph and in general cluster graph.

Given two of the graph nodes $C_i$, $C_j$, the main difference between the definitions of *sepset* in cluster graph and in a general cluster graph is that while in a cluster graph the *sepset* $S_{i,j}$ is defined by $S_{i,j} = C_i \cap C_j$, in a general cluster graph this definition is relaxed to $S_{i,j} \subseteq C_i \cap C_j$. In other words, while in a cluster graph the *sepset* is associated with the variables in the intersection of the two neighboring nodes, in general cluster graph it is associated with a subset of these variables.

*running intersection* in a cluster graph - if $x \in C_i$ and $x \in C_j$ then x is in each cluster in the (unique) path between $C_i$ and $C_j$.
*running intersection* in a general cluster graph - for each $X \in C_i$ and $X \in C_j$, there is exactly one path between $C_i$ and $C_j$ for which $X \in S$ for each subset $S$ along the path.

(a) **[10 points]** Below are two schemes for converting a Bayesian network $\mathcal{B}$ to a cluster graph $\mathcal{K}$. For each of these two schemes, either show (by proving the necessary properties) that it produces a valid general cluster graph for a general Bayesian network, or disprove this result by showing a counter-example. Note that exactly one of these schemes produces a valid general cluster graph for GBP inference.
**Scheme 1:** For each node $X_i$ in $\mathcal{B}$, define a factor $\phi_i$ over the family of $X_i$ in $\mathcal{B}$ (the node $X_i$ and its parents). Connect $\phi_i$ and $\phi_j$ if $X_j$ is a parent of $X_i$ in $\mathcal{B}$. The scope of such an edge is the intersection of the clusters.
**Scheme 2:** For each node $X_i$ in $\mathcal{B}$, define a factor $\phi_i$ over the family of $X_i$ in $\mathcal{B}$ (the node $X_i$ and its parents). Connect $\phi_i$ and $\phi_j$ if $X_j$ is a parent of $X_i$ in $\mathcal{B}$. The scope of such an edge is $\{X_j\}$.
In both cases, the initial potential in each factor $\phi_i$ is simply the CPD corresponding to $X_i$ given its parents in $\mathcal{B}$.

(b) **[10 points]** Construct an alternative scheme to the ones proposed in (a) that uses a spanning tree algorithm. (You can assume the existence of a spanning tree algorithm without defining it, but must be precise on how it is used.) Your scheme must transform any Bayesian network into a valid cluster graph. **Reminder:** A spanning tree algorithm is an algorithm that takes a graph $G$ as an input, and finds a subset of the edges that form a tree which spans all $G$'s nodes. **Hint:** You can show how you use a spanning tree algorithm to fix the result of the scheme in (a) that outputs a non-valid general cluster graph.

(c) **[5 points]** Show an example where the method you suggested in (b) could give a superior result to that of the one legal method of (a). **Hint:** You can actually show an example where (b) can give exact marginals, but the one legal method of (a) is only guaranteed to be an approximation.