# CS 228, Winter 2016
# Final Exam

**This exam is worth 100 points. You have 3 hours to complete it. Good luck!**

## Stanford University Honor Code

The Honor Code is the University's statement on academic integrity written by students in 1921. It articulates University expectations of students and faculty in establishing and maintaining the highest standards in academic work:

- The Honor Code is an undertaking of the students, individually and collectively:
  - that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
  - that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.
- The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.
- While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

## Signature

I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Stanford University Honor Code.

**Name / SUnetID**:

**Signature**:

| **Question** | **Score** | **Question** | **Score** |
|---|---|---|---|
| 1 | / 24 | 5 | / 16 |
| 2 | / 8 | 6 | / 16 |
| 3 | / 10 | | |
| 4 | / 26 | | |

**Total score:** **/ 100**

**Note: Partial credit will be given for partially correct answers. Zero points will be given to answers left blank. -2 points will be given to non-blank solutions that are completely incorrect and include a lot of irrelevant information. This applies to all questions on the exam.**

1. **[24 points] Conceptual Short Answer**

   (a) [**3 points**] Suppose you have an undirected graphical model where the graph is a pseudoforest (A pseudoforest is an undirected graph in which every connected component has at most one cycle–see fig 1). Can you guarantee that inference can be done tractably (i.e. in polynomial time) on this graph? Why or why not?

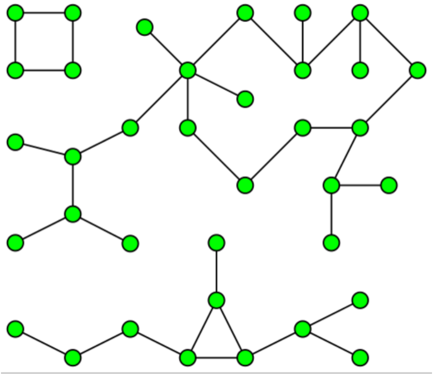


Figure 1: An example of a pseudoforest.

   **Answer:** Yes, the treewidth is bounded.

   (b) [**3 points**] An oil exploration company is using decision trees to predict the presence/absence of oil at several locations, using a set of geological features $f_i$ for each location. After taking CS 228, you think you can improve on their algorithm using a graphical model to jointly predict presence/absence at multiple locations, capturing spatial correlations. You have access to a large amount of labeled data, but you don't have a lot of domain knowledge in oil exploration. Which graphical model is more suitable for this? A Bayes Net, a MRF or a CRF? Why?

   **Answer:** CRF

   (c) [**3 points**] Let $p(X_1, X_2, X_3)$ be a joint probability distribution specified by a graphical model $G$. If $G$ is undirected, is it possible that $X_1$ and $X_3$ are (marginally) independent, but not conditionally independent given $X_2$? What if $G$ is directed?

   **Answer:** No, Yes

   (d) [**3 points**] Suppose that you're sampling from a CRF as in assignment 4 (image denoising), but you're working with very low precision floating point numbers, so values less than some nonzero $\epsilon$ are rounded to 0. What problems might this pose for your Gibbs sampling, if any?

   **Answers** The chain is no longer ergodic: the state space may now be disconnected.

   (e) [**3 points**] You have developed a distributed inference platform to do approximate inference on massive pairwise undirected graphical models. Your approach is to use loopy belief propagation (LBP), distributing the computation over $K$ computing nodes. Specifically, you partition the original graph into $K$ disjoint subsets of variables, and assign each subset to a computing node, which is responsible for all the message updates in that part of the original graph. There is no synchronization between the various computing nodes (updates are based on the messages received most recently), and messages may be delayed due to network latency. State what are the formal guarantees for this parallel LBP algorithm (e.g. convergence to approximate marginals, exact marginals, or no convergence at all). What if the underlying graphical model is a tree? **Answers** no guarantees; convergence to exact marginals

   (f) [**3 points**] Deva wants to estimate $\mathbb{E}_P[f(\mathbf{X})]$ for some distribution $P$ and a function $f$ over a random vector $\mathbf{X}$. Deva draws $M$ samples directly from the distribution $P$ and computes a Monte Carlo estimate $\hat{f}$. Simha comes along and says we can come up with a better estimator

$\tilde{f}$ (better here means less variance) by drawing the same number of $M$ samples from a different distribution $Q$. Deva argues it's not possible since he drew samples directly from the distribution $P$. Who is right? Why?

**Answer:** Simha can use importance sampling to get an estimator with less variance. The proposal distribution can be chosen in such a way (namely by taking into account $f$) that the variance variance is reduced in the importance sampling estimator.

(g) [**3 points**] You want to study the relationship between Smoking ($S$) and Cancer ($C$). You are given a dataset $\mathcal{D} = \{(s_0, c_0), \cdots, (s_m, c_m)\}$ and decide to implement a Bayesian network structure learning algorithm using BIC score. The optimal network you get is $C \to S$, even though you were expecting to get the more intuitive $S \to C$. Can you explain why?

**Answer:** No, $C \to S$ and $S \to C$ are i-equivalent

(h) Suppose you have a *bipartite* undirected Markov random field over disjoint sets of variables $U$ and $V$ (bipartite means that there are only edges between the $U$ and $V$ variables in the graph), specifying a joint probability distribution $p(u, v)$.

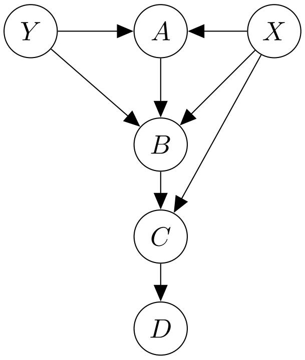
i. [**1.5 points**] You decide to use a mean field approximation for $p(u, v)$. In general, is there any theoretical guarantee for how accurate the mean field approximation will be? Why? Does the initialization of the mean field inference algorithm matter? **Answer:** can be very inaccurate, yes it matters.

ii. [**1.5 points**] You decide to use a mean field approximation for the posterior $p(u|v)$. In general, is there any theoretical guarantee for how accurate the mean field approximation will be? Why? Does the initialization of the mean field inference algorithm matter?

**Answer:** it will be exact. initialization does not matter

2. **[8 points] I-Maps and Conditional Independencies**

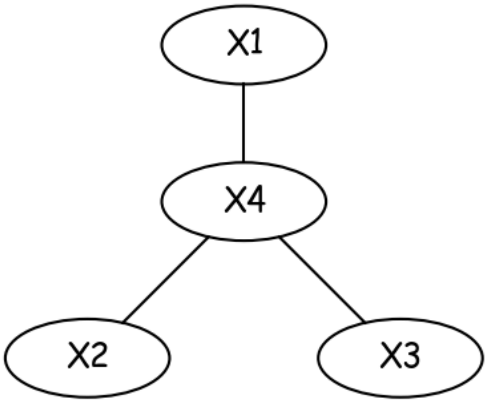
   How many I-equivalent graphs are there to the Bayes Network below? Justify your answer.



**Answers** There are 2 including the one given. The only edge that can be reversed is $A \to B$. The immoralities are $Y \to A \leftarrow X$ and $Y \to B \leftarrow X$.

3. **[10 points] Message Passing**

Suppose we have the following Markov network on 4 binary variables $X_1, X_2, X_3, X_4$:



The joint probabilty can be represented as:

$$P(X_1, X_2, X_3, X_4) = \frac{1}{Z}\phi_1(X_1, X_4)\phi_1(X_2, X_4)\phi_1(X_3, X_4)\phi_2(X_4),$$

where Z is the partition function, and

$\phi_1(X, Y) =$

| $Y$<br>$X$ | 0 | 1 |
|---|---|---|
| 0 | 1 | 2 |
| 1 | 3 | 4 |

that is, $\phi_1(0, 0) = 1$, $\phi_1(0, 1) = 2$, $\phi_1(1, 0) = 3$, $\phi_1(1, 1) = 4$, and

$\phi_2(X) =$

| X | 0 | 1 |
|---|---|---|
| | 1 | 2 |

that is, $\phi_2(0) = 1$ and $\phi_2(1) = 2$.

**Note that the order of the arguments matters**.

Suppose you run belief propagation (sum product) on this network.

(a) [**3 points**] What's the message $M_{1\to4}$? Give a symbolic and a numeric answer.

**Answers**
symbolic: $\sum_{X_1} \phi(X_1, X_4)$
Numeric:[4, 6]

(b) [**5 points**] What's the message $M_{4\to2}$? Give a symbolic and a numeric answer.

**Answers**
symbolic:

$$
\begin{aligned}
M_{4\to 2} &= \sum_{X_4} \phi(X_2, X_4)\phi(X_4) M_{1\to 4} M_{3\to 4} && (1) \\
&= [160, 336] && (2) \\
& && (3)
\end{aligned}
$$

(c) **[2 points]** What's the marginal probability $P(X_2 = 1)$?
**Answers**

$$M_{4\to 2}(1)/(M_{4\to 2}(1) + M_{4\to 2}(0))$$

4. **[26 points] Restricted Boltzmann Machine**

Restricted Boltzmann machines (RBMs) have been widely used as a generative model in many fields of machine learning: image processing, speech recognition and collaborative filtering. Concretely, an RBM is an undirected graphical model defined on variables $(\mathbf{v}, \mathbf{h})$, $\mathbf{v} \in \{0,1\}^m$ and $\mathbf{h} \in \{0,1\}^n$. The joint probability is given by

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(\phi(\mathbf{v}, \mathbf{h}))$$

where

$$\phi(\mathbf{v}, \mathbf{h}) = -\alpha^T \mathbf{v} - \beta^T \mathbf{h} - \mathbf{v}^T W \mathbf{h}$$

is a potential function. Here $\alpha \in \mathbb{R}^m$, $\beta \in \mathbb{R}^n$, $W \in \mathbb{R}^{m \times n}$ and $Z$ is the normalizing constant. You can interpret it as a fully connected *bipartite* network with two layers: one for visible variables $\mathbf{v}$ and one for hidden variables $\mathbf{h}$. In this problem, you will explore inference and learning for RBMs.

(a) [**5 points**] What is the marginal conditional distribution of $P(\mathbf{h}_i \mid \mathbf{v})$ for a single hidden variable $\mathbf{h}_i$? Can it be computed tractably?

**Answer:**

$$P(h_i|v) = \exp(-h_i * (\sum_j W_{ji} v_j) - \beta_i * h_i) / (\sum_{h_i} \exp(-h_i * (\sum_j W_{ji} v_j) - \beta_i * h_i))$$

(b) [**5 points**] What is the conditional distribution $P(\mathbf{h} \mid \mathbf{v})$? Can it be expressed in a compact (factored) form? **Answer:** Yes, product of the above marginals

(c) [**3 points**] Suppose we are given samples for the visible units $\mathcal{D} = \{\mathbf{v}^1, \cdots, \mathbf{v}^K\}$, and we want to learn the parameters of the RBM. Since we don't know the values of hidden variables $\mathbf{h}$, we use the marginal likelihood of each sample $\mathbf{v}$, given by

$$L(\mathbf{v}) = \sum_{\mathbf{h}} P(\mathbf{v}, \mathbf{h})$$

and thus the marginal log-likelihood for a single data sample $\mathbf{v}$ is

$$LL(\mathbf{v}) = \log(L(\mathbf{v})).$$

Can $L(\mathbf{v})$ be evaluated efficiently?

**Answer:** No

(d) [**3 points**] The learning procedure will apply gradient descent to maximize the marginal data log-likelihood $LL(\mathcal{D}; \alpha, \beta, W) = \sum_{\ell=1}^{K} \log(L(\mathbf{v}^\ell))$ with respect to the parameters $W, \alpha, \beta$. Can the gradient of $LL(\mathcal{D}; \alpha, \beta, W)$ be evaluated efficiently?

**Answer:** No

(e) [**4 points**] As an alternative approach, you decide to try EM. Can the E-step be performed efficiently? Why?

**Answer:** Yes, probabilities above

(f) [**3 points**] Can the M-step be performed efficiently? Justify your answer.

**Answer:** No, parameter learning in undirected model

(g) [**3 points**] Is EM guaranteed to find a global optimum for the marginal log-likelihood $LL(\mathcal{D}; \alpha, \beta, W)$?

**Answer:** No

5. **[16 points] Score-Based Structure Learning**

In score-based approaches of structure learning, we first define a score function $score(\mathcal{G}; \mathcal{D})$ that can score each candidate structure $\mathcal{G}$ with respect to the training data $\mathcal{D}$. After the definition of score function, we search in the space of directed acyclic graphs (DAGs) to find the graph structure $\mathcal{G}$ that maximizes the score $score(\mathcal{G}; \mathcal{D})$.

(a) **[10 points]** If the score function $score(\mathcal{G}; \mathcal{D})$ is defined as the log-likelihood $LL(\mathcal{D} \mid \mathcal{G})$, we have seen that a fully-connected graph $\mathcal{G}^{full}$ always maximizes the score. Let $G'$ be another Bayes Net structure. Let $\widehat{p}$ denote the empirical data distribution corresponding to $\mathcal{D}$. Under what conditions on $\widehat{p}$ is $LL(\mathcal{D} \mid \mathcal{G}^{full}) = LL(\mathcal{D} \mid \mathcal{G}')$? You may assume that the topological order of $G'$ and $G^{full}$ is the same. Hint: Your answer should consist of a list of conditional independencies.

**Answer:** For every variable, $X_i \perp \{X_1, \cdots, X_{i-1}\} \setminus \mathrm{Pa}_{X_i} \mid \mathrm{Pa}_{X_i}$ according to $\widehat{p}$

(b) **[2 points]** An additional term is usually added to the score function,

$$score(\mathcal{G}; \mathcal{D}) = LL(\mathcal{D} \mid \mathcal{G}) - \psi(M)\|G\|,$$

where $\|\mathcal{G}\|$ is the number of parameters in $\mathcal{G}$ and $M$ is the number of data samples. When $\psi(M) = 1$, it is AIC score; when $\psi(M) = \log M/2$, it is BIC score. Briefly explain why this additional term will prevent us from obtaining a fully-connected graph.

**Answer:** fully connected graph will be penalized because it requires many parameters to be specified.

(c) **[2 points]** Given a score function, a local search algorithm could help us find a locally optimal graph structure. Concretely, a local search method is an iterative procedure that starts with an initial guess for the best structure. Then, at each step we will consider all the neighbors of the current best graph. If none of them has a higher score than the current guess, we terminate the search. Otherwise, we pick the neighbor with the highest score as our new best guess. The neighbors of the graph are defined as all valid Bayesian networks obtained by a single operation of edge addition, deletion or reversal.

Which of the following are neighbors of the graph in Figure 2?

(a) removal of edge $B \to A$

(b) reversal of edge $D \to A$

(c) addition of edge $E \to C$

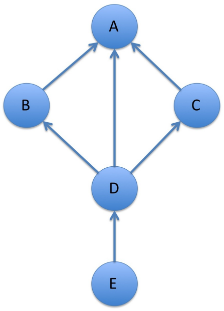


Figure 2: Part (c)

**Answer:** a,c

(d) **[2 points]** If we start the local search with two different initial graphs, as exemplified in Figure 3, will the two local search procedures end up with the same graph? If yes, please explain why. If not, explain how to choose between Graph1 and Graph2.

**Answer:** no, pick the one with highest score

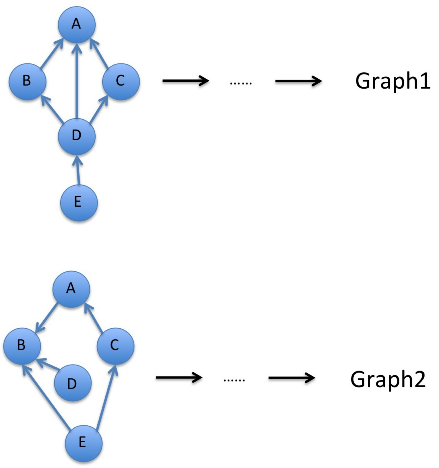


Figure 3: Part (d)

6. **[16 points] Parameter estimation**

A computer program generates pairs of coin flips $(X_1, X_2)$ according to the following generative model:

- With probability $\theta$, it generates two identical coin flips. The outcome is two heads (with probability $p$) or two tails (with probability $1 - p$).
- With probability $1 - \theta$, it generates two coin flips independently with probability of head equal to $q$.

You observe a sequence of pairs of coin flips generated by this program:

- $N_H$ of them are two heads
- $N_T$ of them are two tails
- $N_M$ of them are mixed, one head and one tail

$\theta, p, q$ are unknown, and you would like to estimate them from data. Note that $p$ and $q$ need not be the same.

- **[2 points]** What algorithm can you use to estimate the parameters $\theta, p, q$?
  **Answer:** EM
- **[14 points]** Provide pseudo-code
  **Answer:** Assume we know the hidden variables, i.e., which process generated the coin flips. Let
  - HHP: number of HH events generated by $P$
  - HHQ: number of HH events generated by $Q$
  - TTP: number of TT events generated by $P$
  - TTQ: number of TT events generated by $Q$

  We know that $HT$ and $TH$ flips must have been generated by $Q$.
  M-step:

$$p = \frac{HHP}{HHP + TTP}$$

$$q = \frac{2HHQ + N_M}{2TTQ + 2N_M + 2HHQ}$$

$$\theta = \frac{HHP + TTP}{HHP + TTP + HHQ + TTQ + N_M} = \frac{HHP + TTP}{N}$$

  E-step:

$$HHP = N_H \frac{\theta p}{\theta p + (1 - \theta) q^2}$$

$$HHQ = N_H - HHP$$

$$TTP = N_T \frac{\theta(1-p)}{\theta(1-p) + (1-\theta)(1-q)^2}$$

$$TTQ = N_T - TTP$$