TTIC 31180 — Probabilistic Graphical Models

Final Exam

Due: June 12, 2020

Name: _____

1. _____ / 6
2. _____ / 14
3. _____ / 5
4. _____ / 26
5. _____ / 1

Total: _____ / 52

The exam is open-book and open-notes, however you are not allowed to discuss the exam (or collaborate) with anyone

Clearly mark your answers (e.g., draw a box around them)

Make sure that you show your work and that your writing is legible

1 Probabilistic Graphical Models: A Quick Review [6pts]

Consider the following multiple choice questions, which provide a sampling of topics covered in the course. There is only one correct answer for each question. Circle your answer.

1. Which of the following are NOT true?
   (a) Every distribution has at least one I-map that is a DAG
   (b) Every distribution has at least one minimal I-map as a DAG
   (c) Every distribution has at least one P-map as a DAG
   (d) If a distribution has many P-maps, these P-maps are I-equivalent
2. Which of the following is true regarding graph elimination?
   (a) For a given graph, different elimination orderings yield the same computational complexity
   (b) For a star graph, the best strategy is to first eliminate the center node
   (c) For a tree, the best strategy is to first eliminate the root
   (d) For an Ising model, the best strategy is to first eliminate a row/column
3. Which of the following is true regarding loopy belief propagation?
   (a) There are no convergence guarantees for any graphs
   (b) If it converges for a given graph, it is guaranteed to converge to the correct solution
   (c) The beliefs associated with a subtree of a calibrated cluster graph are valid marginals for a distribution over the tree
   (d) None of the above
4. Which of the following is true regarding variational inference?
   (a) Variational methods formulate inference as an optimization problem
   (b) Mean field provides an outer bound for the marginal polytope
   (c) Belief propagation provides an inner bound for the marginal polytope
   (d) All of the above
5. Which of the following statements regarding maximum likelihood parameter estimation is NOT true?
   (a) MLE is purely data driven, and does not consider any a priori knowledge of parameters
   (b) MLE doesn't provide a measure of confidence in the resulting estimates
   (c) MLE for Bayesian networks requires global optimization over the full set of parameters
   (d) MLE for Markov networks may not have a unique global optimum
6. Which of the following statements about learning in the context of graphical models is true?
   (a) Maximum a posteriori estimation can prevent overfitting due to data sparsity
   (b) Fully observed Bayesian network structure learning is NP-hard since the number of directed graph is super-exponential in the number of random variables.
   (c) Fully observed MLE of Bayesian network structure favors complete graphs
   (d) All of the above

2 Importance Sampling for Robot Localization and Mapping [14pts]

Consider a robot that navigates an environment with n stationary landmarks. Let X_t be the pose (i.e., position and orientation) of the robot at time t and L_k be the pose of landmark k. At each time step, the robot makes noisy observations $O_t = \{O_{t,1}, \ldots, O_{t,k}, \ldots O_{t,n}\}$ of its pose relative to each landmark k.

Neither the robot's pose nor that of the landmarks are known a priori, and thus X_t and L_k are random variables, as is O_t. However, the robot needs an estimate its location in the environment in order to navigate. One solution is use the observations to maintain an estimate of the landmark poses (i.e., build a map) and to use these estimated landmark poses to estimate the robot's pose history (i.e., localize the robot).

Specifically, we are interested in maintaining the posterior distribution

$$P(X_{1:t}, L \mid O_{1:t})$$

at each time step t, where $X_{1:t} = \{X_1, X_2, \ldots, X_t\}$ is the robot's trajectory, $L = \{L_1, L_2, \ldots, L_n\}$ is the pose of the landmarks, and $O_{1:t} = \{O_{1,1}, O_{1,2}, \ldots, O_{1,n}, O_{2,1}, \ldots, O_{t,n}\}$ is the history of observations. We would like to maintain this posterior over time, which amounts to updating the distribution as the robot moves and we get new observations O_{t+1}.

Assume that the robot's pose follows a first-order Markov process, i.e., $P(X_{t+1} \mid X_0, X_1, \ldots, X_t) = P(X_{t+1} \mid X_t)$ (the *motion model*), and that the observations are conditionally independent given the robot and landmark poses $P(O_t \mid X_t, L) = \prod_k P(O_{t,k} \mid X_t, L_k)$ (the *measurement model*).

Unfortunately, the nature of the motion and measurement models means that the full posterior $P(X_{1:t}, L \mid O_{1:t})$ doesn't have a closed-form expression, making exact inference intractable. Instead, we factor the posterior

$$P(X_{1:t}, L \mid O_{1:t}) = P(L \mid X_{1:t}, O_{1:t}) P(X_{1:t} \mid O_{1:t})$$

and (i) employ importance sampling to maintain a sample-based approximation to the distribution over the robot trajectory $P(X_{1:t} \mid O_{1:t})$ and (ii) maintain a closed-form representation of $P(L \mid X_{1:t}, O_{1:t})$.

(a) **[2pts]** Draw the Bayesian network representation of this problem.

(b) **[12pts]** We will use importance sampling to maintain the distribution over the robot's trajectory. At each time step t, we have a set of M weighted particles $\{(X_{1:t}^{(m)}, w_t^{(m)})\}_{m=1}^{M}$, each of which is associated with a closed-form expression for the landmark distribution $P_t^{(m)}(L) = P(L \mid X_{1:t}^{(m)}, O_{1:t})$.

(i) Nominally, the landmark distribution is exponentially large. Can you exploit conditional independencies to compactly represent $P_t^{(m)}(L)$?

(ii) Consider the problem of computing the posterior at time $t+1$ based on the posterior at time t according to the robot's motion and the latest observation O_{t+1}, i.e., $P(X_{1:t}, L \mid O_{1:t}) \longrightarrow P(X_{1:t+1}, L \mid O_{1:t+1})$. Using the motion model $P(X_{t+1} \mid X_t)$ as the proposal distribution,

(a) generate a new set of particles (samples) for time $t+1$;

(b) compute the new weights $w_{t+1}^{(m)}$ of each particle; and

(c) compute the landmark distribution $P_{t+1}^{(m)}(L)$.

(a) Model 1: M_1 (b) Model 2: M_2

Figure 1: Two candidate models for the coin flip domain.

3 Maximum Likelihood Bayesian Model Selection [5pts]

You have two coins and are interested in modeling their likelihood of coming up heads or tails. Let $X_1, X_2 \in \{\text{H}, \text{T}\}$ be binary random variables that denote whether each coin comes up heads or tails. The joint distribution is then $P(X_1, X_2)$.

Consider two candidate models for the structure and parameterization of the Bayesian network. Model 1 (Fig. 1(a)) assumes that X_1 and X_2 are marginally independent, while Model 2 (Fig. 1(b)) assumes a causal relationship between the two coins.

Suppose that we are given the following dataset $\mathcal{D}$ with the results of 10 I.I.D. tosses of the pair of coins.

	1	2	3	4	5	6	7	8	9	10
X_1	T	T	H	H	T	H	T	T	H	H
X_2	T	H	T	T	H	T	H	H	T	T

For each model, compute the parameters that maximize the likelihood of the data. If you were to use maximum likelihood to determine the model structure, which model would you prefer. Why might this not be desirable?

4 Learning: Gaussian Networks [26pts]

This problem investigates parameter and structure learning in the context of Gaussian graphical models.

Suppose that we have a multivariate random vector $\boldsymbol{x}$ governed by a Gaussian probability distribution. A Gaussian distribution can be parameterized in one of two forms. The first is the *standard form* $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ and Σ are the mean vector and covariance matrix, respectively. The second is the *canonical* (*information*) form $\mathcal{N}^{-1}(\boldsymbol{x}; \boldsymbol{\eta}, \Lambda)$, where $\boldsymbol{\eta}$ is the information vector and Λ is the information (precision) matrix. The two parametrizations exhibit the following relationships: $\boldsymbol{\eta} = \Lambda\boldsymbol{\mu}$ and $\Lambda = \Sigma^{-1}$.

The canonical form defines the structure of the Gaussian Markov random field associated with the distribution and in turn, the conditional independencies among the variables. Specifically, the information matrix expresses the following independence relationships

$$\lambda_{ij} = 0 \leftrightarrow (x_i \perp x_j \,|\, \{\boldsymbol{x} - \{x_i, x_j\}\})$$

where $\boldsymbol{x} - \{x_i, x_j\}$ denotes the set of all variables in $\boldsymbol{x}$ except x_i and x_j. Thus, a non-zero entry λ_{ij} in the information matrix implies that there is no edge between the corresponding variables x_i and x_j in the Markov random field. Consequently, parameter estimation in the context of a canonical representation of a multivariate Gaussian also provides an estimate of the structure of the MRF.

Suppose that we have a set $\mathcal{D}$ consisting of M samples $\boldsymbol{x}^{(m)} \in \mathbb{R}^n$ drawn from a multivariate Gaussian distribution. We can group these samples into a single matrix of the form:

$$\boldsymbol{D} = \begin{bmatrix} {\boldsymbol{x}^{(1)}}^\top \\ {\boldsymbol{x}^{(2)}}^\top \\ \vdots \\ {\boldsymbol{x}^{(M)}}^\top \end{bmatrix}$$

(a) **[4pts]** Assume a univariate Gaussian represented in the standard (covariance) form as $\mathcal{N}(x; \mu, \sigma^2)$, where μ and σ^2 are scalars. Derive the expressions for the maximum likelihood estimates of the mean and variance.

(b) **[4pts]** Now, consider the canonical expression for the univariate distribution $\mathcal{N}(x; \mu_x, \lambda_x^{-1})$. Suppose that we know λ_x and are only concerned with estimating the mean μ_x. Suppose that we have a prior over the mean $P(\mu_x) = \mathcal{N}(\mu_x; \mu_{\mu_x}, \lambda_{\mu_x}^{-1})$. Show that the posterior $P(\mu_x \mid \mathcal{D}) = \mathcal{N}(\mu_x; \mu'_{\mu_x}, (\lambda'_{\mu_x})^{-1})$, where

$$\lambda'_{\mu_x} = M\lambda_x + \lambda_{\mu_x}$$
$$\mu'_{\mu_x} = \frac{M\lambda_x}{\lambda'_{\mu_x}}\mathbb{E}_{\mathcal{D}}[x] + \frac{\lambda_{\mu_x}}{\lambda'_{\mu_x}}\mu_{\mu_x}$$

Hint: Begin by proving that $\sum_m (x^{(m)} - \mu_x)^2 = M(\mu_x - \mathbb{E}_D[x]) + c$, where c is a constant that does not depend upon μ_x.

(c) **[6pts]** Let's now consider the complement, where we know the mean μ_x, but do not know the information term λ_x. It is useful to consider the Gamma distribution $y \sim \text{Gamma}(\alpha, \beta)$ for $\alpha, \beta > 0$, where

$$P(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^\alpha e^{-\beta y} \tag{1}$$

for which $\mathbb{E}[Y] = \frac{\alpha}{\beta}$ and $\text{Var}[Y] = \frac{\alpha}{\beta^2}$.

(i) Show that Gamma distributions are a conjugate family for this learning task, i.e., show that if $P(\lambda_x \mid \mu) \sim \Gamma(\alpha, \beta)$, then $P(\lambda_x \mid \mathcal{D}, \mu) \sim \text{Gamma}(\alpha', \beta')$, where

$$\alpha' = \alpha + \frac{1}{2} M$$
$$\beta' = \beta + \frac{1}{2} \sum_m (x^{(m)} - \mu)^2$$

(ii) Derive the mean and variance of λ_x in the posterior. What can you say about beliefs given the data? How do they differ from the MLE estimate?

(d) **[6pts]** Let's return to the multivariate case parameterized in the information form. This question steps you through the process of relating the information matrix to conditional independencies and, in turn, the structure of the graphical model. Without loss of generality, assume that the mean and information vector are zero, i.e., $\boldsymbol{\eta} = \boldsymbol{\mu} = \mathbf{0}$.[1]

(i) Partition $\boldsymbol{x}$ into two subsets $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$. The standard formulation of the Gaussian is then

$$p\left(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}; \mathbf{0}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right) \tag{2}$$

Derive $P(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$ expressed in the standard (covariance) form. (Hint: Express the inverse of a block matrix in terms of the Schur complement.)

(ii) Derive $\text{Var}(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$ in terms of the information matrix $\Lambda = \begin{bmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{bmatrix}$.

[1] This is easily achieved by subtracting the empirical mean from the samples, which does not affect the information (or covariance) matrix.

(iii) Based upon the above, argue why $\Lambda_{ij} = 0$ iff x_i is conditionally independent of x_j given the remaining variables.

(e) **[4pts]** Show that the log-likelihood of the data follows as

$$\ell(\Lambda : \mathcal{D}) = \log p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)} \mid \Lambda) \propto \log \det(\Lambda) - \text{tr}(S\Lambda)$$

where $S = \frac{1}{M} D^\top D$ is the sample covariance matrix.

(f) **[2pts]** Derive the maximum likelihood estimate for Λ.

5 Time Accounting

(a) **[1pts]** How long did you spend on this problem set?

Name: _____ Problem: _____

Name: _____ Problem: _____

Name: _____ Problem: _____