MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science

6.438 ALGORITHMS FOR INFERENCE
Fall 2011

Quiz 1
Tuesday, October 25, 2011
7:00pm–10:00pm

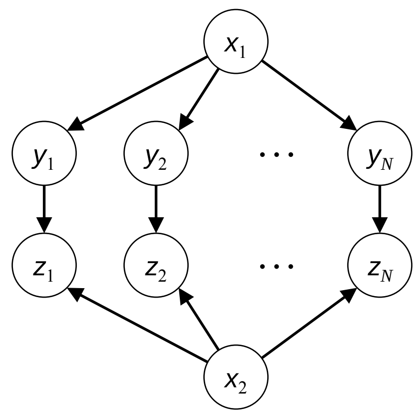
- This is a closed book exam, but two $8\frac{1}{2}'' \times 11''$ sheets of notes (4 sides total) are allowed.
- Calculators are **not** allowed.
- There are **3** problems of approximately equal value on the exam.
- The problems are not necessarily in order of difficulty. We recommend that you read through all the problems first, then do the problems in whatever order suits you best.
- Record all your solutions in the answer booklet provided. **NOTE: Only the answer booklet is to be handed in—no additional pages will be considered in the grading**. You may want to first work things through on the scratch paper provided and then neatly transfer to the answer sheet the work you would like us to look at. Let us know if you need additional scratch paper.
- A correct answer does not guarantee full credit, and a wrong answer does not guarantee loss of credit. You should clearly but concisely indicate your reasoning and **show all relevant work**. Your grade on each problem will be based on our best assessment of your level of understanding as reflected by what you have written in the answer booklet.
- Please be neat—we can't grade what we can't decipher!

Problem 1

Consider random variables $x_1, x_2, \mathbf{y} = (y_1, \dots, y_N), \mathbf{z} = (z_1, \dots, z_N)$ distributed according to

$$p_{x_1,x_2,\mathbf{y},\mathbf{z}}(x_1, x_2, \mathbf{y}, \mathbf{z}) = p_{x_1}(x_1)\, p_{x_2}(x_2) \prod_{n=1}^{N} \big[p_{y|x_1}(y_n|x_1)\, p_{z|y,x_2}(z_n|y_n, x_2)\big],$$

where $x_1, y_1, \dots, y_N, z_1, \dots, z_N$ take on values in $\{1, 2, \dots, K\}$, and x_2 instead takes on a value in $\{1, 2, \dots, N\}$. A minimal directed I-map for the distribution is as follows:



Assume throughout this problem that the complexity of table lookups for p_{x_1}, p_{x_2}, $p_{y|x_1}$, and $p_{z|y,x_2}$ are $O(1)$.

(a) Draw the moral graph over random variables $x_1, x_2, y_1, \dots, y_N$ conditioned on $\mathbf{z}$.

(b) Provide a good elimination ordering. For your elimination ordering:

i) Draw the reconstituted graph over random variables $x_1, x_2, y_1, \dots, y_N$ conditioned on $\mathbf{z}$.

ii) Determine α and β such that complexity of computing $p_{x_1|\mathbf{z}}$ using the associated Elimination Algorithm is $O(N^\alpha K^\beta)$.

For parts (c) and (d), suppose that we also have the following *context-dependent* conditional independencies: $y_i \perp\!\!\!\perp z_i | x_2 = c$ for all $c \neq i$.

(c) For fixed $\mathbf{z}$, x_1, and c, show that

$$p_{\mathbf{z}|x_1,x_2}(\mathbf{z}|x_1, c) = \eta(x_1, c, z_c)\, \lambda(c, \mathbf{z})$$

for some function $\eta(x_1, c, z_c)$ that can be computed in $O(K)$ operations, and some function $\lambda(c, \mathbf{z})$ that can be computed in $O(N)$ operations. Express $\eta(x_1, c, z_c)$ in terms of $p_{y|x_1}$ and $p_{z|y,x_2}$, and $\lambda(c, \mathbf{z})$ in terms of $p_{z|x_2}$.

(d) Provide an expression for $p_{x_1,\mathbf{z}}(x_1, \mathbf{z})$ in terms of p_{x_1}, p_{x_2}, η, and λ. Use your expression to explain how $p_{x_1|\mathbf{z}}(\cdot|\mathbf{z})$ can be computed in $O(N^2 + NK^2)$ operations for a fixed $\mathbf{z}$.

Problem 2

Consider a stochastic process that transitions among a finite set of states $s_1, \ldots, s_k$ over time steps $i = 1, \ldots, N$. The random variables $\mathsf{x}_1, \ldots, \mathsf{x}_N$ representing the state of the system at each time step are generated as follows:

- Sample the initial state s from an initial distribution p_1, and set $i := 1$.
- Repeat the following:
  - Sample a duration d from a duration distribution p_{D} over the integers $\{1, \ldots, M\}$, where M is the maximum duration.
  - Remain in the current state s for the next d time steps, i.e., set

$$\mathsf{x}_i := \mathsf{x}_{i+1} := \ldots := \mathsf{x}_{i+\mathsf{d}-1} := s.$$

  - Sample a successor state s' from a transition distribution $p_{\mathrm{T}}(\cdot|s)$ over the other states $s' \neq s$ (so there are no self-transitions).
  - Assign $i := i + d$ and $s := s'$.

This process continues indefinitely, but we only observe the first N time steps. You need not worry about the end of the sequence to do any of the problems.

As an example calculation with this model, the probability of the sample state sequence $s_1, s_1, s_1, s_2, s_3, s_3$ is

$$p_1(s_1)\, p_{\mathrm{D}}(3)\, p_{\mathrm{T}}(s_2|s_1)\, p_{\mathrm{D}}(1)\, p_{\mathrm{T}}(s_3|s_2) \sum_{2 \leq d \leq M} p_{\mathrm{D}}(d).$$

Finally, we do not directly observe the x_i's, but instead observe emissions y_i at each time step sampled from a distribution $p_{\mathsf{y}_i|\mathsf{x}_i}(y_i|x_i)$.

(a) For this part only, suppose $M = 2$, p_D is uniform over $\{1, 2\}$, and each x_i takes on a value in alphabet $\{a, b\}$. Draw a minimal directed I-map for the first five time steps using the variables $(\mathsf{x}_1, \ldots, \mathsf{x}_5, \mathsf{y}_1, \ldots, \mathsf{y}_5)$. Explain why none of the edges can be removed.

(b) This process can be converted to an HMM using an *augmented state representation.* In particular, the states of this HMM will correspond to pairs (x, t), where x is a state in the original system, and t represents the time elapsed in that state. For instance, the state sequence $s_1, s_1, s_1, s_2, s_3, s_3$ would be represented as $(s_1, 1), (s_1, 2), (s_1, 3), (s_2, 1), (s_3, 1), (s_3, 2)$. The transition and emission distributions for the HMM take the forms

$$\tilde{p}_{\mathsf{x}_{i+1},\mathsf{t}_{i+1}|\mathsf{x}_i,\mathsf{t}_i}(x_{i+1}, t_{i+1}|x_i, t_i) = \begin{cases} \nu(x_i, x_{i+1}, t_i) & \text{if } t_{i+1} = 1 \text{ and } x_{i+1} \neq x_i \\ \xi(x_i, t_i) & \text{if } t_{i+1} = t_i + 1 \text{ and } x_{i+1} = x_i \\ 0 & \text{otherwise} \end{cases}$$

and $\tilde{p}_{\mathsf{y}_i|\mathsf{x}_i,\mathsf{t}_i}(y_i|x_i, t_i)$, respectively. Express $\nu(x_i, x_{i+1}, t_i)$, $\xi(x_i, t_i)$, and $\tilde{p}_{\mathsf{y}_i|\mathsf{x}_i,\mathsf{t}_i}(y_i|x_i, t_i)$ in terms of the parameters p_1, p_{D}, p_{T}, $p_{\mathsf{y}_i|\mathsf{x}_i}$, k, N, and M of the original model.

(c) We wish to compute the marginal probability for the final state x_N given the observations $\mathsf{y}_1, \ldots, \mathsf{y}_N$. If we naively apply the forward-backward algorithm to the construction in part (b), the computational complexity is $O(Nk^2M^2)$. Show that by exploiting additional structure in the model, it is possible to reduce the complexity to $O(N(k^2 + kM))$. In particular, give the corresponding rules for computing the forward messages

$$\alpha(x_i, t_i) = p_{\mathsf{y}_1,\ldots,\mathsf{y}_i,\mathsf{x}_i,\mathsf{t}_i}(y_1, \ldots, y_i, x_i, t_i).$$

Equivalently, if you wish, you may give rules for computing the belief propagation messages $m_{i-1\to i}(x_i, t_i)$ instead of α.

Be sure to justify your answer. You need not worry about the beginning or end of the sequence, i.e., you may restrict attention to $2 \leq i \leq N-1$.

Note: If you cannot fully solve this part of the problem, you can receive substantial partial credit by constructing an algorithm with complexity $O(Nk^2M)$.

Hint: Substitute your solution from part (b) into the standard HMM messages and simplify as much as possible.

Problem 3

The graph $\mathcal{G}$ is a perfect undirected map for some strictly positive distribution p_{x} over a set of random variables $\mathsf{x} = (\mathsf{x}_1, \dots, \mathsf{x}_N)$, each of which takes values in a discrete set $\mathcal{X}$. Choose some variable x_i and let x_A denote the rest of the variables in the model, i.e., $\{\mathsf{x}_i, \mathsf{x}_A\} = \{\mathsf{x}_1, \dots, \mathsf{x}_N\}$.

(a) Construct the graph $\mathcal{G}'$ from $\mathcal{G}$ by connecting all the neighbors of x_i, and then removing x_i and all its edges. Show that $\mathcal{G}'$ is a perfect map for the marginal distribution p_{x_A}.

(b) Construct the graph $\mathcal{G}''$ from $\mathcal{G}$ by removing the node x_i and all its edges. Let some value $c \in \mathcal{X}$ be given. Show that $\mathcal{G}''$ is not necessarily a perfect map for the conditional distribution $p_{\mathsf{x}_A|\mathsf{x}_i}(\cdot|c)$ by giving a counterexample.

(c) Would the graph $\mathcal{G}''$ from part (b) be a perfect map for $p_{\mathsf{x}_A|\mathsf{x}_i}(\cdot|c)$ if the variables x were jointly Gaussian rather than discrete? If so, provide a proof; if not, give a counterexample.

6.438 Algorithms for Inference
Fall 2014