Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2018

Problem Set 1

Issued: Wednesday, Sept. 12, 2018 **Due:** Tuesday, Sept. 18, 2018

Problem 1.1

An important concept that arises frequently in statistical signal processing, control, and machine learning is that of conditional independence. Specifically x and y might represent random variables or vectors of interest — e.g., they might be consecutive samples of a random sequence or one might be observed and the other the quantity that we wish to estimate based on that observation. Typically, there will be some statistical dependency between x and y, and in many cases that statistical dependency is captured through the intermediary of another random variable or vector, z. Formally, x and y are conditionally independent given z if

$$p_{\mathsf{x},\mathsf{y}|\mathsf{z}}(x, y|z) = p_{\mathsf{x}|\mathsf{z}}(x|z)p_{\mathsf{y}|\mathsf{z}}(y|z).$$

We denote this by $\mathsf{x} \perp\!\!\!\perp \mathsf{y}|\mathsf{z}$.

(a) Suppose that z, w_1, and w_2 are mutually independent random variables. Construct two other random variables x and y as a function of z, w_1, and w_2 such that $\mathsf{x} \not\perp\!\!\!\perp \mathsf{y}$, but $\mathsf{x} \perp\!\!\!\perp \mathsf{y}|\mathsf{z}$.

(b) Show that $\mathsf{x} \perp\!\!\!\perp \mathsf{y}|\mathsf{z}$ if and only if the joint distribution for the three variables factors in the following form:

$$p_{\mathsf{x},\mathsf{y},\mathsf{z}}(x, y, z) = h(x, z)g(y, z) \ .$$

Problem 1.2

We consider the problem of generating a sample from a given probability distribution and demonstrate the power of graphical models. We first consider a naive approach to sampling : Consider a finite set $\mathcal{S}$ and a distribution $\mu(\cdot)$ over $\mathcal{S}$. We want to generate a sample from $\mu(\cdot)$ using a uniform random variable $U \sim \text{unif}([0, 1])$.

(a) Describe an algorithm which takes as input a uniform random variable U and the probability mass function $\mu(\cdot)$ and outputs a sample from the distribution after $O(|\mathcal{S}|)$ computations.

Hint: Let $\mathcal{S} = \{0, 1\}$. Note that $\mathbb{P}(U \leq \mu(0)) = \mu(0)$ and $\mathbb{P}(U > \mu(0)) = \mu(1)$. Generalize this to arbitrary finite sets.

Let $\mathcal{X}$ be a finite state space. Consider probability mass function $p(\cdot)$ over $\mathcal{X}^n$. We consider the problem of sampling a random element $(X_1, \ldots, X_n) \in \mathcal{X}^n$ such that:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p(x_1, \ldots, x_n)$$

for all $(x_1, \ldots, x_n) \in \mathcal{X}^n$. If we apply the algorithm in (a) with $\mathcal{S} = \mathcal{X}^n$ and $\mu(\cdot) = p(\cdot)$, it takes $O(|\mathcal{X}|^n)$ time to produce a sample. Therefore, naively applying the algorithm in (a) is highly inefficient. We now show that if we have access to arbitrary marginals of $p(\cdot)$, then we can efficiently generate a sample from $p(\cdot)$. That is, we assume that for any subset of indices $i_1, \ldots, i_r \in \{1, \ldots, n\}$ and set of values $x_{i_1}, \ldots, x_{i_r} \in \mathcal{X}$, we have access to the value $p_{i_1,\ldots,i_r}(x_{i_1}, \ldots, x_{i_r})$ through a black-box. We call this 'Black-box A'. (Here $p_{i_1,\ldots,i_r}$ is the marginal of $p(\cdot)$ over the indices $i_1, \ldots, i_r$).

(b) Describe an algorithm which can make queries to the Black-box A and independently call the algorithm in (a) n times to sequentially generate $X_1, X_2 \ldots X_n$ such that $(X_1, \ldots, X_n) \sim p(\cdot)$. The algorithm should make $O(n)$ calls to the black-box and take $O(n|\mathcal{X}|)$ time for computations.

The assumptions involving Black-box A are too powerful - finding the marginal of a large subset of nodes also involves the calculation of an intractable partition function in most interesting problems. But finding the marginal of a single node is tractable. This is where see the power of graphical models. Suppose $X_1, \ldots X_n$ are each represented by nodes of an undirected graphical model G. Assume that the distribution $p(\cdot) > 0$ as required by the Hammersley-Clifford Theorem.

(c) Suppose we condition on the event $X_1 = x_1$. Show that the conditional distribution on the remaining random variables is another undirected graphical model where all the edges involving node 1 are removed. Conclude by induction that conditioning on any arbitrary subset of nodes

gives another undirected graphical model over the remaining nodes. Find a set of clique potentials for the conditional distribution after conditioning on an arbitrary subset of nodes in terms of the original clique potentials of $p(\cdot)$.

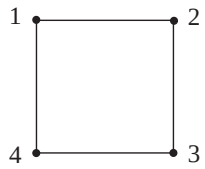
Hint: Use Hammersley-Clifford Theorem and use the definition of conditional distribution.

As we saw in (c), it is easy to describe conditional distribution of an undirected graphical model as another graphical model. We shall see later in the course that finding the marginal of a single node is easy for certain classes of undirected graphical models. Therefore, we construct a different kind of black-box: Black-box B. It takes as input the potentials of any undirected graphical model, and the address of a node (say node i) and returns the marginal distribution of the node i induced by the graphical model.

(d) Suppose the distribution $p(\cdot)$ is given as an undirected graphical model through its clique potentials. Describe an algorithm which uses the Black-box B instead of Black-box A in part (b) to sequentially generate $X_1, X_2, \ldots, X_n$ such that $(X_1, \ldots, X_n) \sim p(\cdot)$. The algorithm should make $O(n)$ calls to the black-box B and take $O(n|\mathcal{X}|)$ time for computations.

Problem 1.3

We'll show by example that the distribution of a graphical model need not have a factorization of the form in the Hammersley–Clifford Theorem if the distribution is not strictly positive. In particular, we'll take a look at a distribution on the following simple 4-node cycle:



where at each node we have a binary random variable, x_i, $i = 1, 2, 3, 4$. Consider a distribution $p(x_1, x_2, x_3, x_4)$ that assigns a probability of 1/8 to

each of the following configurations (ordered set) of values for (x_1, x_2, x_3, x_4):

$$\begin{array}{cccc} (0,0,0,0) & (1,0,0,0) & (1,1,0,0) & (1,1,1,0) \\ (0,0,0,1) & (0,0,1,1) & (0,1,1,1) & (1,1,1,1) \end{array}$$

and a value of 0 to all other configurations of (x_1, x_2, x_3, x_4).

(a) We first show that this distribution has the Markov structure expresed by our graph, which can be proved by showing the following two conditions:

- The pair of variables x_1 and x_3 are conditionally independent given (x_2, x_4)
- The pair of variables x_2 and x_4 are conditionally independent given (x_1, x_3)

(i) Show that if we interchange x_1 and x_4 and interchange x_2 and x_3 we obtain the same distribution, i.e., $p(x_1, x_2, x_3, x_4) = p(x_4, x_3, x_2, x_1)$. This implies that if we can show the first of the conditions listed above, then the other is also true.

(ii) Show that conditioned on any pair of values for (x_2, x_4), the value of either x_1 or x_3 is known with certainty, trivially proving conditional independence.

(b) Next we prove by contradiction that the distribution can't be factored in the way stated in the Hammersley–Clifford Theorem. Note that the maximal cliques in our graph are simply the edges. By absorbing the single node potential functions and the proportionality constant $1/Z$ into the maximal clique terms, if our distribution had the factorization implied by Hammersley–Clifford, it could be written in the following form:

$$p(x_1, x_2, x_3, x_4) = \phi_{12}(x_1, x_2)\phi_{23}(x_2, x_3)\phi_{34}(x_3, x_4)\phi_{41}(x_4, x_1).$$

Show that assuming that our distribution has such a factorization leads to a contradiction by examining the values of $p(0, 0, 0, 0)$, $p(0, 0, 1, 0)$, and $p(0, 0, 1, 1)$, and $p(1, 1, 1, 0)$.

Problem 1.4

Let there be K different coins, each with different biases, $c_1, \ldots, c_K \in [0,1]$. The k^{th} coin comes up heads with probability c_k, and tails with probability $1 - c_k$. Let $\mathsf{t} \in \{1, 2, \ldots K\}$ be a random variable having the mass function $p_{\mathsf{t}}(\cdot)$. Then define the random variable $\mathsf{x} \in \{0, 1\}$ to be 1 if the t^{th} coin comes up heads, and 0 if the t^{th} coin comes up tails. In other words, to generate a sample of x, we first sample t, then toss coin number t, and set x equal to the indicator function of coin t coming up heads.

(a) Write down a graphical model description of the above described mixture distribution involving variables x and t. Provide the diagram of the graphical model in addition to the potential functions.

In addition, we are given K (known) N-dimensional vectors $\Theta^1, \ldots \Theta^K \in \mathbb{R}^N$. If $\mathsf{t} = k$, we generate random variables $\mathbf{y} = (\mathsf{y}_1, \ldots, \mathsf{y}_N) \in \{0, 1\}^N$ according to

$$p_{\mathbf{y}}(\mathbf{y}) \propto \exp\big(\sum_{i=1}^{N} \theta_i^k y_i\big), \tag{1}$$

for $\mathbf{y} = (y_1, \ldots, y_N) \in \{0, 1\}^N$, where we used the notation $\Theta^k = (\theta_1^k, \ldots, \theta_N^k)$.

(b) Write down a graphical model description of the above described mixture distribution involving varables $\mathsf{x}, \mathsf{t}, \mathsf{y}_1, \ldots \mathsf{y}_N$. Provide the diagram of the graphical model in addition to the potential functions.

Problem 1.5 (practice)

Random variables x_1, x_2, and x_3 represent the outcomes of three (independent) fair coin tosses, $\mathsf{x}_4 = \mathbb{1}_{\mathsf{x}_1 = \mathsf{x}_2}$, and $\mathsf{x}_5 = \mathbb{1}_{\mathsf{x}_2 = \mathsf{x}_3}$, where $\mathbb{1}_A = 1$ if A is true, and 0 otherwise.

(a) Specify a directed graphical model (give the directed acyclic graph and local conditionals) that describes the joint probability distribution.

(b) List any conditional independencies that are displayed by this probability distribution but are not implied by the graph.

(c) Specify an *undirected* graphical model (give the graph and clique potentials) that describes the joint probability distribution.

(d) List any conditional independencies that are displayed by this probability distribution but are not implied by the graph.

(e) If the coins were biased, would your answer to part (b) or (d) change?