HW1 (6 problems)

Jamie Morgenstern *CSE515*

1. *[2 points]* In this exercise, you will construct an undirected graphical model for the problem of segmenting foreground and background in an image. Load the image `flower.bmp`. Partial labeling of the foreground and background pixels are given in the mask images `foreground.bmp` and `background.bmp`, respectively. In each mask, the white pixels indicate positions of representative samples of foreground or background pixels in the image.

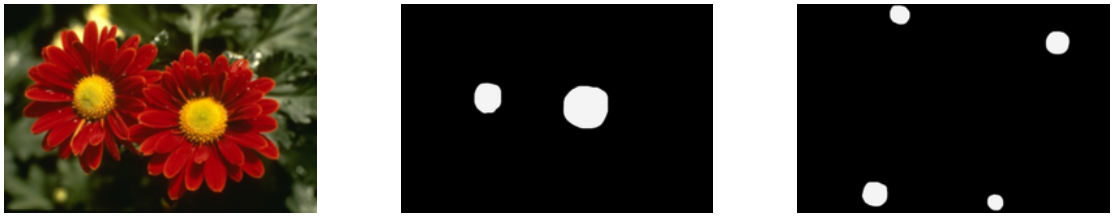


image y in `flower.bmp` `foreground.bmp` `background.bmp`

Figure 1: Given observations for this problem.

Let $y = \{y_i\}$ be an observed color image, so each y_i is a 3-vector (of RGB values between 0 and 1) representing the pixel indexed by i. Let $x = \{x_i\}$, where the unknown variable $x_i \in \{0, 1\}$ is a foreground(1)/background(0) labeling of the image at pixel i. Let us say the probabilistic model for x and y given by their joint distribution can be factored in the form

$$\mu(x, y) \quad = \quad \frac{1}{Z} \prod_i \phi(x_i, y_i) \prod_{(j,k) \in E} \psi(x_j, x_k) \tag{1}$$

where E is the set of all pairs of adjacent pixels in the same row or column as in 2-dimensional grid. Suppose that we choose

$$\psi(x_j, x_k) = \begin{cases} 0.9 & \text{if } x_j = x_k \\ 0.1 & \text{if } x_j \neq x_k \end{cases}$$

This encourages neighboring pixels to have the same label–a reasonable assumption. Suppose further that we use a simple model for the conditional distribution $\phi(x_i, y_i) = \mathbb{P}_{Y_i|X_i}(y_i|x_i)$:

$$\mathbb{P}(y_i|x_i = \alpha) \quad \propto \quad \frac{1}{(2\pi)^{3/2}\sqrt{\det \Lambda_\alpha}} \exp\Big\{ -\frac{1}{2}(y_i - \mu_\alpha)^T \Lambda_\alpha^{-1} (y_i - \mu_\alpha) \Big\} + \epsilon$$

for $y_i \in [0, 1]^3$. That is, the distribution of color pixel values over the same type of image region is a modified Gaussian distribution, where ϵ accounts for outliers. Set $\epsilon = 0.01$ in this problem.

(*a*) Sketch an undirected graphical model that represents $\mu(x, y)$ for a smaller example of an image with 5×5 pixels.

(*b*) Compute $\mu_\alpha \in \mathbb{R}^3$ and $\Lambda_\alpha \in \mathbb{R}^{3\times 3}$ for each $\alpha \in \{0, 1\}$ from the labeled masks by finding the sample mean and covariance of the RGB values of those pixels for which the label $x_i = \alpha$ is known from `foreground.bmp` and `background.bmp`. The sample mean of samples $\{y_1, \ldots, y_N\}$ is $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$ and the sample covariance is $C_y = \frac{1}{N-1}\sum_{i=1}^N (y_i - \bar{y})(y_i - \bar{y})^T$. Write the means μ_0 and μ_1 and covariances Λ_0 and Λ_1. You don't have to submit the code you used to compute these values, but save the code as we will use the model in future homework.

2. *[2 points]* **(Exercise 2.5 in Koller/Friedman)**

Let X, Y, Z be three disjoint subsets of random variables. We say X and Y are conditionally independent given Z if and only if

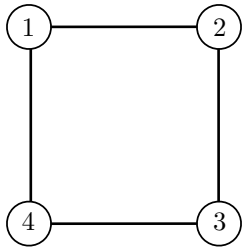
$$\mathbb{P}_{X,Y|Z}(x, y|z) \quad = \quad \mathbb{P}_{X|Z}(x|z)\, \mathbb{P}_{Y|Z}(y|z)\ .$$

Show that X and Y are conditionally independent given Z if and only if the joint distribution for the three subsets of random variables factors in the following form:

$$\mathbb{P}_{X,Y,Z}(x, y, z) \quad = \quad h(x, z)\, g(y, z)\ .$$

3. *[2 points]* **(Exercise 4.1 in Koller/Friedman)**

In this problem, we will 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 will consider a distribution on the following simple 4-cycle where each node is a binary



random variable, X_i, for $i \in \{1, 2, 3, 4\}$. Consider a probability distribution that assigns a probability $1/8$ uniformly to each of the following set of values (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 assigns zero to all other configurations of (X_1, X_2, X_3, X_4).

(a) We first need to show that this distribution is Markov on our graph. To do this, it should not be difficult to see that what we need to show are the following 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).

First, show that if we interchange X_1 and X_4 and interchange X_2 and X_3, we obtain the same distribution, i.e.., $\mathbb{P}(x_1, x_2, x_3, x_4) = \mathbb{P}(x_4, x_3, x_2, x_1)$. This implies that if we can show the first condition, then the other is also true.

(b) Show that whatever pair of values you choose for (X_2, X_4), we then know either X_1 or X_3 with certainty. For example, $(X_2 = 0, X_4 = 0)$ implies that $X_3 = 0$. Since we know either X_1 or X_3 with certainty, then conditioning on the other one of these obviously provides no additional information, trivially proving conditional independence.

(*c*) What we now need to show is that the distribution cannot be factorized in the way stated in the Hammersley-Clifford Theorem. We will do this by contradiction. Noting that the maximal cliques in our graph are just the edges and absorbing the normalization $1/Z$ into any of the pairwise compatibility functions, we know that if our distribution has the factorization implied by the Hammersley-Clifford Theorem, we can write it in the following form:

$$\mathbb{P}(x_1, x_2, x_3, x_4) \quad = \quad \psi_{12}(x_1, x_2)\, \psi_{23}(x_2, x_3)\, \psi_{34}(x_3, x_4)\, \psi_{41}(x_4, x_1)\ .$$

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

4. *[2 points]* Given a graph $G = (V, E)$, an *independent set* of G is a subset $S \subseteq V$ of the vertices, such that no two vertices in S is connected by an edge in E. Precisely, if $i, j \in S$ then $(i, j) \notin E$. We let $\mathrm{IS}(G)$ denote the set of all independent sets of G, and let $Z(G) = |\mathrm{IS}(G)|$ denote its size, i.e. the total number of independent sets in G. The number of independent sets $Z(G)$ is at least $1 + |V|$, since the empty set and all subsets with single vertex are always independent sets. We are interested in the uniform probability measure over S:

$$\mathbb{P}_{\mathrm{IS}(G)}(S) = \frac{1}{Z(G)} \mathbb{I}(S \in \mathrm{IS}(G))\ ,$$

where $\mathbb{I}(A)$ is an indicator function which is one if event A is true and zero if false.

The set S can be naturally encoded by a binary vector $x \in \{0,1\}^{|V|}$ by letting $x_i = 1$ if and only if $i \in S$. Denote by $\mathbb{P}_G(x)$ the probability distribution induced on this vector x according to $\mathbb{P}_{\mathrm{IS}(G)}(S)$. This $\mathbb{P}_G(x)$ is a **pairwise Markov random field** on G as it can be written as

$$\mathbb{P}_G(x) \quad = \quad \frac{1}{Z(G)} \prod_{(i,j) \in E} \underbrace{\mathbb{I}(x_i \times x_j = 0)}_{\psi_{i,j}(x_i, x_j)}\ , \tag{2}$$

where $Z(G) = \sum_{x \in \{0,1\}^{|V|}} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j)$.

Let $G = T_{k,\ell}$ denote the rooted tree with branching factor k and ℓ generations, that is the root has k descendants and each other node has one ancestor and k descendants except for the leaves. The total number of vertices is $(k^{\ell+1} - 1)/(k - 1)$, and $T_{k,\ell=0}$ is the graph consisting only of the root. We let ϕ denote the root of $T_{k,\ell}$.

(*a*) Let $Z_\ell = Z(T_{k,\ell})$ denote the total number of independent sets of $G = T_{k,\ell}$. Let $Z_\ell(0)$ be the number of independent sets in $T_{k,\ell}$ such that the root is $x_\phi = 0$, and $Z_\ell(1)$ be the number of independent sets such that $x_\phi = 1$. It is immediate that $Z_0(0) = Z_0(1) = 1$. Derive a recursion expressing $(Z_{\ell+1}(0), Z_{\ell+1}(1))$ as a function of $(Z_\ell(0), Z_\ell(1))$.

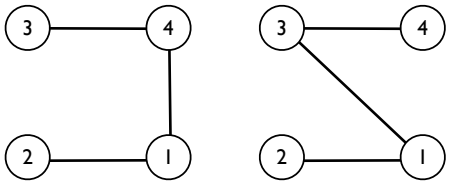
(*b*) Using the above recursion, derive a recursion for the probability that the root belongs to a uniformly random independent set. Explicitly, derive a recursion for

$$p_\ell \quad = \quad \mathbb{P}_{T_{k,\ell}}(\{x_\phi = 1\})\ .$$

(*c*) Program this recursion and plot p_ℓ as a function of $\ell \in \{0, 1, \ldots, 50\}$ for four values of k, e.g. $k \in \{1, 2, 3, 10\}$. Comment on the qualitative behavior of these plots. Print the code you used in your solution pdf file as a part your answer.

(*d*) (optional) Prove that, for $k \leq 3$, the recursion converges to a unique value using Banach's fixed point theorem.

5. *[2 points]* **(I-map)**
In this problem, we will show that when the distribution $\mu(x)$ is not strictly positive (i.e. $\mu(x) = 0$ for some x), then the I-map for this distribution is not unique. Consider a distribution of 4 binary random variables x_1, x_2, x_3, and x_4 such that $\mu(x_1 = x_2 = x_3 = x_4 = 1) = 0.5$ and $\mu(x_1 = x_2 = x_3 = x_4 = 0) = 0.5$. The following two undirected graphical models are both minimal I-maps for this distribution, hence it is not unique.



(*a*) Prove that the two undirected graphical models above are minimal I-maps for the distribution $\mu(x)$. You need to show that both graphs are I-maps for the given distribution $\mu(x)$ and that removing any edge results in introducing independencies that are not implied by the distribution $\mu(x)$.

(*b*) Now, we show that starting with a complete graph and eliminating edges that are pairwise conditionally independent does not always give you an I-map (minimal or not). Start with a complete graph K_4 (K_4 is an undirected graph with 4 nodes and edges between all pairs of nodes). For each pair of nodes, eliminate the edge between this pair if they are conditionally independent given the rest of the nodes in the graph. Continue this procedure for all pairs of nodes and examine the resulting graph. Is this an I-map of the distribution $\mu(x_1, x_2, x_3, x_4)$?

Recall from class, that a distribution over x is (globally) Markov with respect to $G = (V, E)$ if, for any disjoint subsets of nodes A, B, C such that B separates A from C, x_A–x_B–x_C is satisfied. Recall two other notions of Markovity. A distribution is pairwise Markov with respect to G if, for any two nodes i and j not directly linked by an edge in G, the corresponding variables x_i and x_j are independent conditioned on all of the remaining variables, i.e. for all $(i, j) \notin E$,

$$x_i – x_{V \setminus \{i,j\}} – x_j$$

A distribution is locally Markov with respect to G if any node i, when conditioned on the variables on the neighbors of i, is independent of the remaining variables, i.e. for all $i \in V$,

$$x_i – x_{\partial i} – x_{V \setminus \{i, \partial i\}}$$

(*c*) Using the example of distribution on 4 random variables as a counter example, prove that a distribution is pairwise Markov w.r.t. G does not always imply that it is locally Markov w.r.t. the same graph G. (However, if the distribution is positive, pairwise Markovity implies local and global Markovity.)

6. *[2 points]* 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 $X_1, \ldots, X_N$ representing the state of the system at each time step are generated as follows:

   - Sample the initial state $X_1 = s$ 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
     $$X_i := X_{i+1} := \ldots := X_{i+d-1} := s$$
     * Sample a successor state s' from a transition distribution $p_T(\cdot|s)$ 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

   $$\mathbb{P}((X_1, X_2, X_3, X_4, X_5, X_6) = (s_1, s_1, s_1, s_2, s_3, s_3)) \;=\; p_1(s_1)p_D(3)p_T(s_2|s_1)p_D(1)p_T(s_3|s_2) \sum_{2 \geq d \leq M} p_D(d) \; .$$

   Finally, we do not directly observe the X_i's, but instead observe emissions y_i at each step sampled from a distribution $p_{Y_i|X_i}(y_i|x_i)$.

   (*a*) For this part only, suppose $M = 2$, and $p_D(d) = \begin{cases} 0.6 & \text{for } d = 1 \\ 0.4 & \text{for } d = 2 \end{cases}$, and each X_i takes on a value from an alphabet $\{a, b\}$. Draw a minimal directed I-map for the first five time steps using the variables $(X_1, \ldots, X_5, Y_1, \ldots, Y_5)$. Explain why the edges you added cannot be removed.
   [Note: you do not need to solve part (*a*) in order to solve part (*b*) and (*c*).]

   (*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 distribution for the HMM take the forms
   $$\tilde{p}_{X_{i+1}, T_{i+1}|X_i, T_i}(x_{i+1}, t_{i+1}|x_i, t_i) = \begin{cases} \phi(x_i, x_{i+1}, t_i) & \text{if } t_{i+1} = 1 \text{ and } x_{i+1} \neq x_i \\ \xi(x_i, t_i) & \text{if } t_{i+1} = t_i + 1 \text{ and } x_{i+1} = x_i \\ 0 & \text{otherwise} \end{cases}$$
   and $\tilde{p}_{Y_i|X_i,T_i}(y_i|x_i, t_i)$, respectively. Express $\phi(x_i, x_{i+1}, t_i)$, $\xi(x_i, t_i)$, and $\tilde{p}_{Y_i|X_i,T_i}(y_i|x_i, t_i)$ in terms of parameters p_1, p_D, p_T, $p_{Y_i|X_i}$, 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 $Y_1, \ldots, Y_N$. If we naively apply the **sum-product 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 $\nu_{i+1 \to i+2}(x_{i+1}, t_{i+1})$ from the previous message $\nu_{i \to i+1}(x_i, t_i)$. Do not worry about the beginning or the end of the sequence and restrict your attention to $2 \leq i \leq N - 1$.
   [Hint: substitute your solution from part (*b*) into the standard update rule for HMM messages and simplify as much as possible.]
   [Note: If you cannot fully solve this part of the problem, you can receive substantial partial credit by constructing an algorithm with complexity $O(Nk^2M)$.]