TTIC 31180 — Probabilistic Graphical Models

Problem Set #1 Solutions

Due Date: April 30, 2020

1 Probability Review [10pts]

(a) **[2pts]** Suppose that you have a distribution over four random variables $P(A, B, C, D)$ and that we know that $(A \perp B \,|\, C)$. Show that the following statement

$$P(A, B \,|\, C) = P(A \,|\, C)P(B \,|\, C)$$

is equivalent to the following statements (you need to show both directions)

$$P(A \,|\, B, C) = P(A \,|\, C)$$

and

$$P(B \,|\, A, C) = P(B \,|\, C)$$

Solution:

(i) First, let's show that $P(A, B \,|\, C) = P(A \,|\, C)P(B \,|\, C) \Rightarrow P(A \,|\, B, C) = P(A \,|\, C)$:

$$\begin{aligned} P(A \,|\, B, C) &= \frac{P(A, B, C)}{P(B, C)} \\ &= \frac{P(A, B \,|\, C)P(C)}{P(B \,|\, C)P(C)} \\ &= \frac{P(A \,|\, C)P(B \,|\, C)}{P(B \,|\, C)} \\ &= \underline{P(A \,|\, C)} \end{aligned}$$

(ii) Similarly, it follows that $P(A, B \,|\, C) = P(A \,|\, C)P(B \,|\, C) \Rightarrow P(B \,|\, A, C) = P(B \,|\, C)$:

$$\begin{aligned} P(B \,|\, A, C) &= \frac{P(A, B, C)}{P(A, C)} \\ &= \frac{P(A, B \,|\, C)P(C)}{P(A \,|\, C)P(C)} \\ &= \frac{P(A \,|\, C)P(B \,|\, C)}{P(A \,|\, C)} \\ &= \underline{P(B \,|\, C)} \end{aligned}$$

(iii) Now, let's show the other direction, i.e., $P(A \mid B, C) = P(A \mid C) \Rightarrow P(A, B \mid C) = P(A \mid C)P(B \mid C)$, which follows simply as:

$$\begin{aligned} P(A, B \mid C) &= P(A \mid B, C)P(B \mid C) \\ &= \underline{P(A \mid C)P(B \mid C)} \end{aligned}$$

(iv) And similarly, $P(B \mid A, C) = P(B \mid C) \Rightarrow P(A, B \mid C) = P(A \mid C)P(B \mid C)$, which follows simply as:

$$\begin{aligned} P(A, B \mid C) &= P(B \mid A, C)P(A \mid C) \\ &= \underline{P(B \mid C)P(A \mid C)} \end{aligned}$$

■

(b) **[2pts]** Prove the **weak union** property of conditional independence:

$$(X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)$$

Solution:

First, the fact that $(X \perp Y, W \mid Z)$ implies that $P(X \mid Y, W, Z) = P(X \mid Z)$. However, by the decomposition property, i.e., $(X \perp Y, W \mid Z) \Rightarrow (X \perp W \mid Z)$, we have that $P(X \mid W, Z) = P(X \mid Z)$. Thus,

$$\begin{aligned} P(X \mid Y, W, Z) &= P(X \mid Z) \\ &= P(X \mid W, Z), \end{aligned}$$

which implies that $(X \perp Y \mid W, Z)$. ■

(c) **[2pts]** Prove the **contraction** property of conditional independence:

$$(X \perp W \mid Z, Y) \ \&\ (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)$$

Solution:

Consider $P(X \mid W, Y, Z)$:

$$\begin{aligned} P(X \mid W, Y, Z) &= P(X \mid Y, Z) \\ &= P(X \mid Z), \end{aligned}$$

where the first line follows from $(X \perp W \mid Z, Y)$ and the second follows from the statement $(X \perp Y \mid Z)$. Thus, we can conclude that $(X \perp Y, W \mid Z)$. ■

(d) **[2pts]** It is often useful to consider the effect of some specific propositions in the context of some background evidence that remains fixed. The following questions ask you to prove more general versions of Bayes' rule with respect to some background evidence E.

(i) Prove the following conditional version of the general product rule:

$$P(A, B \mid E) = P(A \mid B, E)P(B \mid E)$$

(ii) Prove the conditional version of Bayes' rule:

$$P(A \mid B, E) = \frac{P(B \mid A, E)P(A \mid E)}{P(B \mid E)}$$

Solution:

(i) Per Bayes' rule,

$$\begin{aligned} P(A, B \mid E) &= \frac{P(A, B, E)}{P(E)} \\ &= \frac{P(A \mid B, E)P(B \mid E)P(E)}{P(E)} \\ &= P(A \mid B, E)P(B \mid E) \end{aligned}$$

where the second line follows from the chain rule.

(ii) This equality follows similarly as

$$\begin{aligned} P(A \mid B, E) &= \frac{P(A, B, E)}{P(B, E)} \\ &= \frac{P(B \mid A, E)P(A \mid E)P(E)}{P(B \mid E)P(E)} \\ &= \frac{P(B \mid A, E)P(A \mid E)}{P(B \mid E)} \end{aligned}$$

Thus, yielding the equalities. ■

(e) **[2pts]** This question considers the way in which conditional independence relationships affect the amount of information that is necessary for different probability calculations.

(i) Suppose that we have a distribution over three random variables H, E_1, and E_2, and want to calculate $P(H \mid E_1, E_2)$ without any knowledge of conditional independencies. Which **one** of the following sets of numbers are sufficient to compute this likelihood.

- $P(E_1, E_2)$, $P(H)$, $P(E_1 \mid H)$, $P(E_2 \mid H)$
- $P(E_1, E_2)$, $P(H)$, $P(E_1, E_2 \mid H)$
- $P(E_1 \mid H)$, $P(E_2 \mid H)$, $P(H)$

Solution:

By Bayes' rule, we have

$$
\begin{aligned}
P(H \mid E_1, E_2) &= \frac{P(H, E_1, E_2)}{P(E_1, E_2)} \\
&= \frac{P(E_1, E_2 \mid H)P(H)}{P(E_1, E_2)},
\end{aligned}
$$

which we can not reduce further without additional knowledge of independencies. Therefore, we can compute the desired likelihood from $P(E_1, E_2)$, $P(H)$, and $P(E_1, E_2 \mid H)$. ■

(ii) Now, suppose that we know that $(E_1 \perp E_2 \mid H)$. Which of the above three sets are sufficient?

Solution:

Knowing that E_1 and E_2 are conditionally independent given H, we can then reduce the above expression as

$$
\begin{aligned}
P(H \mid E_1, E_2) &= \frac{P(E_1, E_2 \mid H)P(H)}{P(E_1, E_2)} \\
&= \frac{P(E_1 \mid H)P(E_2 \mid H)P(H)}{P(E_1, E_2)},
\end{aligned}
$$

and can therefore compute the desired likelihood from the following terms $P(E_1, E_2)$, $P(H)$, $P(E_1 \mid H)$, and $P(E_2 \mid H)$. ■

2 Medical Diagnosis [20pts]

Let us consider a simple medical diagnosis problem in which we are interested in the relationship between the flu and dehydration, various symptoms, and the weather. We will represent these problem in terms of the following binary random variables: Winter (W) (we assume that there are only two seasons: winter and summer), Flu (F), Dehydration (D), Muscle Ache (M), Headache (H), Nausea (N), and Lightheadedness (L).

Suppose that you consulted a doctor who explained the causal relationship between these variables upon which you constructed the Bayesian network structure in Figure 1 to model the joint distribution.

2.1 Bayesian Network Independencies [12pts]

An advantage of Bayesian network structures is that they make explicit certain conditional independencies among the random variables that are integral to performing inference (unlike undirected graphical models). Determine whether the graph in Figure 1 encodes the following independencies. Justify your answers.

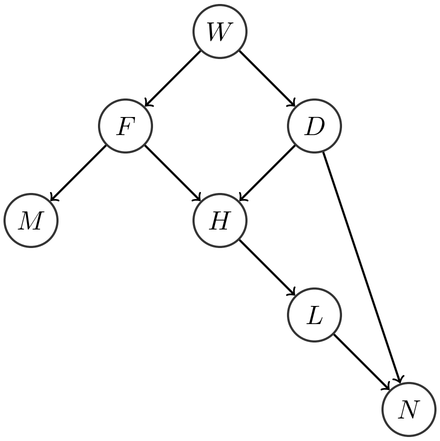


Figure 1: Bayesian network structure for medical diagnosis.

(a) $W \perp M$

Solution:

False. There is an active trail $W \rightarrow F \rightarrow M$ ■

(b) $W \perp M \mid F$

Solution:

True. There is no active trail between W and M when F is observed. ■

(c) $W \perp H \mid F$

Solution:

False. When F is observed, there is an active trail $W \rightarrow D \rightarrow H$. ■

(d) $W \perp H \mid F, D$

Solution:

True. When F and D are observed, there is no active trail between W and H, and thus, they are d-separated. ■

(e) $W \perp N \mid D$

Solution:

False. There is an active trail $W \to F \to H \to L \to N$. ■

(f) $W \perp N \,|\, D, H$

Solution:

True. There is no active trail between W and N when D and H are observed. ■

(g) $F \perp D$

Solution:

False. There is an active trail $F \leftarrow W \to D$. ■

(h) $F \perp D \,|\, W, H$

Solution:

False. There is an active trail as $F \to H \leftarrow D$ is a v-structure. ■

(i) $F \perp D \,|\, W$

Solution:

True. There is no active trail between F and D when only W is observed. ■

(j) $F \perp D \,|\, W, N$

Solution:

False. There is a v-structure as N is a descendent of H. ■

(k) $M \perp N$

Solution:

False. There is an active trail $M \leftarrow F \leftarrow W \to D \ \to N$. ■

(l) $M \perp N \,|\, H$

Solution:

False. There is a v-structure at H, and thus, an active trail $M \leftarrow F \rightarrow H \leftarrow D \rightarrow N$. ■

2.2 Factorized Joint Distribution [2pts]

Write the expression for the joint distribution that factorizes according to the Bayesian network structure in Figure 1.

Solution:

$$P(W, F, D, M, H, J, N) = P(W)P(F \mid W)P(D \mid W)P(M \mid F)P(H \mid F, D)P(L \mid H)P(N \mid L, D)$$

■

2.3 Inference [6pts]

Suppose that the conditional probability distributions for the Bayesian network are given by the following tables.

$P(W)$	$P(\neg W)$
0.5	0.5

	$P(F \mid W)$	$P(\neg F \mid W)$
W	0.4	0.6
$\neg W$	0.1	0.9

	$P(D \mid W)$	$P(\neg D \mid W)$
W	0.1	0.9
$\neg W$	0.3	0.7

	$P(M \mid F)$	$P(\neg M \mid F)$
F	0.8	0.2
$\neg F$	0.1	0.9

		$P(H \mid F, D)$	$P(\neg H \mid F, D)$
F	D	0.9	0.1
$\neg F$	D	0.8	0.2
F	$\neg D$	0.8	0.2
$\neg F$	$\neg D$	0.3	0.7

	$P(L \mid H)$	$P(\neg L \mid H)$
H	0.8	0.2
$\neg H$	0.2	0.8

		$P(N \mid L, D)$	$P(\neg N \mid L, D)$
L	D	0.9	0.1
$\neg L$	D	0.8	0.2
L	$\neg D$	0.6	0.4
$\neg L$	$\neg D$	0.2	0.8

Table 1: Conditional probability distributions for medical diagnosis example.

(a) **[2pts]** What is the (marginal) probability that you have the flu?

Solution:

$$
\begin{aligned}
P(F) &= \sum_{w\in\{0,1\}} P(F, W = w) \\
&= \sum_{s\in\{0,1\}} P(F \mid W = w)P(W = w) \\
&= 0.4 \cdot 0.5 + 0.1 \cdot 0.5 \\
&= \underline{0.25}
\end{aligned}
$$

■

(b) **[2pts]** Given that it is winter (W is TRUE) and you have a headache (H is TRUE), what is the probability that you have a flu?

Solution:

$$
\begin{aligned}
P(F \mid W, H) &= \frac{P(F, W, H)}{P(W, H)} \\
&= \frac{\sum_d P(F, W, H, D)}{\sum_{d,f} P(F, D, W, H)} \\
&= \frac{\sum_d P(H \mid F, D = d)P(F \mid W)P(D = d \mid W)P(W)}{\sum_{d,f} P(H \mid F = f, D = d)P(F = f \mid W)P(D = d \mid W)P(W)} \\
&= \frac{0.9 \cdot 0.4 \cdot 0.1 \cdot 0.5 + 0.8 \cdot 0.4 \cdot 0.9 \cdot 0.5}{0.9 \cdot 0.4 \cdot 0.1 \cdot 0.5 + 0.8 \cdot 0.6 \cdot 0.1 \cdot 0.5 + 0.8 \cdot 0.4 \cdot 0.9 \cdot 0.5 + 0.3 \cdot 0.6 \cdot 0.9 \cdot 0.5} \\
&= \underline{0.61}
\end{aligned}
$$

■

(c) **[2pts]** Given that it is winter (W is TRUE), you know that you are dehydrated ((D is TRUE)), and you have a headache (H is TRUE), what is the probability that you have a flu?

Solution:

$$
\begin{aligned}
P(F \mid W, D, H) &= \frac{P(F, W, D, H)}{P(W, D, H)} \\
&= \frac{P(F, W, D, H)}{\sum_f P(F = f, W, D, H)} \\
&= \frac{P(H \mid F, D) P(F \mid W) P(D \mid W) P(W)}{\sum_f P(H \mid F = f, D) P(F = f \mid W) P(D \mid W) P(W)} \\
&= \frac{0.9 \cdot 0.4 \cdot 0.1 \cdot 0.5}{0.9 \cdot 0.4 \cdot 0.1 \cdot 0.5 + 0.8 \cdot 0.6 \cdot 0.1 \cdot 0.5} \\
&= \underline{0.43}
\end{aligned}
$$

■

3 Restricted Boltzmann Machine [25pts]

Restricted Boltzmann machines (RBMs) are a type of Markov networks that have proven effective at modeling a number of problems in machine learning, including those related to collaborative filtering, dimensionality reduction, and feature learning, among others. As their name implies, restricted Boltzmann machines are a special case of Boltzmann machines (see Koller & Friedman p.p. 126–127) in which the Markov network forms a bipartite graph (Fig. 2) between binary random variables (neurons) in a hidden (latent) layer and neurons in a visible (observed) layer.

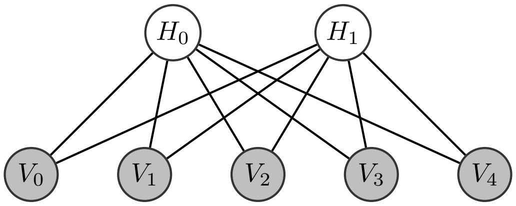


Figure 2: A restricted Boltzmann machine with two hidden neurons and three visible neurons.

The energy function of an RBM takes the form

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{ij} w_{ij} v_i h_j - \sum_i a_i v_i - \sum_j b_j h_j \tag{1}$$

where $\boldsymbol{h} = \{h_0, h_1, \ldots, h_n\}$, $\boldsymbol{v} = \{v_0, v_1, \ldots, v_m\}$, w_{ij} are weight parameters that relate the hidden and visible units, and a_i and b_j are offset (unary) parameters for the visible and hidden units, respectively.

The joint distribution is then defined in terms of the energy function as

$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h})} \tag{2}$$

where the partition function Z takes the usual form.

For a concrete example, consider a simple version of the Netflix Prize[1] in which the objective is

[1] See `https://www.netflixprize.com`.

to determine the genre of movies that a particular user likes. Let v_i for $i \in \{0, \ldots, 4\}$ be a binary rating (like vs. not like) that the user assigns to five movies (0: "Die Hard," 1: "Casablanca," 2: "Batman," 3: "Jaws," 4: "The Notebook") and let h_j for $j \in \{0, 1\}$ indicate the user's unknown preference for two different genres (0: action movies, 1: romance movies).

(a) **[8pts]** Restricted Boltzmann machines allow us to infer the user's hidden preference by observing their movie ratings. Show from Equation 2 that the conditional distribution over hidden units given visible units factorizes as

$$P(\boldsymbol{H} \,|\, \boldsymbol{V}) = \prod_{j=0}^{n} P(H_j \,|\, \boldsymbol{V}),$$

where $\boldsymbol{H} = \{H_0, H_1\}$ and $\boldsymbol{V} = \{V_0, V_1, V_2, V_3, V_4\}$ in this example, and

$$P(H_j = 1 \,|\, \boldsymbol{V} = \boldsymbol{v}) = \sigma\left(b_j + \sum_{i=0}^{m} w_{ij} v_i)\right) \tag{3}$$

with $\sigma(x) = \frac{e^x}{1+e^x}$ being the sigmoid function. In other words, the hidden variables H_j are conditionally independent given the set of random variables $\boldsymbol{V}$ representing the visible layer. Note that this is a simple neural network model that expresses the activation of neurons based on the states of individual units.

Solution:

$$\begin{aligned}
P(\boldsymbol{H} = \boldsymbol{h} \,|\, \boldsymbol{V} = \boldsymbol{v}) = \frac{P(\boldsymbol{h}, \boldsymbol{v})}{P(\boldsymbol{v})} &= \frac{P(\boldsymbol{h}, \boldsymbol{v})}{\sum\limits_{h_0} \cdots \sum\limits_{h_n} P(h, \boldsymbol{v})} \\
&= \frac{\exp\left(\sum_i a_i v_i\right) \exp\left(\sum_{ij} w_{ij} v_i h_j + \sum_j b_j h_j\right)}{\exp\left(\sum_i a_i v_i\right) \sum_{h_0} \cdots \sum_{h_n} \exp\left(\sum_{ij} w_{ij} v_i h_j + \sum_j b_j h_j\right)} \\
&= \frac{\prod_j \exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right)}{\sum_{h_0} \cdots \sum_{h_n} \prod_j \exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right)} \\
&= \frac{\prod_j \exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right)}{\prod_j \sum_{h_j} \exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right)} \\
&= \prod_j \frac{\exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right)}{1 + \exp\left(\sum_i w_{ij} v_i + b_j\right)} \\
&= \prod_j P(h_j \,|\, \boldsymbol{v})
\end{aligned}$$

where (i) we are free to swap the order of the summation and product in the denominator in the fourth line since, letting $f(h_j) = \exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right)$,

$$\begin{aligned}
\sum_{h_0} \cdots \sum_{h_n} \prod_j \exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right) &= \sum_{h_0} \cdots \sum_{h_n} f(h_0) f(h_1) \cdots f(h_n) \\
&= \sum_{h_0} f(h_0) \cdot \sum_{h_1} f(h_1) \cdots \sum_{h_n} f(h_n) \\
&= \prod_j \sum_{h_j} f(h_j) \\
&= \prod_j \sum_{h_j} \exp\left(\sum_i w_{ij} v_i h_j + b_j h_j\right)
\end{aligned}$$

and (ii) $P(H_j = 0 \,|\, \boldsymbol{V}) = 1 - P(H_j = 1 \,|\, \boldsymbol{V}) = \frac{1}{1+\exp(b_j + \sum\limits_i w_{ij} v_i))}$. ■

(b) **[3pts]** Similarly, determine the factorized form for the distribution over visible units conditioned on the hidden units, $P(\boldsymbol{V} \,|\, \boldsymbol{H})$.

Solution:

This follows readily from the above derivation via symmetry

$$P(\boldsymbol{V} = \boldsymbol{v} \,|\, \boldsymbol{H} = \boldsymbol{h}) = \prod_i P(v_i \,|\, \boldsymbol{h}),$$

where

$$P(V_i = 1 \,|\, \boldsymbol{H} = \boldsymbol{h}) = \sigma\left(a_i + \sum_j w_{ij} h_j\right)$$

and $P(V_i = 0 \,|\, \boldsymbol{H} = \boldsymbol{h}) = 1 - P(V_i = 1 \,|\, \boldsymbol{H} = \boldsymbol{h})$. ■

(c) **[3pts]** Is it possible to factorize the marginal distribution over hidden units $P(\boldsymbol{H})$ and if so, what is the factorization? If not, justify your answer.

Solution:

The marginal distribution $P(\boldsymbol{H})$ follows as

$$\begin{aligned}
P(\boldsymbol{H}=\boldsymbol{h}) &= \sum_{v} P(\boldsymbol{v},\boldsymbol{h}) \\
&\propto \sum_{v} \exp\left(-E(\boldsymbol{v},\boldsymbol{h}\right) \\
&= \sum_{v} \exp\left(\sum_{ij} w_{ij} v_i h_j + \sum_{i} a_i v_i + \sum_{j} b_j h_j\right) \\
&= \exp\left(\sum_{j} b_j h_j\right) \sum_{v} \exp\left(\sum_{ij} w_{ij} v_i h_j + \sum_{i} a_i v_i\right) \\
&= \exp\left(\sum_{j} b_j h_j\right) \sum_{v} \exp\left(\sum_{ij} w_{ij} v_i h_j + \sum_{i} a_i v_i\right) \\
&= \exp\left(\sum_{j} b_j h_j\right) \sum_{v} \exp\left(\sum_{i}\left(\sum_{j} w_{ij} v_i h_j + \sum_{i} a_i v_i\right)\right) \\
&= \left(\prod_{j} \exp\left(b_j h_j\right)\right)\left(\sum_{v}\prod_{i} \exp\left(v_i \sum_{j}\left(w_{ij} h_j + \sum_{i} a_i\right)\right)\right) \\
&= \left(\prod_{j} \exp\left(b_j h_j\right)\right)\left(\prod_{i}\sum_{v_i} \exp\left(v_i \left(\sum_{j} w_{ij} h_j + \sum_{i} a_i\right)\right)\right) \\
&= \left(\prod_{j} \exp\left(b_j h_j\right)\right)\left(\prod_{i}\sum_{v_i} \exp\left(v_i \left(\sum_{j} w_{ij} h_j + \sum_{i} a_i\right)\right)\right) \\
&= \left(\prod_{j} \exp\left(b_j h_j\right)\right)\left(\prod_{i} 1 + \exp\left(\sum_{j} w_{ij} h_j + \sum_{i} a_i\right)\right)
\end{aligned}$$

While the first term is a function of only the hidden units, the second term is a product over visible units i and not hidden unit j. Therefore, $P(\boldsymbol{H})$ does not factorize. ■

(d) **[2pts]** Does the distribution in Equation 2 adhere to the independence relationships expressed in the Markov network in Figure 2? Is the Markov network an I-map for the distribution (i.e., are there independencies that are expressed in the Markov network that are not valid under the distribution)?

Solution:

Yes. As a bipartite graph between the hidden and visible units, the undirected graph implies that the hidden units are conditionally independent given the visible units (and vice versa). These are exactly the independencies that we identified above. ■

(e) **[8pts]** Our ability to update the states of neurons (perform inference) relies upon knowledge of the pairwise and unary weights. We can learn these parameters from a collection of training data (observations of $\boldsymbol{V}$) using a form of approximate gradient descent known as “contrastive divergence.” To that end, determine the gradient of the log-likelihood objective with respect to the weights w_{ij} by demonstrating the following:

$$\begin{aligned}\frac{\partial \log P(\boldsymbol{V}=\boldsymbol{v})}{\partial w_{ij}} &= \sum_{\boldsymbol{h}} P(\boldsymbol{H}=\boldsymbol{h} \mid \boldsymbol{V}=\boldsymbol{v}) v_i h_j - \sum_{\boldsymbol{v},\boldsymbol{h}} P(\boldsymbol{V}=\boldsymbol{v}, \boldsymbol{H}=\boldsymbol{h}) v_i h_j \\ &= \mathbb{E}[V_i H_j \mid \boldsymbol{V}=\boldsymbol{v}] - \mathbb{E}[V_i H_j]\end{aligned}$$

where $\sum_{\boldsymbol{h}}(\cdot) = \sum_{h_0}\sum_{h_1}\cdots\sum_{h_n}(\cdot)$ (and similarly for $\sum_{\boldsymbol{h},\boldsymbol{v}}$), and we can arrive $\mathbb{E}[V_i H_j \mid \boldsymbol{V} = \boldsymbol{v}]$ via Equation 3, however the expectation in the expression for $\mathbb{E}[V_i H_j]$ is less straightforward as it is taken over both V_i and H_j.

Hint: The partition function Z is a function of w_{ij}.

Hint: For the sake of space, don’t expand $E(\boldsymbol{v},\boldsymbol{h})$ until you have $\frac{E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}}$.

Solution:

$$\begin{aligned}\frac{\partial \log P(\boldsymbol{v})}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}} \log \frac{\sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v},\boldsymbol{h}))}{Z} \\ &= \frac{Z}{\sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v},\boldsymbol{h}))} \cdot \frac{\partial}{\partial w_{ij}} \left(\frac{\sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v},\boldsymbol{h}))}{Z} \right) \\ &= \frac{Z}{\sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v},\boldsymbol{h}))} \cdot \left\{ \frac{1}{Z} \sum_{\boldsymbol{h}} \frac{\partial}{\partial w_{ij}} \{\exp(-E(\boldsymbol{v},\boldsymbol{h}))\} - \frac{\sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v},\boldsymbol{h}))}{Z^2} \frac{\partial Z}{\partial w_{ij}} \right\} \\ &= -\sum_{\boldsymbol{h}} \frac{\exp(-E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{h}'} \exp(-E(\boldsymbol{v},\boldsymbol{h}'))} \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} - \frac{1}{Z} \frac{\partial}{\partial w_{ij}} \sum_{\boldsymbol{v},\boldsymbol{h}} \exp(-E(\boldsymbol{v},\boldsymbol{h})) \\ &= \sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}) v_i h_j + \sum_{\boldsymbol{v},\boldsymbol{h}} \frac{\exp(-E(\boldsymbol{v},\boldsymbol{h}))}{Z} \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} \\ &= \sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}) v_i h_j - \sum_{\boldsymbol{v},\boldsymbol{h}} P(\boldsymbol{v},\boldsymbol{h}) v_i h_j \\ &= \mathbb{E}[V_i H_j \mid \boldsymbol{V}=\boldsymbol{v}] - \mathbb{E}[V_i H_j]\end{aligned}$$

■

(f) **[1pts]** Suppose that $H_0 = 1$ corresponds to the action and $H_1 = 1$ corresponds to the romance genre. Which w_{ij} would you expect to be positive after training, where i and j index the visible and hidden units, respectively?

Solution:

$w_{00}, w_{11}, w_{02}, w_{03}, w_{14}$ ■

4 Markov Networks [15pts]

Given a distribution P over $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$, the Markov network H is a minimal I-map for P if $I(H) \subseteq I(P)$ and removing any edge from H will invalidate this property.

This problem explores the ability to generate a Markov network that is a minimal I-map for a given distribution. Recall that the Markov blanket $\text{MB}_P(X)$ for a variable X in a distribution P is the minimal set of variables $\boldsymbol{U}$ such that $X \notin \boldsymbol{U}$ and

$$(X \perp \boldsymbol{X} - \{X\} - \boldsymbol{U} \mid \boldsymbol{U}) \in I(P).$$

Let us construct a Markov network H using the following procedure: Add an edge between each variable X_i and all the variables X_j in its Markov blanket $X_j \in \text{MB}_P(X_i)$.

(a) **[3pts]** Let's consider a particular P over four binary random variables and suppose that $P(0,0,0,0) = P(1,1,1,1) = 0.5$. Show that the following Markov network follows from the above algorithm, but that the graph is not a minimal I-map for P. Explain why this is, given the discussion above.

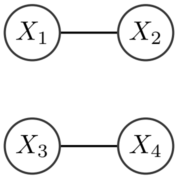


Figure 3: Graph for a distribution over four binary random variables for which $P(0,0,0,0) = P(1,1,1,1) = 0.5$.

Solution:

According to the distribution, X_1 defines the Markov blanket for X_2 (and vice versa), and similarly, X_3 defines the Markov blanket for X_4 (and vice versa) since knowing one specifies the value of the other. Thus, the procedure described above results in the undirected graph in Figure 3. However, the undirected graph implies that X_1 and X_2 are each independent of X_3 and X_4, which is clearly not the case as knowing X_3 or X_4 specifies the value of X_1 and X_2. The reason for this discrepancy is that this distribution is not positive (e.g., $P(0,1,0,1) = 0$), while the theorems that relate a graph constructed according to a distribution's Markov blanket as being a (unique, minimal) I-map assume that the distribution is positive. ■

(b) **[3pts]** Show that for any distribution P and Markov network H, $P \models I_l(H) \Rightarrow P \models I_p(H)$.

Solution:

This proof is straightforward since $P \models I_l(H)$ implies the following independence relationship is valid under P: $(X \perp \boldsymbol{X} - \{X\} - \text{MB}_P(X) \mid \text{MB}_P(X)) \in I(P)$. The pairwise independencies $I_p(H) = \{(X \perp Y \mid \boldsymbol{X} - \{X, Y\}) : X - Y \notin H\}$ involve conditioning on not just the Markov blanket, but also everything else except X and Y. Thus, we can conclude that $P \models I_p(H)$ via the weak union property. ■

(c) **[9pts]** As opposed to Bayesian networks, where the minimal I-map for a distribution is not unique, we discussed the following theorem (given in Koller and Friedman) that provides a procedure for constructing a Markov network structure that is the unique minimal I-map for a given positive distribution P.

Theorem 4.6: For a given positive distribution P, let $\text{MB}_P(X)$ be a minimal set of nodes $\boldsymbol{U}$ satisfying Eqn. 4. Construct a Markov network H by introducing an edge $\{X, Y\}$ for all X and $Y \in \text{MB}_P(X)$. The resulting Markov network is the *unique* minimal I-map for P.

$$(X \perp \boldsymbol{X} - \{X\} - \boldsymbol{U} \,|\, \boldsymbol{U}) \in I(P) \tag{4}$$

This question asks you to prove that this construction is indeed the unique minimal I-map for P. To that end, it is useful to consider a particular node X, let $\mathcal{U}$ be the set of all subsets $\boldsymbol{U}$ that satisfy Equation 4, and define $\boldsymbol{U}^*$ to the intersection of all $\boldsymbol{U} \in \mathcal{U}$.

(i) Prove that $\boldsymbol{U}^* \in \mathcal{U}$ and thus, that $\text{MB}_P(X) = \boldsymbol{U}^*$

(ii) Prove that $P \models (X \perp Y \,|\, \boldsymbol{X} - \{X, Y\}) \Rightarrow Y \notin \text{MB}_P(X)$

(iii) Prove that $Y \notin \text{MB}_P(X) \Rightarrow P \models (X \perp Y \,|\, \boldsymbol{X} - \{X, Y\})$

(iv) Conclude that $\text{MB}_P(X)$ is exactly the set of neighbors of X in H as defined above, and thus that the above construction produces a minimal I-map for P.

Solution:

(i) Consider sets $\boldsymbol{U}_i \in \mathcal{U}$ and $\boldsymbol{U}_j \in \mathcal{U}$, and let $\boldsymbol{U}' = \boldsymbol{U}_i \cap \boldsymbol{U}_j$ be the intersection of these two sets. Let $\tilde{\boldsymbol{U}}_i = \boldsymbol{U}_i \backslash \boldsymbol{U}'$ and $\tilde{\boldsymbol{U}}_j = \boldsymbol{U}_j \backslash \boldsymbol{U}'$.

Since $\boldsymbol{U}_i, \boldsymbol{U}_j \in \mathcal{U}$, the independencies expressed in Equation 4 hold for $\boldsymbol{U} = \boldsymbol{U}_i$ and $\boldsymbol{U} = \boldsymbol{U}_j$. Since $\tilde{\boldsymbol{U}}_i \subseteq \{\boldsymbol{X} - \{X\} - \boldsymbol{U}_j\}$ and $\tilde{\boldsymbol{U}}_j \subseteq \{\boldsymbol{X} - \{X\} - \boldsymbol{U}_i\}$, the following also hold

$$P \models (X \perp \tilde{\boldsymbol{U}}_j \,|\, \tilde{\boldsymbol{U}}_i \cup \boldsymbol{U}')$$
$$P \models (X \perp \tilde{\boldsymbol{U}}_i \,|\, \tilde{\boldsymbol{U}}_j \cup \boldsymbol{U}')$$

where $\boldsymbol{U}_i = \tilde{\boldsymbol{U}}_i \cup \boldsymbol{U}'$ (and similarly for $\boldsymbol{U}_j$). By the intersection property, it follows that

$$P \models (X \perp \{\boldsymbol{X} - \{X\} - \boldsymbol{U}_i \tilde{\boldsymbol{U}}_i \cup \tilde{\boldsymbol{U}}_j\} \,|\, \boldsymbol{U}')$$

Additionally, we have

$$P \models (X \perp \{\boldsymbol{X} - \{X\} - \boldsymbol{U}_1\} \,|\, \overbrace{\tilde{\boldsymbol{U}}_i \cup \boldsymbol{U}'}^{\boldsymbol{U}_i})$$
$$P \models (X \perp \tilde{\boldsymbol{U}}_1) \,|\, \boldsymbol{U}')$$

By the contraction property (letting $W = \{\boldsymbol{X} - \{X\} - \boldsymbol{U}_1\}$, $Y = \tilde{\boldsymbol{U}}_1$, and $Z = \boldsymbol{U}'$),

$$P \models (X \perp \{\boldsymbol{X} - \{X\} - \boldsymbol{U}'\} \,|\, \boldsymbol{U}')$$

Thus, $\boldsymbol{U}' \in \mathcal{U}$. Extending this recursively, we see that $\boldsymbol{U}^* \in \mathcal{U}$.

Now, we need to prove that $\boldsymbol{U}^* = \mathrm{MB}_P(X)$. Suppose that there was a set $\tilde{\boldsymbol{U}} \in \mathcal{U}$ such that $\tilde{\boldsymbol{U}} \subset \boldsymbol{U}^*$, which would imply that $\boldsymbol{U}^* \backslash \tilde{\boldsymbol{U}} \subset \boldsymbol{U}^*$. However, this in violation of the definition of $\boldsymbol{U}^*$ as the intersection of all sets in $\mathcal{U}$. Thus, $\boldsymbol{U}^*$ is the minimal set satisfying the above independencies and is therefore the Markov blanket for X.

(ii) We will prove this by contradiction. Assume that $Y \in \mathrm{MB}_P(X)$, and let $\overline{\mathrm{MB}}_P(X) = \mathrm{MB}_P(X) \backslash Y$. We have $(X \perp Y \mid (\boldsymbol{X} \backslash \overline{\mathrm{MB}}_P(X)) \cup \overline{\mathrm{MB}}_P(X) - \{X\})$ and, by the definition of the Markov blanket $(X \perp \boldsymbol{X} - \{X, Y\} - \overline{\mathrm{MB}}_P(X) \mid \overline{\mathrm{MB}}_P(X) \cup Y)$. Per the intersection property (letting $W = \boldsymbol{X} - \{X, Y\} - \overline{\mathrm{MB}}_P(X)$ and $Z = \overline{\mathrm{MB}}_P(X)$), we have $(X \perp \boldsymbol{X} - \{X\} - \overline{\mathrm{MB}}_P(X) \mid \overline{\mathrm{MB}}_P(X))$. This implies that there is a Markov blanket $\overline{\mathrm{MB}}_P(X) \subset \mathrm{MB}_P(X)$, which violates the assumption that the Markov blanket is the minimal subset.

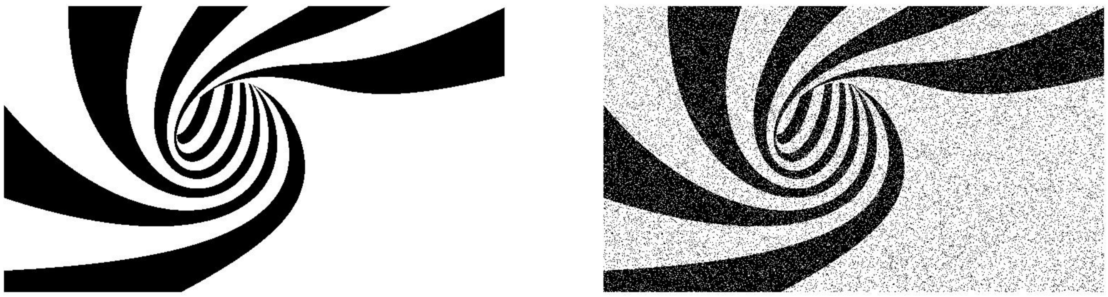
(iii) This statement follows from the definition of pairwise independencies (since we are conditioning on the Markov blanket for X as well as all other nodes that are not in the Markov blanket (except for X and Y)).

(iv) Per the above two statements (ii) and (iii), any node that is not connected to X by an edge is not in its Markov blanket. The Markov blanket is thus formed by the set of neighbors of X and per (i), this set is minimal.

Thus, the construction is unique and minimal. ■

5 Image Denoising [30pts]

Markov random fields have proven effective for modeling a number of problems in computer vision, including image segmentation and image denoising. This problem asks you to implement an MRF for image denoising. Consider the case in which images are binary, represented as a two-dimensional array of binary pixel values $X_i \in \{-1, +1\}$. Suppose that we have access to a noisy copy of the true image, where the value of each pixel is flipped with a 10% probability (Fig. 4).



(a) Original Image

(b) Noisy Image

Figure 4: A (a) binary image and (b) a noisy version in which the value of each pixel is flipped with a 10% probability.

Given observations of the noisy pixel values, the goal of denoising is to recover the original, noise-free image. We model this problem using a pair-wise Markov network with random variables

X_i that represent the (latent) original intensity of each pixel and Y_i that denotes the (observed) noise-corrupted values. The Markov network (Fig. 5) consists of a lattice over $\boldsymbol{X}$ with edges between each pixel X_i and the four pixels above, below, left, and right, as well as an edge between X_i and the observed value Y_i.

We specify the energy associated with this MRF as a combination of terms that express pairwise consistency (smoothing) and unary priors

$$E(\boldsymbol{x}, \boldsymbol{y}) = -\alpha \sum_i x_i - \beta \sum_{i,j} x_i x_j - \nu \sum_i x_i y_i \tag{5}$$

where, for simplicity, we assume that the weight parameters $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^+$, and $\nu \in \mathbb{R}^+$ are constant. You will choose settings for these parameters.

Given a noise-corrupted image, we ask you implement a function (in Python) that infers the (latent) values of each pixel intensity in the original image by minimizing the energy in Equation 5. To do so, initialize each X_i to its noisy value and then iterate through each X_i, setting it to whichever $\{-1, +1\}$ value yields lower energy. Repeat this process over the entire image until the total energy has converged. You will measure the accuracy of the resulting estimate relative to the original image in terms of error rate, i.e., the percentage of pixels that are incorrectly estimated.

(a) Implement an algorithm that minimizes the energy in the MRF as a Python file `denoise.py` that can be called as

```
$ python denoise.py images.mat
```

or

```
$ python denoise.py image-noisy.pgm image-original.pgm
```

This call should print the error rate to the terminal and visualize and/or save the denoised image. If there are multiple files, please submit an archive (e.g., tar, zip, etc.).

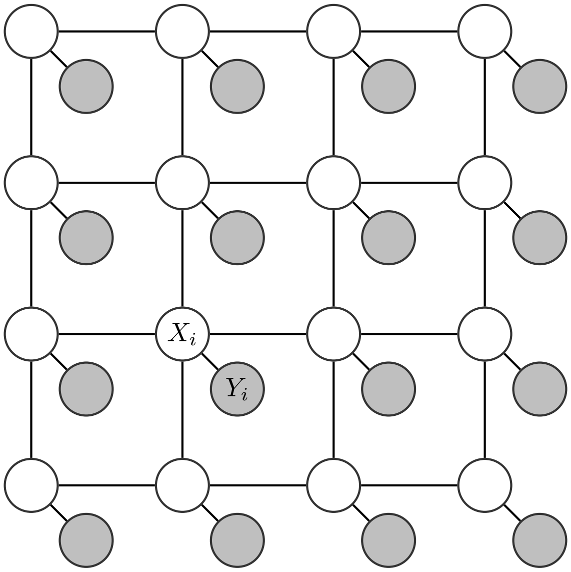


Figure 5: Markov random field for the image denoising problem in which X_i denotes the unknown original pixel intensity and Y_i denotes the noisy observation.

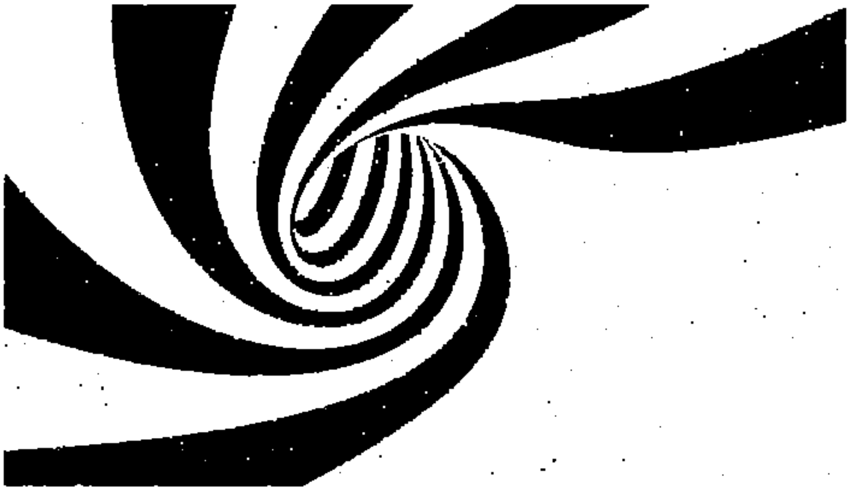


Figure 6: The denoised image resulting from (approximate) inference over an MRF with $\alpha = 1.0, \beta = 100.0, \nu = 1.0$.

(b) Report the error rates for five settings of the weight parameters $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^+$, and $\nu \in \mathbb{R}^+$ that yield noticeably different results. Explain the differences in behavior.

(c) Provide a figure depicting the denoised image corresponding to the lowest error rate.

What is included: A `ps1-images.tgz` file that contains the original and noisy images as separate pgm files,[2] as well as a single mat file. You are welcome to use either, though the mat file may be easier to import in Python (e.g., via `scipy.io.loadmat`). We also include a template `denoise.py` file that includes a function for loading images.

Solution:

For $\alpha = 1.0, \beta = 100.0, \nu = 1.0$, we get the denoised image depicted in Figure 6. ■

[2] Note that the intensity values in the pgm images will need to be converted from $\{0, 255\}$ to $\{-1, +1\}$.