CSE 250a. Assignment 3

Out: *Tue Apr 19*
Due: *Tue Apr 26* (beginning of class)

3.1 Inference in a chain

Consider the belief network shown above with random variables $X_t \in \{1, 2, \ldots, n\}$. Suppose that the CPT at each non-root node is given by the same $n \times n$ transition matrix; that is, for all $t \geq 1$, we have:

$$U_{ij} \;=\; P(X_{t+1}=j|X_t=i).$$

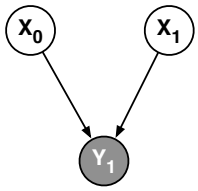
(a) Show that $P(X_{t+1}=j|X_1=i) = [U^t]_{ij}$, where U^t is the t^{th} power of the transition matrix.

(b) Consider the computational complexity of this inference. Devise a simple algorithm, based on matrix-vector multiplication, that scales as $O(n^2 t)$.

(c) Show alternatively that the inference can also be done in $O(n^3 \log_2 t)$.

(d) Suppose that the transition matrix U_{ij} is sparse, with at most $m \ll n$ non-zero elements per row. Show that in this case the inference can be done in $O(mnt)$.

3.2 More inference in a chain



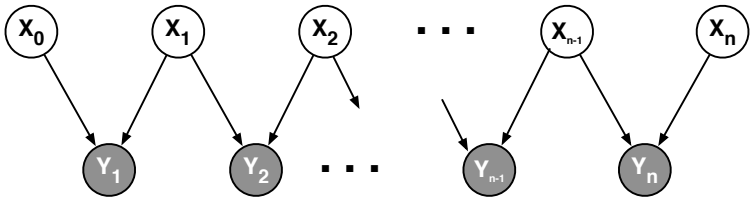
Consider the simple belief network shown to the right, with nodes X_0, X_1, and Y_1. To compute the posterior probability $P(X_1|Y_1)$, we can use Bayes rule:

$$P(X_1|Y_1) \;=\; \frac{P(Y_1|X_1)\,P(X_1)}{P(Y_1)}$$

(a) Show how to compute the conditional probability $P(Y_1|X_1)$ that appears in the numerator of Bayes rule from the CPTs of the belief network.

(b) Show how to compute the marginal probability $P(Y_1)$ that appears in the denominator of Bayes rule from the CPTs of the belief network.

Next you will show how to generalize these simple computations when the basic structure of this DAG is repeated to form an extended chain. Like the previous problem, this is another instance of efficient inference in polytrees.



Consider how to efficiently compute the posterior probability $P(X_n|Y_1, Y_2, \ldots, Y_n)$ in the above belief network. One approach is to derive a recursion from the *conditionalized* form of Bayes rule

$$P(X_n|Y_1, Y_2, \ldots, Y_n) \;=\; \frac{P(Y_n|X_n, Y_1, Y_2, \ldots, Y_{n-1})\, P(X_n|Y_1, Y_2, \ldots, Y_{n-1})}{P(Y_n|Y_1, \ldots, Y_{n-1})}$$

where the nodes $Y_1, Y_2, \ldots, Y_{n-1}$ are treated as background evidence. In this problem you will express the conditional probabilities on the right hand side of this equation in terms of the CPTs of the network and the probabilities $P(X_{n-1}{=}x|Y_1, Y_2, \ldots, Y_{n-1})$, *which you may assume have been computed at a previous step of the recursion*. Your answers to (a) and (b) should be helpful here.

(c) Simplify the term $P(X_n|Y_1, Y_2, \ldots, Y_{n-1})$ that appears in the numerator of Bayes rule.

(d) Show how to compute the conditional probability $P(Y_n|X_n, Y_1, Y_2, \ldots, Y_{n-1})$ that appears in the numerator of Bayes rule. Express your answer in terms of the CPTs of the belief network and the probabilities $P(X_{n-1}{=}x|Y_1, Y_2, \ldots, Y_{n-1})$, which you may assume have already been computed.

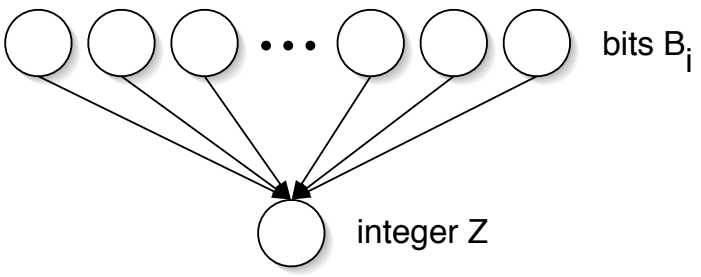
(e) Show how to compute the conditional probability $P(Y_n|Y_1, Y_2, \ldots, Y_{n-1})$ that appears in the denominator of Bayes rule. Express your answer in terms of the CPTs of the belief network and the probabilities $P(X_{n-1}{=}x|Y_1, Y_2, \ldots, Y_{n-1})$, which you may assume have already been computed.

3.3 Stochastic simulation

Consider the belief network shown below, with n binary random variables $B_i \in \{0, 1\}$ and an *integer* random variable Z. Let $f(B) = \sum_{i=1}^{n} 2^{i-1} B_i$ denote the nonnegative integer whose binary representation is given by $B_n B_{n-1} \ldots B_2 B_1$. Suppose that each bit has prior probability $P(B_i{=}1) = \frac{1}{2}$, and that

$$P(Z|B_1, B_2, \ldots, B_n) \;=\; \left(\frac{1-\alpha}{1+\alpha}\right) \alpha^{|Z - f(B)|}$$

where $0 < \alpha < 1$ is a parameter measuring the amount of noise in the conversion from binary to decimal. (Larger values of α indicate greater levels of noise.)



(a) Show that the conditional distribution for binary to decimal conversion is normalized; namely, that $\sum_z P(Z=z|B_1, B_2, \ldots, B_n) = 1$, where the sum is over all integers $z \in [-\infty, +\infty]$.

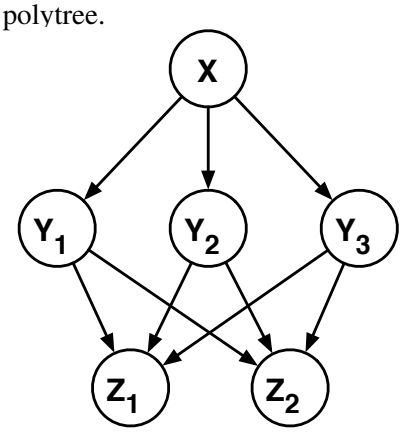
(b) Use the method of *likelihood weighting* to estimate the probability $P(B_8=1|Z=128)$ for a network with $n=10$ bits and noise level $\alpha=0.25$.

(c) Plot your estimate in part (b) *as a function of the number of samples*. You should be confident from the plot that your estimate has converged to a good degree of precision (say, at least two significant digits).

(d) Submit your source code (electronically). You may program in the language of your choice, and you may use any program at your disposal to plot the results.

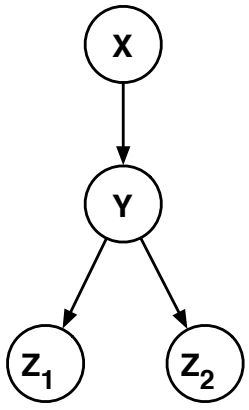
3.4 Node clustering

Consider the belief network shown below over binary variables X, Y_1, Y_2, Y_3, Z_1, and Z_2. The network can be transformed into a polytree by clustering the nodes Y_1, Y_2, and Y_3 into a single node Y. From the CPTs in the original belief network, fill in the missing elements of the CPTs for the polytree.



X	$P(Y_1=1\|X)$	$P(Y_2=1\|X)$	$P(Y_3=1\|X)$
0	0.1	0.3	0.5
1	0.8	0.6	0.4

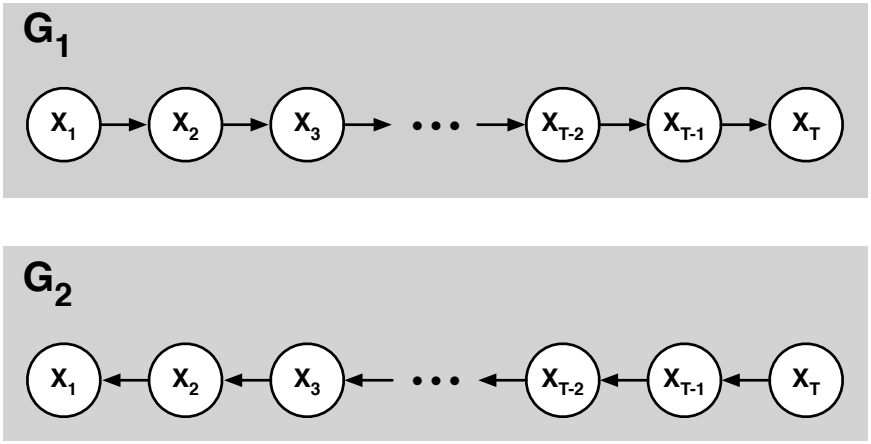
Y_1	Y_2	Y_3	$P(Z_1=1\|Y_1,Y_2,Y_3)$	$P(Z_2=1\|Y_1,Y_2,Y_3)$
0	0	0	0.1	0.9
1	0	0	0.2	0.8
0	1	0	0.3	0.7
0	0	1	0.4	0.6
1	1	0	0.5	0.5
1	0	1	0.6	0.4
0	1	1	0.7	0.3
1	1	1	0.8	0.2



Y_1	Y_2	Y_3	Y	$P(Y\|X=0)$	$P(Y\|X=1)$	$P(Z_1=1\|Y)$	$P(Z_2=1\|Y)$
0	0	0	1				
1	0	0	2				
0	1	0	3				
0	0	1	4				
1	1	0	5				
1	0	1	6				
0	1	1	7				
1	1	1	8				

3.5 Maximum likelihood estimation

Consider the two DAGs shown below, G_1 and G_2, over the same nodes $\{X_1, X_2, \ldots, X_T\}$, that differ only in the direction of their edges.



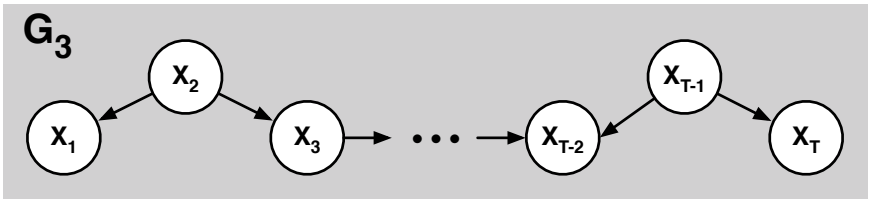
Suppose that the CPTs for these DAGs are obtained by maximum likelihood estimation from a "fully observed" data set in which each example provides a complete instantiation of the nodes in the DAG. In this data set, let $\text{COUNT}_t(x)$ denote the number of examples in which $X_t = x$, and let $\text{COUNT}_t(x, x')$ denote the number of examples in which $X_t = x$ and $X_{t+1} = x'$.

(a) Express the maximum likelihood estimates for the CPTs in G_1 in terms of these counts.

(b) Express the maximum likelihood estimates for the CPTs in G_2 in terms of these counts.

(c) Using your answers from parts (a) and (b), show that the maximum likelihood CPTs for G_1 and G_2 from this data set give rise to the same joint distribution over the nodes $\{X_1, X_2, \ldots, X_T\}$.

(d) Suppose that some but not all of the edges were reversed, as in the graph G_3 shown below. Would the maximum likelihood CPTs for G_3 also give rise to the same joint distribution? (*Hint*: does G_3 imply all the same statements of conditional independence as G_1 and G_2?)



3.6 Statistical language modeling

In this problem, you will explore some simple statistical models of English text. Download and examine the data files on Piazza for this assignment. (Start with the `readme.txt` file.) These files contain unigram and bigram counts for 500 frequently occurring tokens in English text. These tokens include actual words as well as punctuation symbols and other textual markers. In addition, an "unknown" token is used to represent all words that occur outside this basic vocabulary. Submit (electronically) your results and **source code** for the following problems. (Don't forget the code!) As usual you may program in the language of your choice.

(a) Compute the maximum likelihood estimate of the unigram distribution $P_u(w)$ over words w. Print out a table of all the tokens (i.e., words) that start with the letter "B", along with their numerical unigram *probabilities* (not counts). (You do not need to print out the unigram probabilities for all 500 tokens.)

(b) Compute the maximum likelihood estimate of the bigram distribution $P_b(w'|w)$. Print out a table of the ten most likely words to follow the word "ONE", along with their numerical bigram probabilities.

(c) Consider the sentence **"The stock market fell by one hundred points last week."** Ignoring punctuation, compute and compare the log-likelihoods of this sentence under the unigram and bigram models:

$$\begin{aligned}\mathcal{L}_u &= \log\Big[P_u(\textbf{the})\,P_u(\textbf{stock})\,P_u(\textbf{market})\,\ldots\,P_u(\textbf{points})\,P_u(\textbf{last})P_u(\textbf{week})\Big]\\ \mathcal{L}_b &= \log\Big[P_b(\textbf{the}|\langle\textbf{s}\rangle)\,P_b(\textbf{stock}|\textbf{the})\,P_b(\textbf{market}|\textbf{stock})\,\ldots\,P_b(\textbf{last}|\textbf{points})\,P_b(\textbf{week}|\textbf{last})\Big]\end{aligned}$$

In the equation for the bigram log-likelihood, the token $\langle s\rangle$ is used to mark the beginning of a sentence. Which model yields the highest log-likelihood?

(d) Consider the sentence **"The fourteen officials sold fire insurance."** Ignoring punctuation, compute and compare the log-likelihoods of this sentence under the unigram and bigram models:

$$\begin{aligned}\mathcal{L}_u &= \log\Big[P_u(\textbf{the})\,P_u(\textbf{fourteen})\,P_u(\textbf{officials})\,\ldots\,P_u(\textbf{sold})\,P_u(\textbf{fire})\,P_u(\textbf{insurance})\Big]\\ \mathcal{L}_b &= \log\Big[P_b(\textbf{the}|\langle\textbf{s}\rangle)\,P_b(\textbf{fourteen}|\textbf{the})\,P_b(\textbf{officials}|\textbf{fourteen})\,\ldots\,P_b(\textbf{fire}|\textbf{sold})\,P_b(\textbf{insurance}|\textbf{fire})\Big]\end{aligned}$$

Which pairs of adjacent words in this sentence are not observed in the training corpus? What effect does this have on the log-likelihood from the bigram model?

(e) Consider the so-called *mixture* model that predicts words from a weighted interpolation of the unigram and bigram models:

$$P_m(w'|w) = (1-\lambda)P_u(w') + \lambda P_b(w'|w),$$

where $\lambda \in [0,1]$ determines how much weight is attached to each prediction. Under this mixture model, the log-likelihood of the sentence from part (d) is given by:

$$\mathcal{L}_m = \log\Big[P_m(\textbf{the}|\langle\textbf{s}\rangle)\,P_m(\textbf{fourteen}|\textbf{the})\,P_m(\textbf{officials}|\textbf{fourteen})\,\ldots\,P_m(\textbf{fire}|\textbf{sold})\,P_m(\textbf{insurance}|\textbf{fire})\Big].$$

Compute and plot the value of this log-likelihood $\mathcal{L}_m$ as a function of the parameter $\lambda \in [0,1]$. From your results, deduce the optimal value of λ to two significant digits.