CSE 150A/250A Assignment 5 *(64 pts)*

Out: *Tues May 13*
Due: *Mon May 19* (by 11:59 PM, Pacific Time, via gradescope)
Grace period: 24 hours

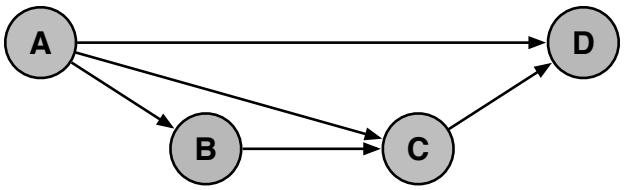
5.1 Data collection for movie recommendation system *(2 pts)*

In Homework 7, we will use the EM algorithm to build a simple movie recommendation system. Please finish the survey for movie rating:

`http://bit.ly/4j4LjYc`

5.2 EM algorithm *(19 pts)*

(a) **Complete data** *(3 pts)*



Consider a complete data set of *i.i.d.* examples $\{a_t, b_t, c_t, d_t\}_{t=1}^{T}$ drawn from the joint distribution of the above belief network. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Express your answers in terms of equality-testing functions, such as:

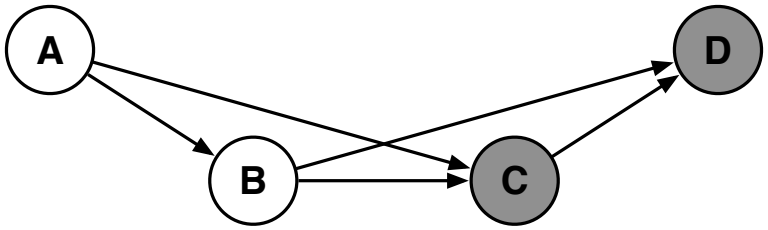
$$I(a, a_t) = \begin{cases} 1 \text{ if } a = a_t, \\ 0 \text{ if } a \neq a_t. \end{cases}$$

For example, in terms of this function, the maximum likelihood estimate for the CPT at node A is given by $P(A = a) = \frac{1}{T} \sum_{t=1}^{T} I(a, a_t)$. Complete the numerators and denominators in the below expressions.

$$P(B{=}b|A{=}a) \quad = \quad \text{———————————}$$

$$P(C{=}c|A{=}a, B{=}b) \quad = \quad \text{———————————}$$

$$P(D{=}d|A{=}a, C{=}c) \quad = \quad \text{———————————}$$



(b) **Posterior probability** *(3 pts)*

Consider the belief network shown above, with observed nodes C and D and hidden nodes A and B. Compute the posterior probability $P(a, b|c, d)$ in terms of the CPTs of the belief network—that is, in terms of $P(a)$, $P(b|a)$, $P(c|a, b)$ and $P(d|b, c)$.

(c) **Posterior probability** *(2 pts)*

Compute the posterior probabilities $P(a|c, d)$ and $P(b|c, d)$ in terms of your answer from part (b); that is, for this problem, you may assume that $P(a, b|c, d)$ is given.

(d) **Log-likelihood** *(3 pts)*

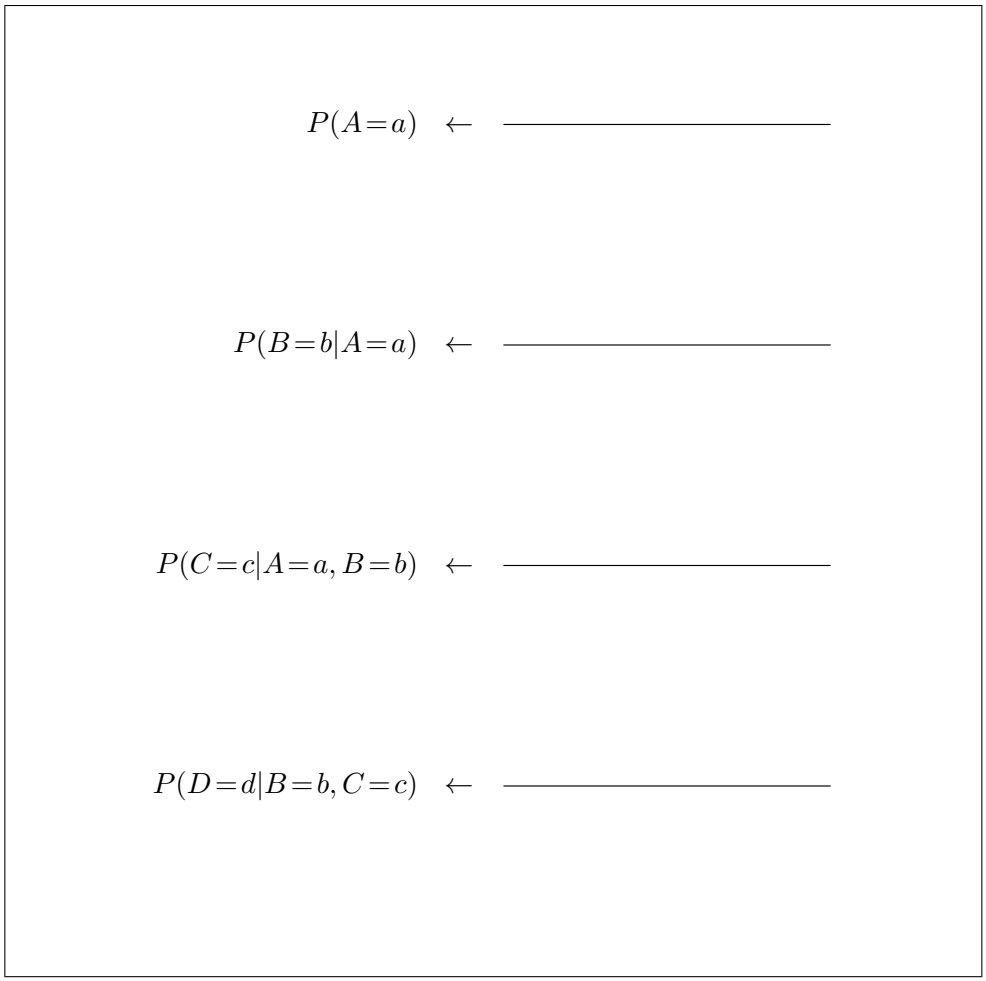
Consider a partially complete data set of *i.i.d.* examples $\{c_t, d_t\}_{t=1}^{T}$ drawn from the joint distribution of the above belief network. The log-likelihood of the data set is given by:

$$\mathcal{L} = \sum_t \log P(C = c_t, D = d_t).$$

Compute this log-likelihood in terms of the CPTs of the belief network. You may re-use work from earlier parts of the problem.

(e) **EM algorithm** *(8 pts)*

Give the EM updates to estimate CPTs that maximize the log-likelihood in part (c); in particular, complete the numerator and denominator in the below expressions for the update rules. Simplify your answers as much as possible, expressing them in terms of the posterior probabilities $P(a, b|c_t, d_t)$, $P(a|c_t, d_t)$, and $P(b|c_t, d_t)$, as well as the functions $I(c, c_t)$, and $I(d, d_t)$.

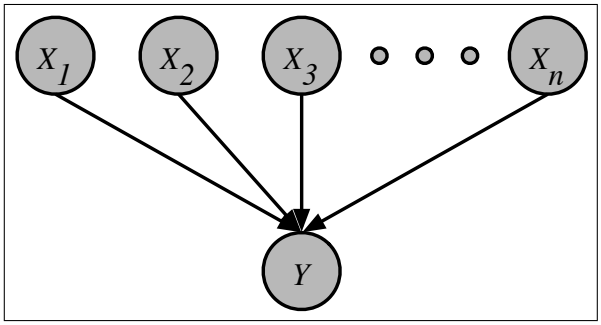

$P(A=a)$ $\leftarrow$
$P(B=b|A=a)$ $\leftarrow$
$P(C=c|A=a,B=b)$ $\leftarrow$
$P(D=d|B=b,C=c)$ $\leftarrow$

5.3 EM algorithm for noisy-OR *(24 pts)*

Consider the belief network on the right, with binary random variables $X \in \{0,1\}^n$ and $Y \in \{0,1\}$ and a noisy-OR conditional probability table (CPT). The noisy-OR CPT is given by:

$$P(Y=1|X) \;=\; 1 - \prod_{i=1}^{n} (1-p_i)^{X_i},$$

which is expressed in terms of the noisy-OR parameters $p_i \in [0,1]$.



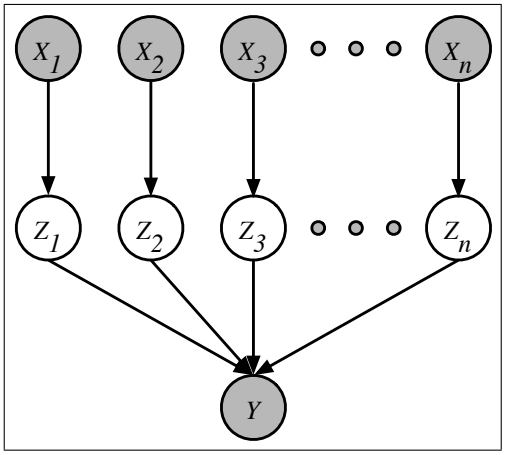
In this problem, you will derive and implement an EM algorithm for estimating the noisy-OR parameters p_i. It may seem that the EM algorithm is not suited to this problem, in which all the nodes are observed, and the CPT has a parameterized form. In fact, the EM algorithm can be applied, but first we must express the model in a different but equivalent form.

Consider the belief network shown to the right. In this network, a binary random variable $Z_i \in \{0,1\}$ intercedes between each pair of nodes X_i and Y. Suppose that:

$$\begin{aligned} P(Z_i=1|X_i=0) &= 0, \\ P(Z_i=1|X_i=1) &= p_i. \end{aligned}$$

Also, let the node Y be *determined* by the logical-OR of Z_i. In other words:

$$P(Y=1|Z) \;=\; \begin{cases} 1 \text{ if } Z_i=1 \text{ for any } i, \\ 0 \text{ if } Z_i=0 \text{ for all } i. \end{cases}$$



(a) *(4 pts)* Show that this "extended" belief network defines the same conditional distribution $P(Y|X)$ as the original one. In particular, starting from

$$P(Y=1|X) \;=\; \sum_{Z \in \{0,1\}^n} P(Y=1, Z|X),$$

show that the right hand side of this equation reduces to the noisy-OR CPT with parameters p_i. To perform this marginalization, you will need to exploit various conditional independence relations.

(b) *(4 pts)* Consider estimating the noisy-OR parameters p_i to maximize the (conditional) likelihood of the observed data. The (normalized) log-likelihood in this case is given by:

$$\mathcal{L} \;=\; \frac{1}{T} \sum_{t=1}^{T} \log P(Y=y^{(t)}|X=\vec{x}^{(t)}),$$

where $(\vec{x}^{(t)}, y^{(t)})$ is the tth joint observation of X and Y, and where for convenience we have divided the overall log-likelihood by the number of examples T. From your result in part (a), it follows that we can estimate the parameters p_i in either the original network or the extended one (since in both networks they would be maximizing the same equation for the log-likelihood).

Notice that in the extended network, we can view X and Y as observed nodes and Z as hidden nodes. Thus in this network, we can use the EM algorithm to estimate each parameter p_i, which simply defines one row of the "look-up" CPT for the node Z_i.

Compute the posterior probability that appears in the E-step of this EM algorithm. In particular, for joint observations $x \in \{0,1\}^n$ and $y \in \{0,1\}$, use Bayes rule to show that:

$$P(Z_i=1, X_i=1 | X=x, Y=y) \;=\; \frac{y x_i p_i}{1 - \prod_j (1-p_j)^{x_j}}$$

(c) *(3 pts)* For the data set $\{\vec{x}^{(t)}, y^{(t)}\}_{t=1}^{T}$, show that the EM update for the parameters p_i is given by:

$$p_i \;\leftarrow\; \frac{1}{T_i} \sum_t P\left(Z_i=1, X_i=1 | X=x^{(t)}, Y=y^{(t)}\right),$$

where T_i is the number of examples in which $X_i=1$. (You should derive this update as a special case of the general form presented in lecture.)

(d) *(5 pts)* Download the data files on the course web site, and use the EM algorithm to estimate the parameters p_i. The data set[1] has $T = 267$ examples over $n = 23$ inputs. To check your solution, initialize all $p_i = 0.05$ and perform 256 iterations of the EM algorithm. At each iteration, compute the log-likelihood shown in part (b). (If you have implemented the EM algorithm correctly, this log-likelihood will always increase from one iteration to the next.) Also compute the number of mistakes $M \leq T$ made by the model at each iteration; a mistake occurs either when $y_t = 0$ and $P(y_t=1|\vec{x}_t) \geq 0.5$ (indicating a false positive) or when $y_t=1$ and $P(y_t=1|\vec{x}_t) \leq 0.5$ (indicating a false negative). The number of mistakes should generally decrease as the model is trained, though it is not guaranteed to do so at each iteration. Complete the following table:

iteration	number of mistakes M	log-likelihood $\mathcal{L}$
0	175	-0.95809
1	56	
2		-0.40822
4		
8		
16		
32		
64	37	
128		
256		-0.31016

You may use the already completed entries of this table to check your work.

(e) *(8 pts)* Turn in your source code. As always, you may program in the language of your choice.

[1] For those interested, more information about this data set is available at *http://archive.ics.uci.edu/ml/datasets/SPECT+Heart*. However, be sure to use the data files provided on Canvas, as they have been specially assembled for this assignment.

5.4 Identifying and Formulating a Problem for Hidden Markov Modeling *(7 pts)*

Hidden Markov Models (HMMs) are powerful tools for modeling systems that evolve over time, where the underlying state is not directly observable, but influences a sequence of observable events. Recall the key components of an HMM:

- A sequence of hidden states, $S_1, S_2, ..., S_T$, that follow the Markov property: $P(S_{t+1}|S_t, S_{t-1}, ...) = P(S_{t+1}|S_t)$.
- A sequence of observable emissions, $O_1, O_2, ..., O_T$, where each observation O_t depends only on the corresponding hidden state S_t: $P(O_t|S_t, S_{t-1}, ..., O_{t-1}, ...) = P(O_t|S_t)$.
- An initial state distribution $\pi = P(S_1)$.
- A state transition probability distribution $A = P(S_{t+1}|S_t)$.
- An emission probability distribution $B = P(O_t|S_t)$.

Furthermore, in typical HMM applications, both the state transitions and the emissions are non-deterministic (probabilistic).

Your task for this question is to:

(a) *(1 pt)* **Identify a different real-world or plausible scenario** (one not discussed in detail as a primary HMM example in course materials, other than perhaps brief mentions) that you believe can be appropriately modeled using a Hidden Markov Model. Think about processes in various domains such as biology, finance, user interaction, system monitoring, weather, etc., where there's an underlying, unobserved state influencing observable events over time.

(b) *(1 pt)* **Briefly explain *why* your chosen scenario is suitable for HMM modeling**, highlighting the sequential nature, the presence of hidden states, and the dependence of observations on these states. Discuss why the Markov assumption for the hidden states is a reasonable (even if simplifying) approximation in your chosen scenario.

(c) **Formulate the problem for your chosen scenario using the components of a Hidden Markov Model.** For your specific problem, define the following based on your scenario:

  i *(1 pt)* **Hidden States (S_t)**: Define the set of all possible hidden states in your model.

  ii *(1 pt)* **Observations (O_t)**: Define the set of all possible observable emissions in your model.

  iii *(1 pt)* **Initial State Distribution (π)**: Describe conceptually what this distribution represents in your scenario.

  iv *(1 pt)* **Transition Probabilities ($P(S_{t+1}|S_t)$)**: Describe conceptually what these probabilities represent. Explain *why* these transitions are non-deterministic in your scenario.

v *(1 pt)* **Emission Probabilities ($P(O_t|S_t)$)**: Describe conceptually what these probabilities represent. Explain *why* these emissions are non-deterministic (i.e., why observing an O_t does not perfectly reveal the corresponding S_t) in your scenario.

Crucially, you do *not* need to provide any specific numerical probabilities for π, $P(S_{t+1}|S_t)$, or $P(O_t|S_t)$. Your focus is on clearly defining the states and observations and explaining the structure of the model in the context of your chosen problem.

Consider a formulation of the **Part-Of-Speech tagging problem** below as a reference. Your formulation should follow the structure below, but need not be as detailed.

(a) **Identify a different real-world or plausible scenario** The scenario is the task of Part-of-Speech (POS) tagging a sentence. Given a sequence of words $W = w_1, w_2, ..., w_T$, we want to assign the most likely sequence of grammatical tags (like Noun, Verb, Adjective) $T = t_1, t_2, ..., t_T$, where each t_i is the POS tag for word w_i.

(b) **Briefly explain *why* your chosen scenario is suitable for HMM modeling** This scenario is suitable for an HMM because a sentence is a sequence $(w_1, ..., w_T)$ where each observable word $O_t = w_t$ is influenced by an unobservable, underlying Part-of-Speech tag $S_t = t_t$. The sequence of hidden tags is assumed to follow a Markov property, meaning the next tag S_{t+1} depends primarily on the current tag S_t. Both tag transitions ($P(S_{t+1}|S_t)$) and word emissions from a tag ($P(O_t|S_t)$) are probabilistic (explained more later).

(c) **Formulate the problem for your chosen scenario using the components of a Hidden Markov Model.**

i **Hidden States (S_t)**: The set of possible hidden states is the set of all possible Part-of-Speech tags defined for the language (e.g., {Noun, Verb, Adjective, Adverb, Determinant, Preposition, etc.}). The state S_t represents the true, unobservable POS tag of the t-th word in the sentence. These are hidden because we only observe the words, not their grammatical function directly without context.

ii **Observations (O_t)**: The set of possible observable emissions is the vocabulary of the language (all possible words). The observation O_t is the actual t-th word in the sentence that we observe.

iii **Initial State Distribution (π)**: This is a probability distribution over the set of POS tags, representing the likelihood that a sentence will start with a word having a particular POS tag ($P(S_1 = \text{tag})$). For instance, sentences often start with Determinants or Nouns.

iv **Transition Probabilities ($P(S_{t+1}|S_t)$)**: This is a matrix of probabilities where each entry $P(\text{tag}_j|\text{tag}_i)$ represents the likelihood that a word with POS tag tag_i will be followed by a word with POS tag tag_j. This models the grammatical structure of the language (e.g., a Determinant is likely followed by an Adjective or a Noun; a Verb might be followed by a Preposition or an Adverb). These transitions are non-deterministic because a given POS tag can be validly followed by multiple different POS tags with varying probabilities.

v **Emission Probabilities ($P(O_t|S_t)$)**: This is a matrix of probabilities where each entry $P(\text{word}_k|\text{tag}_i)$ represents the likelihood that if the hidden state (POS tag) at time t is tag_i, the observed word O_t will be word_k. This captures which words typically belong to which grammatical categories

(e.g., P("cat"|Noun) would be high, P("jumped"|Verb) would be high). These emissions are non-deterministic because a single word can often belong to multiple parts of speech (e.g., "well" as an adverb ("You played well"), adjective ("I am feeling well today.") or noun ("Bring some water from the well."))

5.5 Conditional independence *(12 pts)*

Consider the hidden Markov model (HMM) shown below, with hidden states S_t and observations O_t for times $t \in \{1, 2, \ldots, T\}$. State whether the following statements of conditional independence are true or false.

_____ $P(S_t|S_{t-1}) = P(S_t|S_{t-1}, S_{t+1})$

_____ $P(S_t|S_{t-1}) = P(S_t|S_{t-1}, O_{t-1})$

_____ $P(S_t|S_{t-1}) = P(S_t|S_{t-1}, O_t)$

_____ $P(S_t|O_{t-1}) = P(S_t|O_1, O_2, \ldots, O_{t-1})$

_____ $P(O_t|S_{t-1}) = P(O_t|S_{t-1}, O_{t-1})$

_____ $P(O_t|O_{t-1}) = P(O_t|O_1, O_2, \ldots, O_{t-1})$

_____ $P(S_2, S_3, \ldots, S_T|S_1) = \prod_{t=2}^{T} P(S_t|S_{t-1})$

_____ $P(S_1, S_2, \ldots, S_{T-1}|S_T) = \prod_{t=1}^{T-1} P(S_t|S_{t+1})$

_____ $P(S_1, S_2, \ldots, S_T|O_1, O_2, \ldots, O_T) = \prod_{t=1}^{T} P(S_t|O_t)$

_____ $P(S_1, S_2, \ldots, S_T, O_1, O_2, \ldots, O_T) = \prod_{t=1}^{T} P(S_t, O_t)$

_____ $P(O_1, O_2, \ldots, O_T|S_1, S_2, \ldots, S_T) = \prod_{t=1}^{T} P(O_t|S_t)$

_____ $P(O_1, O_2, \ldots, O_T) = \prod_{t=1}^{T} P(O_t|O_1, \ldots, O_{t-1})$