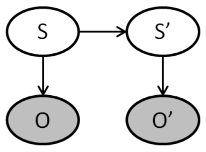
CS 228, Winter 2011-2012 Problem Set #1

This assignment is due at 12 noon on January 23. Submissions should be placed in the filing cabinet labeled "CS228 Homework Submission Box" located in the lobby outside Gates 187"

A Hidden Markov Model (HMM) is a dynamic Bayesian network with two variables for each time slice t: a state variable $S^{(t)}$ and an output variable $O^{(t)}$. In a standard HMM, the output variable is always observed for all time slices t, while the hidden state variable is never observed. The state at each time slice depends only on the state at the previous time slice (i.e., $S^{(t)}$ depends only on $S^{(t-1)}$), and the output at each time slice only depends on the state at that time slice ($O^{(t)}$ depends only on $S^{(t)}$). For the purposes of this exercise, we will assume that the variables $S^{(t)} \in \{s^1, \dots, s^K\}$ and $O^{(t)} \in \{o^1, \dots, o^N\}$, for all t, where K denotes the number of states, and N denotes the number of possible observations.

HMMs are defined by their **transition model** $P(S' \mid S)$ and **observation model** $P(O \mid S)$. We will use the variables f and g to represent these where necessary, i.e., $P(S' = s^j \mid S = s^i) = f_{ij}$ and $P(O = o^j | S = s^i) = g_{ij}$. In this question, we will consider only **stationary** transition and observation models, i.e., these models f and g are the same for all time steps t. We can thus represent a HMM with the following 2-TBN, where shaded nodes denote observed variables:



Despite their simplicity, HMMs are used extensively in real-world applications in which we suspect that there is an underlying sequence of hidden states which are generating the observed outcomes. For example, HMMs are the method of choice for speech recognition, with the hidden states representing the actual word that the speaker is saying, and the observed states representing the audio recording of the word. In this case, the transition model would be based on the language and context (e.g., in English, the word "San" might be very likely to transition to the word "Francisco").

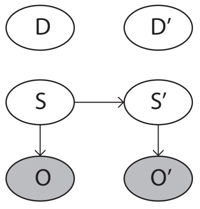
However, standard HMMs are often unable to represent more complex distributions. In this exercise, we will investigate a series of extensions to the standard Hidden Markov Model that allows it to encode a richer class of distributions, and apply these more expressive models to the problem of sequence alignment.

Problem 1

a) **State duration in HMMs. [5 points]** Consider an HMM with transition model f and observation model g. Assuming that the process is at state s^i at time t, what is the distribution over the number of steps until it first transitions out of state s^i (that is, the smallest number d_i such that $S^{(t+d_i)} \neq s^i$)?

b) **Duration HMMs. [10 points]** It can be restrictive if the distribution over state durations is always fixed to the form you calculated in the previous question. For some applications, we want a DBN model that allows us to incorporate an arbitrary distribution over the duration d_i that a process stays in state s^i after it first transitions to s^i. These models are called *duration HMMs.*

Modify the standard HMM to obtain a duration HMM. Your model should allow the distribution over d_i to depend on s^i. To ease notation, let $P(d_i = j | S = s^i) = h_{ij}$, and use f and g as before. Draw the 2-TBN corresponding to your model, and write out the CPDs for all variables with incoming edges. Do not worry about the distribution over starting states / the ground Bayesian network. To help you, we've provided a skeletal 2-TBN strucutre that is missing some edges.



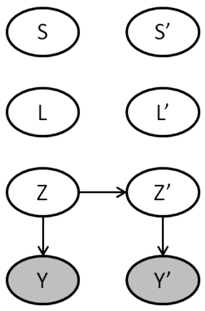
Add the appropriate edges, and describe the CPDs of all variables with incoming edges.

Note: In the above graph each time slice (instance of the variables O, S, and D) corresponds to the emission of a single observation o_i. Each of these observations is an emission of exactly one token.

c) **Segment HMMs. [10 points]** A *segment HMM* is a Markov chain that transitions over state $S^{(t)}$, but where each state emits not a single symbol as output, but rather a string of unknown length. Consider if our model starts to emit a new string at time t. In this case, it starts with some state $S^{(t)} = s$ and selects a segment length $L^{(t)}$, using a distribution that can depend on s. The model then emits a segment $Y^{(1)}, \dots, Y^{(L^{(t)})}$ of length $L^{(t)}$. Thus, the observed sequence can be thought of as a sequence of symbol subsequences:

$$Y^{(1,1)} \dots, Y^{(1,L^{(1)})} \dots Y^{(k,1)} \dots, Y^{(k,L^{(k)})}$$

where neither the L's nor the number k of total subsequences is given. In this exercise, we assume that the distribution on the output segment for a given state s is modeled by a separate HMM $\mathcal{H}_s$, with state variable Z and output variable Y. Note that the observed sequence is generated by this output variable Y, where each instance $Y^{(t)}$ represents a single symbol of output. Construct a 2-TBN model that encodes this model. For ease of notation, use f and g as required, and let $P(L^{(t)} = j | S^{(t)} = s^i) = h_{ij}$. To help you, we have provided a skeletal 2-TBN structure that is missing some of the edges:



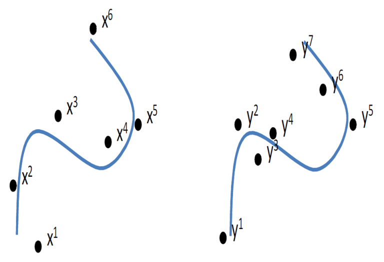
Add the appropriate edges, and describe the CPDs of the two variables S' and L'.

Note: The variables L and L' may not correspond exactly to L^t, consider how to use these variables to achieve the desired behavior.

d) **Duration HMMs vs. Segment HMMs. [5 points]** Let D be the set of distributions that duration HMMs can represent, and S be the same for segment HMMs. Is $D \subseteq S$, where "$\subseteq$" means "a subset of or equivalent to"? If so, describe how to represent a duration HMM as a segment HMM. If not, give a counter example.

e) **Sequence Alignment. [20 points]** Now, consider the following problem of *sequence alignment*. You are given two sequences $x^{(1)}, \ldots, x^{(n)}$ and $y^{(1)}, \ldots, y^{(m)}$, where $\boldsymbol{x}$ is a template and $\boldsymbol{y}$ is a noisy instance of this template. In this case, "noisy" means that elements of $\boldsymbol{x}$ can be duplicated, omitted completely, and/or included with noise added in $\boldsymbol{y}$. Additionally, we can have some spurious observations in $\boldsymbol{y}$ that are not even in $\boldsymbol{x}$. Thus, we can describe the influence of noise as two types of variations: temporal variations – X might move faster than Y through one part of the sequence or slower in another part; and state variations – even when two time points correspond to the same part of the maneuver, the values are unlikely to be exactly the same (since different people move differently).

We would like to *align* these sequences to each other, that is, to determine which $x^{(i)}$ corresponds to which $y^{(j)}$. Note that, due to temporal variation, a point $x^{(i)}$ may correspond to 0, 1, or more points $y^{(j)}$ (although it may be reasonable to have a prior on how large the rate discrepancy can get). Similarly, more than 1 point $x^{(i)}$ might correspond to the same point $y^{(j)}$. For example, in the figure below, the blue shows the average observed trajectory (irrelevant to the model), and the points show the actual X and Y trajectories. Here, we might want to infer a correspondence: $x^{(1)}$ to $y^{(1)}$; $x^{(2)}$ is unaligned; $x^{(3)}$ to all of $y^{(2)}, y^{(3)}, y^{(4)}$; $x^{(4)}$ to $y^{(5)}$; $x^{(5)}$ to $y^{(5)}$; $y^{(6)}$ is unaligned; and $x^{(6)}$ to $y^{(7)}$.



We assume that the relative orderings within each sequence are respected, i.e., if a particular $x^{(i)}$ is aligned to a particular $y^{(j)}$, then there is no alignment between any of $x^{(i+1)}, x^{(i+2)}, \ldots, x^{(n)}$

and any of $y^{(1)}, y^{(2)}, \ldots, y^{(j-1)}$, and likewise there is no alignment between any of $x^{(1)}, x^{(2)}, \ldots, x^{(i-1)}$ and any of $y^{(j+1)}, y^{(j+2)}, \ldots, y^{(m)}$.

Construct a probabilistic model for this problem using the model in part (c) as a starting point , i.e., come up with a segment HMM-like model that can assign a probability to any *alignment* between a given pair of sequences $x^{(1)}, \ldots, x^{(n)}$ and $y^{(1)}, \ldots, y^{(m)}$. Your model should include the following priors on the extent of both temporal and state variations, i.e., how likely is it that X goes through a segment K times as fast as Y, and a distribution on the differences in values between two aligned points $x^{(i)}$ and $y^{(j)}$:

- **Temporal Variation:** The number of points aligned to a single point in the X sequence should be a binomial distribution with fixed mean K and fixed variance p, where K represents how many times faster we are moving through X than Y. This captures the intuition that if we are moving through X K times as fast as Y then we expect K points from Y to align to a single point from X.
- **State variation:** Your model should also include a prior that encodes our belief that points with similar values (ie. positions in 2-D space) are more likely to be matched. This prior on whether $x^{(i)}$ matches $y^{(j)}$ should have the form of a Gaussian over the distance between $x^{(i)}$ and $y^{(j)}$, $d(x^{(i)}, y^{(j)})$ with some pre-determined variance σ^2.

Your answer should explicitly state all the variables in the your model, with a brief (2-3 sentence) description of each. Draw the 2-TBN for your model, and provide the CPDs for all variables with incoming edges. Explain how you can calculate the probability of a given pair of sequences and their alignment with your model. As before, you do not need to worry about the ground network / the initial state of the model.

Note: In devising your solution, you should use one of the sequences ($\boldsymbol{x}$) to define an HMM-style probabilistic model (like the one used in part (c)) that generates the other sequence ($\boldsymbol{y}$). In doing so, you may want to abstract-away the actual values of the $\boldsymbol{x}$ and $\boldsymbol{y}$ and focus on generation of matches between the indices in each sequence rather than the values directly. It is possible that not all nodes found in part (c) will be necessary to create such a model.