Graphical Models

As explained in the note *Conditional probabilities and graphical*, a graphical model is a graphical notation to describe the dependency relationships when specifying a joint probability.

From graph to joint probability

**Exercise 1**: For the following four graphs, write down the joint probability of the random variables.

Solution:

\[
p(X)p(Y)p(Z \mid X, Y)
\]

\[
p(X)p(Y \mid X)p(Z \mid X, Y)
\]

\[
p(X)p(Y \mid X)p(Z \mid X)p(W \mid Y, Z)
\]

\[
p(Z_1) \prod_{i=1}^{5} p(X_i \mid Z_i) \prod_{i=2}^{5} p(Z_i \mid Z_{i-1})
\]

From joint probability to graph
**Exercise 2:** Draw the following four joint probabilities as dependency graphs:

\[ p(X)p(Y)p(Z) \]

\[ p(X)p(Y \mid X)p(Z \mid X) \]

\[ p(X)p(Y \mid X)p(Z \mid Y)p(W \mid X, Z) \]

\[ p(Z_1)p(X_1 \mid Z_1) \prod_{i=2}^{5} p(X_i \mid Z_i, X_{i-1}) \prod_{i=2}^{5} p(Z_i \mid Z_{i-1}) \]

**Solutions:**
Hidden Markov Models
**Exercise 3:** Questions to slides *Hidden Markov Models - Terminology, Representation and Basic Problems:*

1. How much time does it take to compute the joint probability $P(X, Z|\Theta)$ in terms of $N$ and $K$, where $X = x_1, \ldots, x_N$, $Z = z_1, \ldots, z_N$, and $K$ is the number of hidden states in the hidden Markov model $\Theta$?

   **Solution:**
   
   The computation consists of $O(N)$ multiplications of factors that we can look up in constant time, i.e. the running time would be $O(N)$.

1. How many terms are there in the sum on slide 34 for computing $P(X|\Theta)$? Why?

   **Solution:**
   
   We sum over all possible sequences of hidden states $Z = z_1, \ldots, z_N$, where each $z_i$ can have $K$ values, so there are $K^N$ terms in the sum.

1. How many terms are there in the maximization on slide 38 for computing the Viterbi decoding $Z^*$? Why?

   **Solution:**
   
   We maximize over all possible sequences of hidden states $Z = z_1, \ldots, z_N$, where each $z_i$ can have $K$ values, so there are $K^N$ terms in maximization.

1. How many terms are there in the maximization on slide 39 for computing $z_n^*$, i.e. the $n$th state in a posterior decoding? Why?

   **Solution:**
   
   We maximize over the possible values of $z_n$, so we maximize over $K$.

**Exercise 4:** Questions to slides *Hidden Markov Models - Algorithms for decoding:*

1. Where in the derivation of $\omega(z_n)$ on slide 7 do we use that the fact that we are working with hidden Markov models? And how do we use it?

   **Solution:**
   
   We use it to rewrite the joint probability $p(x_1, \ldots, x_n, z_1, \ldots, z_n)$ as $p(z_1) \prod_{i=2}^{n} p(z_i|z_{i-1}) \prod_{i=1}^{n} p(x_i|z_i)$.

1. Where in the derivation of $p(z_n | x_1, \ldots, x_N)$ on slide 16 do we use the fact that we are working with hidden Markov models? And how do we use it?

   **Solution:**
We use it to rewrite/simplify the probability $p(x_{n+1}, \ldots, x_N|z_n, x_1, \ldots, x_n)$ to the probability $p(x_{n+1}, \ldots, x_N|z_n)$, i.e. to remove $x_1, \ldots, x_n$ from what we condition on. We can do this because $x_1, \ldots, x_n$ and $x_{n+1}, \ldots, x_N$ become independent when $X$ and $Z$ depend on the each other as they do in an HMM and we condition on $z_n$.

1. Where in the derivation of $\alpha(z_n)$ and $\beta(z_n)$ on slide 20 and 26 do we use that the fact that we are working with hidden Markov models? And how do we use it?

**Solution:**

On slide 20, we use it to rewrite the joint probability $p(x_1, \ldots, x_n, z_1, \ldots, z_n)$ as $p(z_1) \prod_{i=2}^{n} p(z_i|z_{i-1}) \prod_{i=1}^{n} p(x_i|z_i)$.

On slide 26, we use it to rewrite the joint probability $\sum_{z_{n+1},\ldots,z_N} p(x_{n+1}, \ldots, x_N, z_n, z_{n+1}, \ldots, z_N)$ as $\sum_{z_{n+1},\ldots,z_N} p(z_n) \prod_{i=n+1}^{N} p(z_i|z_{i-1}) \prod_{i=n+1}^{N} p(x_i|z_i)$.

1. Why is $P(X) = \sum_{z_n} \alpha(z_n)\beta(z_n) = \sum_{z_N} \alpha(z_N)$ as stated on slide 31?

**Solution:**

$\alpha(z_n)\beta(z_n) = p(x_1, \ldots, x_n, z_n)p(x_{n+1}, \ldots, x_N|z_n) = p(x_1, \ldots, x_n, x_{n+1}, \ldots, x_N, z_n) = p(X, z_n)$. Summing this probability over all $K$ possible values of $z_n$ yields $p(X)$. Similarly, $\alpha(z_N) = p(x_1, \ldots, x_N, z_N) = p(X, z_N)$, and summing over all $K$ possible values of $z_N$ yields $p(X)$.

1. Algorithmic question: Slide 35 shows how to compute $P(X)$ from $\alpha(z_N)$ in time $O(K^2 N)$, i.e. the time it takes to compute the last (rightmost) column in the $\alpha$-table. How much space do you need to compute this column? Do you need to store the entire $\alpha$-table?

**Solution:**

In the forward algorithm, we compute column $n$ in the $\alpha$-table from column $n - 1$. If we in the end only need access to column $N$, then we only need to keep two columns in memory when we compute the $\alpha$-table column by column from left to right, namely the current column $n$, and the previous column $n - 1$. 

In [ ]: