10-701 Introduction to Machine Learning

Homework 5

Due Dec 1, 11:59 am

Rules:

1. Homework submission is done via CMU Autolab system. Please compile your writeup and code in a single pdf file and submit to https://autolab.cs.cmu.edu/courses/10701-f15. Please let us know asap if you have trouble accessing Autolab.
2. Like conference websites, repeated submission is allowed. So please feel free to refine your answers. We will only grade the latest version.
3. Autolab may allow submission after the deadline, note however it is because of the late day policy. Please see course website for policy on late submission.
4. Please typeset your homework using appropriate software such as LaTeX. We will not accept scanned copies of handwritten papers.
5. You are allowed to collaborate on the homework, but you should write up your own solution and code. Please indicate your collaborators in your submission.

---

1 Gaussian Graphical Model (10 Points) (Hao)

In this problem, we will explore the connections between multivariate Gaussian distribution and undirected graphical models.

Given $\boldsymbol{X} = \{X_1, \ldots, X_d\} \in \mathbb{R}^d$ and $P(\boldsymbol{X}|\mu, \Sigma) = \mathcal{N}(\mu, \Sigma)$, and let $\Omega = \Sigma^{-1}$ be the precision matrix (inverse covariance matrix). We can represent the distribution as an undirected graph. Specifically, let $G = (V, E)$ be a (complete) graph with $V = \{1, 2, \ldots, d\}$.

1. Following the notations of undirected graphical models, we can write

$$P(\mathbf{X}|\mu, \Sigma) \propto \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j) \prod_{i \in V} \psi_i(x_i)$$

Derive $\psi_{ij}(x_i, x_j)$ and $\psi_i(x_i)$.

Answer:

$$\begin{aligned} \psi_{ij}(x_i, x_j) &= \exp(-\frac{1}{2}\Omega_{ij}x_i x_j) \\ \psi_i(x_i) &= \exp(-\frac{1}{2}\Omega_{ii}x_i^2 + x_i \sum_j \mu_j \Omega_{ij}) \end{aligned} \tag{1}$$

2. Also, we can rewrite the above as a product of factors with respect to edges, i.e.,

$$P(\mathbf{X}|\mu, \Sigma) \propto \prod_{(i,j)\in E} \psi'_{ij}(x_i, x_j)$$

where $\psi'_{ij}(x_i, x_j) = \psi_{ij}(x_i, x_j) \times \psi_i(x_i)^{\frac{1}{n(i)}} \times \psi_j(x_j)^{\frac{1}{n(j)}}$ and $n(i)$ be the number of neighbors of i. Use this formulation to argue the following equivalence,

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j | X_{V\setminus\{i,j\}}$$

Answer:

Since we take the product over all edges and the product over all nodes, we can incorporate the product over all nodes in the product over all edges, while being careful for "double-counting". Note that if a node has n neighbors (i.e., n edges involving that node), that means that if we incorporate the term $\psi(x_i)$ corresponding to that node in the edges product, we would need to take it's nth root, because it would appear in n terms in that product (i.e., one term for each edge in which it appears). Therefore, we can see that:

$$P(\mathbf{X}|\mu, \Sigma) \propto \prod_{(i,j)\in E} \psi_{i,j}(x_i, x_j)\psi_i(x_i)^{\frac{1}{n(i)}}\psi_j(x_j)^{\frac{1}{n(j)}}$$

which is the form that is provided to us in the problem sheet. Now, we see that if $(i, j) \in E$, then in order for the value of this density to remain unchanged we need to have that $\Omega_{ij} = 0$ (i.e., so that all the terms of the product term corresponding to that edge are equal to 1). It is not difficult to see then that given $X_{V\setminus\{i,j\}}$ and $\Omega_{ij} = 0$, there is not "coupling" between the terms involving X_i and X_j and the conditional probability density function factorizes with respect to these two variables, implying that $X_i \perp X_j | X_{V\setminus\{i,j\}}$. Furthermore, following the same way of reasoning, if $X_i \perp X_j | X_{V\setminus\{i,j\}}$ we need to have no coupling between the terms involving X_i and X_j in the above equations. This means that (i, j) must not be in E. Thus, we have argued (but not proved formally as this was not required from the problem statement) that $(i, j) \notin E \Leftrightarrow X_i \perp X_j | X_{V\setminus\{i,j\}}$.

2 Sampling Methods (35 Points)(Hao)

2.1 Inverse sampling

1. One of the most widely used sampling methods is the *inverse sampling*. To sample from a distribution $p(y)$, we first sample a random variable z from the uniform distribution over $(0, 1)$, then transform z using $y = h^{-1}(z)$ where $h(y)$ is defined as,

$$h(y) = \int_{-\infty}^{y} p(\hat{y})d\hat{y}$$

Prove that this yields a sample y which follows the distribution $p(y)$. Give one drawback of the inverse sampling method.

Answer:

As $h(x)$ is a cdf, it must be non-decreasing in $[0, 1]$. Then,

$$\begin{aligned} P(X \leq x) &= P(h^{-1}(z) \leq x) \\ &= P(z \leq h(x)) \\ &= h(x) \end{aligned}$$

which indicates that x follows the cdf $h(x)$.

This method won't work well in the following two situations:

- It's hard to get the CDF function $F(x)$ while it may be easy to know pdf $f(x)$.
- Even when we know $F(x)$, it may be quite difficult to get its inverse function $F^{-1}(x)$.

2. Given a random variable z which is uniformly distributed over $(0, 1)$, find a transformation $y = g(z)$ such that y has a Cauchy distribution given by $y \sim \frac{1}{\pi}\frac{1}{1+y^2}$.

$$g(y) = h^{-1}(y) = \tan(\pi y - \frac{\pi}{2})$$

2.2 Rejection sampling

Sometimes directly drawing samples from the desired distribution $p(z)$ is hard, but the value of $\tilde{p}(x)$ can readily be evaluated, where $p(x) = \frac{1}{Z_{\tilde{p}}}\tilde{p}(x)$ with $Z_{\tilde{p}}$ as a normalization constant. In this case, we refer to another sampling method called *rejection sampling*.

The procedures of rejection sampling are described as follows: we first introduce a proposal distribution $q(z)$, from which we can readily draw samples. We next introduce a (smallest) constant k whose value is chosen such that $kq(z) \geq \tilde{p}(z)$ for all values of z. To sample from $p(z)$, we first generate a number z_0 from the distribution $q(z)$. Next, we generate a number u_0 from the uniform distribution over $[0, kq(z_0)]$. Finally, if $u_0 > \tilde{p}(z_0)$ then the sample is rejected, otherwise z_0 is retained. Write down the probability that given z_0 that z_0 is accepted, and further prove that the above procedure yields a sample z_0 which follows the distribution $p(z)$. Also, give one drawback of this method.

Answer:

$$P(\text{accept}|z_0) = \frac{\tilde{p}(z_0)}{kq(z_0)}$$

$$
\begin{aligned}
p(z = z_0|\text{accept } z) &= \frac{p(z = z_0, \text{accept } z)}{p(\text{accept } z)} \\
&= \frac{p(\text{accept}|z_0)p(z = z_0)}{\int_z p(z, \text{accept } z)} \\
&= \frac{\frac{\tilde{p}(z_0)}{kq(z_0)}q(z_0)}{\int_z \frac{\tilde{p}(z_0)}{kq(z_0)}q(z_0)} \\
&= \frac{Z_{\tilde{p}}p(z_0)}{Z_{\tilde{p}}\int_z p(z)dz} \\
&= p(z_0)
\end{aligned}
$$

Drawback: To get a desired sample, one usually needs to draw many unwanted samples.

2.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps.

1. *Metropolis-Hastings* (MH) algorithm is a MCMC method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. This sequence can be used to approximate the distribution. Similar to the rejection sampling, MH assumes that $\tilde{p}(x)$, the unnormalized target distribution, is easy to evaluate, and samples from the target distribution $p(x)$ using the proposal distribution q. In particular, at step τ, in which the current state is $z^{(\tau)}$, we draw z^* from the distribution $q(z|z^{(\tau)})$ and accept it with probability

$$
A = \min\left(1, \frac{\tilde{p}(z^*)q(z^{(\tau)}|z^*)}{\tilde{p}(z^{(\tau)})q(z^*|z^{(\tau)})}\right)
$$

Repeating the above procedures thus forms a Markov chain. Show that $p(x)$ is the stationary distribution of the Markov chain. Once again, describe a major drawback of MH algorithm. (Hint: See the pp.29 in the slides of lecture 18 about how to prove $p(x)$ is a stationary distribution of a Markov chain).

Answer: The transition kernel of the Markov chain is defined as:

$$
K(z^*|z^{(\tau)}) = q(z^*|z^{(\tau)}) \min(1, \frac{\tilde{p}(z^*)q(z^{(\tau)}|z^*)}{\tilde{p}(z^{(\tau)})q(z^*|z^{(\tau)})}) \tag{2}
$$

We investigate $p(z^{(\tau)})K(z^*|z^{(\tau)})$:

$$
\begin{aligned}
p(z^{(\tau)})K(z^*|z^{(\tau)}) &= p(z^{(\tau)})q(z^*|z^{(\tau)}) \min(1, \frac{\tilde{p}(z^*)q(z^{(\tau)}|z^*)}{\tilde{p}(z^{(\tau)})q(z^*|z^{(\tau)})}) \\
&= p(z^{(\tau)})q(z^*|z^{(\tau)}) \min(1, \frac{p(z^*)q(z^{(\tau)}|z^*)}{p(z^{(\tau)})q(z^*|z^{(\tau)})}) \\
&= \min(p(z^{(\tau)})q(z^*|z^{(\tau)}), p(z^*)q(z^{(\tau)}|z^*)) \\
&= p(z^*)q(z^{(\tau)}|z^*) \min(\frac{p(z^{(\tau)})q(z^*|z^{(\tau)})}{p(z^*)q(z^{(\tau)}|z^*)}, 1) \\
&= p(z^*)q(z^{(\tau)}|z^*) \min(\frac{\tilde{p}(z^{(\tau)})q(z^*|z^{(\tau)})}{\tilde{p}(z^*)q(z^{(\tau)}|z^*)}, 1) \\
&= p(z^*)p(z^{(\tau)}|z^*)
\end{aligned} \tag{3}
$$

For the discrete case, we have:

$$\sum_{z^{(\tau)}} K(z^{(\tau)})p(z^*|z^{(\tau)}) = \sum_{z^{(\tau)}} K(z^*)p(z^{(\tau)}|z^*) = p(z^*) \tag{4}$$

For the continuous case, we have:

$$\int_{z^{(\tau)}} K(z^{(\tau)})p(z^*|z^{(\tau)}) = \int_{z^{(\tau)}} K(z^*)p(z^{(\tau)}|z^*) = p(z^*) \tag{5}$$

Proved.

The drawback is that the accepted probability may be quite slow every time, need many trials to converge.

2. In class we showed the procedures of Gibbs sampling for target distribution $p(\boldsymbol{x}) = p(x_1, \cdots, x_d)$: for each $j \in \{1, \cdots, d\}$, draw $t \sim p(x_j|\text{rest})$ and set $x_j = t$. Show that $p(\boldsymbol{x})$ is the stationary distribution of the Markov chain defined by this procedure.

   Answer:

   The transition kernel of Markov chain in Gibbs sampling is defined as:

$$K(\boldsymbol{x}'|\boldsymbol{x}) = p(x_1'|x_2, \ldots, x_d) \times p(x_2'|x_1', \ldots, x_d) \cdots \times p(x_d'|x_1', \cdots, x_{d-1}') \tag{6}$$

   Investigating $\int p(\boldsymbol{x}'|\boldsymbol{x})p(x)d\boldsymbol{x}$,

$$\begin{aligned}
\int K(\boldsymbol{x}'|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} &= \int p(x_1'|x_2, \ldots, x_d) \cdots \times p(x_2, \cdots, x_d)p(x_1|x_2, \cdots, x_d)dx_1 \cdots dx_d \\
&= \int p(x_2'|x_1', \cdots, x_d) \cdots \times p(x_1', x_2, \cdots, x_d)dx_2 \cdots dx_d \\
&= \int p(x_2'|x_1', \cdots, x_d) \cdots \times p(x_2|x_1', \cdots, x_d)dx_2 p(x_1', \cdots, x_d) \cdots dx_d \\
&= \cdots \\
&= p(\boldsymbol{x}')
\end{aligned} \tag{7}$$

   Same for the discrete case.

   Proved.

3. Finally, prove that Gibbs sampling is a special case of the MH algorithm. Will Gibbs sampling suffer the drawback you described for MH algorithm?

   Answer:

   Gibbs sampling is obviously a special case of MH algorithm by setting the acceptance probability to 1. It does not suffer the drawback because in every iteration Gibbs sampling will approach the stationary distribution as the acceptance probability is 1.

3 Expectation Maximization (EM) and Variational Inference (VI) (35 Points) (Zhiting)

Here we give a general treatment of the EM algorithm, and introduce VI which can be derived similarly [1].

Hint: the Kullback-Leibler (KL) divergence between two probability distributions $q(\boldsymbol{x})$ and $p(\boldsymbol{x})$ is defined as $KL(q\|p) = -\int_{\boldsymbol{x}} q(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$; and we have $KL(q\|p) \geq 0$ and $=$ holds if and only if $q = p$.

[1] PGM (10708) will cover more advanced topics on sampling and VI. You are welcome to take the course :)

3.1 EM

Let $\boldsymbol{x}$ be observed variables, $\boldsymbol{z}$ latent variables, and θ the parameters. The EM algorithm maximizes the likelihood function $p(\boldsymbol{x}|\theta) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$ through a two-stage iterative optimization.

1. Introduce a distribution $q(\boldsymbol{z})$ over the latent variables. Show that for any choice of $q(\boldsymbol{z})$, the following decomposition holds

$$\ln p(\boldsymbol{x}|\theta) = \mathcal{L}(q, \theta) + KL(q\|p), \tag{8}$$

where

$$\mathcal{L}(q, \theta) = \int_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \left\{ \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z})} \right\},$$
$$KL(q\|p) = -\int_{\boldsymbol{z}} q(\boldsymbol{z}) \ln \left\{ \frac{p(\boldsymbol{z}|\boldsymbol{x}, \theta)}{q(\boldsymbol{z})} \right\},$$

so that $\mathcal{L}(q, \theta)$ is the lower bound on $\ln p(\boldsymbol{x}|\theta)$.

Answer

We have $\ln p(\boldsymbol{x}, \boldsymbol{z}|\theta) = \ln p(\boldsymbol{z}|\boldsymbol{x}, \theta) + \ln p(\boldsymbol{x}|\theta)$. Plug this into the expression for $\mathcal{L}(q, \theta)$. This give rise to two terms, one of which cancels $KL(q\|p)$ while the other gives the required log likelihood $\ln p(\boldsymbol{x}|\theta)$ after noting that $q(\boldsymbol{z})$ is a normalized distribution that sums to 1.

2. In the E-step, the lower bound $\mathcal{L}(q, \theta)$ is maximized w.r.t $q(\boldsymbol{z})$ while fixing θ. Show that $\mathcal{L}(q, \theta)$ is maximized when $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta)$.

Answer

From Eq (8) we have $\mathcal{L}(q, \theta) \leq \ln p(\boldsymbol{x}|\theta)$ because $KL(q\|p) \geq 0$. Note that $\ln p(\boldsymbol{x}|\theta)$ does not depend on $q(\boldsymbol{z})$ and so $\mathcal{L}(q, \theta)$ is maximized when $KL(q\|p) = 0$, which occurs when $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta)$.

3. In the subsequent M-step, the lower bound $\mathcal{L}(q, \theta)$ is maximized w.r.t θ while fixing $q(\boldsymbol{z})$. This step will necessarily increase the log likelihood $\ln p(\boldsymbol{x}|\theta)$. Explain why.

Answer

(As long as we have not reached the maximum), the lower bound $\mathcal{L}(q, \theta)$ is increased given a new θ. On the other hand, the distribution q is fixed while the posterior $p(\boldsymbol{z}|\boldsymbol{x}, \theta)$ is changed due to the updated θ, so there will be a nonzero $KL(q\|p)$. According to Eq (8), $\ln p(\boldsymbol{x}|\theta)$ is increased.

3.2 VI

For simplicity, here we omit the parameter θ and only consider the latent variables $\boldsymbol{z}$ and the observed variables $\boldsymbol{x}$. We repeat Eq(8) as follows

$$\ln p(\boldsymbol{x}) = \mathcal{L}(q) + KL(q\|p).$$

From the above we know the lower bound $\mathcal{L}(q)$ is maximized when $q(\boldsymbol{z})$ equals the posterior $p(\boldsymbol{z}|\boldsymbol{x})$. However sometimes the true posterior is intractable so we have to resort to some approximation, one popular technique of which is the variational inference. The main idea is to restrict the family of distributions $q(\boldsymbol{z})$ and find the member of this family for which the lower bound $\mathcal{L}(q)$ is largest.

Let $\boldsymbol{z} = \{z_1, \ldots, z_K\}$. Here we restrict the q family by assuming the distribution factorizes as

$$q(\boldsymbol{z}) = \prod_{k=1}^{K} q_k(z_k).$$

To find the optimal q in this family, we maximize $\mathcal{L}(q)$ w.r.t each of the factors in turn.

1. Show that for any $k \in \{1, \ldots, K\}$,

$$\mathcal{L}(q) = \int q_k(z_k)\mathbb{E}_{j\neq k}[\ln p(\boldsymbol{x}, \boldsymbol{z})]dz_k - \int q_k(z_k)\ln q_k(z_k)dz_k + const,$$

where $\mathbb{E}_{j\neq k}[\ln p(\boldsymbol{x}, \boldsymbol{z})]$ denotes the expectation of $\ln p(\boldsymbol{x}, \boldsymbol{z})$ w.r.t the q distribution over all z_j for $j \neq k$ (i.e., $\prod_{j\neq k} q_j$); and *const* is some constant irrelevant to q.

Answer

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_j q_j \left\{\ln p(\boldsymbol{x}, \boldsymbol{z}) - \sum_j \ln q_j\right\} d\boldsymbol{z} \\ &= \int q_k \left\{\int \ln p(\boldsymbol{x}, \boldsymbol{z}) \prod_{j\neq k} q_j dz_j\right\} dz_k - \int q_k \ln q_k dz_k + const \\ &= \int q_k \mathbb{E}_{j\neq k}[\ln p(\boldsymbol{x}, \boldsymbol{z})]dz_k - \int q_k \ln q_k dz_k + const,\end{aligned} \tag{9}$$

2. Keeping $\{q_j\}_{j\neq k}$ fixed, show that the optimal $q_k(z_k)$ is given by

$$\ln q_k^*(z_k) = \mathbb{E}_{j\neq k}[\ln p(\boldsymbol{x}, \boldsymbol{z})] + const.$$

Answer

Let $\ln \tilde{p} = \mathbb{E}_{j\neq k}[\ln p(\boldsymbol{x}, \boldsymbol{z})] + const$. Note that Eq (9) is a negative KL divergence between $q_k(z_k)$ and $\tilde{p}$. So $q_k^* = \tilde{p}$, and $\ln q_k^* = \mathbb{E}_{j\neq k}[\ln p(\boldsymbol{x}, \boldsymbol{z})] + const$.

4 HMM (20 Points) (Zhiting)

1.1 Consider a HMM with 6 states (plus a start and end states) and an alphabet $\{A, C, G, T\}$. Table 1 lists the transition and emission probabilities, and Figure 1 shows the state diagram.

	0	S_1	S_2	S_3	S_4	S_5	S_6	f	A	C	G	T
0	0	1	0	0	0	0	0	0				
S_1	0	0	1	0	0	0	0	0	0.2	0.3	0	0.5
S_2	0	0	0	0.3	0	0.7	0	0	0.6	0.1	0.2	0.1
S_3	0	0	0	0	1	0	0	0	0.7	0	0.1	0.2
S_4	0	0	0	0	0	0	0	1	0.2	0.3	0.4	0.1
S_5	0	0	0	0	0	0	1	0	0.3	0.3	0.3	0.1
S_6	0	0	0	0	0	0	0	1	0.5	0.3	0	0.2

Table 1: The transition and emission probabilities.

Let z denote latent variables and x denote observed variables. Place $<$, $>$, or $=$ between the two components of each of the following pairs. Justify your answer.

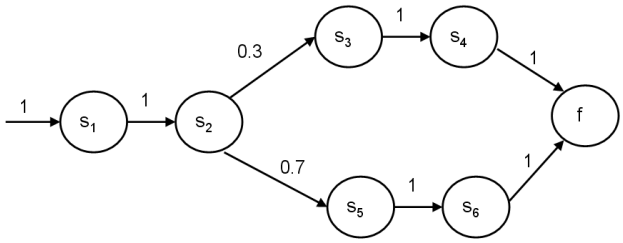

Figure 1: The state diagram of the HMM.

1. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_3 = S_3, z_4 = S_4)$
   $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4)$

   Answer: $<$

$$P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_3 = S_3, z_4 = S_4)$$
$$= P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4) P(z_3 = S_3, z_4 = S_4),$$
$$P(z_3 = S_3, z_4 = S_4) = 0.3 < 1.$$

2. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T)$
   $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4)$

   Answer: $<$

$$P(x_1 = T, x_2 = C, x_3 = A, x_4 = T)$$
$$= P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_3 = S_3, z_4 = S_4) + P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_3 = S_5, z_4 = S_6)$$
$$= P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4) * P(z_3 = S_3, z_4 = S_4)$$
$$+ P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_5, z_4 = S_6) * P(z_3 = S_5, z_4 = S_6)$$
$$= (0.7 * 0.1) * 0.3 + (0.3 * 0.2) * 0.7.$$

   Note that

$$P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4) = 0.07 > P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_5, z_4 = S_6) = 0.06.$$

   From the above we conclude that

$$P(x_1 = T, x_2 = C, x_3 = A, x_4 = T) < P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_3 = S_3, z_4 = S_4)$$

3. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_1 = S_1, z_2 = S_2)$
   $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_1 = S_1, z_2 = S_2)$

   Answer: $=$

$$P(x_1 = T, x_2 = C, x_3 = A, x_4 = T, z_1 = S_1, z_2 = S_1)$$
$$= P(x_1 = T, x_2 = C, x_3 = A, x_4 = T | z_1 = S_1, z_2 = S_2) P(z_1 = S_2, z_1 = S_2),$$
$$P(z_1 = S_2, z_1 = S_2) = 1.$$

4. $P(x_1 = T, x_2 = C, x_3 = A, x_4 = T)$
   $P(x_1 = T, x_2 = A, x_3 = A, x_4 = G)$

Answer: $<$

Since the first and third letters are the same, we only need to worry about the second and fourth. The left hand side is: $0.1 * (0.3 * 0.1 + 0.7 * 0.2) = 0.017$ while the right hand side is: $0.6(0.7 * 0 + 0.3 * 0.4) = 0.072$

1.2 Prove that $p(x_1, \ldots, x_i, z_i) = p(x_i|z_i) \sum_{z_{i-1}} p(x_1, \ldots, x_{i-1}, z_{i-1}) p(z_i|z_{i-1})$.

Answer:

From the definition of the hidden Markov model (HMM) we have that:

$$P(x_1, \ldots, x_i, z_1, \ldots, z_i) = P(x_i \mid z_i) P(z_i \mid z_{i-1}) P(x_1, \ldots, x_{i-1}, z_1, \ldots, z_{i-1}).$$

We can now marginalize out the variables $z_1, \ldots, z_{i-1}$ to obtain the following:

$$\begin{aligned}
P(x_1, \ldots, x_i, z_i) &= \sum_{z_1, \ldots, z_{i-1}} P(x_1, \ldots, x_i, z_1, \ldots, z_i), \\
&= \sum_{z_1, \ldots, z_{i-1}} P(x_i \mid z_i) P(z_i \mid z_{i-1}) P(x_1, \ldots, x_{i-1}, z_1, \ldots, z_{i-1}), \\
&= P(x_i \mid z_i) \sum_{z_1, \ldots, z_{i-1}} P(z_i \mid z_{i-1}) P(x_1, \ldots, x_{i-1}, z_1, \ldots, z_{i-1}), \\
&= P(x_i \mid z_i) \sum_{z_{i-1}} P(z_i \mid z_{i-1}) \sum_{z_1, \ldots, z_{i-2}} P(x_1, \ldots, x_{i-1}, z_1, \ldots, z_{i-1}), \\
&= P(x_i \mid z_i) \sum_{z_{i-1}} P(z_i \mid z_{i-1}) P(x_1, \ldots, x_{i-1}, z_{i-1}),
\end{aligned}$$

The proof completes.

5 Bayesian Networks (Bonus: 10 Points) (Zhiting)

Given n random variables labeled as $1, 2, \ldots, n$, how many Bayesian networks (which is DAG) can these variables form? Give a lower bound and upper bound, respective, as tight as you can. Justify your answer.

Example answers: (Note: you have to explain your answer rather than just give a number or a reference...)

1. Combinatorial enumeration of DAG:
   https://en.wikipedia.org/wiki/Directed_acyclic_graph#Combinatorial_enumeration

Upper bounds

1. $3^{n(n-1)/2}$: $n(n-1)/2$ pairs; for each pair (i, j), there are at most 3 possible link states, i.e., (1) no link; (2) $i \to j$; (3) $i \leftarrow j$.

Lower bounds

1. $2^{n(n-1)/2}$: $n(n-1)/2$ pairs; for each pair (i, j), there are at least 2 possible link states, i.e., (1) no link; (2) $i \to j$ or $i \leftarrow j$ (it can be proved by contradiction that given a DAG over the n variables, if adding $i \to j$ leads to circles, then adding $i \leftarrow j$ will not lead to circles).