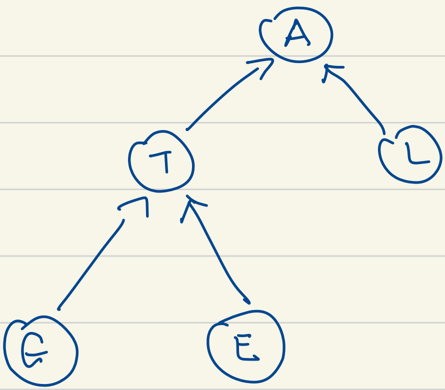
# Question 1

1) A: Boolean (discrete) T: Continuous. L: discrete.
G: discrete E: Continuous.




2) Factorization:

$P(A, T, L, G, E)$
$= P(G) P(E) P(L) P(T \mid G, E) P(A \mid T, L)$

3) In my graph, G, L are discrete random variables with no parents. Use educated guess.

$P(L=M) = 0.4$ $P(G=l) = 0.5$.

$P(L=0) = 0.6$. $P(G=d) = 0.5$.

E is a continuous random variable with no parents. Could be described using linear Gaussians.

NB: In 2018 BC's GDP is 295,401 million CAD.

So I guess $N(295, 20)$ measured in billion CAD.

T is a continuous random variable with one discrete parent G, one continuous parent E.

Using $p(x_i | pa_i) = \mathcal{N}(x_i | \sum_{j \in pa_i} w_{ij} x_j + b_i, v_i)$ (8.11)

$$P(T=t | G=l, E=e) = \mathcal{N}(t | w_g l + b_g + w_e e + b_e, v_t)$$

$$P(T=t | G=d, E=e) = \mathcal{N}(t | w_g d + b_g + w_e e + b_e, v_t).$$

A is a discrete R.V with one continuous parent T and one discrete parent L.

Could be described using sigmoid.

$$P(y=1 | x_1 \cdots x_M) = \sigma\left(w_0 + \sum_{j=1}^{M} w_i x_i\right) = \sigma(w^T x) \quad (8.10)$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

$$P(A | L, T, E) = \frac{1}{1 + \exp(-w^T x)}$$

4) $X \in \{ X_1, X_2, X_3 \dots X_N \}$

$x_n = (a_n, l_n, g_n, e_n, t_n)$

The likelihood function can be denoted as

$$L(X|\theta) = P(X_1|\theta) \cdot P(X_2|\theta) \cdots P(X_N|\theta)$$

$$= \prod_{n=1}^{N} P(X_n|\theta)$$

To calculate each $P(X_n|\theta), n \in N$ we need to use the factorization formula.

$$p(x) = \prod_{k=1}^{K} p(x_k | pa_k) \qquad (8.5)$$

Therefore, the likelihood function has become.

$$L(X|\theta) = \prod_{k=1}^{K} \prod_{n=1}^{N} P(X_n^k | pa_n^k, \theta)$$

NB: K here is correspond to (a, l, g, e, t)

We can convert expression of node a. l. g. e. t to integers as we saw in the text book $x_1$ $x_2$ $x_3$ ....

Hence the maximum likelihood becomes.

$$\arg\max_{\theta} \prod_{n=1}^{N} p(x_n|\theta)$$

$$\arg\max_{\theta} \prod_{k=1}^{K} \prod_{n=1}^{N} p(x_n^k \,|\, pa_n^k, \theta).$$

As probability "p" can not be negative. To maximize. the product, we can just maximize each factor in produce. In other words

$$\arg\max_{\theta_k} \prod_{n=1}^{N} p(x_n^k \,|\, pa_n^k, \theta_k)$$

This way can also makes it easier for us to learn parameters for $p(A|pa_A)$.
We just need to keep data related to nodes A, T, L. which are $a_n$. $l_n$, $t_n$, since here $\theta$. $\xi$ wouldn't affect T (since T is given as input).

Also, according to "ancestral sampling" from the textbook.
" To obtain a sample from some marginal distribution corresponding to a subset of the variables, we simply take the sampled values from the requested

nodes and ignore the sampled values from the remaining nodes."

## Question 2.

1) No, it's not divergence symmetric.

2) $$D_{KL}(P\|P) = \int P(x) \ln \frac{P(x)}{P(x)} dx$$

$$= \int P(x) (\ln P(x) - \ln P(x)) dx. = \int 0 \, dx = 0.$$

3) $\ln(1+x) \leq x.$

$$\ln(1+(P(x)-1)) \leq P(x) - 1$$

$$\ln(P(x)) \leq P(x) - 1 \qquad (eq.1).$$

$$-D_{KL}(P\|Q) = -\int P(x) \ln \frac{P(x)}{Q(x)} dx$$

$$= \int P(x) \ln\left(\left(\frac{P(x)}{Q(x)}\right)^{-1}\right) dx.$$

$$= \int P(x) \ln\left(\frac{Q(x)}{P(x)}\right) dx$$

$$\leq \int P(x) \left(\frac{Q(x)}{P(x)} - 1\right) dx. \qquad (eq.2)$$

$$= \int (Q(x) - P(x)) dx = \int Q(x) dx - \int P(x) dx = 1 - 1 = 0$$

Therefore, $-D_{KL}(P\|Q) \leq 0$

$$D_{KL}(P\|Q) \geqslant 0.$$

KL Divergence is always non-negative.

4). $D_{KL}(P\|Q) = \ln \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}.$

when $\mu_p = \mu_q$, we have.

$$= \ln \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{1}{2}$$

$$= \ln \frac{\sigma_q}{\sigma_p} + \frac{1}{2}\left(\frac{\sigma_p}{\sigma_q}\right)^2 - \frac{1}{2}.$$

Similarly, we have. $D_{KL}(Q\|P) = \ln \frac{\sigma_p}{\sigma_q} + \frac{1}{2}\left(\frac{\sigma_q}{\sigma_p}\right)^2 - \frac{1}{2}.$

As we are given. $\frac{x}{2} > \ln(x) + \frac{1}{2x}$ for $x > 1$.

When $x = \left(\frac{\sigma_p}{\sigma_q}\right)^2 > 1$.

$$\frac{1}{2}\left(\frac{\sigma_p}{\sigma_q}\right)^2 > \ln\left(\frac{\sigma_p}{\sigma_q}\right)^2 + \frac{1}{2\left(\frac{\sigma_p}{\sigma_q}\right)^2}$$

$$\frac{1}{2}\left(\frac{\sigma_p}{\sigma_q}\right)^2 > 2\ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{1}{2}\left(\frac{\sigma_q}{\sigma_p}\right)^2$$

$$\frac{1}{2}\left(\frac{\sigma_p}{\sigma_q}\right)^2 + \ln\left(\frac{\sigma_q}{\sigma_p}\right) > 2\ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{1}{2}\left(\frac{\sigma_q}{\sigma_p}\right)^2 + \ln\left(\frac{\sigma_q}{\sigma_p}\right)$$

$$\frac{1}{2}\left(\frac{\sigma_p}{\sigma_q}\right)^2 + \ln\left(\frac{\sigma_q}{\sigma_p}\right) > 2\ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{1}{2}\left(\frac{\sigma_q}{\sigma_p}\right)^2 \underline{\underline{- \ln\left(\frac{\sigma_p}{\sigma_q}\right)}}$$

$$\frac{1}{2}\left(\frac{\sigma_p}{\sigma_q}\right)^2 + \ln\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2} > \ln\frac{(\sigma_p)}{\sigma_q} + \frac{1}{2}\left(\frac{\sigma_q}{\sigma_p}\right)^2 - \frac{1}{2}.$$

$$= \ln\left(\left(\frac{\sigma_p}{\sigma_q}\right)^{-1}\right)$$

$$= \ln\left(\frac{\sigma_q}{\sigma_p}\right)$$

$$D_{KL}(P \| Q) > D_{KL}(Q \| P).$$

In conclusion. when $\frac{\sigma_p}{\sigma_q} > 1$. i.e $\sigma_p > \sigma_q$.

$$D_{KL}(P \| Q) > D_{KL}(Q \| P)$$

when $\frac{\sigma_q}{\sigma_p} > 1$. i.e $\sigma_q > \sigma_p$.

$$D_{KL}(Q \| P) > D_{KL}(P \| Q)$$

## Question 3

1) $h_j^{(t)} = z_j h_j^{(t-1)} + (1 - z_j)\hat{h}_j^{(t)}$.

if we $h = h^{(t-1)}$ then we need $z$ close to 1

2). when $r_j$ and $z_j$ are both close to "0"

$$h_j^{(t)} = \tilde{h}_j^{(t)}.$$

$$\tilde{h}_j^{(t)} = \phi([Wx]_j + [U(r \odot h^{(t-1)})])$$

$$= \phi([Wx]_j)$$

The hidden state is reset with the current input. but not totally reset (ignore all the previous hidden states) since $r_j \in r$, $r_{i \neq j}$ could be non-zero. and some parts of the previous hidden states will be remained.

## Question 4.

1) The purpose of the sinusoidal positional encoding is As our model contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, we must inject some

$\nearrow$ PE

information about the relative or absolute position of the tokens in the sequence.

The advantage of one-hot encoding is:

We can add PE to the input embeddings are the bottom of the encoder and decoder stacks, and those two can also be summed since they have same dimensions with one-hot encoding schema.

2). If pos are integers. they can't be equal.

Proof: Assume $PE(pos\ 1) = PE(pos\ 2)$
by contradiction.

$$\sin\left(\frac{pos\ 1}{10000^{\frac{2i}{d_{model}}}}\right) = \sin\left(\frac{pos\ 2}{10000^{\frac{2i}{d_{model}}}}\right)$$

Then by sin's property we have.

$$\frac{pos1}{10000^{\frac{2i}{d_{model}}}} = \frac{pos2}{10000^{\frac{2i}{d_{model}}}} + 2\pi \cdot n.$$

$$\underbrace{pos1 - pos2}_{\text{int - int = int.}} = \underbrace{10000^{\frac{2i}{d_{model}}} \cdot 2\pi \cdot n.}_{\text{can not be an integer}} \quad \boxed{\lightning}$$

(told by TA, I tried to prove this using Lemma but didn't succeed).

Thus, two PE can't be equal if pos are integers.