Nattapat Juthaprachakul, Student id: 301350117, Computing id: njuthapr

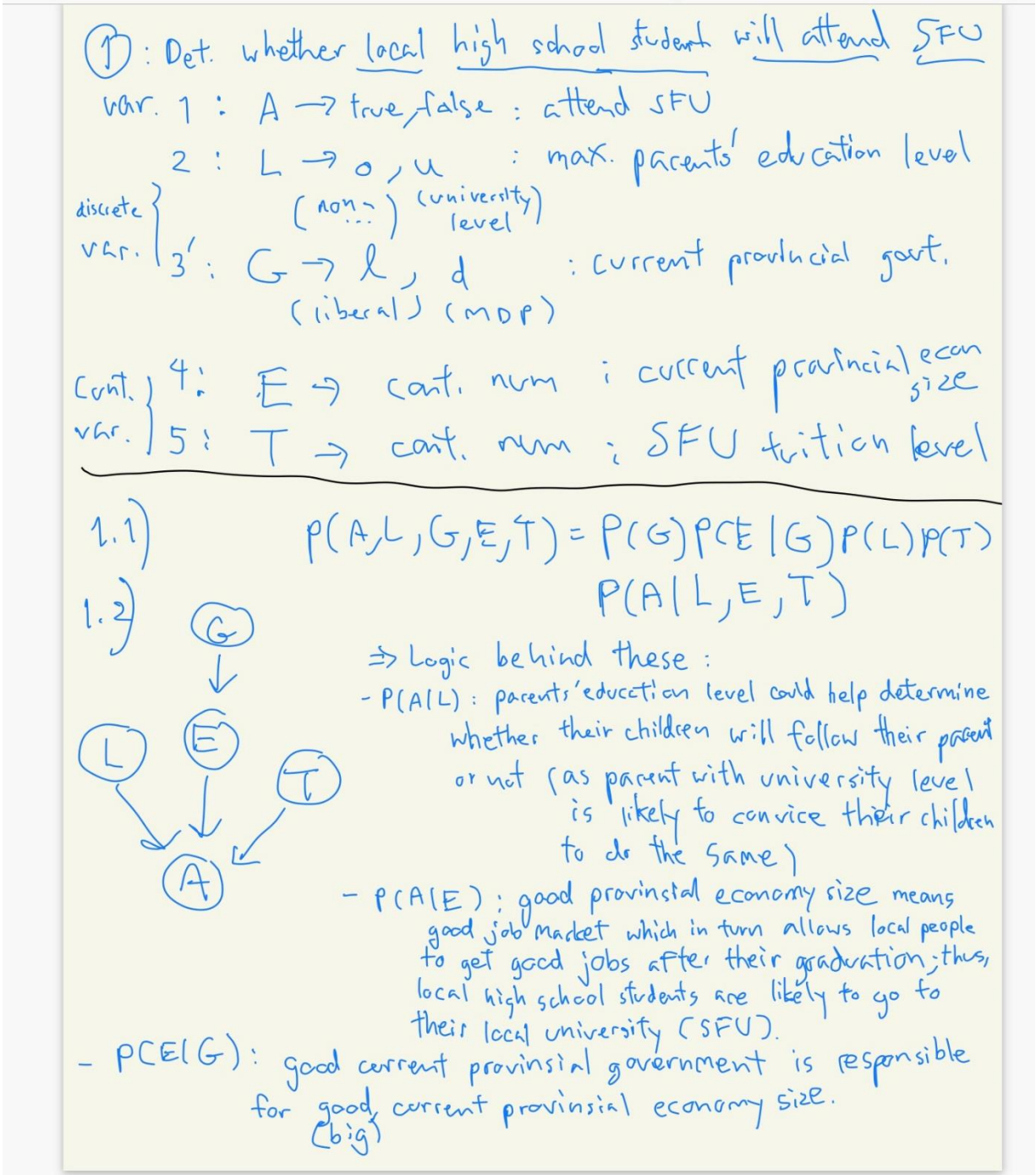
CMPT-726

Assignment 3: Graphical Models / Recurrent Neural Networks

Due November 15 at 11:59pm

1. **Graphical Models (22 marks)**
   1.1. Draw a simple Bayesian network for this domain.
   1.2. Write the factored representation for the joint distribution p(A, L, G, E, T) that is described by your Bayesian network.

ANS:



1.3. Supply all necessary conditional distributions. Provide the type of distribution that should be used and give rough guidance / example values for parameters (do this by hand, educated guesses).

ANS:

1.3) ⇒ P(E|G) & p(T) : continuous rv.

I use Gaussian distribution with mean (μ) and variance (σ^2) as parameter because it has common occurance in many natural phenomena (so it is common and safe to assume the unknown distributions to be Gaussian distribution.

⇒ P(L), P(G), P(A|L,E,T) : discrete random var.

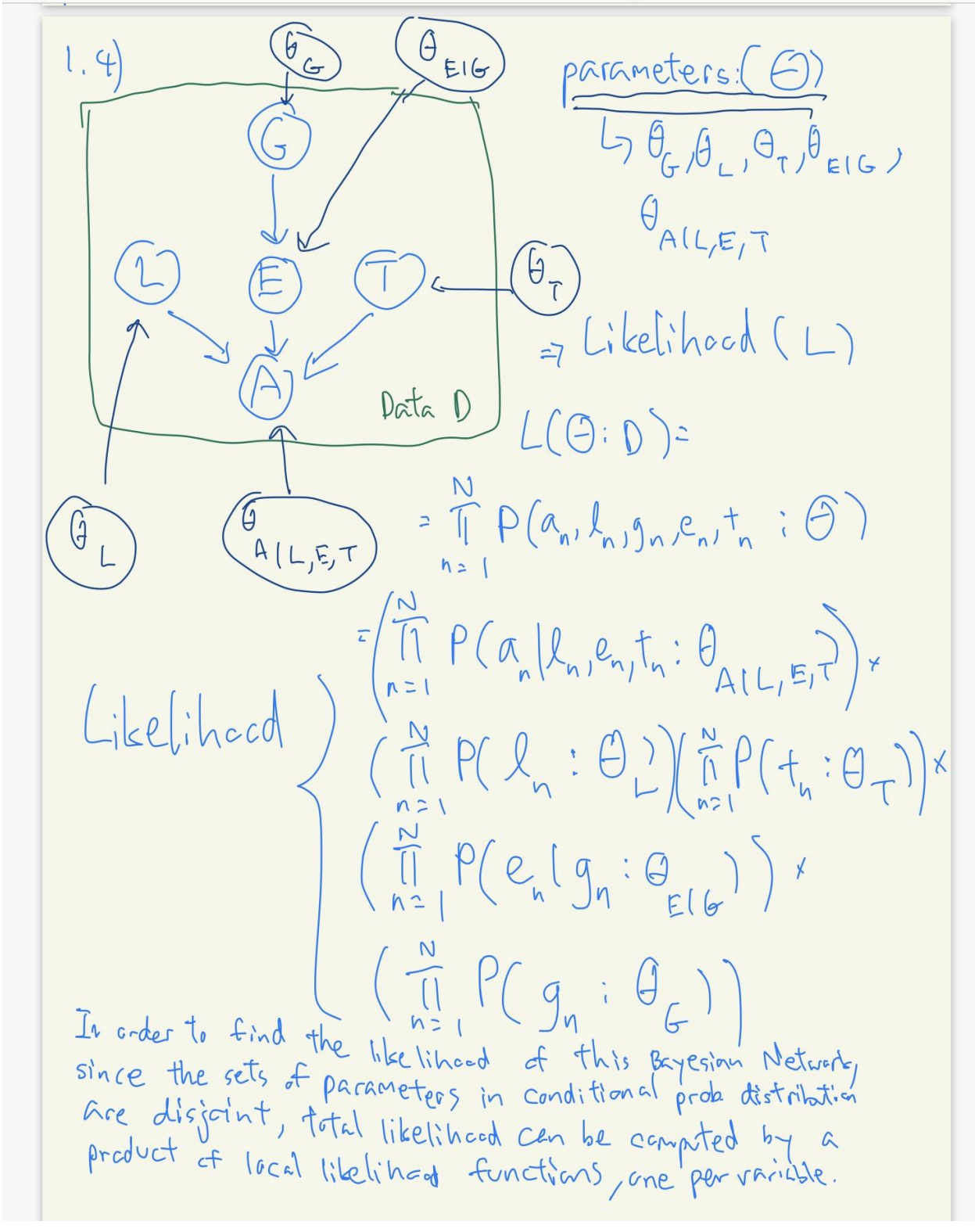
I use Bernoulli distribution.

parameter p is probability of event 1 occurs while 1-p is a probability of event 2 occurs.

For example, P(G) with p = 0.8 is probability that current provinsial government is MPP w/ 80% prob while it is Liberal w/ prob of 20%.

1.4.Suppose we had a training set and wanted to learn the parameters of the distributions using maximum likelihood. Denote each of the N examples with its values for each random variable by xn = (an, ln, gn, en, tn). The training set is {x1, x2, . . . , xN }. Which elements of the training data are needed to learn the parameters for p(A|paA)? Why?1 Start by writing down the likelihood and argue from there.

ANS:



2. KL Divergence (20 marks)

2.1. ANS:

2.2. ANS:

② KL Divergence

2.1) Ans: KL divergence is not symmetric

$D_{KL}(P||Q) - D_{KL}(Q||P)$ may not be zero.

Even though the KL divergence measures the difference between 2 distributions, it is not a distance measure. This is because that the KL divergence is not a metric measure. ✱

2.2) Ans Show $D_{KL}(P||P) = 0$

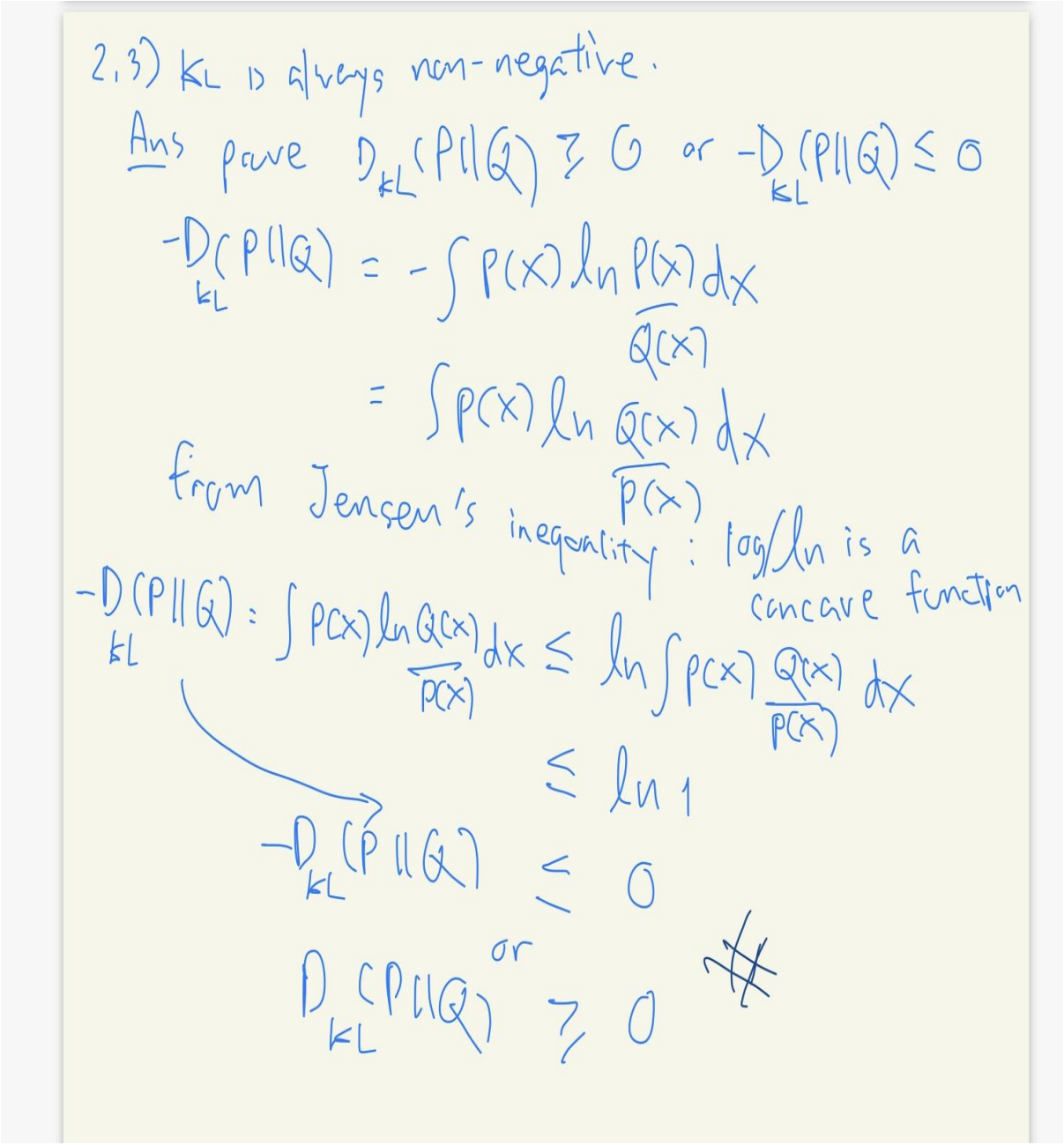
from $D_{KL}(P||P) = \int P(x) \ln \frac{P(x)}{P(x)} dx$

$= \int p(x) [\ln p(x) - \ln p(x)] dx$

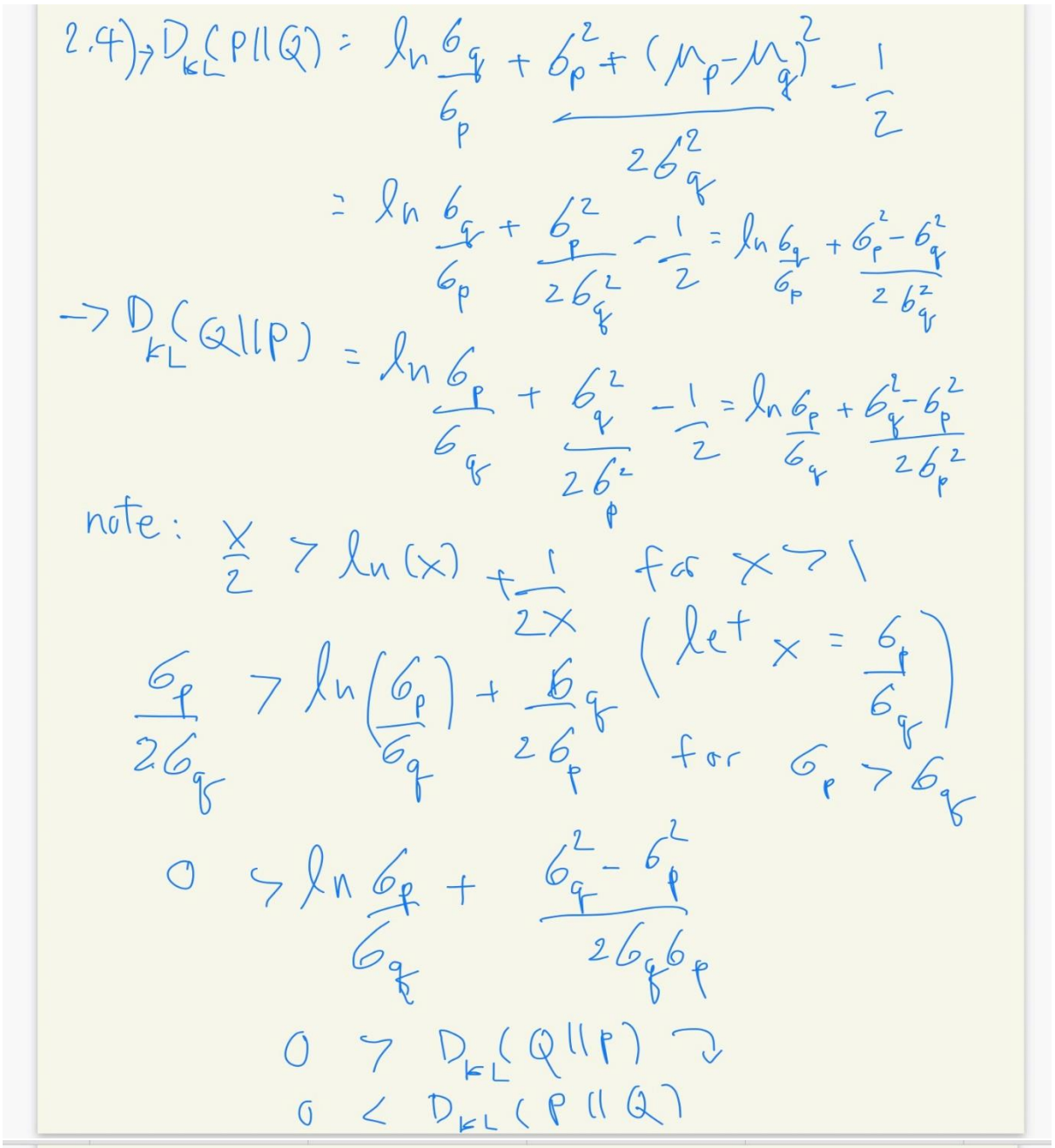
$= \int p(x) \ln p(x) dx - \int p(x) \ln p(x) dx$

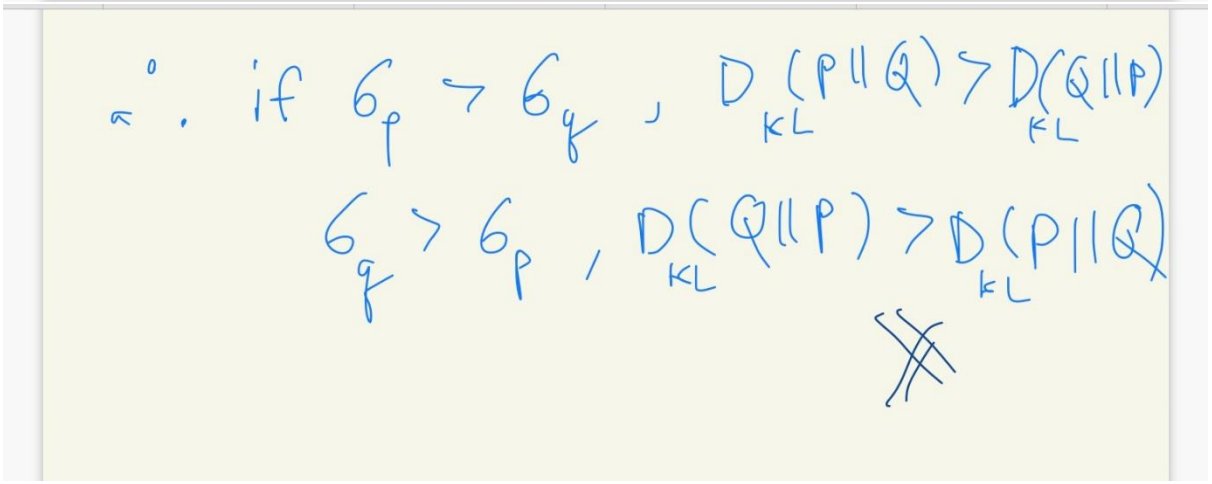
$= 0$ ✱

2.3. ANS:


2.3) KL is always non-negative.
Ans prove $D_{KL}(P||Q) \geq 0$ or $-D_{KL}(P||Q) \leq 0$
$-D_{KL}(P||Q) = -\int P(x) \ln \frac{P(x)}{Q(x)} dx$
$= \int P(x) \ln \frac{Q(x)}{P(x)} dx$
from Jensen's inequality : log/ln is a concave function
$-D_{KL}(P||Q) = \int P(x) \ln \frac{Q(x)}{P(x)} dx \leq \ln \int P(x) \frac{Q(x)}{P(x)} dx$
$\leq \ln 1$
$-D_{KL}(P||Q) \leq 0$
$D_{KL}(P||Q) \geq 0$ or

2.4. ANS:


2.4) → $D_{KL}(P||Q) = \ln\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p-\mu_q)^2}{2\sigma_q^2} - \frac{1}{2}$
$= \ln\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{1}{2} = \ln\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2-\sigma_q^2}{2\sigma_q^2}$
→ $D_{KL}(Q||P) = \ln\frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2}{2\sigma_p^2} - \frac{1}{2} = \ln\frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2-\sigma_p^2}{2\sigma_p^2}$
note: $\frac{x}{2} > \ln(x) + \frac{1}{2x}$ for $x > 1$
$\frac{\sigma_p}{2\sigma_q} > \ln\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{\sigma_q}{2\sigma_p}$ $\left(\text{let } x = \frac{\sigma_p}{\sigma_q}\right)$ for $\sigma_p > \sigma_q$
$0 > \ln\frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2-\sigma_p^2}{2\sigma_q\sigma_p}$
$0 > D_{KL}(Q||P)$
$0 < D_{KL}(P||Q)$


∴ if $\sigma_p > \sigma_q$, $D_{KL}(P||Q) > D_{KL}(Q||P)$
$\sigma_q > \sigma_p$, $D_{KL}(Q||P) > D_{KL}(P||Q)$

3. Gated Recurrent Unit (10 marks)

3.1.ANS:

3.2.ANS:

(3) $r_j = \sigma\left([W_r x]_j + [U_r h_{\langle t-1\rangle}]_j\right)$

$z_j = \sigma\left([W_z x]_j + [U_z h_{\langle t-1\rangle}]_j\right)$

$h_j^{\langle t\rangle} = z_j h_j^{\langle t-1\rangle} + (1 - z_j)\,\tilde{h}_j^{\langle t\rangle}$

$h_j^{\langle t\rangle} = z_j h_j^{\langle t-1\rangle} + (1 - z_j)\,\phi\left([Wx]_j + [U(r \odot h_{\langle t-1\rangle})]_j\right)$

3.1) we want $h_j^{\langle t\rangle} = h_j^{\langle t-1\rangle}$

$\therefore$ z_j must be 1 or close to 1.

while if z_j is not equal to 1,

r_j must be some number that is close to 1 ($r \odot h_{\langle t-1\rangle}$ as to preserve $h_{\langle t-1\rangle}$) ##

3.2) if r_j, z_j are zero.

$h_j^{\langle t\rangle} = 0 + (1)\,\phi\left([Wx]_j\right)$

$h_j^{\langle t\rangle} = \phi\left([Wx]_j\right)$

current content

-> the hidden will use just current content.
(incoming new input x).

note: z => update gate : helps model determine how much the past information (from previous time steps) is needed to be passed to the future.

∴ high z -> preserve much past info ($h_{<t-1>}$)

note: r => reset gate : helps model determine how much the past information to forget.

∴ high r -> ⊙ product will preserve ($r \odot h_{<t-1>}$) past information.

4. Attention Models (10 marks)

4.1.ANS:

4) 4.1) The purpose of the sinusoidal encoding is because the Transformer model does not require word inputs that will be fed into network to have specific orders or positions. Therefore, it needs some technique (sinusoidal encoding) to help the model incorporates the order of words by adding a position-dependent signal to each word-embedding.

→ The sinusoidal encoding does not require the sentence to be fix length compared to one-hot encoding scheme. It uses properties of sin(x) and cos(x) [cyclic] functions to return information of the position of a word in a sentence.

4.2.ANS:

4.2) $PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$

$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$

Formula for positional encoding

→ same position

$\frac{PE(pos1)}{PE(pos1)} = \frac{PE(pos2)}{PE(pos2)}$

$\frac{\sin(pos1)}{\cos(pos1)} = \frac{\sin(pos2)}{\cos(pos2)}$

$\tan(pos1) = \tan(pos2)$ ✗