Q1

a) i) local posterior

$$p(\theta|x_k) = \frac{\overset{\text{prior}}{p(\theta)}\, p(x_k|\theta)}{p(x_k)}$$

← probability of data given θ or likelihood of θ

$p(x_k)$ ← normalising constant

$$\propto p(\theta)\, p(x_k|\theta)$$

ii) global posterior

$$p(\theta|x_1, x_2 \cdots x_k) \propto p(\theta) \prod_{k=1}^{K} p(x_k|\theta)$$

$$\propto p(\theta) \prod_{k=1}^{K} \frac{p(\theta|x_k)}{p(\theta)}$$

$$= \left[\frac{1}{p(\theta)}\right]^{K-1} \prod_{k=1}^{K} p(\theta|x_k)$$

iii) Unlikely to happen if we have the correct model & sufficient data.

However it is likely that each hospital is slightly different & so a model like this would be more appropriate

global parameters

local parameters (will each be different in general)

$$p(\theta) \prod_{k=1}^{K} p(\theta_k|\theta)\, p(x_k|\theta_k)$$

If the data actually come from this model & yet we fit the global model instead, fitting the local models will perform better when the local parameters θ_k are fairly different from one another.

b)

$$p(C=1 \mid LFT=1, PCR=0) = \frac{p(C=1)\,p(LFT=1 \mid C=1)\,p(PCR=0 \mid C=1)}{p(C=1)\,p(LFT=1 \mid C=1)\,p(PCR=0 \mid C=1) + p(C=0)\,p(PCR=0 \mid C=0)\,p(LFT=1 \mid C=0)}$$

$$= \frac{0.01 \times 0.85 \times 0.05}{0.01 \times 0.85 \times 0.05 + 0.99 \times 0.001 \times 1}$$

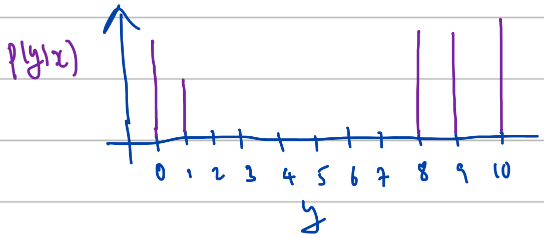
$$= \frac{1}{1 + \frac{0.99}{0.85} \times \frac{0.1}{0.05}} \approx 0.3$$

i.e. still 30% chance of having covid even though the PCR came back negative.

Q2

a) i) $p(y=k \mid \underline{x}) \propto \exp(\underline{w}_k^T \underline{x})$

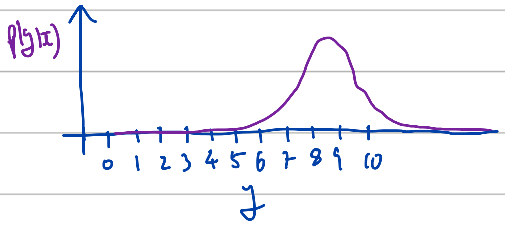
- models data as discrete - good as data are discrete
- does not accont for ordering of data i.e. that for strong students all high scores are likely, all low scores less likely & vice versa for weak students (linked to the fact that it has weights for each score $\{\underline{w}_k\}_{k=0}^{10}$)



Can output distributions like this

$p(y \mid \underline{x}) = N(y; \underline{w}^T \underline{x}, \sigma_y^2)$

- models data as continuous which is inappropriate
- captures the fact that a strong student is likely to get a high score, but not a low score. (Linked to having a single weight vector $\underline{w}$)



all distributions have a single peak

b) Lots of possible options

one idea let $p(z \mid \underline{x}) = N(z; \underline{w}^T \underline{x}, 1)$

$$
\begin{aligned}
&0 \text{ if } && z \leq 0 \\
&1 \text{ if } && 0 < z \leq \mu_1 \\
&2 \text{ if } && \mu_1 < z \leq \mu_2 \\
& && \vdots \\
&K \text{ if } && \mu_{K-1} < z
\end{aligned}
$$

This is called the "ordered probit model" & the general problem is called probit regression.

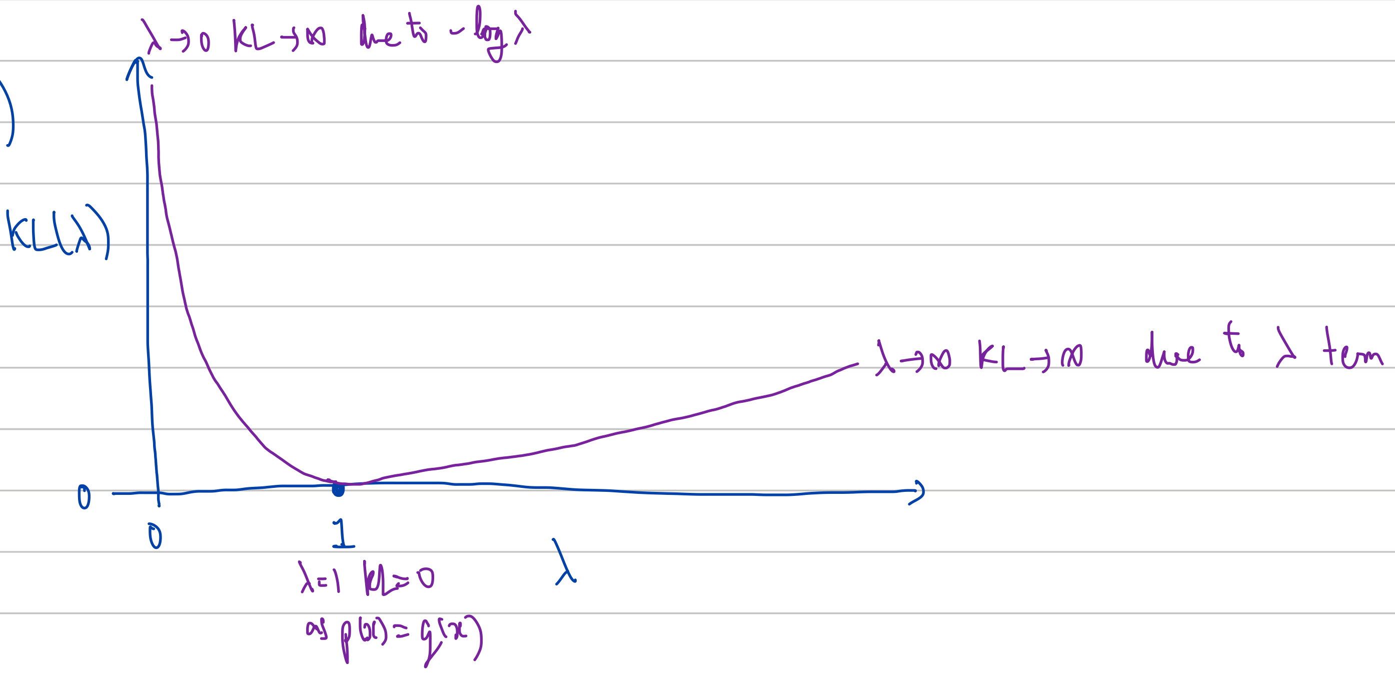
b) i) $$KL(q(x)\,||\,p(x)) = \mathbb{E}_{q(x)}\left[\log p(x) - \log q(x)\right]$$

$$= \mathbb{E}_{q(x)}\left[-\log\lambda - x/\lambda + x\right]$$

$$= -\log\lambda - 1 + \lambda$$

$\mathbb{E}_{q}(x) = \lambda$

ii)



NB. do not need to prove properties of KL - can just use them directly

Q3

a) $p(x_c|N_c,\mu) = \mathcal{N}(x_c;\, N_c\mu,\, N_c)$ [means & variances add for iid variables]

b) $$p(N_c|x_c,\mu) \propto \overset{\text{Uniform}(1,10)}{p(N_c)}\;\overset{\{}{p(x_c|N_c,\mu)} \Rightarrow p(N_c|x_c,\mu) = \frac{\mathcal{N}(x_c;\mu N_c, N_c)}{\sum_{N_c=1}^{10} \mathcal{N}(x_c;\mu N_c', N_c')}$$

c) i) EM will find maximum likelihood setting of parameters

$$\mu = \arg\max_{\mu} \log p(x_{1:C}|\mu)$$

E-Step $$\{q^{(new)}(N_c)\}_{c=1}^{C} = \arg\max_{\{q(N_c)\}_{c=1}^{C}} \mathcal{F}(\mu, \{q(N_c)\}_{c=1}^{C})$$
$$= p(N_c|x_c,\mu)$$

M-Step $$\mu^{(new)} = \arg\max_{\mu} \mathcal{F}(\mu, \{q(N_c)\}_{c=1}^{C})$$

ii) $$\mu^{(new)} = \arg\max_{\mu} \sum_{c=1}^{C}\sum_{N_c=1}^{10} q(N_c)\log p(x_c|\mu,N_c)$$
$$= \arg\max_{\mu} \sum_{c=1}^{C} \mathbb{E}_{q(N_c)}\left[-\frac{1}{2N_c}(x_c - N_c\mu)^2 - \log N_c - \frac{1}{2}\log 2\pi\right]$$

Take derivatives & set to zero:

$$\sum_{c=1}^{C} \mathbb{E}_{q(N_c)}\left[\frac{1}{2\cancel{N_c}}(x_c - \mu^* N_c)\cancel{N_c}\right] = 0 \Rightarrow \mu^* = \frac{\overset{\text{total weight of all bags}}{\sum_{c=1}^{C} x_c}}{\underbrace{\sum_{c=1}^{C} \mathbb{E}_{q(N_c)}(N_c)}_{\text{sum of expected number of oranges in all bags}}}$$

Q4

a) $1\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 2\ 1\ 1\ 1 \,\vdots\, 3\ 3\ 3\ 3\ 3 \,\vdots\, 2\ 2\ 2$ ↙ $y_{1:20}$ (from lectures)

↑ 50% 1s & 2s ↑ all 3s ↑ 50:50

$1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1 \,\vdots\, 2\ 2\ 2\ 2\ 2 \,\vdots\, 1\ 1\ 1$ ↖ $x_{1:20}$

b) $$p(y_t \mid y_{t-1}) = \sum_{x_t, x_{t-1}} p(y_t \mid x_t)\, p(x_t \mid x_{t-1})\, p(x_{t-1} \mid y_{t-1})$$

if $y_{t-1} = 1$ or 2 then $x_{t-1} = 1$

if $y_{t-1} = 3$ then $x_{t-1} = 2$

$$= \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} 0.9 & 0.9 & 0.1 \\ 0.1 & 0.1 & 0.9 \end{bmatrix} = \begin{bmatrix} 0.45 & 0.45 & 0.05 \\ 0.45 & 0.45 & 0.05 \\ 0.1 & 0.1 & 0.9 \end{bmatrix}$$

c) in general

note difference here

$$p(y_t \mid y_{1:t-1}) = \sum_{x_t, x_{t-1}} p(y_t \mid x_t)\, p(x_t \mid x_{t-1})\, p(x_{t-1} \mid y_{1:t-1})$$

However although these two appear different note in this specific case:

if you know y_{t-1} you know x_{t-1}

ie. $p(x_{t-1} \mid y_{1:t-1}) = p(x_{t-1} \mid y_{t-1})$

$$\therefore\ p(y_t \mid y_{1:t-1}) = p(y_t \mid y_{t-1}) = \begin{bmatrix} 0.45 & 0.45 & 0.05 \\ 0.45 & 0.45 & 0.05 \\ 0.1 & 0.1 & 0.9 \end{bmatrix} \quad \& \quad p(y_1) = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/2 \end{bmatrix}$$

d) Not true in general that

$$p(x_{t-1} \mid y_{1:t-1}) = p(x_{t-1} \mid y_{t-1})$$

Summary of Exam Marks

The examination was taken by 132 candidates in total. The marks had an average of 65.7% and standard deviation 11.5% with the top at 97% and bottom at 35%.

Q1 Fundamental Inference Concepts

118 attempts, Ave. raw mark 13.3/20, Stan. Dev. 3.0, Maximum 20, Minimum 7.

A popular question. Generally well answered. In part (aii) many solutions did not give the global posterior as a function of the local posteriors, but rather as a product of the prior and the local likelihood functions which is incorrect. Most people failed to substitute numbers into Bayes' rule correctly in part (b) resulting in incorrect numerical answers.

Q2 Classification and KL divergence

89 attempts, Ave. raw mark 12.5/20, Stan. Dev. 2.7, Maximum 20, Minimum 7.

Generally well answered. In part (a) many failed to spot that the softmax model does not capture the fact that if a student is strong, they are likely to have a high probability for any high score and a low probability of all low scores, and vice versa for weak students. Part (b) was answered more successfully. A significant minority of answers to (bii) sketched KL divergences that went negative, which is impossible by definition.

Q3 The EM Algorithm

84 attempts, Ave. raw mark 12.9/20, Stan. Dev. 3.4, Maximum 19, Minimum 3.

This question is on a challenging topic, but was well answered in general. Curiously many students did not realise that variances of independent variables add and therefore answered the first part of the question incorrectly. In the last part, many people differentiated the free-energy correctly, but then failed to rearrange the expression for \mu correctly.

Q4 Discrete Hidden Markov Models

105 attempts, Ave. raw mark 13.5/20, Stan. Dev. 2.811, Maximum 20, Minimum 9.

The first two parts of this question were well done. Part (c) involved computing the transition matrix and initial state distribution of the equivalent bigram model was not successfully completed by about 1/3 of candidates. Most candidates correctly stated for part (d) that it was not usually possible to write an HMM as a bigram model in this way, but they didn't spot the reason that it is possible here is because, for the example considered in the question, the latent state is deterministic given the observed state at the same time point