Q1

a) i) $$x^{(MAP)} = \arg\max_x p(x|y) = \arg\max_x \log p(x,y)$$

$$\log p(x,y) = -\frac{1}{2}\left(y - \frac{1}{2}kx^2\right)^2 - \frac{1}{2\sigma^2}x^2 + c$$ ← don't depend on x

let $u = x^2$

$$\frac{d}{du}\log p(x,y) = \frac{d}{du}\left(-\frac{\cancel{1}}{\cancel{2}}\left(y - \frac{1}{2}ku\right)^2 - \frac{\cancel{1}}{\cancel{2}\sigma^2}u\right) = 0$$

See note below

NB if $y = 0 \Rightarrow x_{MAP} = 0$

$$0 = \cancel{2}\cdot\left(y - \frac{1}{2}ku\right)\cdot\left(-\frac{\cancel{1}}{\cancel{2}}k\right) + 1/\sigma^2$$

$$0 = -ky + \frac{1}{2}k^2u + \frac{1}{\sigma^2}$$

$$u = -\frac{2}{k^2\sigma^2} + \frac{2y}{k}$$

I was happy if people did not spot that the result only applied if this condition held though

$$\Rightarrow \quad x_{MAP} = \pm\sqrt{\frac{2y}{k} - \frac{2}{k^2\sigma^2}} \quad \underbrace{\text{if } y > \frac{1}{k\sigma^2} \text{ else } x_{MAP} = 0}$$

ii) $$\sigma^2 \to \infty \quad \Rightarrow \quad x_{MAP} \to \pm\sqrt{\frac{2y}{k}}$$ c.f. $y = \frac{1}{2}kx^2 + \text{noise}$

inverts this

This is the maximum likelihood estimate of x as the prior is uninformative b i)

$$p(d \mid y_1, y_2) \propto \underbrace{p(d)}_{N(d;0,1)} \underbrace{p(y_1, y_2 \mid d)}_{N\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}; \begin{bmatrix} d \\ d \end{bmatrix}, \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}^{-1}\right)} \propto N(d; \mu_{d|y_1,y_2}, \sigma^2_{d|y_1,y_2})$$

ii)

$$\log p(d \mid y_1, y_2) = c - \frac{1}{2} d^2 - \frac{1}{2} [y_1 - d, y_2 - d] \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \begin{bmatrix} y_1 - d \\ y_2 - d \end{bmatrix}$$

$$= c - \frac{1}{2} d^2 - \frac{1}{2} (y_1 - d)^2 - \frac{1}{2} (y_2 - d)^2 - (y_1 - d)(y_2 - d)\alpha$$

$$= c' - \frac{1}{2} d^2 - \frac{1}{2} d^2 + y_1 d - \frac{1}{2} d^2 + y_2 d - \alpha d^2 + (y_1 + y_2) d \alpha$$

$$= c' - \frac{1}{2} (3 + 2\alpha) d^2 + (y_1 + y_2)(1 + \alpha) d$$

$$\Rightarrow \sigma^2_{d|y_1,y_2} = \frac{1}{3 + 2\alpha} \qquad \mu_{d|y_1,y_2} = \frac{1}{3 + 2\alpha} \overbrace{(1+\alpha)(y_1 + y_2)}$$

iii)

$$\sigma^2_{d|y_1,y_2} = \frac{1}{3} \quad \text{when } \alpha = 0$$

when $\alpha > 0$ $\quad \sigma^2_{d|y_1,y_2} < \frac{1}{3}$ $\quad$ ie more certain

note

negative correlation

$$\begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}^{-1} = \frac{1}{1 - \alpha^2} \begin{pmatrix} 1 & -\alpha \\ -\alpha & 1 \end{pmatrix}$$

$\Rightarrow$ $\alpha = 0 \Rightarrow$ independent y_1, y_2

$\alpha > 0 \Rightarrow$ negative correlation $\Rightarrow$ forming $y_1 + y_2$ removes some of the noise

$\Rightarrow$ can be more certain

($\alpha < 0 \Rightarrow$ positive correlation in y_1 & y_2 $\Rightarrow \sigma_{d|y_1,y_2} > \frac{1}{3} \Rightarrow$ less certain)

Q2

a i) $\mathcal{L}(w,\sigma^2) = \log \prod_{n=1}^{N} p(y_n | x_n, w, \sigma^2)$ (log likelihood)

ii) $\frac{d}{dw} \mathcal{L}(w,\sigma^2) = 0 = \frac{d}{dw}\left(-\frac{1}{2\sigma^2} \sum_n (y_n - w x_n)^2 - \frac{N}{2}\log\sigma^2\right)$

$\Rightarrow 0 = \sum_{n=1}^{N} (y_n - w x_n) x_n$

$\Rightarrow w = \dfrac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} x_n^2}$

iii) let $YX_N = \sum_{n=1}^{N} y_n x_n$ $\quad XX_N = \sum_{n=1}^{N} x_n^2$

$YX_{N+1} = YX_N + y_{N+1} x_{N+1}$ $\quad XX_{N+1} = XX_N + x_{N+1}^2$

$w_{N+1} = YX_{N+1} / XX_{N+1}$

b) i) $\sigma^2_{MAP} = \underset{\sigma^2}{\arg\max}\ p(\{y_n\}_{n=1}^N \mid \{x_n\}_{n=1}^N, w_{MAP}, \sigma^2)\, p(\sigma^2)$

$$= \underset{\sigma^2}{\arg\max} \left[\sum_{n=1}^N \log p(y_n \mid x_n, w_{MAP}, \sigma^2) + \log p(\sigma^2) \right]$$

$$= \underset{\sigma^2}{\arg\max} \left(-\frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - w_{MAP} x_n)^2 - \frac{N}{2} \log \sigma^2 - \alpha \log \sigma^2 - \beta/\sigma^2 \right)$$

$$= \underset{\sigma^2}{\arg\max} \left(-\frac{1}{\sigma^2} \underbrace{\left(\beta + \frac{1}{2} \sum_{n=1}^N (y_n - w_{MAP} x_n)^2 \right)}_{\beta'} - \underbrace{\left(\frac{N}{2} + \alpha\right)}_{\alpha'} \log \sigma^2 \right)$$

$$\frac{d}{d\sigma^2} \left(-\frac{1}{\sigma^2} \beta' - \alpha' \log \sigma^2 \right) = 0$$

$$+\frac{1}{\sigma^4_{MAP}} \beta' - \frac{\alpha'}{\sigma^2_{MAP}} = 0$$

$$\Rightarrow \sigma^2_{MAP} = \frac{\beta'}{\alpha'} = \frac{\beta + \frac{1}{2} \sum_{n=1}^N (y_n - w_{MAP} x_n)^2}{\frac{N}{2} + \alpha}$$

ii) $\alpha = N'/2$ where N' = # of pseudopoints

$\beta = r'/2$ where r' = sum of squared errors (noise error) on pseudopoints

Q3

a) E-Step

current setting of parameters

$$\{q^{(new)}(s_n)\}_{n=1}^{N} = \underset{\{q(s_n)\}_{n=1}^{N}}{\arg\max}\ \mathcal{F}(\theta_t, \{q(s_n)\}_{n=1}^{N})$$

$$\Rightarrow \quad q^{(new)}(s_n) = p(s_n \mid \underline{x}_n, \theta_t) \propto p(s_n|\theta_t)p(\underline{x}_n|s_n,\theta_t)$$

In this case

$$p(s_n=k) = 1/K \qquad p(\underline{x}_n|s_n=k) = \prod_{d=1}^{D} \pi_{kd}^{x_{dn}} (1-\pi_{kd})^{1-x_{dn}}$$

let $$u_{nk} = 1/K \cdot \prod_{d=1}^{D} \pi_{kd}^{x_{dn}} (1-\pi_{kd})^{1-x_{dn}}$$

$$q^{(new)}(s_n=k) = \frac{u_{nk}}{\sum_{k'} u_{nk'}}$$

b) Hard E-Step will perform $\underset{k}{\arg\max}\ p(s_n=k \mid \underline{x}_n) = \underset{k}{\arg\max}\ \log p(s_n=k|x_n)$

$$= \underset{k}{\arg\max} \sum_{d=1}^{D} \left(x_{dn} \log \pi_{kd} + (1-x_{dn}) \log(1-\pi_{kd})\right)$$

$$= \underset{k}{\arg\max}\ KL(\underline{x}_n \,\|\, \underline{\pi}_k)$$

i.e. you take each data point $\underline{x}_n$ & compute how close it is to each $\underline{\pi}_k$ according to the KL divergence. This is equivalent to the $s_n = \underset{k}{\arg\min} \|\underline{x}_n - \underline{m}_k\|^2$ assignment step in K-means

c) M-Step

$$\theta^{(new)} = \arg\max_{\theta} \mathcal{F}\left(\theta, \{q^{(old)}(s_n)\}_{n=1}^{N}\right)$$

$$= \arg\max_{\theta} \sum_n \mathbb{E}_{q^{(old)}(s_n)}\left[\log p(s_n, \underline{x}_n | \theta)\right]$$

$$\frac{\partial \mathcal{F}}{\partial \pi_{l,e}} = \sum_n \mathbb{E}_{q^{(old)}(s_n)}\left[\log p(\underline{x}_n | s_n, \theta)\right]$$

$$= \sum_n \sum_{k,d} q^{(old)}(s_n=k)\left(x_{nd}\log \pi_{kd} + (1-x_{nd})\log(1-\pi_{kd})\right)$$

$$= \sum_n q^{(old)}(s_n=l)\left(\frac{x_{n,e}}{\pi_{le}} - \frac{(1-x_{ne})}{1-\pi_{l,e}}\right) = 0$$

$$= \sum_n q^{(old)}(s_n=l)\left(x_{ne} - \pi_{l,e}\right)$$

$$\Rightarrow \quad \pi_{l,e} = \frac{\sum_n q^{(old)}(s_n=l)\, x_{n,e}}{\sum_n q^{(old)}(s_n=l)}$$

d) For hard assignments $\pi_{k,d}$ would be updated to be the average of $x_{n,d}$ for all the points assigned to cluster k.

$$\pi_{k,d} = \text{mean}(x_{n,d} : s_n = k)$$

Q4

a) $\quad x_1 \sim N(\mu_1, \sigma_1^2) \qquad x_t = \lambda x_{t-1} + \sigma \varepsilon_t \qquad \varepsilon_t \sim N(0,1)$

b)
$$
\begin{aligned}
z_t &= z_{t-1} + x_t \\
&= z_{t-1} + \lambda x_{t-1} + \sigma \varepsilon_t \\
&= z_{t-1} + \lambda (z_{t-1} - z_{t-2}) + \sigma \varepsilon_t \\
&= (1+\lambda) z_{t-1} - \lambda z_{t-2} + \sigma \varepsilon_t \qquad = \lambda_1 z_{t-1} + \lambda_2 z_{t-2} + \sigma \varepsilon_t
\end{aligned}
$$

$\Rightarrow$ AR(2) proven with $\lambda_1 = 1+\lambda$ & $\lambda_2 = -\lambda$

c) $\quad \underline{s}_1 \sim N(\underline{\mu}_1, \underline{\underline{\Sigma}}_1) \qquad \overbrace{\underline{s}_t \mid \underline{s}_{t-1}}^{t=2\dots T} \sim N(\underline{\underline{A}}\,\underline{s}_{t-1}, \underline{\underline{Q}}) \qquad \overbrace{\underline{y}_t \mid \underline{s}_t}^{t=1\dots T} \sim N(\underline{\underline{C}}\,\underline{s}_t, \underline{\underline{R}})$

d) one option: $\underline{s}_t = \begin{bmatrix} x_t \\ z_t \end{bmatrix} \qquad \underline{\underline{C}} = \underbrace{\underline{\underline{I}}}_{2\times 2} \qquad \underline{\underline{R}} = \underbrace{\underline{\underline{I}}}_{2\times 2} \qquad \underline{\underline{A}} = \begin{bmatrix} \lambda & 0 \\ \lambda & 1 \end{bmatrix} \qquad \underline{\underline{Q}} = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

perfectly correlated state noise

as $\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \underline{\underline{A}} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \underline{\underline{Q}}^{1/2} \begin{pmatrix} n_t^{(1)} \\ n_t^{(2)} \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ \lambda & 1 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ z_{t-1} \end{pmatrix} + \frac{\sigma}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} n_t^{(1)} \\ n_t^{(1)} \end{pmatrix} \overset{\text{dist.}}{=} \begin{bmatrix} \lambda x_{t-1} + \sigma \varepsilon_t \\ \lambda x_{t-1} + z_{t-1} + \sigma \varepsilon_t \end{bmatrix}$

e) Kalman Smoother will return $\quad p(x_t, z_t \mid \underline{y}_{1:T}) = N\left(\begin{bmatrix} x_t \\ z_t \end{bmatrix} ; \begin{bmatrix} \bar{x}_t \\ \bar{z}_t \end{bmatrix}, \begin{bmatrix} \sigma_{x,t}^2 & \rho_t \\ \rho_t & \sigma_{z,t}^2 \end{bmatrix} \right)$

$\bar{z}_{1:T} = \underset{z_{1:t}}{\arg\max}\; p(z_{1:T} \mid \underline{y}_{1:T})$ $\quad$ since ① deterministic relation between x & z

② mode = mean

Summary of Exam Marks

The examination was taken by 120 candidates in total. The raw marks (from those who had taken IB) had an average of 65.5% and standard deviation 13.0% with the top candidate scoring 92% and bottom candidate scoring 30%

Q1 Fundamental Inference Concepts

112 attempts, Ave. raw mark 13.7/20, St.Dev. 2.7, Maximum 20, Minimum 10.

A popular question. Generally well answered. Many people failed to solve correctly for the MAP estimate in part a (i). Very few candidates realised that the depth sensors in b (iii) are negatively correlated when alpha is positive.

Q2 Classification and KL divergence

115 attempts, Ave. raw mark 13.6/20, St.Dev. 4.0, Maximum 20, Minimum 6.

Generally well answered. A surprising number of candidates could not derive the standard linear regression expression for the weights in part a (ii) which is simple book work, but all other parts were well handled.

Q3 The EM Algorithm

84 attempts, Ave. raw mark 11.4/20, St.Dev. 3.0, Maximum 19, Minimum 6.

This question is on a challenging topic, but was answered reasonably well in general. In part (b) many candidates realised that the hard E-step could be interpreted as minimising a distance, but none identified that this distance is the KL divergence between the binary data vector x and the cluster prior parameter. Many candidates failed to get to the correct analytic expressions for the E- and M-steps, but generally the attempts got close and used the right method.

Q4 Auto-regressive Models and Linear Gaussian State Space Models

49 attempts, Ave. raw mark 11.5/20, Stan. Dev. 3.5, Maximum 19, Minimum 6.

This was a challenging question and answers were patchy. Many candidates failed to solve part (b) and actually used spurious methods. Consequently most candidates did not realise that the new process on z is AR(2). Very few candidates constructed a linear Gaussian state space model that correctly captured the correlations between the two variables x and z.