MLMI 1 Exam Crib

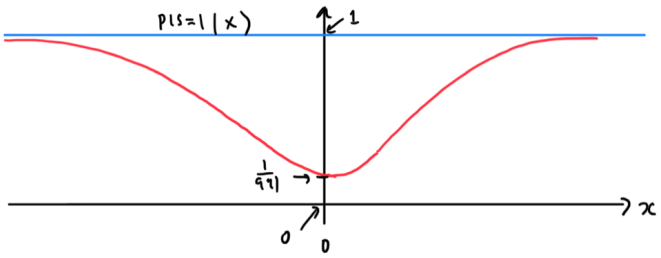
November 2020

Question 1

a) Standard book work. Put a prior on s. Define likelihood $p(x|s)$. Apply Bayes' theorem to get $p(s|x) = \frac{p(x|s)p(s)}{p(x)}$.

b) i) $p(s_n = 1|x_n) = \frac{1}{1+\frac{p(x_n|s_n=0)p(s_n=0)}{p(x_n|s_n=1)p(s_n=1)}}$. Plugging in the values, we get $\phi(x_n) = x_n^2$, $\alpha = -\frac{99}{200} = -0.495$, $\beta = \log 990$.

ii) If $x_n = 0$, $p(s_n = 1|x_n) = \frac{1}{991}$. As $|x_n| \to \infty$, $p(s_n = 1|x_n) \to 1$. This results in a symmetric curve that asymptotes to 1 at $x = \pm\infty$, and dips to $\frac{1}{991}$ at $x_n = 0$.



Question 2

a)

$$
\begin{aligned}
p(x_n|\sigma^2) &= \mathcal{N}(x_n; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} x_n^2\right) \\
p(\{x_n\}_{n=1}^N|\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N x_n^2\right) \\
p(\sigma^2|\{x_n\}_{n=1}^N) &\propto p(\sigma^2|\alpha, \beta) p(\{x_n\}_{n=1}^N|\sigma^2) \\
&= p(\sigma^2|\alpha, \beta) \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N x_n^2\right) \\
&\propto (\sigma^2)^{-\alpha/2} (\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{n=1}^N x_n^2 + \beta\right]\right) \\
&= (\sigma^2)^{-(\alpha+N)/2} \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{n=1}^N x_n^2 + \beta\right]\right)
\end{aligned}
$$

Recognising that $\alpha' = \alpha + N$ and $\beta' = \sum_{n=1}^N x_n^2 + \beta$, shows the posterior has the same form as the prior with normalising constant $Z(\alpha', \beta')$.

b) α is the number of pseudo-observations in the prior. β/α is the sample second moment of the pseudo-observations, i.e., the empirical value of σ^2 for the pseudo-observations.

Question 3

a) Given a posterior distribution $p(\sigma^2|\{x_n\}_{n=1}^N)$, the MAP estimate is the value of σ^2 that maximises the posterior. The ML estimate is the value that maximises the likelihood $p(\{x_n\}_{n=1}^N|\sigma^2)$. They are both point estimates. MAP estimate also depends on the prior.

b) Take logarithms and set the derivative to zero, ignoring constant terms. We get:

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

$$\sigma^2=\frac{\beta'}{\alpha'}$$

c) Need

$$\frac{\beta'}{\alpha'}=\frac{\sum_{n=1}^N x_n^2}{N}$$

$$\frac{\sum_{n=1}^N x_n^2+\beta}{\alpha+N}=\frac{\sum_{n=1}^N x_n^2}{N}$$

Which is true if $\beta=0$ and $\alpha=0$, or in the limit of infinite data $N\to\infty$ where the data overwhelm the prior.

Question 4

a) Compute the log posterior and recognise the quadratic form, dropping terms that don't depend on m. We get:

$$-\frac{1}{2}\left[m^2\left(\sum_{n=1}^{N}\frac{x_n^2}{1+x_n^4}+1\right)-m\left(\sum_{n=1}^{N}\frac{x_n y_n}{1+x_n^4}\right)\right]+\text{constant}$$

Completing the square, we get that the posterior is Gaussian with mean and variance given by:

$$\mu=\frac{\sum_{n=1}^{N}\frac{x_n y_n}{1+x_n^4}}{\sum_{n=1}^{N}\frac{x_n^2}{1+x_n^4}+1} \tag{1}$$

$$\sigma^2=\frac{1}{\sum_{n=1}^{N}\frac{x_n^2}{1+x_n^4}+1} \tag{2}$$

b) We seek the value of x that is a maximum of $x^2/(1+x^4)$, since that leads to a minimum of the posterior variance (minimum parameter uncertainty after the observation). Taking derivatives with respect to x^2, we get $x^2=1$ or $x=\pm 1$.

Question 5

a) Non-negative. Zero if $q = p$. Asymmetric.

b) i)

$$\begin{aligned}
\mathrm{KL}(q\|p) &= \mathbb{E}_q\left[\log q - \log p\right] \\
&= -\frac{1}{2}\mathbb{E}_q\left[(x-\mu)^2\right] + \frac{1}{2}\mathbb{E}_q\left[(x-3)^2\right] \\
&= \frac{1}{2}\mu^2 - 3\mu + \frac{9}{2} \\
&= \frac{1}{2}(\mu - 3)^2
\end{aligned}$$

ii) Convex quadratic with minimum of zero at $\mu = 3$.

Question 6

a) Set the variational posterior $q(s_n)$ to $p(s_n|x_n)$, and compute $\mathcal{F}(\theta, \{p(s_n|x_n)\}_{n=1}^N)$.

$$\begin{aligned} p(s_n = k|x_n) &= p(x_n|s_n = k)p(s_n = k)/p(x_n) \\ &= \frac{\frac{1}{\lambda_k}\exp(-x_n/\lambda_k)}{\frac{1}{\lambda_0}\exp(-x_n/\lambda_0) + \frac{1}{\lambda_1}\exp(-x_n/\lambda_1)} \end{aligned}$$

or using the logistic rather than softmax form

$$p(s_n = 1|x_n) = \frac{1}{1 + \frac{\lambda_1}{\lambda_0}\exp(-x_n(1/\lambda_0 - 1/\lambda_1))}$$

b) Holding $\{q(s_n)\}_{n=1}^N$ fixed, maximise $\mathcal{F}(\theta, \{q(s_n)\}_{n=1}^N)$ with respect to θ.

$$\begin{aligned} \mathcal{F}(\theta, \{q(s_n)\}_{n=1}^N) &= \sum_{n=1}^N \sum_{k=0}^1 q(s_n = k) \log p(s_n = k, x_n) + \text{constant} \\ &= \sum_{n=1}^N \sum_{k=0}^1 q(s_n = k) \left(-\log \lambda_k - x_n/\lambda_k\right) + \text{constant} \\ \frac{\partial \mathcal{F}}{\partial \lambda_k} &= \sum_{n=1}^N q(s_n = k) \left(-\frac{1}{\lambda_k} + \frac{x_n}{\lambda_k^2}\right) = 0 \\ \lambda_k &= \frac{\sum_{n=1}^N q(s_n = k) x_n}{\sum_{n=1}^N q(s_n = k)}. \end{aligned}$$

Question 7

a) A bigram model has the form:

$$p(y_{1:T}) = p(y_1) \prod_{t=1}^{T-1} p(y_{t+1}|y_t)$$

where $p(y_{t+1}|y_t)$ has the same functional form (viewed as a function of the arguments y_t, y_{t+1} for all t). The distribution $p(y_1)$ is the initial state probability, and $p(y_{t+1}|y_t)$ are the transition probabilities.

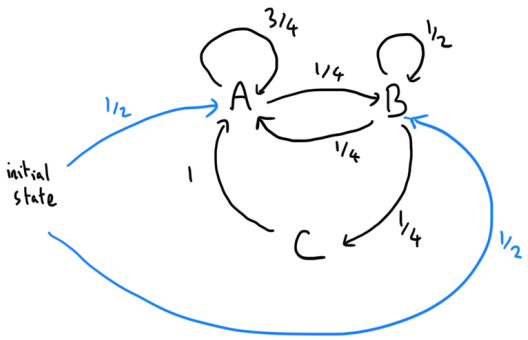
b) Let $p(y_1 = k) = \pi_k$ and $p(y_{t+1} = j|y_t = k) = T_{jk}$.

$$\log p(y_{1:T}^{(1)}, y_{1:T}^{(2)}) = \log p(y_{1:T}^{(1)}) + \log p(y_{1:T}^{(2)})$$

Use standard trick to write the log probability as a sum of logs with indicator functions. Optimise π_k with a Lagrange multiplier to obtain:

$$\pi_k \propto \mathbb{1}[y_1^{(1)} = k] + \mathbb{1}[y_1^{(2)} = k]$$

Hence $\pi_A = 0.5, \pi_B = 0.5, \pi_C = 0$. Do the same with T_{jk} to get the following state-transition probabilities.



c) The probability of the sequence under this model is 0 since $A \to C$ never occurs in the training data. To improve, could put a prior on π and T and do MAP estimation or go the whole-hog and perform Bayesian inference instead of using point estimates.

Question 8

a) Kalman filter. This is a hidden Markov model which has linear Gaussian observation likelihoods and a linear Gaussian hidden state transition probability.

b) Idea is to modify λ and σ^2 in the AR model so that a single transition is equal in distribution to two transitions under the original AR model. Let $\epsilon \sim \mathcal{N}(0, 1)$

$$\begin{aligned} x_{t+2} &= \lambda x_{t+1} + \sigma\epsilon_{t+1} \\ &= \lambda(\lambda x_t + \sigma\epsilon_t) + \sigma\epsilon_{t+1} \\ &= \lambda^2 x_t + \lambda\sigma\epsilon_t + \sigma\epsilon_{t+1}. \end{aligned}$$

We can achieve this behaviour by setting $\lambda' = \lambda^2$ and $\sigma'^2 = \sigma^2(\lambda^2 + 1)$. The likelihood is unchanged.