PML Theory Questions:

Q 1: (a) If we consider the product of the likelihood & prior:

$$p(x|c) = \frac{p(c|x)p(x)}{p(c)} \propto p(c|x)p(x)$$

$$= \left(\frac{N!}{\prod^{K} n_e!} \prod_{k=1}^{K} x_k^{n_k}\right) \cdot \left(\frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod_k x_k^{\alpha_k - 1}\right)$$

$$\propto \prod_k x_k^{n_k + \alpha_k - 1}$$

which we recognize as the functional form of the dirichlet $(\underline{x}|\underline{\alpha}+\underline{n})$ whence Dirichlet is conjugate prior of the multinomial $\boxed{\text{Rn}}$ (i.e. $\underline{X}|\underline{C} \sim Dir(-, \underline{\alpha}+\underline{n})$ with $\underline{n} := (n_1, \dots, n_K)$)

(b) WLOG let $i = 1$, $j = 2$

we aggregate by marginalizing out $j=2$ & summing $x_1 + x_2$:

So let $y := x_1 + x_2$. Then the pdf is given by:

$$p(y, x_3, x_4, \dots, x_K) = \int_0^y \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \zeta^{\alpha_1 - 1}(y-\zeta)^{\alpha_2 - 1}\left(\prod_{j=3}^{K} x_j^{\alpha_j - 1}\right) d\zeta$$

$$\propto \left(\underbrace{\int_0^y \zeta^{\alpha_1 - 1}(y-\zeta)^{\alpha_2 - 1} d\zeta}_{=: I}\right) \prod_{k=3}^{K} x_k^{\alpha_k - 1}$$

the above integral is evaluated using the substitution $Q := \frac{\zeta}{y}$

$$\Rightarrow I = \int_0^1 y^{\alpha_1 + \alpha_2 - 1} Q^{\alpha_1 - 1}(1-Q)^{\alpha_2 - 1} dQ$$

which we recognize as the beta integral $\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}$

$$\Rightarrow p(x_1 + x_2, x_3, \dots, x_K) \propto (x_1 + x_2)^{\alpha_1 + \alpha_2 - 1} x_3^{\alpha_3 - 1} \cdots x_K^{\alpha_K - 1}$$

$$\Rightarrow (X_1 + X_2, X_3, \dots, X_K) \sim Dir(\alpha_1 + \alpha_2, \alpha_3, \dots, \alpha_K)$$

$\boxed{\text{Rn}}$

•) The cross entropy loss is given by : (part (c))

$$CE = \underset{\underline{x}}{\operatorname{argmax}} \log p(C|\underline{x}) = \underset{\underline{x}}{\operatorname{argmax}} \log \text{Likelihood}$$

we have

$$\log p(C|\underline{x}) = \log N! + \sum_{k=1}^{K} [\log(x_k)] n_k - \sum_{k=1}^{K} \log n_k$$

in order to find the argmax, one must employ Lagrange multipliers and optimize $\log p(C|x) + \lambda \left(1 - \sum_{k=1}^{K} x_k\right) =: \mathcal{F}(\underline{x}, \lambda)$

then $\partial_{x_k} \mathcal{F} = \frac{n_k}{x_k} - \lambda \quad \forall k \in \{1, .., K\}$

optimizing $(\partial_{x_k} \mathcal{F} = 0) \Rightarrow x_k = \frac{n_k}{\lambda} \quad \forall k$

$$\Rightarrow \sum x_k = \frac{\sum n_k}{\lambda} \; \forall k \Rightarrow \lambda = \frac{\sum n_k}{\sum x_k} = \sum n_k = N$$

thus $\underline{x}_{MLE} = \frac{1}{N}(n_1, ..., n_K)$

•) Our posterior is $\mathcal{D}ir(-, \underline{\alpha} + \underline{n})$

so $p(\underline{x}|C) = \frac{\Gamma(\sum \alpha_k + n_k)}{\prod \Gamma(\alpha_k + n_k)} \prod x_k^{\alpha_k + n_k - 1}$

we use the same (Lagrange multiplier) trick to solve the max

$$\log p(\underline{x}|C) = \log(\sum \alpha_k + n_k) - \sum \log(\alpha_k + n_k) + \sum (\alpha_k + n_k - 1) \log x_k$$

$$\mathcal{G}(\underline{x}, \lambda) := \log p(\underline{x}|C) + \lambda(1 - \sum x_k)$$

$\partial_{x_k} \mathcal{G} = \frac{\alpha_k + n_k - 1}{x_k} - \lambda$. Setting equal to zero gives

$x_k = \frac{\alpha_k + n_k - 1}{\lambda}$. By the same logic as above; $\longrightarrow$

$\lambda = \sum_i \alpha_k + \sum_i n_k - K$ (summing k expressions)

where the max is given by

$$\hat{x}_k = \frac{\alpha_k + n_k - 1}{\sum_i \alpha_k + n_k - K} \quad \forall k \in \{1, \dots, K\}$$

The two forms are very similar but the posterior takes into account the prior information (parametrized by $\underline{\alpha}$)

(d) (i) we have $MLE(\underline{x} \mid C_1) = (0,0,0,0,1)$

$MLE(\underline{x} \mid C_2) = \frac{1}{410}(1, 0, 12, 43, 354)$

(ii) $p(\underline{x} \mid C_1, \underline{\alpha}) = D(\underline{x} \mid (1,1,1,1,4))$

$p(\underline{x} \mid C_2, \underline{\alpha}) = D\left(\underline{x} \,\middle|\, \frac{2}{415}, \frac{1}{415}, \frac{13}{415}, \frac{44}{415}, \frac{355}{415}\right)$

we have $p(C_{N+1,i} = (0,0,0,0,1) \mid \underline{x}, C, \underline{\alpha}) = x_5 \quad \forall i \in \{1,2\}$

(likelihood of single event is just probability of that event)

$$\Rightarrow p((0,0,0,0,1) \mid C, \alpha, \underline{x}) = \int_{[0,1]^5} p(C_{N+1} \mid \underline{x})\, p(\underline{x} \mid C, \alpha, \underline{x})\, d\underline{x}$$

$$= \int_{[0,1]^5} x_5 \, D(\underline{x} \mid \underline{\alpha} + \underline{n})\, d\underline{x} = \mathbb{E}_{x \sim D(\underline{n} + \underline{\alpha})}(x_5)$$

$=$ if we apply the aggregation property to indices 1-4 of $\underline{x} = (x_1, \dots, x_5)$

we get $(x_1 + x_2 + x_3 + x_4, x_5) \sim Dir(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4, \alpha_5) = Beta\left(\sum_{j=1}^{4} \alpha_j, \alpha_5\right)$

our probability then reduces to $\frac{\alpha_5}{\sum_i \alpha_j}$ (Beta integral has known solution)
(and for 4-stars, $\frac{\alpha_4}{\sum^i \alpha_j}$)

Thus for C_1, we have:

$$p(C_{N+1} = 5\star s \mid \text{data, prior}) = \frac{4}{8} = \frac{1}{2}$$

$$p(C_{N+1} \in \{4\star s, 5\star s\} \mid \text{info}) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$$

and for C_2:

$$p(C_{N+1} = 5\star s \mid \text{info}) = \frac{355}{415}$$

$$p(C_N \in \{4\star s, 5\star s \mid \text{info}) = \frac{355}{415} + \frac{44}{415} = \frac{399}{415}$$

So C_2 seems a better choice with a reasonable prior!