1. a) The vector that would lead to $-\inf$ is $a = [-1000, -1000]$

The vector that would lead to $\inf$ is $b = [1000, 1000]$

b) $RHS = \log\left(\sum_{i=0}^{k} \exp(a_i - \max_{j=0}^{k}\{a_j\})\right) + \max_{j=0}^{k}\{a_j\}$

Let c denote $\max_{j=0}^{k}\{a_j\}$

$$= \log\left(\sum_{i=0}^{k} \exp(a_i - c)\right) + c$$

$$= \log\left(\sum_{i=0}^{k} \exp(a_i - c)\right) + \log(\exp(c))$$

$$= \log\left(\exp(c) \cdot \sum_{i=0}^{k} \exp(a_i - c)\right)$$

$$= \log\left(\sum_{i=0}^{k} \exp(a_i - c + c)\right)$$

$$= \log\left(\sum_{i=0}^{k} \exp(a_i)\right) = LHS$$

The calculation is robust to overflow because

$\max_{i=0}^{K}\{a_i - c\} = c - c = 0$ and e^k will not cause

overflow when $k \leq 0$.

The calculation is robust to underflow because underflow scenario only rises when all elements in the vector are small. However, we know that the value $\max_{i=0}^{k}\{a_i - c\} = 0$ will always exist and this ensures the term inside log is always greater or equal to $e^0 = 1$, preventing underflow.

2.

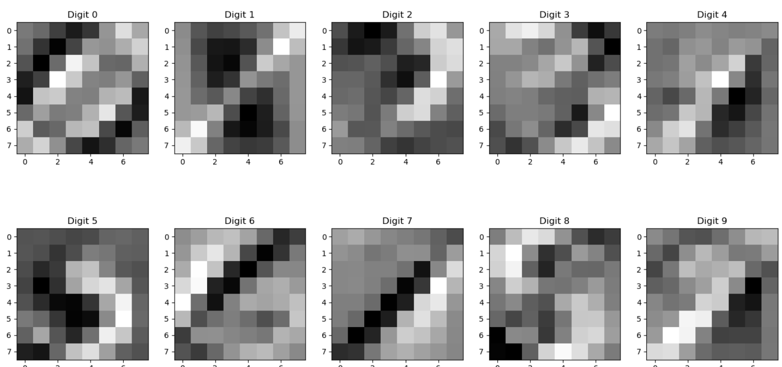
a) Train avg log likelihood: -2.53

Test avg log likelihood: -2.60

b) Train accuracy: 98.14%

Test accuracy: 97.28%

c)



3.

a) By Bayes' rule:

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)} \propto P(D \mid \theta) \cdot P(\theta)$$

By assumptions of independence

$$P(D \mid \theta) = \prod_{i=1}^{N} P(x^{(i)} \mid \theta)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{K} \theta_j^{x_j^{(i)}}$$

Therefore $P(\theta \mid D) \propto \prod_{i=1}^{N} \prod_{j=1}^{K} \theta_j^{x_j^{(i)}} \cdot \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} = \prod_{j=1}^{K} \theta_j^{N_j + \alpha_j - 1}$

b) $\log(P(\theta \mid D))$

$$= \log\left(\prod_{i=1}^{N} \prod_{j=1}^{K} \theta_j^{x_j^{(i)}} \cdot \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \right)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} x_j^{(i)} \log(\theta_j) + \sum_{j=1}^{K} (\alpha_j - 1) \log(\theta_j)$$

Substitute $\theta_K = 1 - \sum_{p=1}^{K-1} \theta_p$ into $\log(P(\theta \mid D))$

$$= \sum_{i=1}^{N} \left[\sum_{j=1}^{K-1} x_j^{(i)} \log(\theta_j) + x_K^{(i)} \log\left(1 - \sum_{p=1}^{K-1} \theta_p\right) \right] + \sum_{j=1}^{K-1} (\alpha_j - 1) \log(\theta_j)$$

$$+ (\alpha_K - 1) \log\left(1 - \sum_{p=1}^{K-1} \theta_p\right)$$

Case 1:

$\forall j \neq K$,

$$\frac{\partial \log(P(\theta \mid D))}{\partial \theta_j} = \sum_{i=1}^{N} \left[\frac{x_j^{(i)}}{\theta_j} + \frac{-x_K^{(i)}}{1 - \sum_{p=1}^{K-1} \theta_p} \right] + \frac{\alpha_j - 1}{\theta_j} + \frac{-(\alpha_K - 1)}{1 - \sum_{p=1}^{K-1} \theta_p}$$

Substitute $\theta_k = 1 - \sum_{p=1}^{k-1} \theta_p$ back to the f.o.c

$$= \frac{\sum_{i=1}^{N} x_j^{(i)}}{\theta_j} - \frac{\sum_{i=1}^{N} x_k^{(i)}}{\theta_k} + \frac{\alpha_j - 1}{\theta_j} - \frac{\alpha_k - 1}{\theta_k} = 0$$

By definition $\sum_{i=1}^{N} x_j^{(i)} = N_j$, $\sum_{i=1}^{N} x_k^{(i)} = N_k$

$$= \frac{N_j + \alpha_j - 1}{\theta_j} - \frac{N_k + \alpha_k - 1}{\theta_k} = 0$$

$$\hat{\theta}_j = \frac{\hat{\theta}_k (N_j + \alpha_j - 1)}{N_k + \alpha_k - 1}$$

Case 2: $\hat{\theta}_k$

$$1 = \sum_{j=1}^{K} \hat{\theta}_j = \sum_{j=1}^{K-1} \hat{\theta}_j + \hat{\theta}_k = \sum_{j=1}^{K-1} \frac{\hat{\theta}_k (N_j + \alpha_j - 1)}{N_k + \alpha_k - 1} + \frac{\hat{\theta}_k (N_k + \alpha_k - 1)}{N_k + \alpha_k - 1}$$

$$= \frac{\hat{\theta}_k}{N_k + \alpha_k - 1} \left(\sum_{j=1}^{K} N_j + \sum_{j=1}^{K} \alpha_j - \sum_{j=1}^{K} 1 \right)$$

$$= \frac{\hat{\theta}_k}{N_k + \alpha_k - 1} \left(N - K + \sum_{j=1}^{K} \alpha_j \right)$$

$$\therefore \hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N - K + \sum_{j=1}^{K} \alpha_j}$$

Substitute $\hat{\theta}_k$ back into $\hat{\theta}_j$ we get

$$\hat{\theta}_j = \frac{N_j + \alpha_j - 1}{N - K + \sum_{j=1}^{K} \alpha_j}$$

and this can be generalized for $\hat{\theta}_k$ as well.

c) $P(x^{(N+1)} \mid D) = \int p(x^{(N+1)} \mid \theta)\, p(\theta \mid D)\, d\theta$

$\Rightarrow P(x^{(N+1)} = k \mid D) = \int P(x^{(N+1)} = k \mid \theta)\, p(\theta \mid D)\, d\theta$

We can express $P(\theta_i \mid D) = \int p(\theta_i, \theta_{\neq i} \mid D)\, d\theta_{\neq i}$

where $\theta_{\neq i}$ denotes a vector $(\theta_1, \theta_2, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots \theta_K)$

that excludes θ_i. by properties of convolution.

$$\Rightarrow P(x^{(N+1)} = k \mid D) = \int_{\theta_k} \int_{\theta_{\neq k}} p(X^{(N+1)} = k \mid \theta, D)\, p(\theta \mid D)\, d\theta_{\neq k}\, d\theta_k$$

Also, we know that $P(X^{(N+1)} = k \mid \theta, D) = \hat{\theta}_k$

$$\Rightarrow P(x^{(N+1)} = k \mid D) = \int_{\theta_k} \int_{\theta_{\neq k}} \theta_k\ p(\theta \mid D)\, d\theta_{\neq k}\, d\theta_k$$

$$= \int_{\theta_k} \theta_k \int_{\theta_{\neq k}} p(\theta \mid D)\, d\theta_{\neq k}\, d\theta_k$$

$$= \int_{\theta_k} \theta_k \underline{\int_{\theta_{\neq k}} P(\theta_k, \theta_{\neq k} \mid D)\, d\theta_{\neq k}}\, d\theta_k$$

$$\downarrow$$

$$= \int_{\theta_k} \theta_k\, p(\theta_k \mid D)\, d\theta_k$$

Which is the definition of $E(\theta_k \mid D)$ in probability theory

From (a), We know that $p(\theta \mid D) \propto \prod_{i=1}^{N} \prod_{j=1}^{K} \theta_j^{x^{(i)}} \cdot \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$

$$= \prod_{j=1}^{K} \theta_j^{N_j + \alpha_j - 1}$$

Therefore $E(\theta_K \mid D) = \dfrac{N_K + \alpha_K}{\sum_{j=1}^{K} N_j + \alpha_j}$

$\therefore \quad P(x^{(n+1)} = K \mid D) = \dfrac{N_K + \alpha_K}{N + \sum_{j=1}^{K} \alpha_j}$