Q1.

(a)

```
Computing the log-sum-exp of a=[-1000 -1000 -1000]
        Unstable: -inf
        Stable: -998.9013877113318
Computing the log-sum-exp of b=[1000 1000 1000]
        Unstable: inf
        Stable: 1001.0986122886682
```

(b)

$$\log\left(\sum_{i=0}^{k}\exp\left(a_i-\max_{j=1}^{k}\{a_j\}\right)\right)+\max_{j=1}^{k}\{a_j\}$$

$$=\log\left[\sum_{i=0}^{k}\left(\frac{\exp(a_i)}{\exp(\max_{j=1}^{k}\{a_j\})}\right)\right]+\max_{j=1}^{k}\{a_j\}$$

$$=\log\left[\left(\sum_{i=0}^{k}\exp(a_i)\right)/\exp\left(\max_{j=1}^{k}\{a_j\}\right)\right]+\max_{j=1}^{k}\{a_j\}$$

$$=\log\left[\sum_{i=0}^{k}\exp(a_i)\right]-\log\circ\exp\left(\max_{j=1}^{k}\{a_j\}\right)+\max_{j=1}^{k}\{a_j\}$$

$$=\log\left[\sum_{i=0}^{k}\exp(a_i)\right]-\max_{j=1}^{k}\{a_j\}+\max_{j=1}^{k}\{a_j\}$$

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

Thus, the stable implementation should 'theorticolly' equal to the log-sum-exp: $\log\sum e^{a_i}$.

When all $a_i$ are very small (underflow in unstable), we know $a_i-\max_{j=1}^{k}\{a_j\}\approx 0$ (their values are so small that its difference can't be larger), thus the exp value $\approx 1$, leading to avoid a very small input for logarithm.

Similarly, when $a_i$ is very large, use $(a_i - \max_{j=1}^{k}\{a_j\})$ can guarantee its exponential value smaller or equal to 1.

Therefore, the stable version is robust.

## Q2.

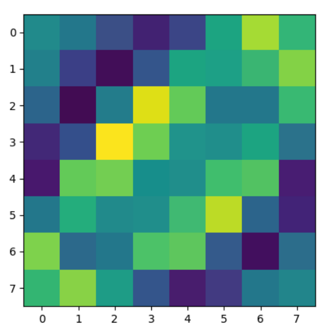
(a)

```
avg cond loglikelihood on training set = -0.12462443666863024.
avg cond loglikelihood on test set = -0.19667320325525564.
```

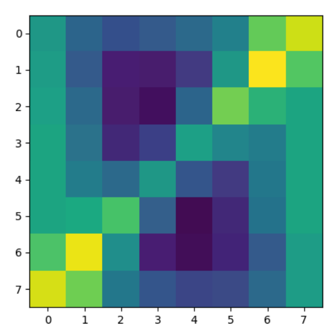
(b)

```
Training accuracy = 0.9814285714285714.
Test accuracy = 0.97275.
```

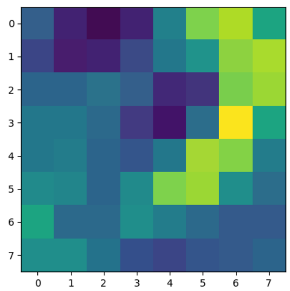
(c)



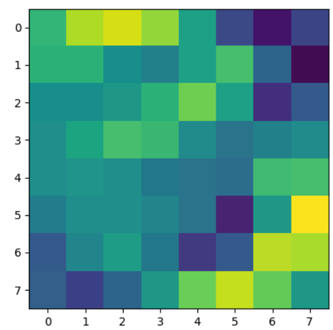

0



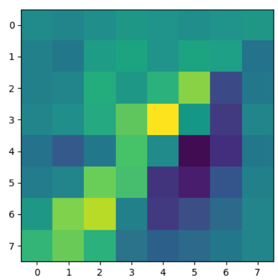

1


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

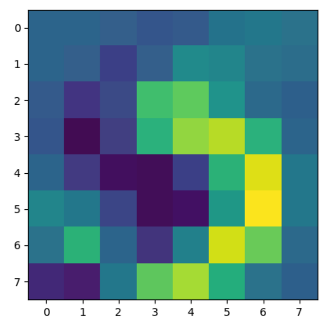
2


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

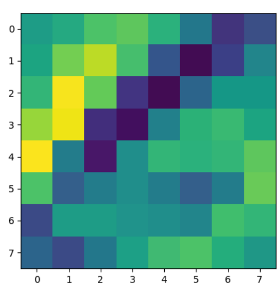
3


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

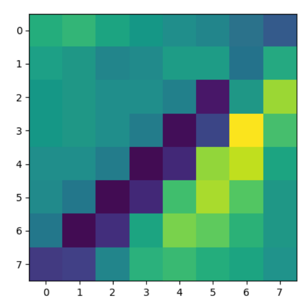
4


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

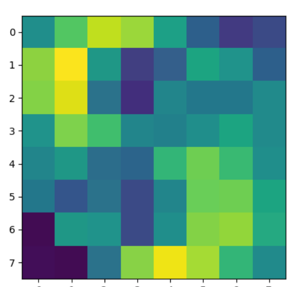
5


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

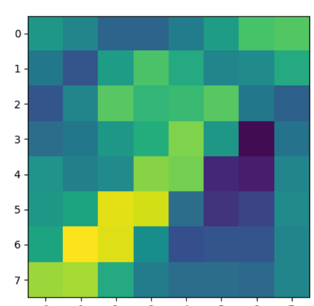
6


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

7


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

8


0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7

9

Q3.

(a)
$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)}$$

$$\Rightarrow p(\theta|D) \propto p(D|\theta)\,p(\theta)$$
$$\propto \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1} \cdot \prod_{i=1}^{N} p(x^{(i)}|\theta)$$
$$= \prod_{j=1}^{K} \theta_j^{a_j-1} \cdot \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}}$$
$$= \prod_{j=1}^{K} \theta_j^{a_j-1} \cdot \prod_{k=1}^{K} \theta_k^{\sum_{i=1}^{N} x_k^{(i)}}$$
$$= \prod_{k=1}^{K} \theta_k^{a_k-1+N_k} \qquad \left(\sum_{i=1}^{N} x_k^{(i)} = N_k\right)$$

(b)
$$\hat{\theta}_{MAP} = \underset{\theta}{\arg\max}\ p(D|\theta)\,p(\theta)$$
$$= \underset{\theta}{\arg\max} \prod_{k=1}^{K} \theta_k^{a_k-1+N_k}$$
$$= \underset{\theta}{\arg\max} \sum_{k=1}^{K} \log \theta_k^{a_k-1+N_k}$$

Define $f(\theta) = \sum_{k=1}^{K} \log \theta_k^{a_k-1+N_k}$, use Lagrange Multiplier

$$\begin{cases} \text{Maximize } f(\theta) \\ \text{subject to } \sum_k \theta_k - 1 = 0 \end{cases}$$

$$\Rightarrow \text{solve} \begin{cases} \nabla_\theta f - \lambda \nabla_\theta \left(\sum_k \theta_k - 1\right) = 0 \\ \sum_k \theta_k - 1 = 0 \end{cases} \Rightarrow \frac{N_k + a_k - 1}{\theta_k} + \lambda = 0$$

$$\theta_k = -\frac{N_k + a_k - 1}{\lambda} \Rightarrow$$

$$\sum_k \theta_k - 1 = -\frac{1}{\lambda} \sum_k \left(N_k + a_k - 1\right) - 1 = 0 \Rightarrow$$

$$\lambda = -N - \sum_k a_k + K$$

Thus $$\theta_j = -\frac{N_j + a_j - 1}{-N - \sum_k a_k + K}$$

$$= \frac{N_j + a_j - 1}{N + \sum_k a_k - K}, \quad \forall j$$

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

$$= \int \theta_k \int p(\theta_k, \theta_j \mid D)\, d\theta_j\, d\theta_k, \quad j \in \{1, \cdots, k\} \setminus \{k\}$$

$$= \int \theta_k\, p(\theta_k \mid D)\, d\theta_k$$

$$= E[\theta_k \mid D]$$

$$= \frac{a_k}{\sum_{k'} a_{k'}}$$

$(\theta \sim \text{Dirichlet}(a_1, \cdots, a_K))$