Question 1

a) It can be shown that the gaussian parameters μ_k, Σ_k that maximize the likelihood function for a *given class* k is the mean and sample variance of the observed points of the class. As per the specifications of the question statement, we assume a prior for each class that is uniform. Ordinarily, we might instead assume a multinomial distribution and take the prior as the number of occurrences of a class divided by total number of occurrences of all classes.

Average log-likelihood for training set:
28.6865797126
Average log likelihood for test:
27.4861999423

I also report the log-likelihood averaged over each class for each class. This is a more useful metric because it shows that the classifier is assigning a higher probability to the correct class relative to the others. As one might expect, the diagonals have the highest number, i.e. the highest probability, so the classifier is working correctly.

Average log-likelihood for training set (ith row is average over all points with ith label):

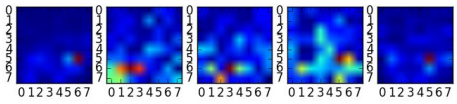
```
[[ 33.92402221 -88.45111834 -16.8932622  -25.83617441 -95.99459913, -28.0469699  -33.43930375 -128.00676019 -40.12341711 -120.7505451 ]
 [ -68.22979333  43.8330528   -8.62313018 -41.79892253  -6.16921495, -42.54976231 -22.89991813  -33.06720503   4.20662396  -19.78080586]
 [ -48.25937897 -55.92309386  16.88110064 -28.9959108  -65.50319452, -47.8188899  -54.71585182  -84.2803214  -35.97571408  -79.85123569]
 [ -37.76535226 -60.50596233 -11.52881147  25.34749836 -80.23781057, -5.942891   -61.89235394  -74.20703712  -7.77750941  -63.81771045]
 [ -63.57874438 -38.1128814  -27.37434988 -43.54955558  24.20650818, -37.62118033 -38.11452022  -49.30644691 -19.02665868  -3.93941094]
 [ -31.92250461 -45.89294551 -23.0214228   0.2051562  -56.8635482, 24.12660812 -36.96891002  -84.4180825   -2.6016538  -53.5127435 ]
 [ -16.08499354 -65.52399532 -17.22797111 -38.51566436 -59.26386047,  -19.90746009  29.24659472 -165.40709904 -29.97466814 -126.96745943]
 [ -47.34810481 -22.74497808   0.17304175 -13.15842334 -22.22716495,  -39.09215712 -80.81237908   38.53051509  -7.68591751   9.77505843]
 [ -51.89004965 -33.31890348 -15.94135661 -21.54837236 -45.81703868, -18.8929108  -47.39728619  -66.1336154   20.02956548  -37.4736449 ]
 [ -65.1618034  -46.64166345 -24.55400204 -32.3243134   -3.74088864,  -39.22988767 -85.57166051   -7.67267147  -5.99624188   30.74033154]]
```

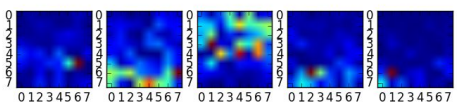
Average log likelihood for test:

```
[[ 33.21271943 -85.10474932 -14.92668303 -25.03919668 -92.24055178,  -28.41301758 -31.98166934 -121.11671514 -38.9435702  -115.68602049]
 [ -65.98273597  43.75667838  -7.89988668 -40.78305048  -4.21465107, -41.85439817 -19.49559598  -27.26307905   5.51581709  -17.43312634]
 [ -48.40011136 -56.79648782  15.76987273 -32.02961238 -64.05056948,-48.58909001 -52.90835733  -87.94971095 -36.4708241   -81.13238684]
 [ -36.61811894 -59.0998393  -10.81252291  23.18843659 -79.32152079,-5.49772638  -60.31175001  -75.53668363  -8.21771127  -64.68566461]
 [ -63.6224916  -38.31251625 -27.68532145 -42.79967361  23.16501442,-36.98680559 -38.8213785   -48.45629329 -19.54241646  -2.7608024 ]
 [ -28.61602964 -40.69825848 -20.36105698   3.672266   -55.92112179, 24.03785909 -32.49333123  -81.85722903  -1.63104965  -53.09672336]
 [ -17.94182886 -70.13173546 -19.32409691 -40.17621616 -61.330969, -20.45038953  26.35234161 -169.74575806 -30.4058043  -128.24679529]
 [ -48.50416378 -22.52592109  -0.48298009 -13.97483349 -20.82250406, -38.86921001 -81.1159767   37.69449875  -8.12088509   9.78420061]
 [ -53.05155379 -34.54053281 -17.60860408 -22.40460407 -46.75337517,-19.61290705 -49.00099401  -68.0970773   18.6596498  -38.65401494]
 [ -64.02644638 -46.67289562 -22.92041583 -30.60657673  -4.95359756,-38.41795983 -86.48303501   -6.37184554  -5.88093645   29.02492862]]
```

b) Accuracy on training set: 0.981429
Accuracy on test set: 0.972750

c)





Question 2

2a) $P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)} \sim P(D|\Theta)P(\Theta)$

As given, $P(\Theta) := \Theta_1^{a_1-1} \cdots \Theta_K^{a_K-1} = \prod_{i=1}^{K} \Theta_i^{a_i-1}$

$P(x_1, \ldots, x_N|\Theta) = P(x_1|\Theta) \cdots P(x_N|\Theta)$ assuming independence

$= \left(\prod_{k=1}^{K} \Theta_k^{x_1^k}\right) \cdots \left(\prod_{k=1}^{K} \Theta_k^{x_N^k}\right)$

$= \prod_{n=1}^{N} \prod_{k=1}^{K} \Theta_k^{x_n^k} = \Theta_1^{N_1} \cdots \Theta_K^{N_K}$

$$\therefore P(\Theta|D) = \left(\prod_{n=1}^{N} \prod_{k=1}^{K} \Theta_k^{x_n^k}\right)\left(\prod_{i=1}^{K} \Theta_i^{a_i-1}\right) = P(\Theta_1, \ldots, \Theta_K|D)$$

$$P(D'|D) = \int P(\Theta|D) P(D'|\Theta) d\Theta$$

$\rightarrow \Theta_i \in (0,1)$. But $\sum_{i=1}^{K} \Theta_i = 1$.

How to do $\int d\Theta$?

$$P(D'|\Theta) = P(X_b^{N+1} = 1|\Theta) = \Theta_b$$

$$\therefore P(D'|D) = \int P(\Theta|D)\, \Theta_b\, d\Theta =$$

$$\int \cdots \int \left(\prod_{n=1}^{N} \prod_{k=1}^{K} \Theta_k^{x_n^k}\right)\left(\prod_{i=1}^{K} \Theta_i^{a_i-1}\right) \Theta_b\, d\Theta_1 \cdots d\Theta_K$$

$$= \int \cdots \int \Theta_1^{N_1} \cdots \Theta_K^{N_K} \cdot \Theta_1^{a_1-1} \cdots \Theta_K^{a_K-1} \cdot \Theta_b\, d\Theta_1 \cdots d\Theta_K$$

$$= \int \cdots \int \underbrace{\Theta_1^{N_1+a_1-1} \cdots \Theta_K^{N_K+a_K-1}}_{\text{Dirichlet} \sim (N_1+a_1, \ldots, N_K+a_K)} \Theta_b\, d\Theta_1 \cdots d\Theta_K$$

$E[\Theta_b]$ given $(\Theta_1, \ldots, \Theta_K) \sim \text{Dirichlet}(N_1+a_1, \ldots, N_K+a_K)$

$$P(X_b^{N+1} = 1|\Theta) = P(D'|D) = E[\Theta_b] = \frac{N_b + a_b}{\sum_{i=1}^{K} (N_i + a_i)}$$

where $N_i = \sum_{n=1}^{N} x_i^{(n)}$

2b) $P(\theta|D) \sim P(D|\theta)\,P(\theta) \sim \log P(D|\theta) + \log P(\theta)$

Recall: $P(D|\theta) = \theta_1^{N_1} \cdots \theta_K^{N_K}$

$P(\theta) = \theta_1^{a_1-1} \cdots \theta_K^{a_K-1}$

$\therefore P(\theta|D) \sim N_1 \log\theta_1 + \ldots + N_K \log\theta_K + (a_1-1)\log\theta_1 + \ldots + (a_K-1)\log\theta_K$

$(N_1+a_1-1)\log\theta_1 + \ldots + (N_K+a_K-1)\log\theta_K$

$\frac{dP(\theta|D)}{d\theta_i} = \frac{N_i+a_i-1}{\theta_i} = 0$ X. Need lagrange again.

Maximize $P(\theta|D) \quad :=f$

subject to $1-(\theta_1+\ldots+\theta_K)=0 \quad :=g$

Lagrangian Method: $\nabla_\theta f = \lambda \nabla_\theta g$ — K equations

$g = 0$ — 1 equation

$\theta_1, \ldots, \theta_K, \lambda \rightarrow K+1$ unknowns

$\nabla_{\theta_i} f = \lambda \nabla_\theta g \Rightarrow \frac{N_i+a_i-1}{\theta_i} = \lambda(-1)$

$\theta_i = \frac{1-a_i-N_i}{\lambda}$

$\therefore 1 - \sum_{i=1}^{K} \theta_i = 0$

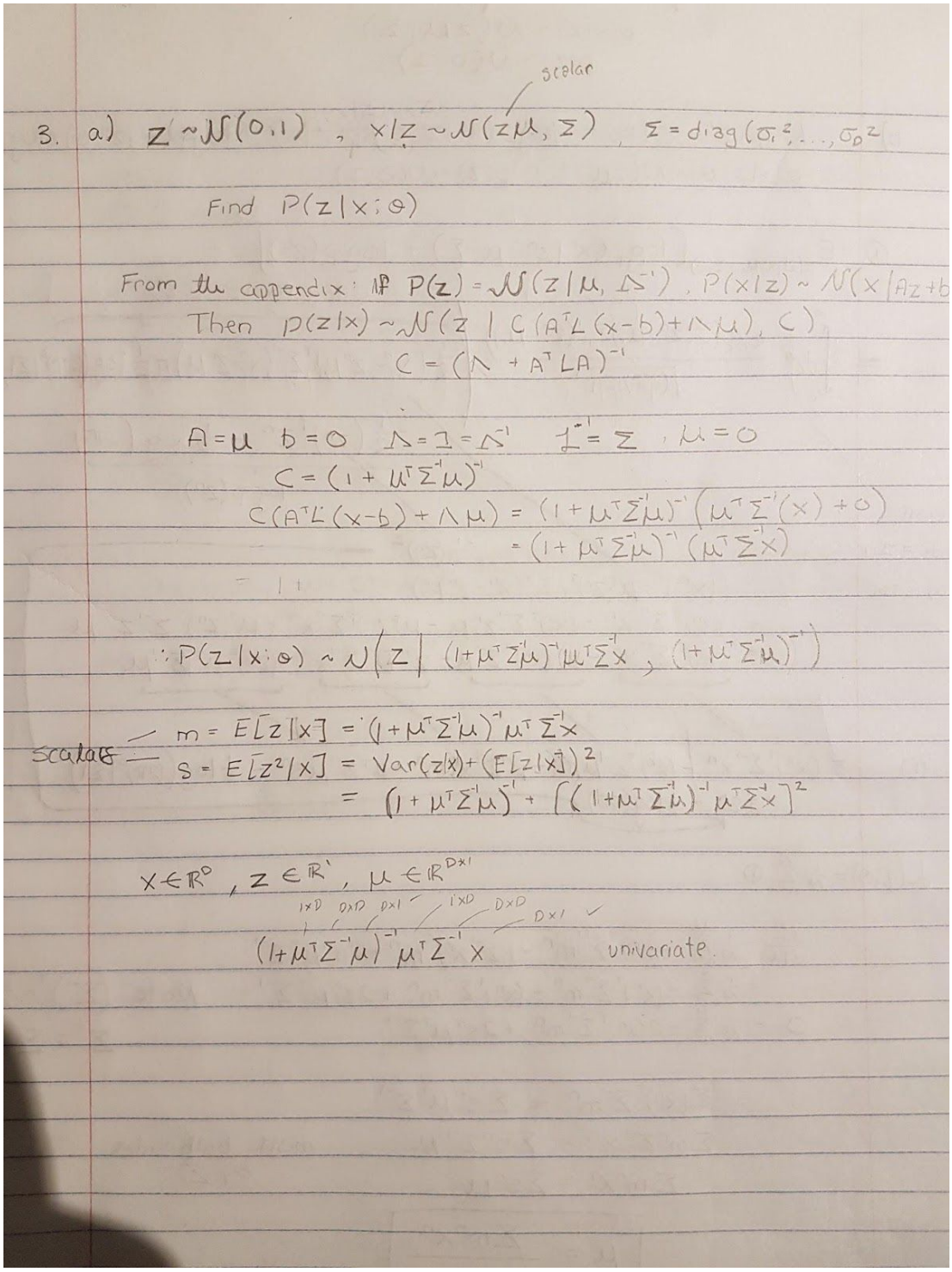
$1 - \sum_{i=1}^{K} \frac{1-a_i-N_i}{\lambda} = 0$

$1 = \frac{1}{\lambda}(\Sigma 1 - \Sigma a_i - \Sigma N_i)$

$\lambda = K - \Sigma a_i - N$

$$\therefore \boxed{\theta_i = \frac{1-a_i-N_i}{K-(\Sigma a_i)-N}} = \boxed{\frac{N_i+a_i-1}{N+\Sigma a_i - K}}$$

Question 3



3. a) $z \sim \mathcal{N}(0,1)$, $x|z \sim \mathcal{N}(z\mu, \Sigma)$ ($z\mu$: scalar) , $\Sigma = diag(\sigma_1^2, \ldots, \sigma_D^2)$

Find $P(z|x;\theta)$

From the appendix: If $P(z) = \mathcal{N}(z|\mu, \Lambda^{-1})$, $P(x|z) \sim \mathcal{N}(x|Az+b$

Then $p(z|x) \sim \mathcal{N}(z \mid C(A^T L(x-b)+\Lambda\mu), C)$

$C = (\Lambda + A^T L A)^{-1}$

$A = \mu \quad b = 0 \quad \Lambda = I = \Lambda^{-1} \quad I^{-1} = \Sigma \quad , \mu = 0$

$C = (1 + \mu^T \Sigma^{-1} \mu)^{-1}$

$C(A^T L(x-b) + \Lambda\mu) = (1 + \mu^T \Sigma^{-1}\mu)^{-1} (\mu^T \Sigma^{-1}(x) + 0)$

$= (1 + \mu^T \Sigma^{-1}\mu)^{-1} (\mu^T \Sigma^{-1} x)$

$\therefore P(z|x;\theta) \sim \mathcal{N}\left(z \mid (1+\mu^T\Sigma^{-1}\mu)^{-1}\mu^T\Sigma^{-1}x \;,\; (1+\mu^T\Sigma^{-1}\mu)^{-1}\right)$

scalar — $m = E[z|x] = (1+\mu^T\Sigma^{-1}\mu)^{-1}\mu^T\Sigma^{-1}x$

scalar — $s = E[z^2|x] = Var(z|x) + (E[z|x])^2$

$= (1 + \mu^T\Sigma^{-1}\mu)^{-1} + \left[(1+\mu^T\Sigma^{-1}\mu)^{-1}\mu^T\Sigma^{-1}x\right]^2$

$x \in \mathbb{R}^D$, $z \in \mathbb{R}^1$, $\mu \in \mathbb{R}^{D\times 1}$

$1\times D \quad D\times D \quad D\times 1 \quad 1\times D \quad D\times D \quad D\times 1$ ✓

$(1+\mu^T\Sigma^{-1}\mu)^{-1}\mu^T\Sigma^{-1}x$ univariate.

$$q = P(z|x) \sim \mathcal{N}\Big(\overbrace{(1+\mu^T\Sigma^{-1}\mu)^{-1}\mu^T\Sigma^{-1}x}^{\mu_1}, \overbrace{(1+\mu^T\Sigma^{-1}\mu)^{-1}}^{\Sigma_1}\Big)$$

$$p(x|z) \sim \mathcal{N}(z\mu, \Sigma)$$

$$p(z) \sim \mathcal{N}(0, I)$$

b) $\log p(z^n, x^n|\theta) = \log p(x^n|z^n,\theta)\cdot p(z^n|\theta) = \log p(x^n|z^n,\theta) + \log p(z^n|\theta)$

$p(x|z;\theta) \sim \mathcal{N}(z\mu, \Sigma)$ $\quad p(z) \sim \mathcal{N}(0, I)$

$$① = E_{p(z^n|x^n,\mu_{old},\Sigma_{old})}\left[\log p(x^n|z^n;\mu,\Sigma) + \log p(z^n)\right]$$

$$= \int_{-\infty}^{\infty} dz^n \; \frac{\exp\left(\frac{-1}{2}(z^n-\mu_i^{old})^T\Sigma_i^{-1}(z^n-\mu_i)\right)}{\sqrt{(2\pi)^k|\Sigma_i|}} \left(\overbrace{\frac{-1}{2}\left(x^n - z^n\mu\right)^T\Sigma^{-1}(x^n-z^n\mu) + \frac{1}{2}\log\left((2\pi)^D|\Sigma|\right)}^{\log p(x^n|z^n;\mu,\Sigma)} \underbrace{- \frac{1}{2}(z^n)^T z^n - \frac{1}{2}\log(2\pi)}_{\log p(z^n)}\right)$$

Recall: $\int_v f(x)dx = 1$ if $f(x)$ is PDF

$-\frac{1}{2}(z^n)^2$

$$\left((x^n)^T - \mu^T(z^n)^T\right)\Sigma^{-1}(x^n - z^n\mu)$$

$$= (x^n)^T\Sigma^{-1}x^n - (x^n)^T\Sigma^{-1}z^n\mu - \mu^T(z^n)^T\Sigma^{-1}x^n + \mu^T(z^n)^T\Sigma^{-1}z^n\mu$$

$$= \underbrace{(x^n)^T\Sigma^{-1}x^n}_{} \underbrace{- z^n(x^n)^T\Sigma^{-1}\mu}_{} \underbrace{- z^n\mu^T\Sigma^{-1}x^n}_{} \underbrace{+ (z^n)^2\mu^T\Sigma^{-1}\mu}_{}$$

$$① = (x^n)^T\Sigma^{-1}x^n - (x^n)^T\Sigma^{-1}\mu m^n - \mu^T\Sigma^{-1}x^n m^n + \mu^T\Sigma^{-1}\mu s^n - \frac{1}{2}\log\left((2\pi)^D|\Sigma|\right) - \frac{1}{2}s^n - \frac{1}{2}\log(2\pi)$$

$$\mathcal{L}(q,\theta) = \frac{1}{N}\sum_{n=1}^{N} ①$$

$$\frac{d\mathcal{L}}{d\mu} = \frac{1}{N}\sum_{n=1}^{N} -(x^n)^T\Sigma^{-1}m^n - \left(\Sigma^{-1}x^n\right)^T m^n + 2s^n\mu^T\Sigma^{-1}$$

$$= \frac{1}{N}\sum_{n=1}^{N} -(x^n)^T\Sigma^{-1}m^n - (x^n)^T\Sigma^{-1}m^n + 2s^n\mu^T\Sigma^{-1}$$

$$0 = \frac{1}{N}\sum_{n=1}^{N} -2(x^n)^T\Sigma^{-1}m^n + 2s^n\mu^T\Sigma^{-1}$$

Note: $\left(\Sigma^{-1}\right)^T = \Sigma^{-1}$, $\Sigma^T = \Sigma$

$$\sum (x^n)^T\Sigma^{-1}m^n = \sum s^n\mu^T\Sigma^{-1}$$

$$\sum m^n\Sigma^{-1}x^n = \sum s^n\Sigma^{-1}\mu$$

mult. both sides by Σ

$$\sum m^n x^n = \sum s^n\mu$$

$$\boxed{\mu = \frac{\sum m^n x^n}{\sum s^n}}$$

Note: it's a sum at the end, not the covariance. I dropped the bounds for convenience.

Proof of some derivatives used above: