Question 1.

i)

- Bernoulli Probabilistic mixture could have generated this data ,Since The Given data-set consists of only 0s and 1s.

- **X** be the data set in which each row X_i is a data point.
- **D** be the number of dimensions or features.
- λ_{ik} is the responsibility of data point X_i for the kth mixture.
- P_{ki} be the Bernoulli parameter of ith dimension for kth mixture.

$$\lambda_{n,k} = \frac{\pi_k \prod_{i=1}^{D} P_{k,i}^{x_{n,i}} (1 - P_{k,i})^{1-x_{n,i}}}{\sum_{m=1}^{K} \pi_m \prod_{i=1}^{D} P_{m,i}^{x_{n,i}} (1 - P_{m,i})^{1-x_{n,i}}}$$

$$\pi_m = \frac{\sum_{n=1}^{N} \lambda_{n,m}}{N}$$

$$P_m = \frac{1}{\sum_{n=1}^{N} \lambda_{n,m}} \sum_{n=1}^{N} \lambda_{n,m} X_n$$

Derivations :

$$P(X_i;\mu) = \sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \mu_{k,d}^{x_{i,d}} (1-\mu_{k,d})^{(1-x_{i,d})}$$

$$L(X;\pi,P,\lambda) = \sum_{i=1}^{N} log\left(\sum_{k=1}^{K} \lambda_{i,k} \frac{\pi_k \prod_{d=1}^{D} P_{k,d}^{x_{i,d}} (1-P_{k,d})^{(1-x_{i,d})}}{\lambda_{i,k}}\right)$$

Taking partial derivative and equating to zero

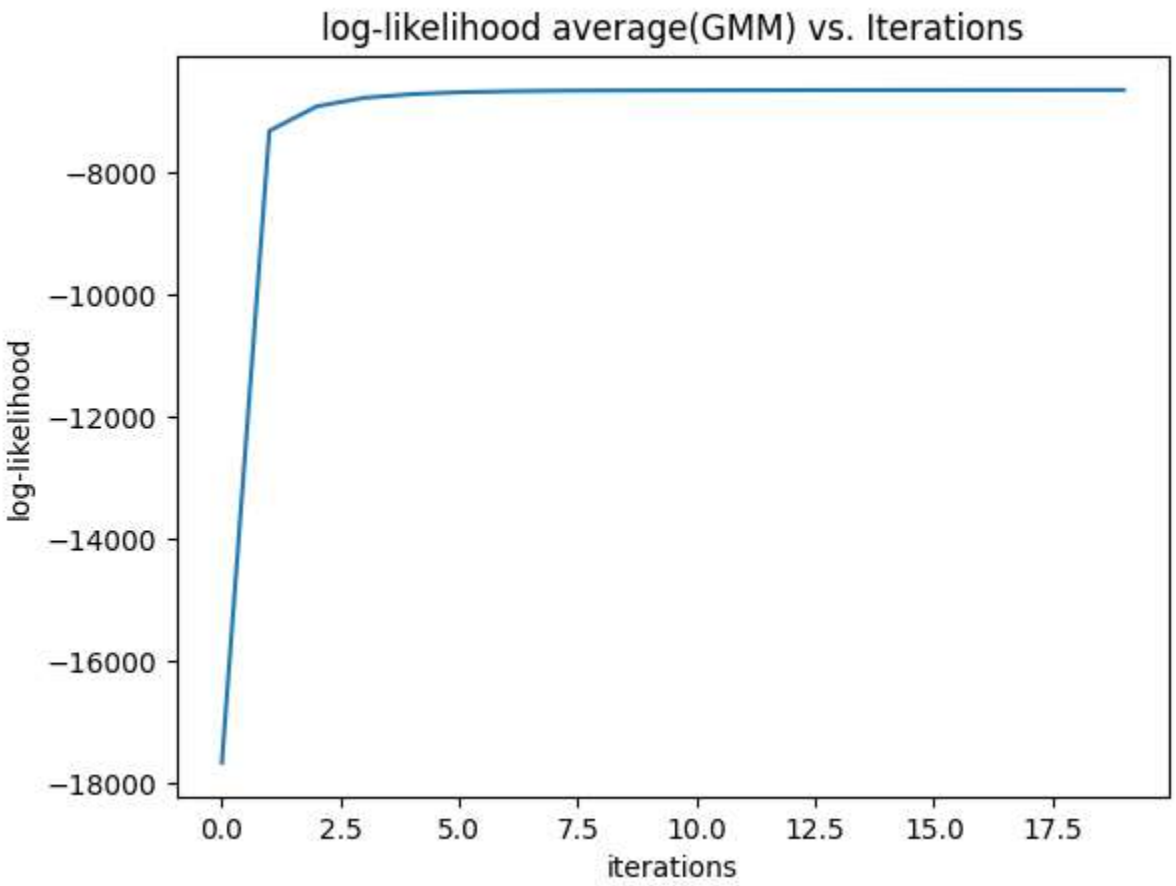
$$\sum_{i=1}^{N} \lambda_{i,k} x_{i,d}(1-P_{k,d}) - \lambda_{i,k}(1-x_{i,d})P_{k,d} = 0$$

$$P_{k,d}\left(\sum_{i=1}^{N} \lambda_{i,k} x_{i,d} + \sum_{i=1}^{N} \lambda_{i,k} - \sum_{i=1}^{N} \lambda_{i,k} x_{i,d}\right) = \sum_{i=1}^{N} \lambda_{i,k} x_{i,d}$$

$$P_{k,d} = \frac{\sum_{i=1}^{N} \lambda_{i,k} x_{i,d}}{\sum_{i=1}^{N} \lambda_{i,k}}$$

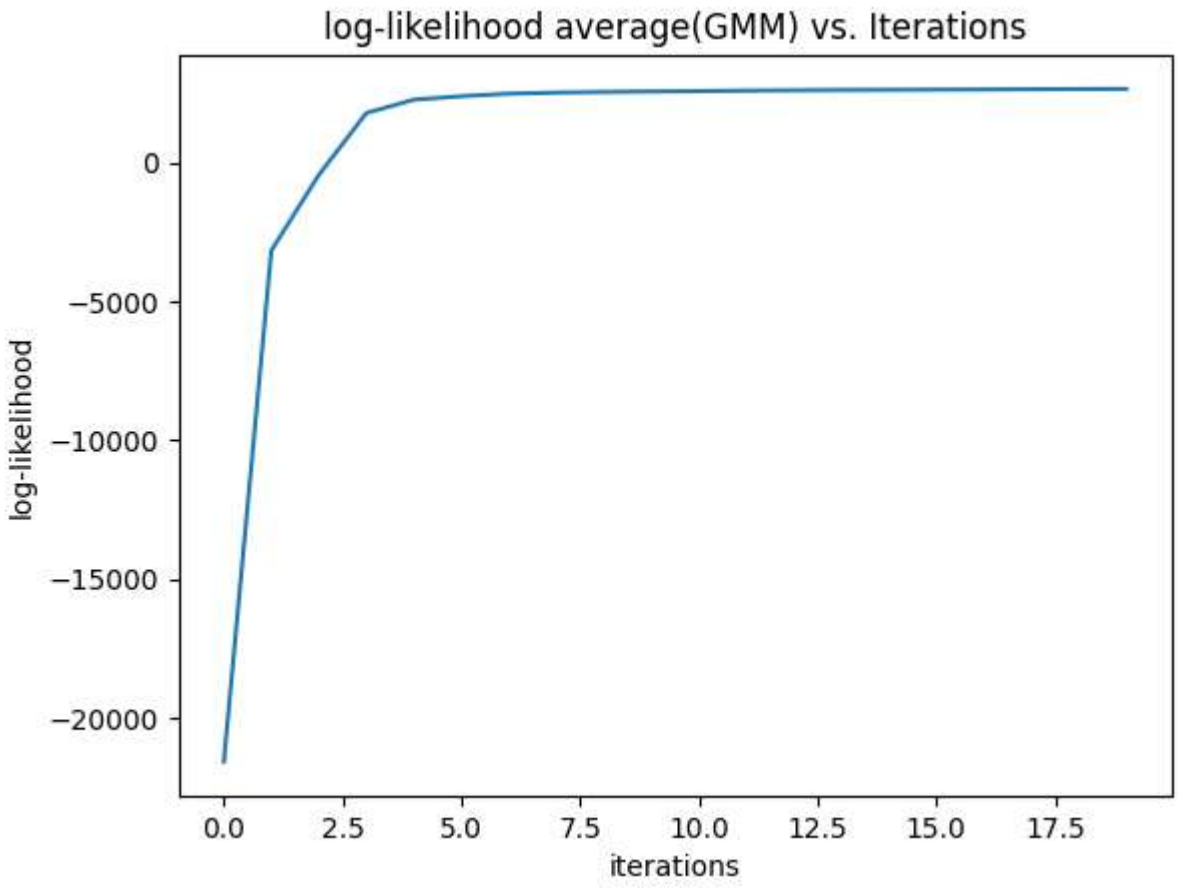
$$P_k = \frac{1}{\sum_{i=1}^{N} \lambda_{i,m}} \sum_{i=1}^{N} \lambda_{i,k} X_i$$

- Below is Plot between log-likelihood averaged over 100 random Initialisation vs number of iterations performed in Bernoulli.


log-likelihood average(GMM) vs. Iterations
log-likelihood
−8000
−10000
−12000
−14000
−16000
−18000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
iterations

ii)

- Below is Plot between log-likelihood averaged over 100 random Initialisation vs number of iterations performed in GMM.



Multivariate Gaussian Model.

- **X** be the data set in which each row X_i is a data point.
- **D** be the number of dimensions or features.
- **μ** is K x D matrix (where μ_k is the mean of k th mixture)
- **π** is K x 1 matrix(where π_k is the k th mixture probability)
- Σ_k is the covariance matrix of kth mixture.
- λ_{ik} is the responsibility of data point X_i for the kth mixture.

$$P(X|\mu,\Sigma) = \frac{1}{\sqrt[D]{(2\pi)}\sqrt{|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(X-\mu)\mathbf{\Sigma}^{-1}(X-\mu)^T\right)$$

$$P(X_i|k) = \frac{1}{\sqrt[D]{(2\pi)}\sqrt{|\mathbf{\Sigma}_k|}} \exp\left(-\frac{1}{2}(X_i-\mu_k)\mathbf{\Sigma}_k{}^{-1}(X_i-\mu_k)^T\right)$$

E-Step :

$$\begin{aligned}\lambda_k(X) &= \frac{p(X|k)p(k)}{\sum_{k=1}^{K} p(k)p(X|k)} \\ &= \frac{p(X|k)\pi_k}{\sum_{k=1}^{K} \pi_k p(X|k)}\end{aligned}$$

$$\begin{aligned}\lambda_{ik} &= \frac{p(X_i|k)p(k)}{\sum_{k=1}^{K} p(k)p(X|k)} \\ &= \frac{p(X_i|k)\pi_k}{\sum_{k=1}^{K} \pi_k p(X_i|k)}\end{aligned}$$

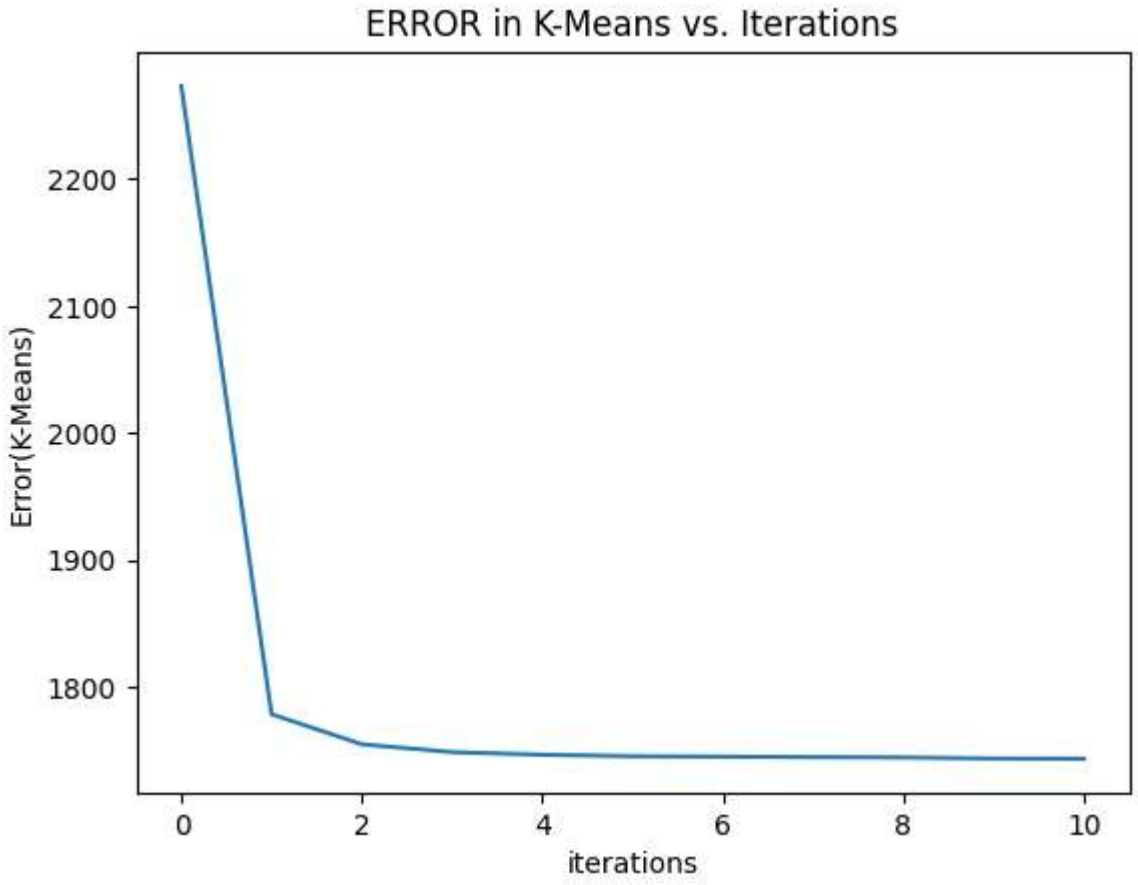
M-Step :

$$\Sigma_k = \frac{\sum_{n=1}^{N} nk(X_n-\mu_k)^T(X_n-\mu_k)}{\sum n = 1^N \lambda_{nk}}$$

$$\mu_k = \frac{1}{\sum_{i=1}^{N} \lambda_{i,m}} \sum_{i=1}^{N} \lambda_{i,k} X_i$$

iii)

- Below is the Plot for Objective Function vs number of iterations of K-means for the same data.



iv)

- Between mixture models and K-means,mixture models provide a better insight into the structure of data.
- Computationally GMM takes more time than Bernoulli.

- J
- Log Likelihood increases more rapidly in bernoulli compared to GMM because they both capture data in a different way.
- Bernoulli is better among the two.

- The above also suggest that we have to choose an appropriate probabilistic model for analyzing the data.
-

Question 2.

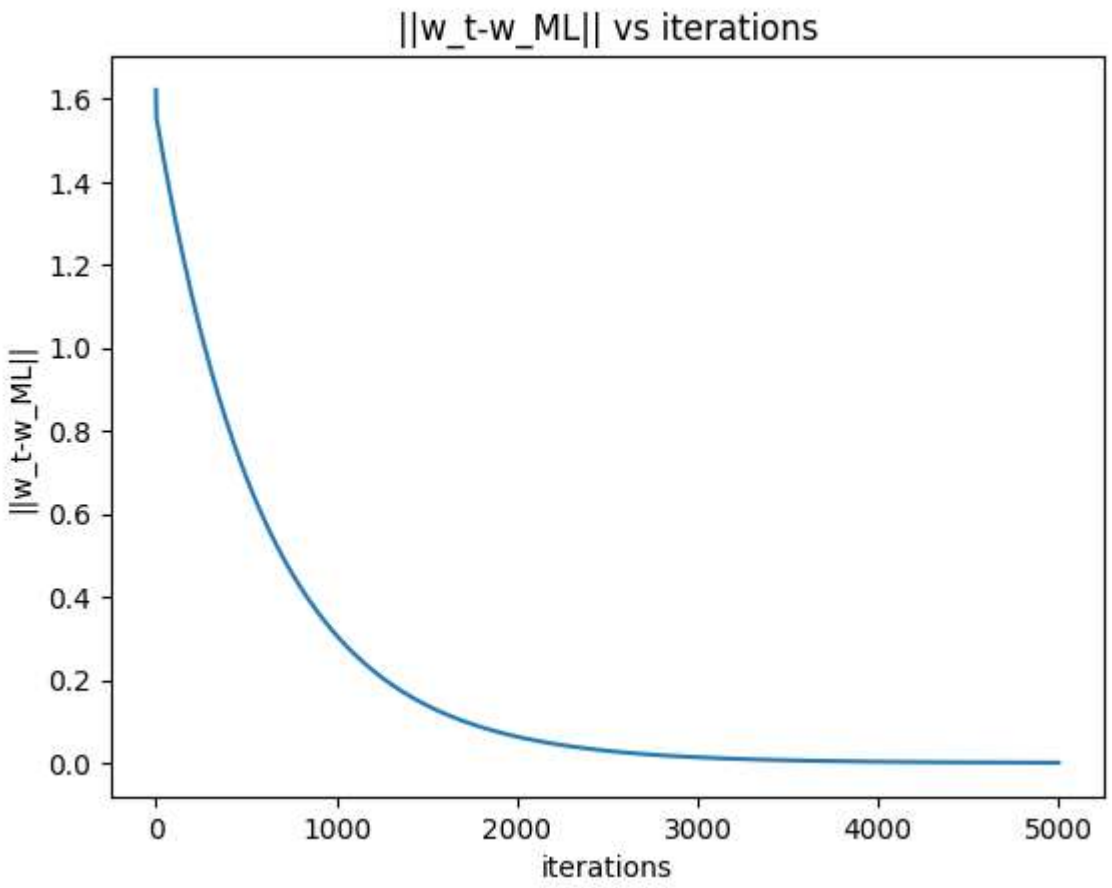
i)

- The least squares solution w_{ML} to the regression problem obtained by the analytical solution is given by

$$w_{ML} = (XX^T)XY$$

- The Train data error and Test data error w.r.t the above w_{ML} are 397 and 185 respectively.

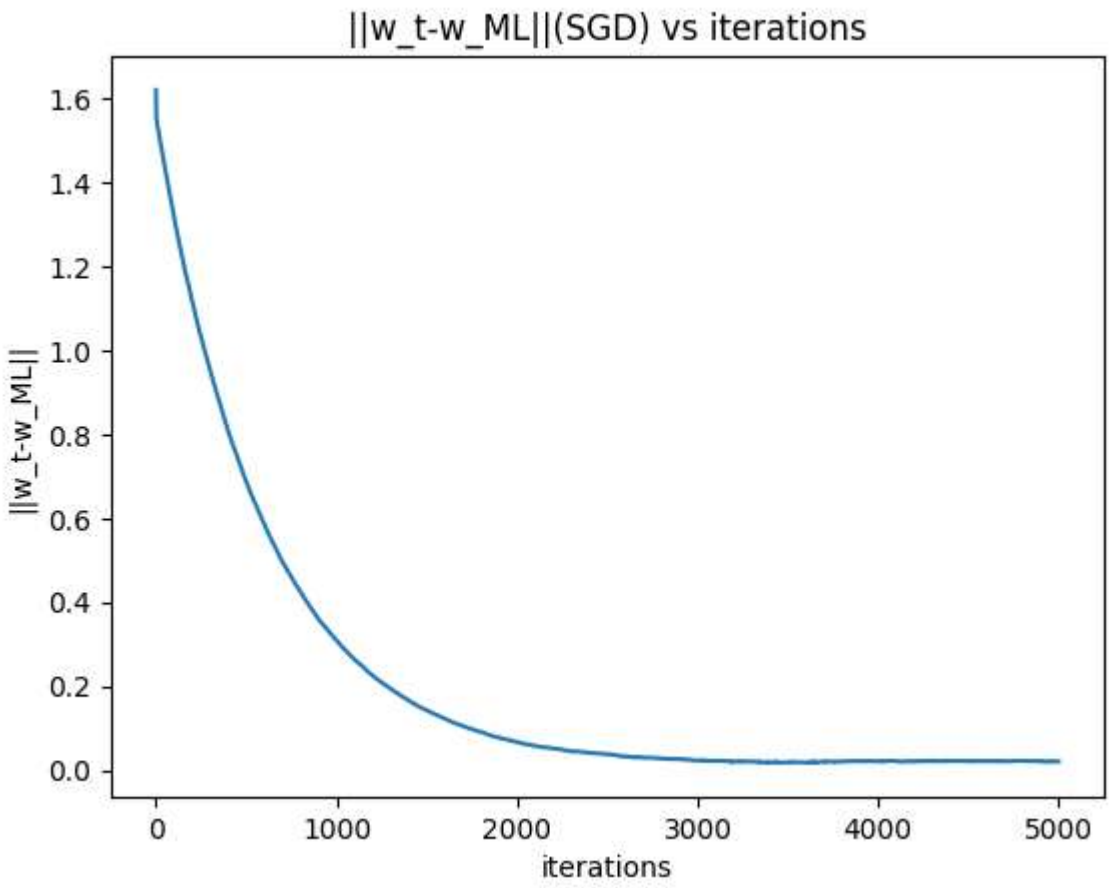
ii) Below is the for $||wt - wM||_2$ vs t(iterations) for gradient descent algorithm.



Observations :

- The graph(w^t)will converge to that of the analytical solution (w_{ML})with varying speeds depending on the step sizes.
- By choosing appropriate step size the graph (w^t) will converge rapidly to the analytical solution.

iii) Below is the for $||wt - wM||_2$ vs t(iterations) for Stochastic gradient descent algorithm.

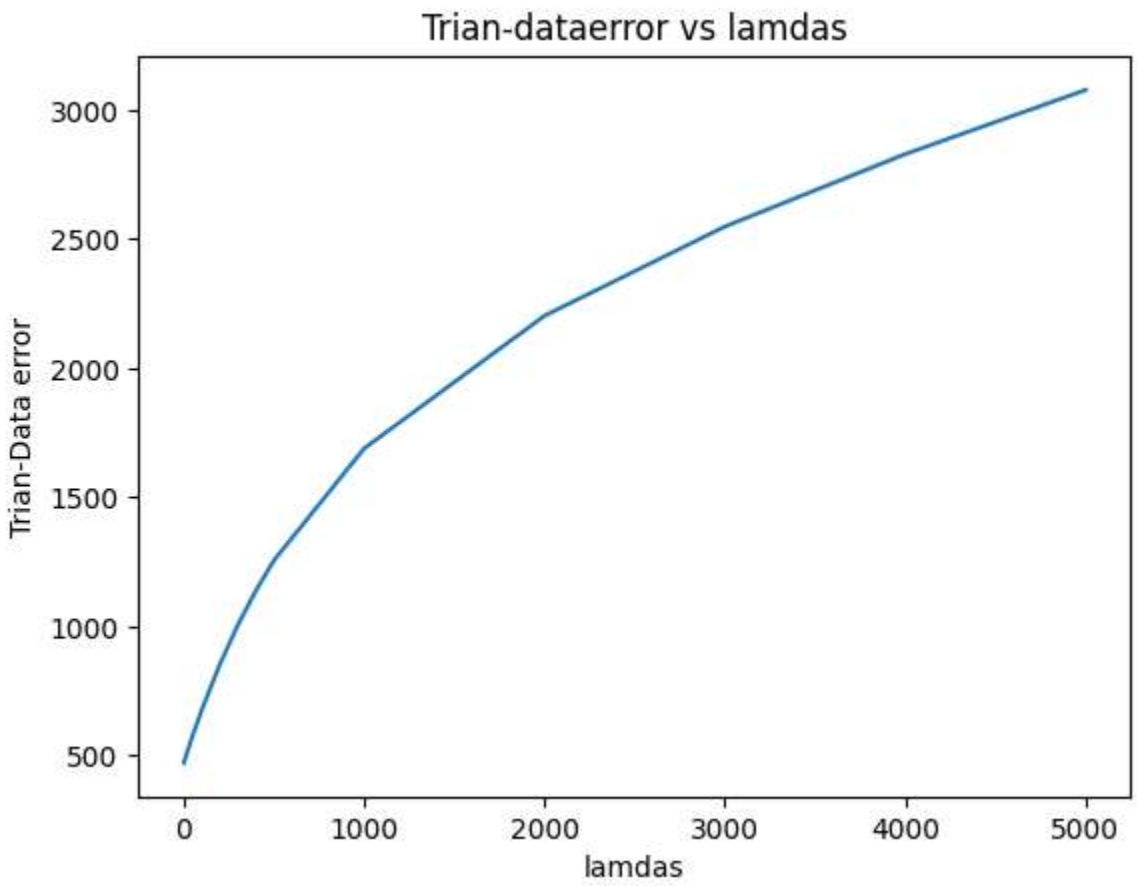


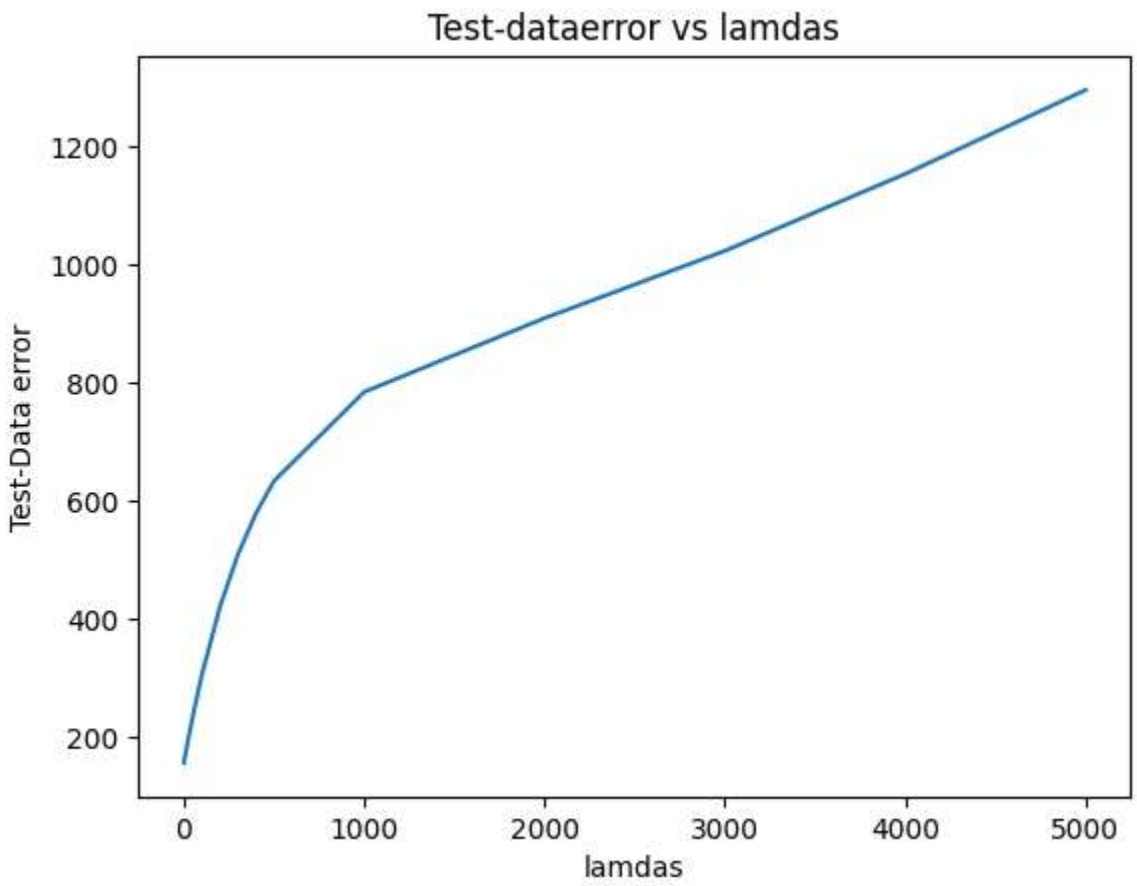
Observations :

- By choosing appropriate step size the graph (w^t) will converge rapidly to the analytical solution.
- The Graph is not that smooth compared to the gradient descent due to the stochastic nature of the algorithm.
- It will converge to analytical solution in lesser number of iterations compared to the normal gradient descent algorithm

iv)

The following are the plots for test and train data error with respect to λ .


Trian-dataerror vs lamdas
Trian-Data error
lamdas
3000
2500
2000
1500
1000
500
0
1000
2000
3000
4000
5000



- For λ = 0 the errors are minimum .
- So Ridge regression with λ = 0 is better than w_ML because the test data error for the former is lower.