**1.Undertake an initial exploratory analysis of the training data and summarise. [5 marks]**

**Your comment on the likely relevance of the variables for predicting “Heating Load”**

How to determine the importance of variable \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

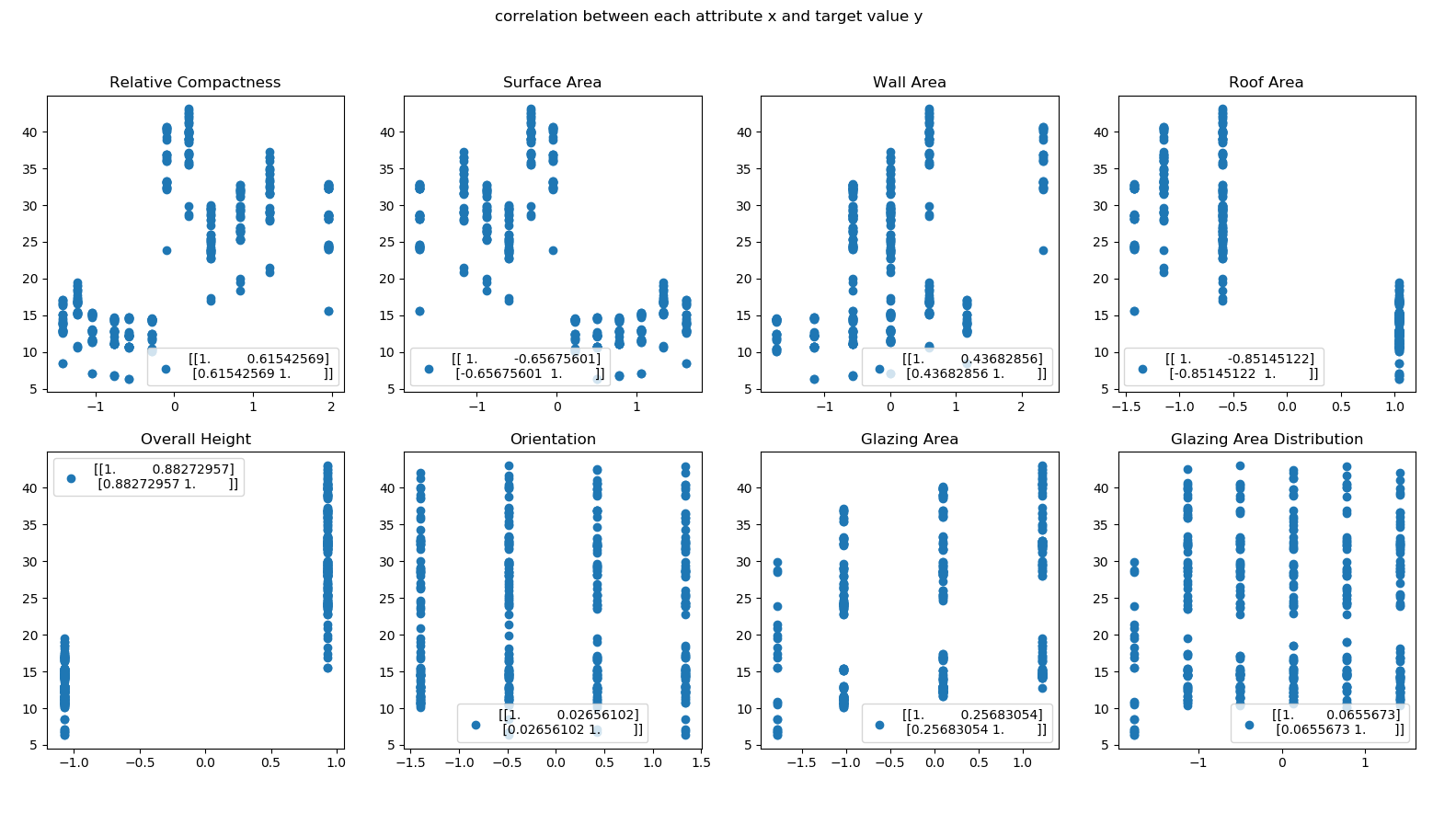
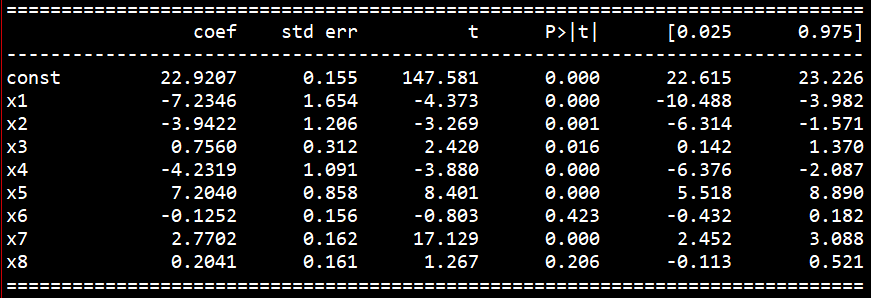
What is correlation \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

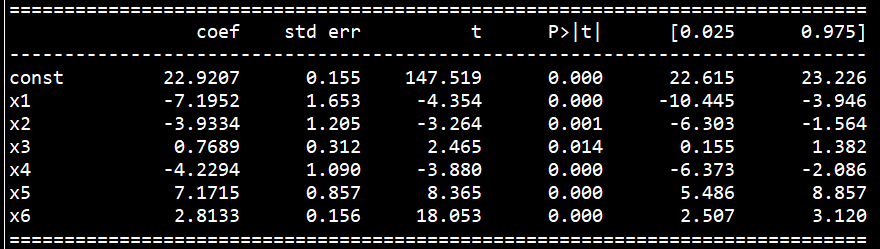
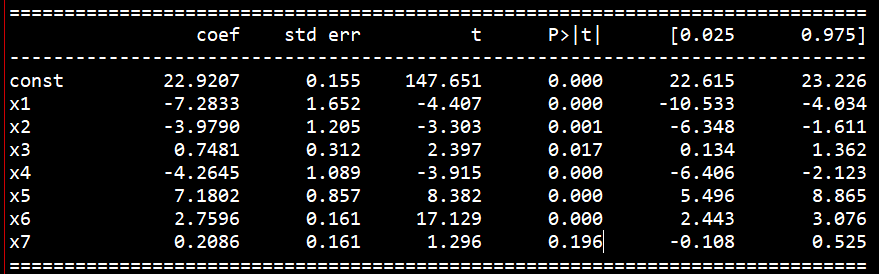
From correlation plot, the most un-important variable are orientation, glazing area, and glazing area distribution as their correlation to dependent variable y is very small such that the absolute value is less than 0.3

What is P-value and backward elimination \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

From backward elimination with significant level of 0.05, the first variable to be remove was orientation that have P-value of 0.423 which is higher than significant level. The second variable was glazing area orientation with P-value of 0.196. After these 2 variables have been remove¸ no variable have p-value more than significant level.

Perform forward elimination and stepwise elimination\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

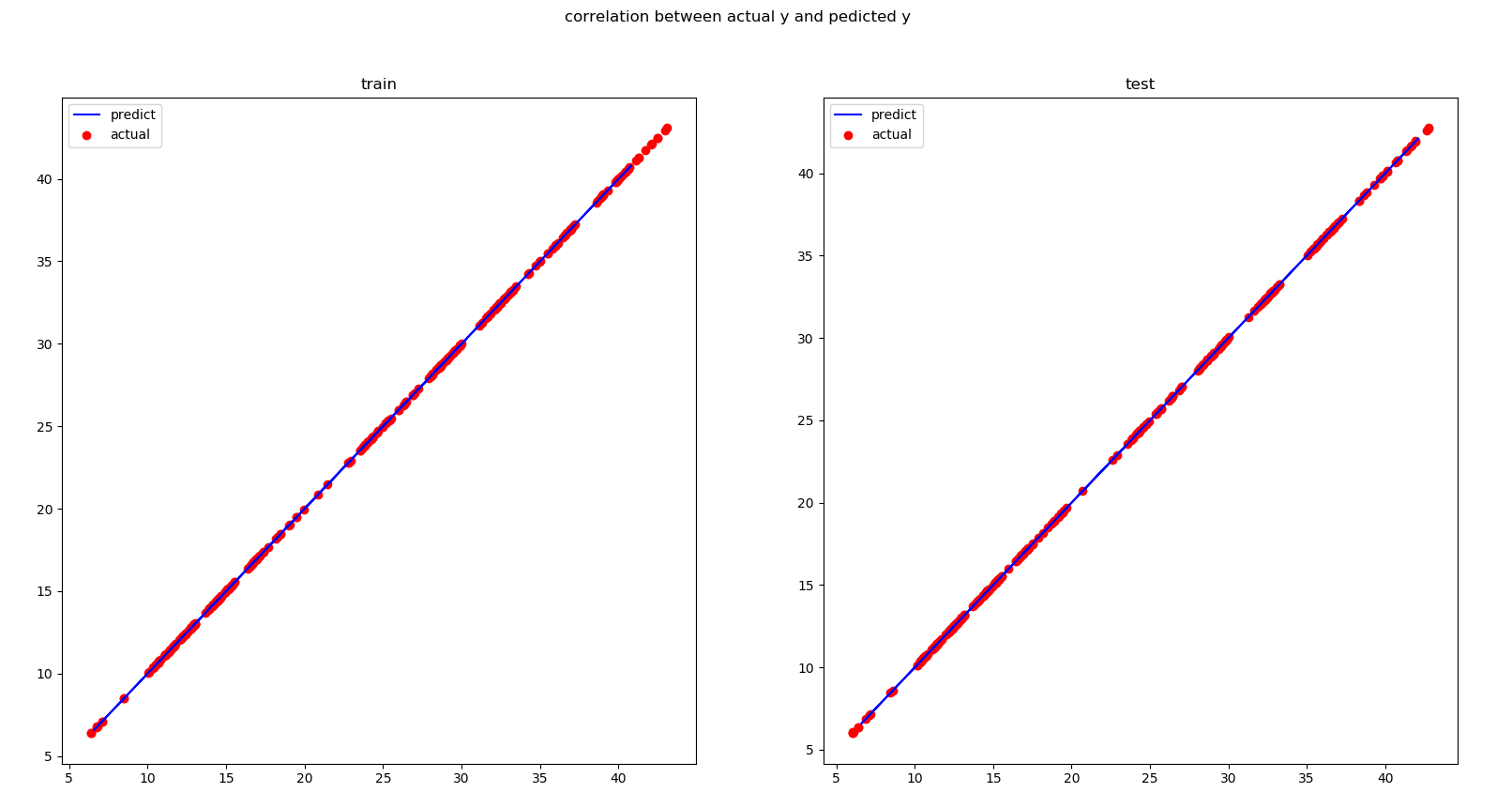
 



**Detail of the accuracy of the least-squares linear model, on both train and test sets, in terms of root-mean-square-error (RMSE)**

If all attributes are considered, train error is 3.0115517876503612 and test error is 3.0958865845448686. If consider only important variable train error is 3.020837151333591 and test error is 3.110716297089577

Since the data have more than 2 dimension, the train and test plot accuracy are generated in terms of correlation between value of dependent variable y.



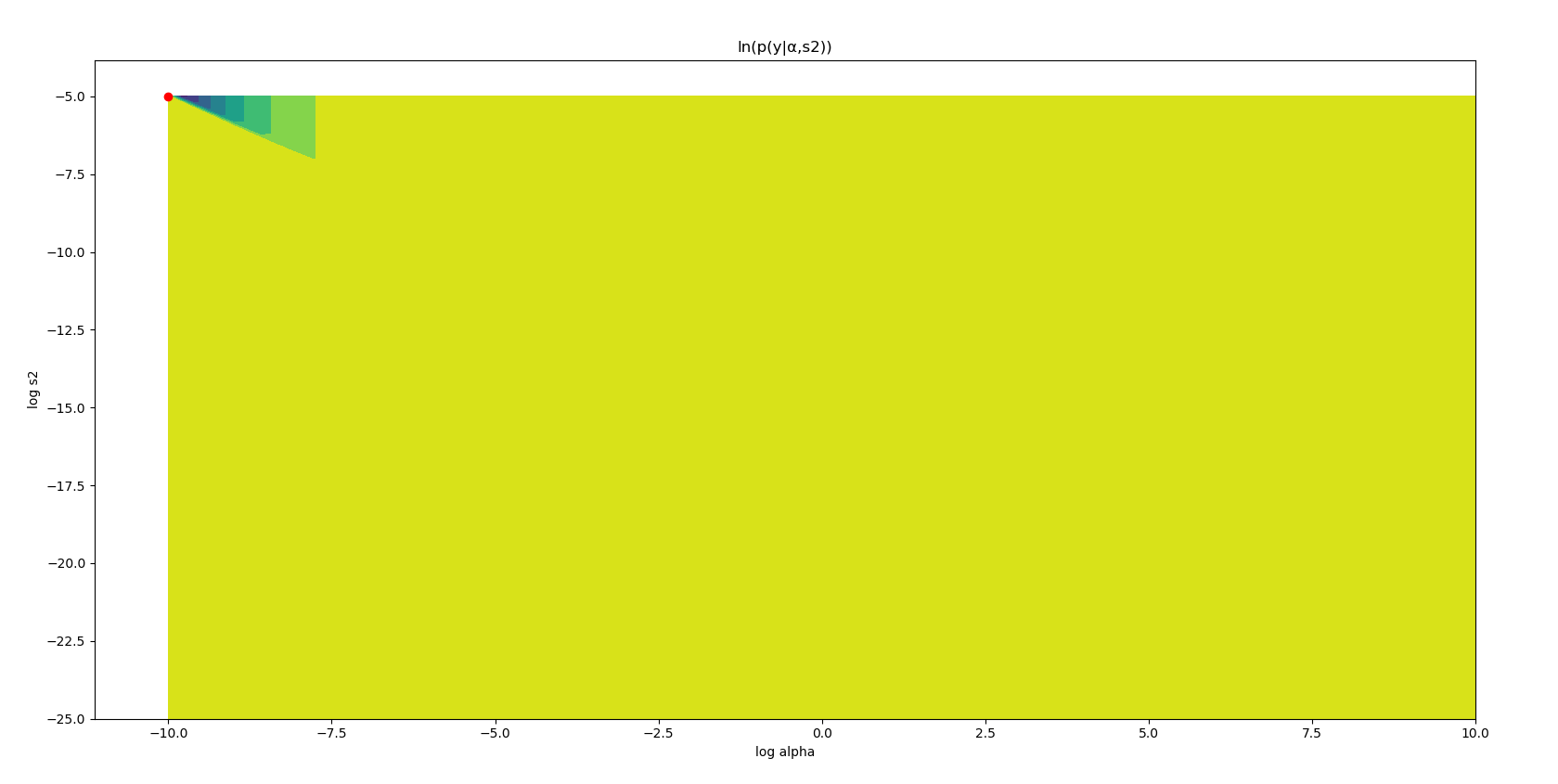
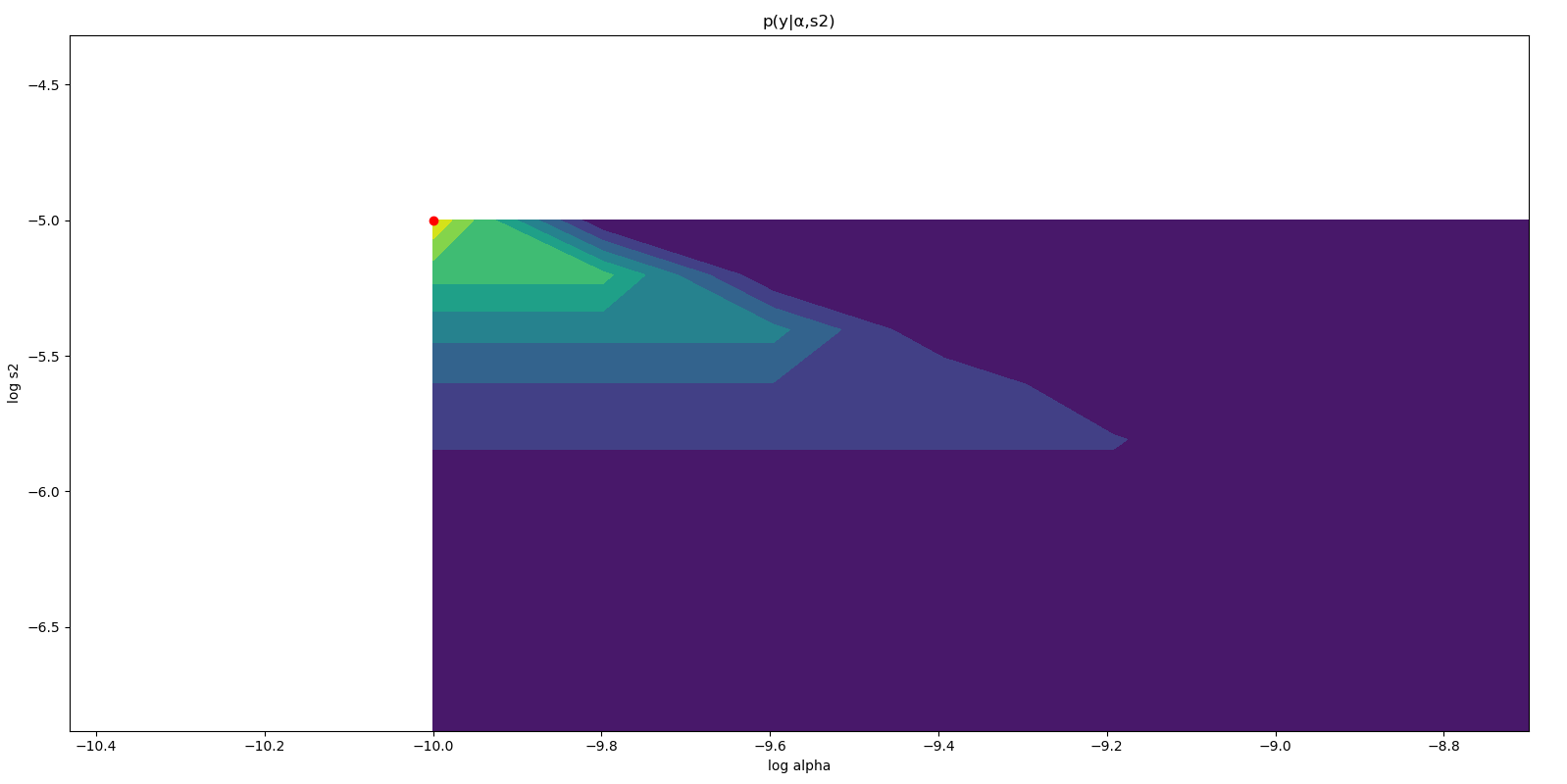
**2(a).using Type-II maximum likelihood (Lecture 04) to estimate “most probable” values for hyper-parameters.[4 marks]**

Since the full posterior that we want is p(w,α,s2|D) which cannot be compute analytically. The type 2 maximum likelihood can be used to approximate or infer the value of hyper-parameter α and s2. First the full posterior can be re-written as a product of weight posterior and probability of α and s2 given the dataset as shown in Eq…. .

p(w,α,s2|D) = p(w|α,s2,D)p(α,s2|y)

The weight posterior is normally distributed with mean of (xTx+s2 αI)-1xTy and covariance of s2(xTx+s2 αI)-1. However, the second probability again cannot be analytically computed. Using the concept of type 2 maximum likelihood¸ the best value of α and s2 can be approximate by maximizing the marginal likelihood p(y| α,s2) by assuming flat uninformative prior over log α and log s2. The term marginal likelihood is the probability of dataset given specific hyper-parameter α and s2 where this likelihood is normally distributed with zero mean and covariance of s2I + α-1xxT . To summarize, the procedure of type 2 maximum likelihood is to try a range of value of hyper-parameter α and s2 and the best hyper-parameters are the combination of value that maximize marginal likelihood probability.

After performing type 2 maximum likelihood¸ the max log likelihood obtained is -49.9791556301691 where best log alpha is-5.0 ( alpha is 0.006737946999085467) and best log s2 is -10.0 (s2 is 4.5399929762484854e-05 )

Since regularization parameter (λ) is the product of α and s2, the affect of regularization parameter is shown in Figure…. to distinguish between good and bad hyper-parameter selection. If the value of λ is too large, this will lead to underfitting and too small λ lead to over fit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| case | α | s2 | RMSE train | RMSE test |
| Best case | 0.0673 | 4.53999e-05 | 3.0116 | 3.0959 |
| Over fit | 0 | 0 | 3.0828 | 3.0977 |
| Under fit | 10 | 10 | 5.9431 | 4.81124 |

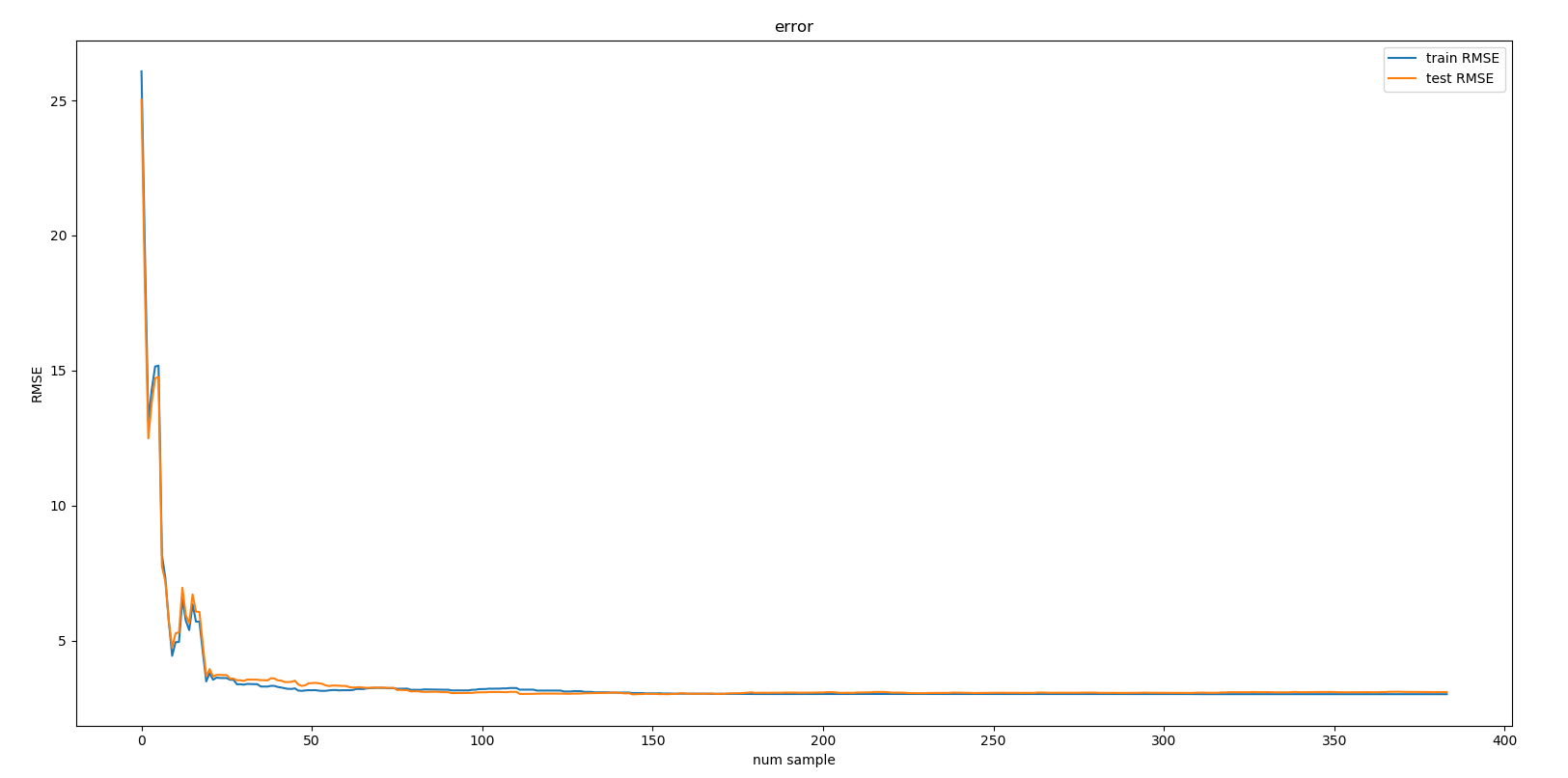
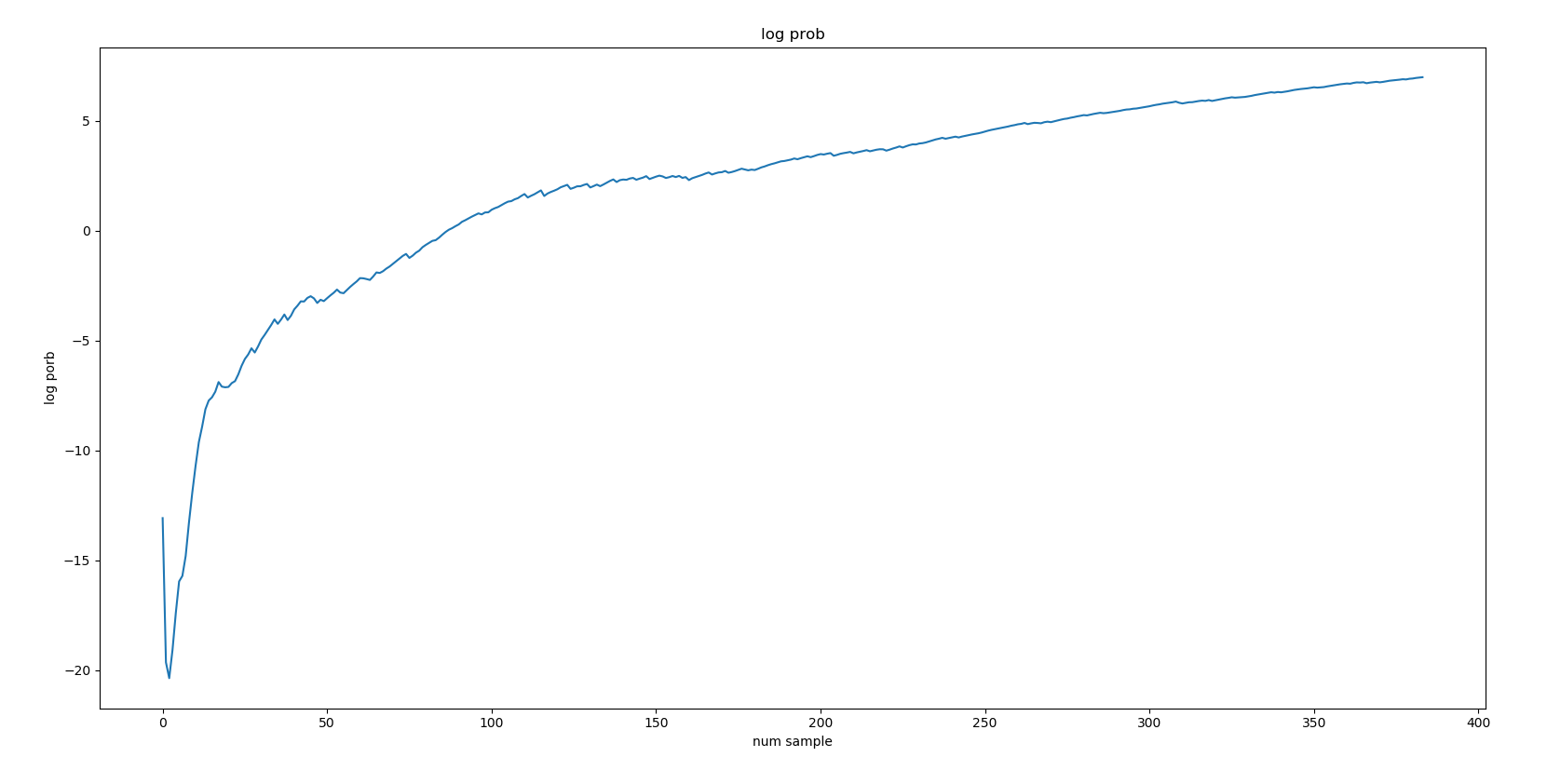
Demonstrate this table as bar chart as well \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

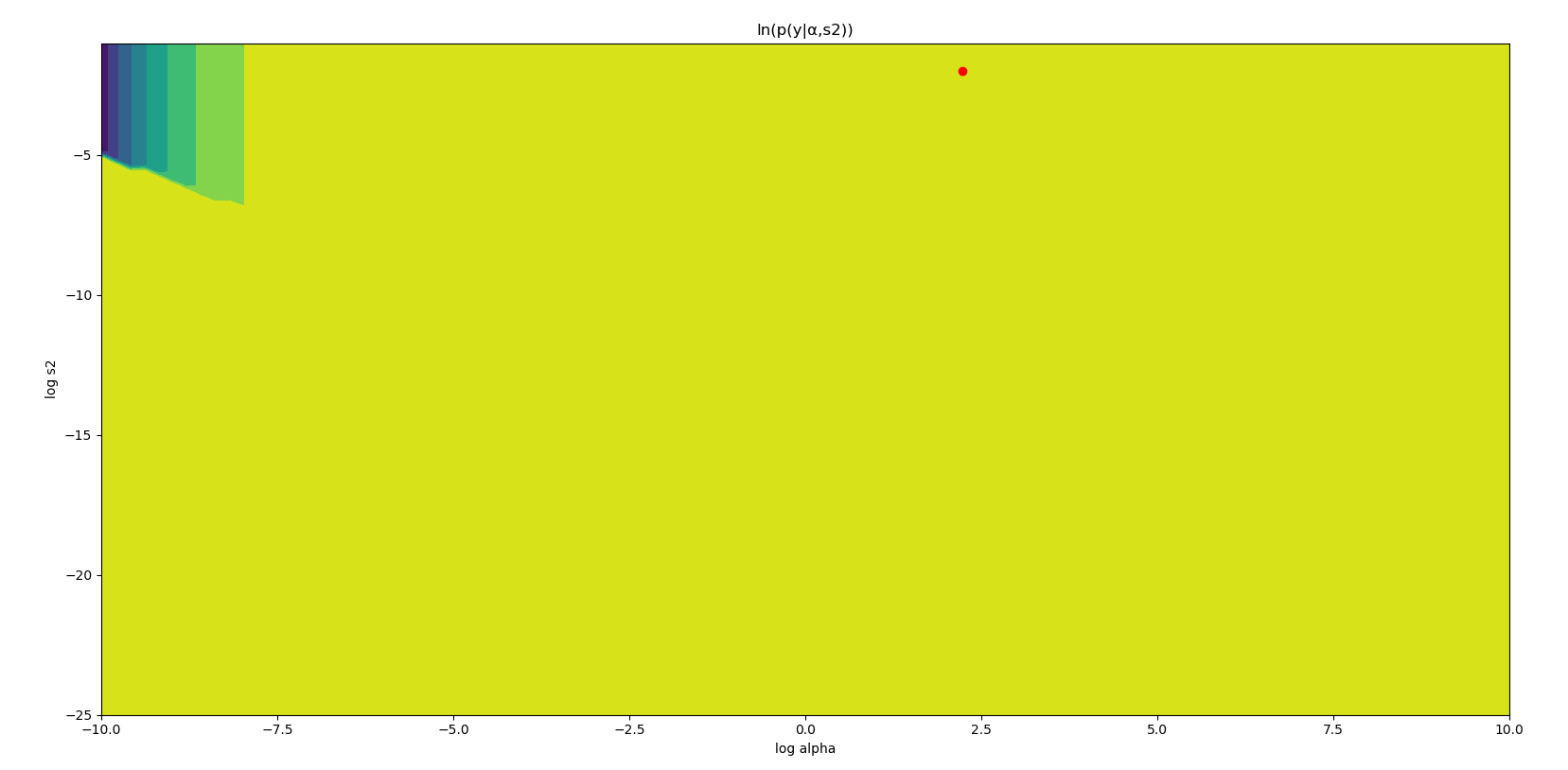
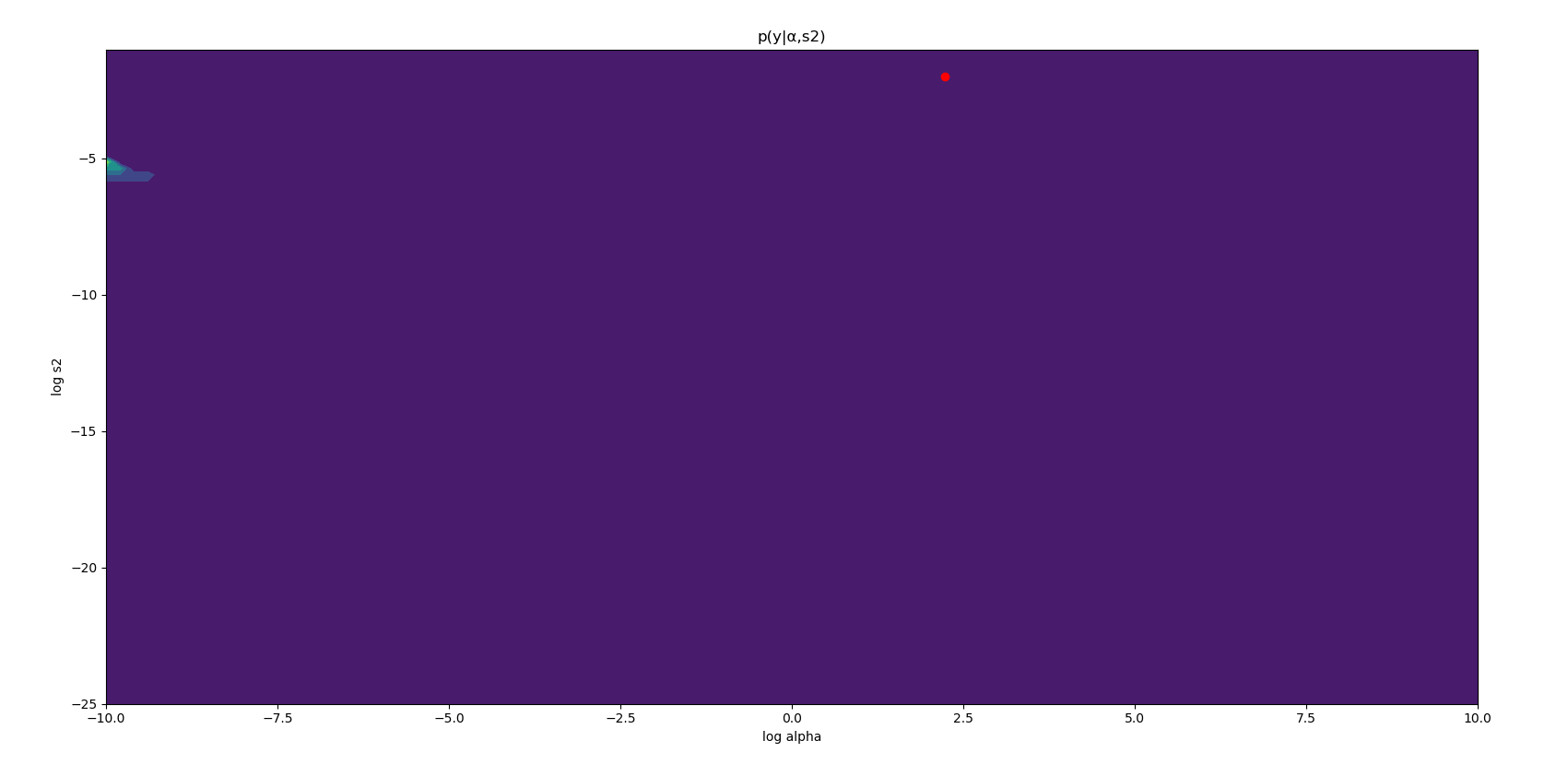
**2(b).using Variational Inference (Lecture 14) with simple ’Mean-Field Theory’ factorisation (Lecture 15) to estimate “most probable” values for the hyper-parameters. [4 marks]**

Variational method is one of deterministic approximation method. The aim is to find good proposal distribution Q(θ) that is most similar to unknown posterior distribution P(θ). To determine the similarity or difference between these two distributions, KL (Kullback-Leibler) divergence, also known as relative entropy, is used to measure the difference between two distribution based on the observation sample from both distributions. The best case is to have KL divergence to be zero. Latter, due to the constant term and since proposal distribution Q(θ) depends on set of hyper-parameters, minimizing KL divergence is equal to maximizing ELBO (Evidence Lower Bound) which is equal to finding best hyper-parameter set that maximize ELBO.

Mean field theory is applied to avoid complex joint distribution since sometime the problem contains more than one hyper-parameter to consider by breaking down one joint distribution to many small independent distributions where each distribution depends only on single hyper-parameter.

The optimal proposal can be obtained using Eq…. where this update rule is similar to Gibb sampling in a sense that the update of particular hyper-parameter depends on all hyper-parameters except itself. Finally, the mean of proposal distribution is the estimate for the hyper-parameter.

The final RMSE of train and test set are 3.011767383050821 and 3.0918273471103173 ;respectively, with the optimal value of alpha of 0.13462209254849022 and s2 of 9.260917427758185.

The variational inference in this project follow procedure described in Murphy…, where prior of w, λ, and α are N(w|0,( λ α)-1), Gam(λ|a0 λ b0 λ), and Gam(α |a0 α b0 α). The proposal distribution is factorized in the way as shown in Eq…. . Note that λ is 1/s2

q(w, α, λ)= q(w, λ) q(α)

Therefore,

q(w)=N(w|0,( λ α)-1)

q(λ )=Gam(λ|a0 λ ,b0 λ)

q(α )=Gam(α |a0 α ,b0 α).

Where

VN= aN λ / bN λ I + xTx

WN=VNxTy

aN λ = a0 λ + (N/2)

bN λ = b0 λ 0.5( ||y-xWN||2 + WNT(aN λ / bN λ) WN

aN α = a0 α + (K/2)

bN α = b0 α 0.5( (aN λ / bN λ) WNT WN + tr(VN))

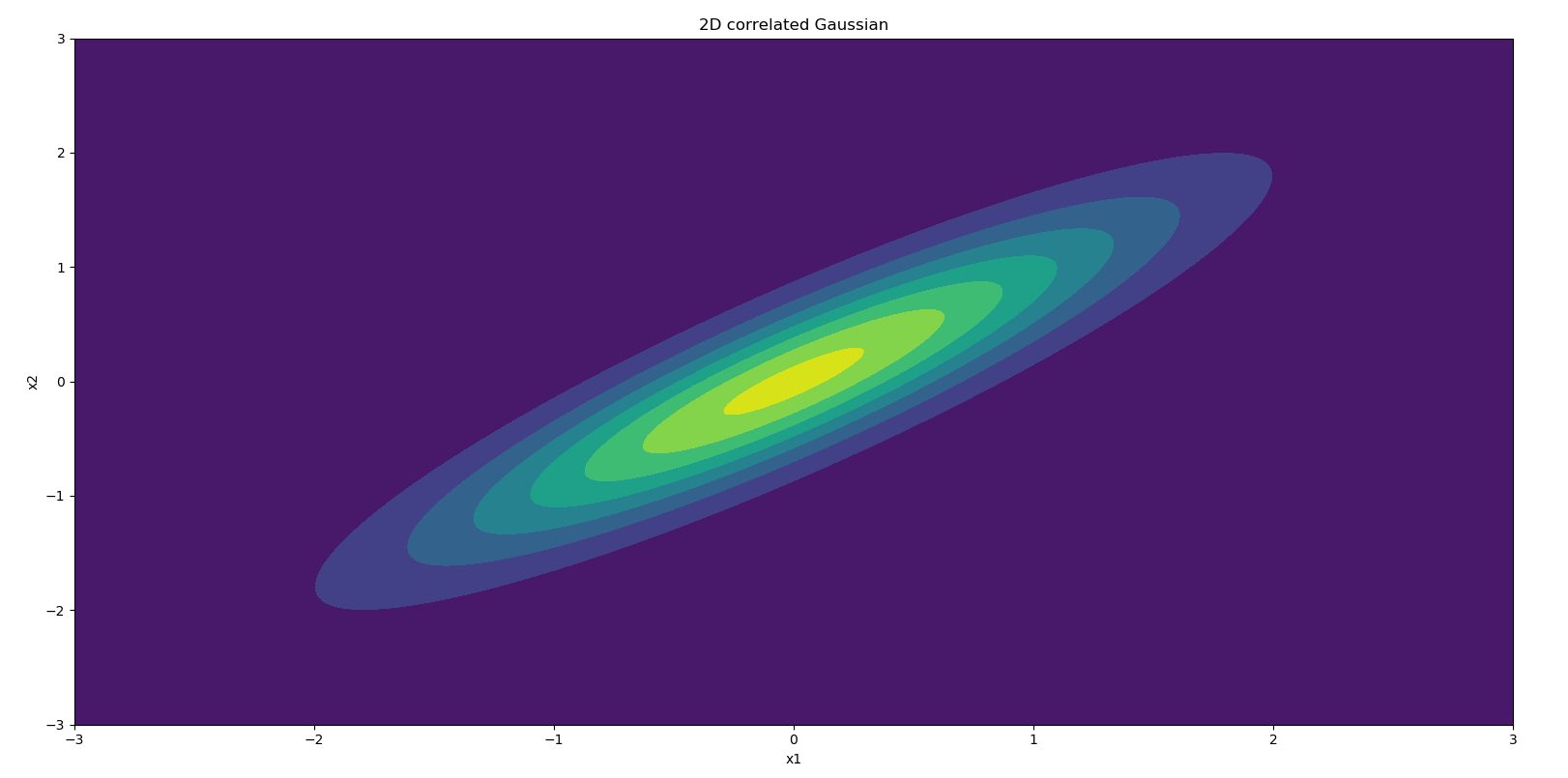
The variable a0 α , b0 α ,a0 λ ,and b0 λ are the hyper-parameter for and λ

**2(c).along with task 2(b), derive the corresponding variational approximation of the joint posterior distribution for the hyper-parameters. [4 marks]**

**3. Familiarise yourself with the use of the Hamiltonian Monte Carlo (HMC) algorithm (Lecture 07), initially verifying the HMC implementation on a simple Gaussian example. [5 marks]**

What is HMC \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

The example design in this problem is correlated bivariate normal with mean and variance of both variables are 0 and 1 and Correlation is 0.9 as shown in Figure ….



Energy function input the value of x1 and x2 and output energy value that is the negative log pdf as shown in Eq… . The energy function is validated by compare with the value from stats library.

P(x1,x2) = (2pi)-1 (1-ρ)-1 exp[ (-x12+2ρx1x2 - x22)/(2(1- ρ2)]

Energy = -ln(P(x1,x2))

= ln(2pi) + 0.5ln(1- ρ2) + [ (x12-2ρx1x2 +x22)/(2(1- ρ2)]

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coding | Stats library | Absolute difference |
| P(x1,x2) | 0.00032165685084216075 | 0.00032165685084216075 | 0.0 |

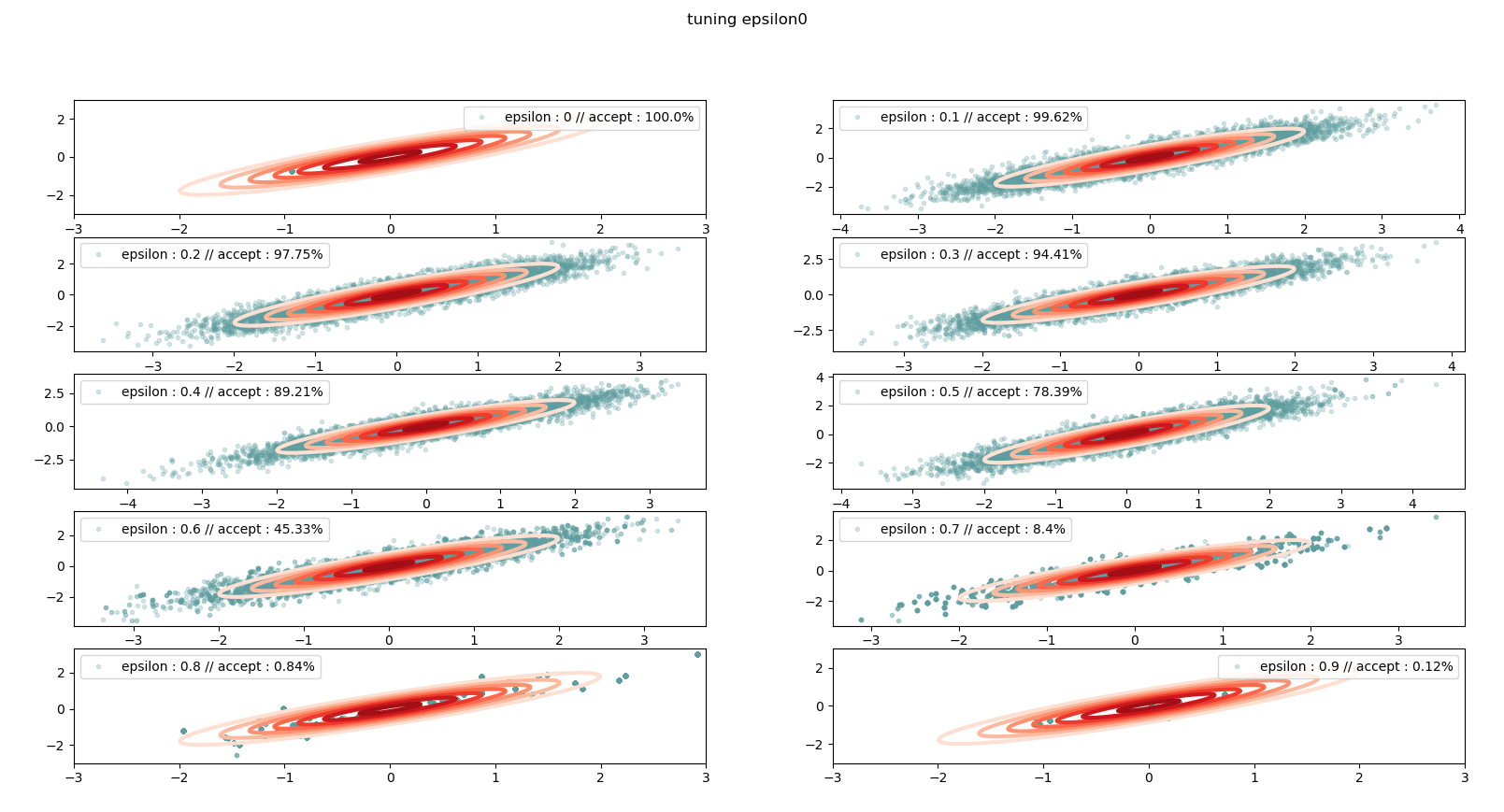
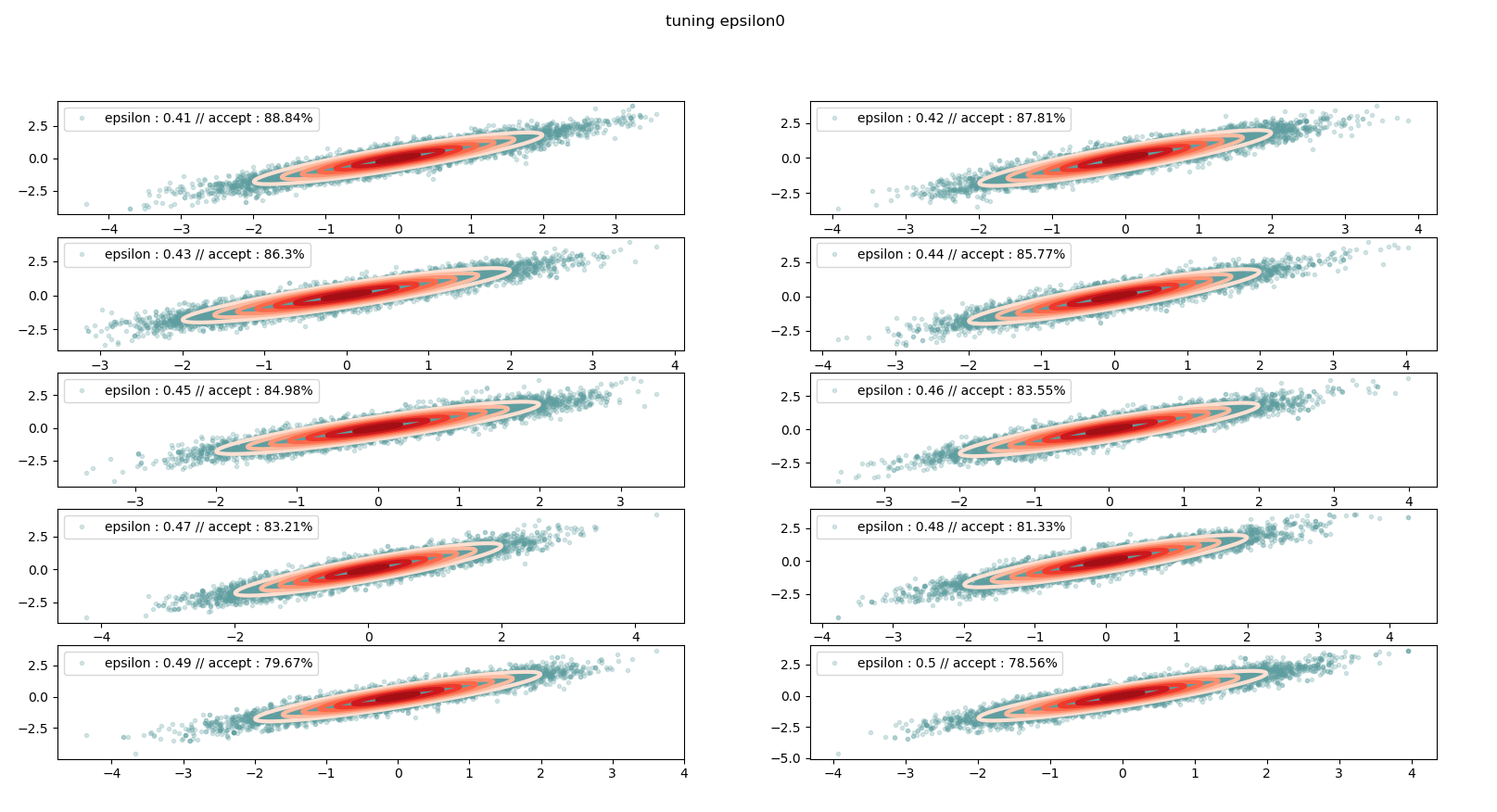
For gradient function, the function takes the input of x1 and x2 and return the corresponding gradient for each variable. Gradient can be determined by taking the derivative of energy function with respect to each variable as shown in Eq… .The gradient function is validated using check gradient function, the result is shown in table…

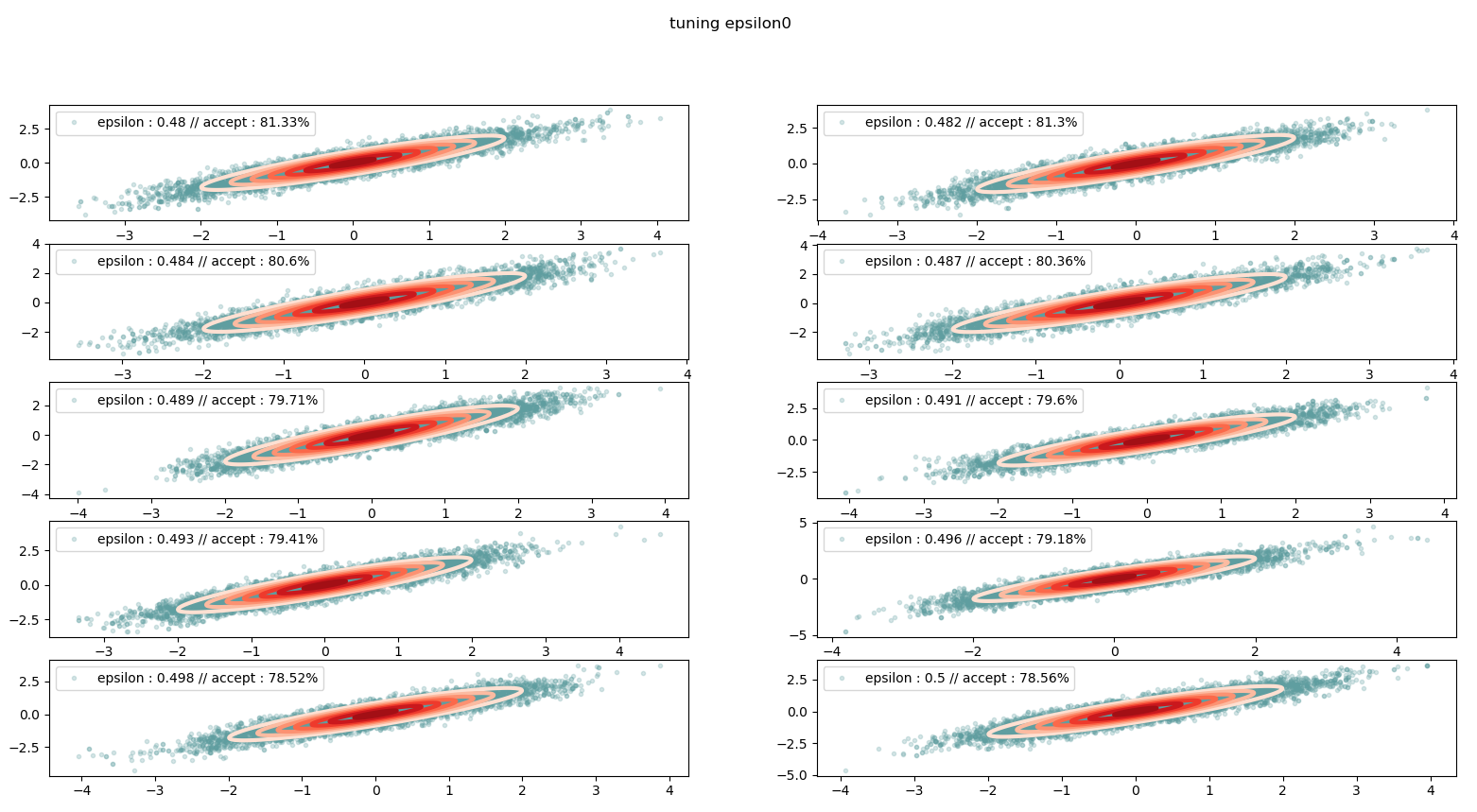
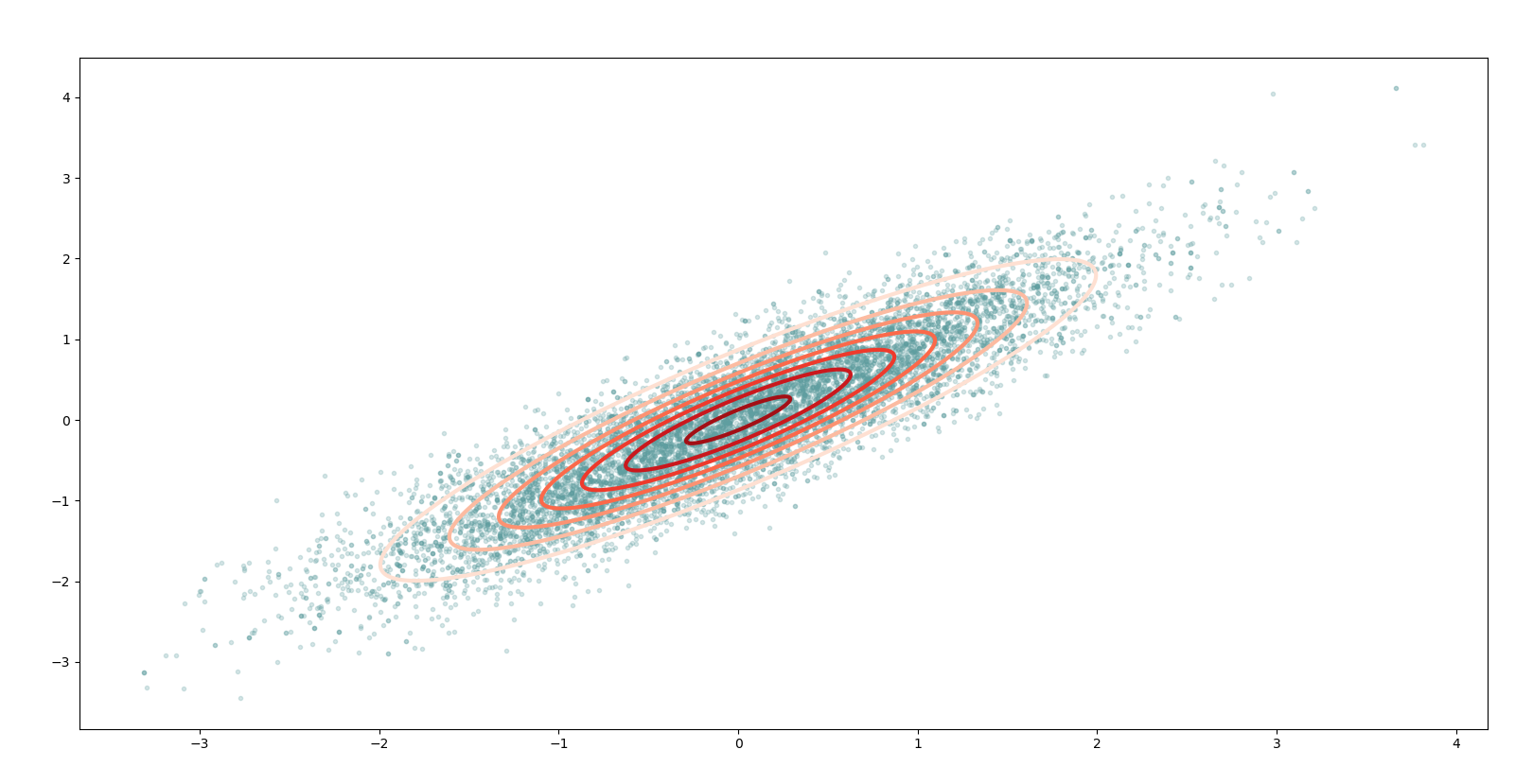
Grad x1 = = (2x1 - 2ρx2)/(2(1- ρ2)

Grad x2 = = (2x2 - 2ρx1)/(2(1- ρ2)

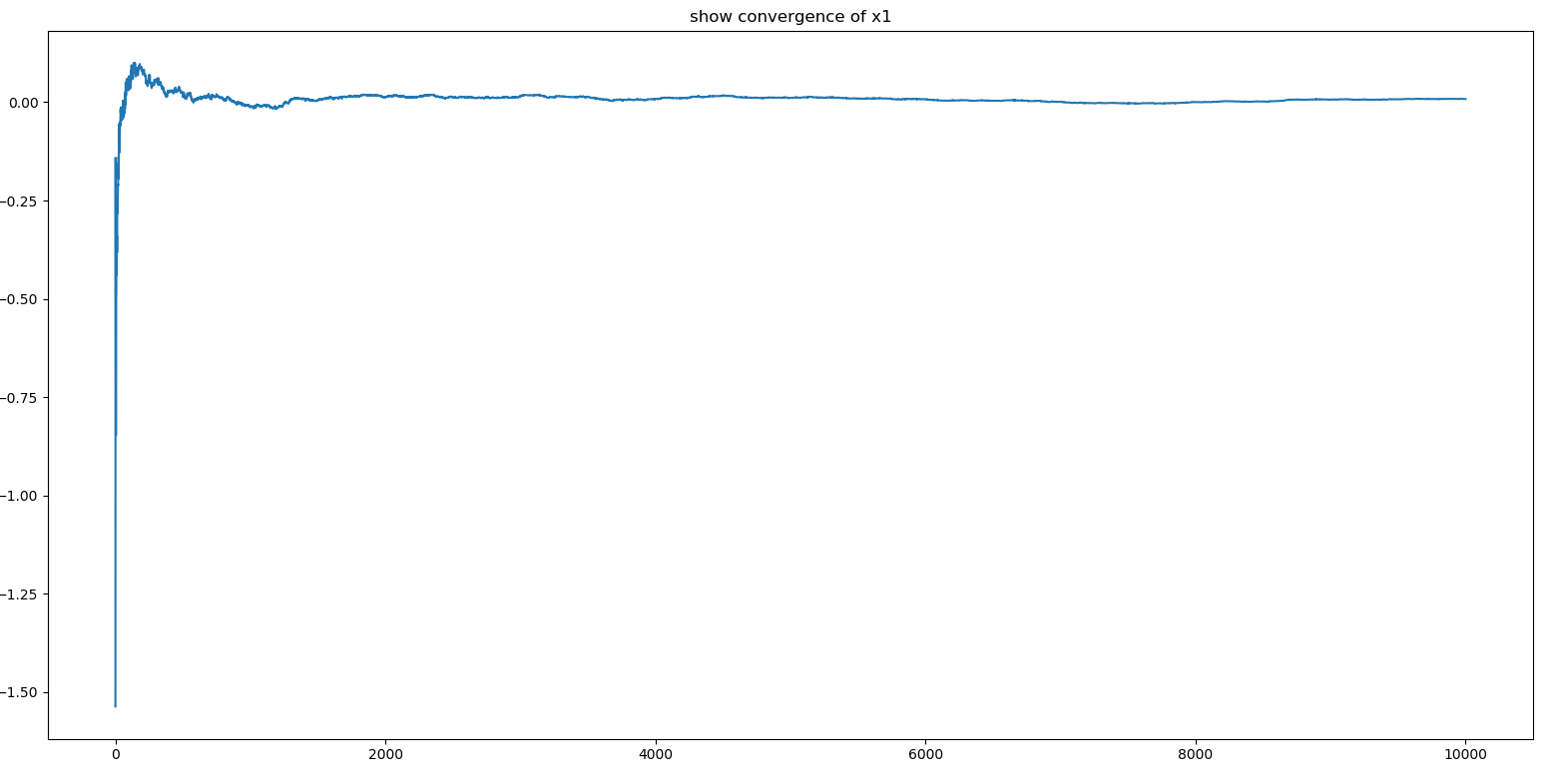
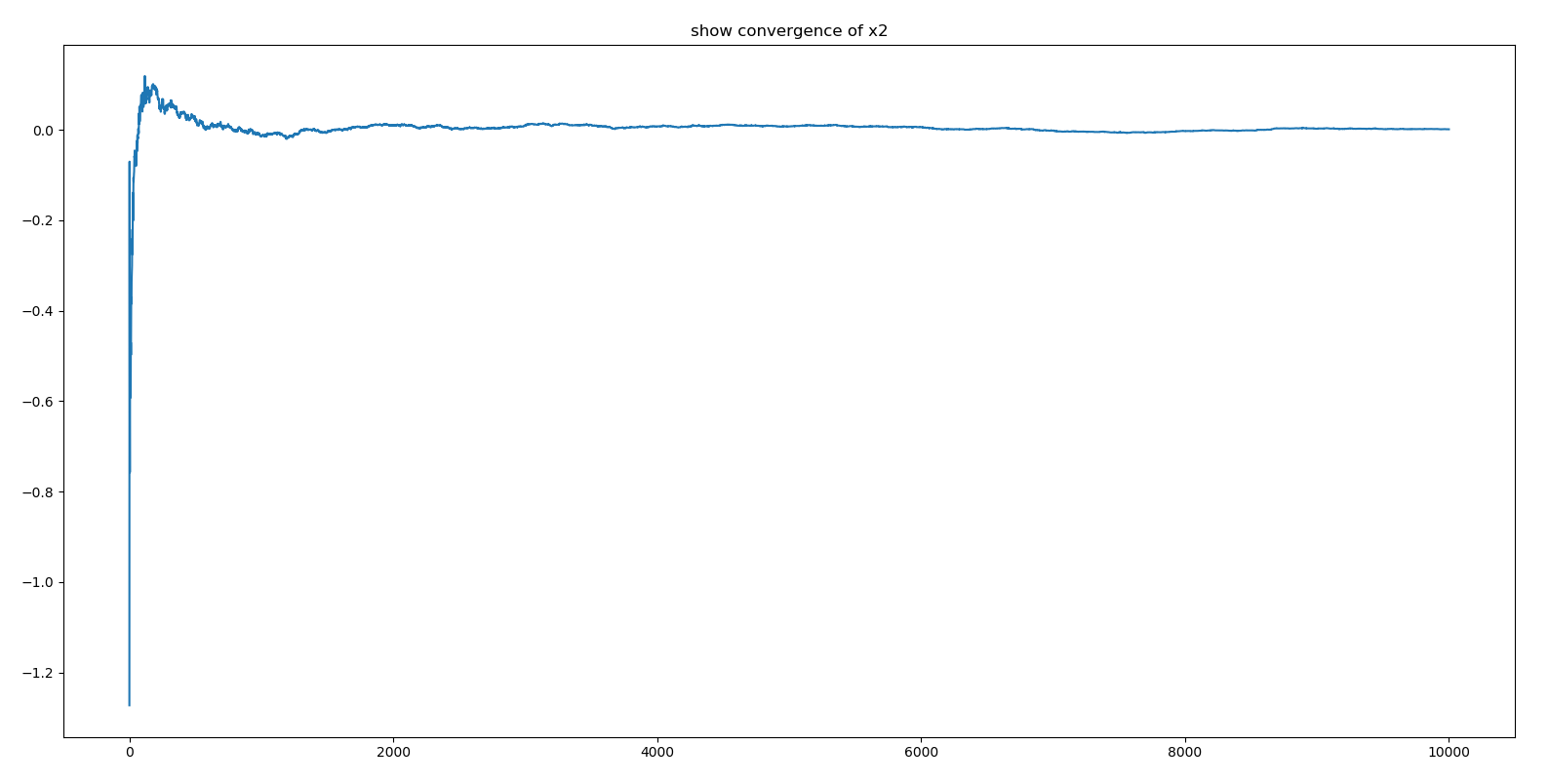
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Calculated | Numeric | Delta | Accuracy |
| X1 | 11.2954 | 11.2954 | -4.116973e-10 | 11 |
| X2 | -11.8463 | -11.8463 | -1.740030e-10 | 11 |

The value of hyper-parameter L used is 25 as it is sufficient for small example while R is chosen to be 10000. The hyper-parameter epsilon refers to \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. With this value of L, the hyper-parameter epsilon is tuned such the way that the percent acceptance is close to 80 percent as possible. This is done by keep increasing the value of epsilon until the percent acceptance is lower than 80 percent for the first time, then the value of epsilon when the percent acceptance is higher than 80% for the last time is selected. As shown in Fig… , Figure, and Figure…. , the procedure is done by narrowing down the range of epsilon from 0 to 1¸ 0.41 to 0.5, and 0.48 to 0.5. The value of epsilon chosen is 0.487 with percent acceptance of 80.36% because if the epsilon is increased to 0.489, the percent acceptance will fall to 79.41%

The HMC algorithm is validate as shown in Figure…. , the sample are located around the pattern of bivariate gaussian distribution with zero mean. Also, the estimated value also converge to true value which is zero as shown in Figure.. and Figure … . The estimations are done by averaging each sample with all the sample before this sample using cumulative sum.

**4.Apply HMC to sample weights and the hyper-parameters of the standard Bayesian regression model. [8 marks]**

Demonstrate the accuracy of energy function

Since P(w, α,s2 | data) is proportional to P(w|data,α,s2)\* P(data | α,s2), the weight posterior, the energy is negative log of posterior that is ln (P(w, α,s2 | data)) α ln(P(w|data,α,s2))+ln( P(data | α,s2)). Assume α=5 and s2=5

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coding | Stats library | Absolute difference |
| ln(P(w|data,α,s2)) | 8.405338709792451 | 8.405338709792455 | 3.55271367e-15 |
| ln(P(data | α,s2)) | -2429.4152919400444 | -2429.415291940045 | 4.5474735081e-13 |
| ln(P(w, α,s2 | data)) | -2421.009953230252 | -2421.0099532302524 | 4.54747350886e-13 |

Demonstrate the accuracy of gradient function

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Calculated | Numeric | Delta | Accuracy |
| S2 | -75127.7 | -75127.7 | -3.951091e-04 | 9 |
| α | -397.292 | -397.292 | -8.374301e-06 | 8 |
| W0 | -1136.19 | -1136.19 | 6.247463e-06 | 9 |
| W1 | 1249.59 | 1249.59 | -1.078964e-06 | 10 |
| W2 | -1280.86 | -1280.86 | 1.211076e-05 | 9 |
| W3 | 1831.72 | 1831.72 | 2.052312e-06 | 9 |
| W4 | -1939.33 | -1939.33 | -1.011282e-06 | 10 |
| W5 | -59.0401 | -59.0401 | 2.504216e-06 | 8 |
| W6 | -476.426 | -476.426 | 7.867835e-07 | 9 |
| W7 | -372.347 | -372.347 | -6.207000e-06 | 8 |
| W8 | -7265.4 | -7265.4 | -3.975402e-06 | 10 |

Derive the energy and gradient equation

The posterior that we are sample from is P(w, α,s2|y) which is equal to P(w|y,α,s2)P(α,s2|y) where P(α,s2|y) is proportional to P(y|α,s2).

P(w|y,α,s2) follow normal distribution with mean of (xTx+s2 αI)-1xTy and covariance of s2(xTx+s2 αI)-1 with pdf of Eq…and P(y|α,s2) follow normal distribution with zero mean and covariance of s2 α+ α-1xxT with pdf of Eq…

P(w|y,α,s2)=(2pi)-k/2det(cov1)-1/2exp(-1/2(w-mu1)Tcov1-1(w-mu1)

P(y|α,s2)=(2pi)-N/2det(cov2)-1/2exp(-1/2yTcov2-1y)

Therefore, energy equation can be derived by taking log on pdf and multiply by -1 which is shown in Eq….

Energy=0.5(ln(det cov1)+(w-mu1)Tcov1-1(w-mu1)+Kln(2pi)+ln(det cov2) +yTcov2y+Nln(2pi))

Where K is dimension of x and N number of samples, each term of equation is simplified to simple alphabet shown in Eq….

Energy=0.5(a+b+c+d+e+f)

The following section will demonstrate the derivative of each term respect to 3 variables (α ,s2,w) which represent the gradient. There are 11 gradient to be compute consist of 9 for weight and other 2 for α and s2. Term c and f are constant.

Firstly, consider the derivative respect to α.

=tr[(adj cov1 (-s2s2I) (xTx+s2 αI)-2)]/(det cov1)

=tr[-adj cov2 xxT/ α2]/(det cov2)

=yT(-cov2)-1(-xxT)cov2-1y/ α2

Where term b is separate into sub term which are b1 : (w-mu1)T, b2 : cov-1, and b3 : (w-mu1)

=yTx b11-1(s2IT)b11-1 where b11 = xTx+s2 αIT

=-cov1-1 s2(-b12-1)(s2I)(b12-1) cov1-1 where b12= xTx+s2 αI

=b13(s2I)b13xTy where b13 = (xTx+s2 αI)-1

Moving to derivative respect to s2.

=tr[adj cov2 ]

= yT(-cov2)-1I cov2-1y

=tr[(adj cov1 term /(det cov1) where term=[s2(-term2-1) α I (-term2-1)]+ (-term2-1)

and term2=xTx+s2 αI

=yTx b11-1(α IT)b11-1 where b11 = xTx+s2 αIT

= -cov1-1 [s2(-b12-1)(s2I)(b12-1) + b12-1]cov1-1 where b12= xTx+s2 αI

=b13(αI)b13xTy where b13 = (xTx+s2 αI)-1

Finally, to derivative respect to w.

=0

=0

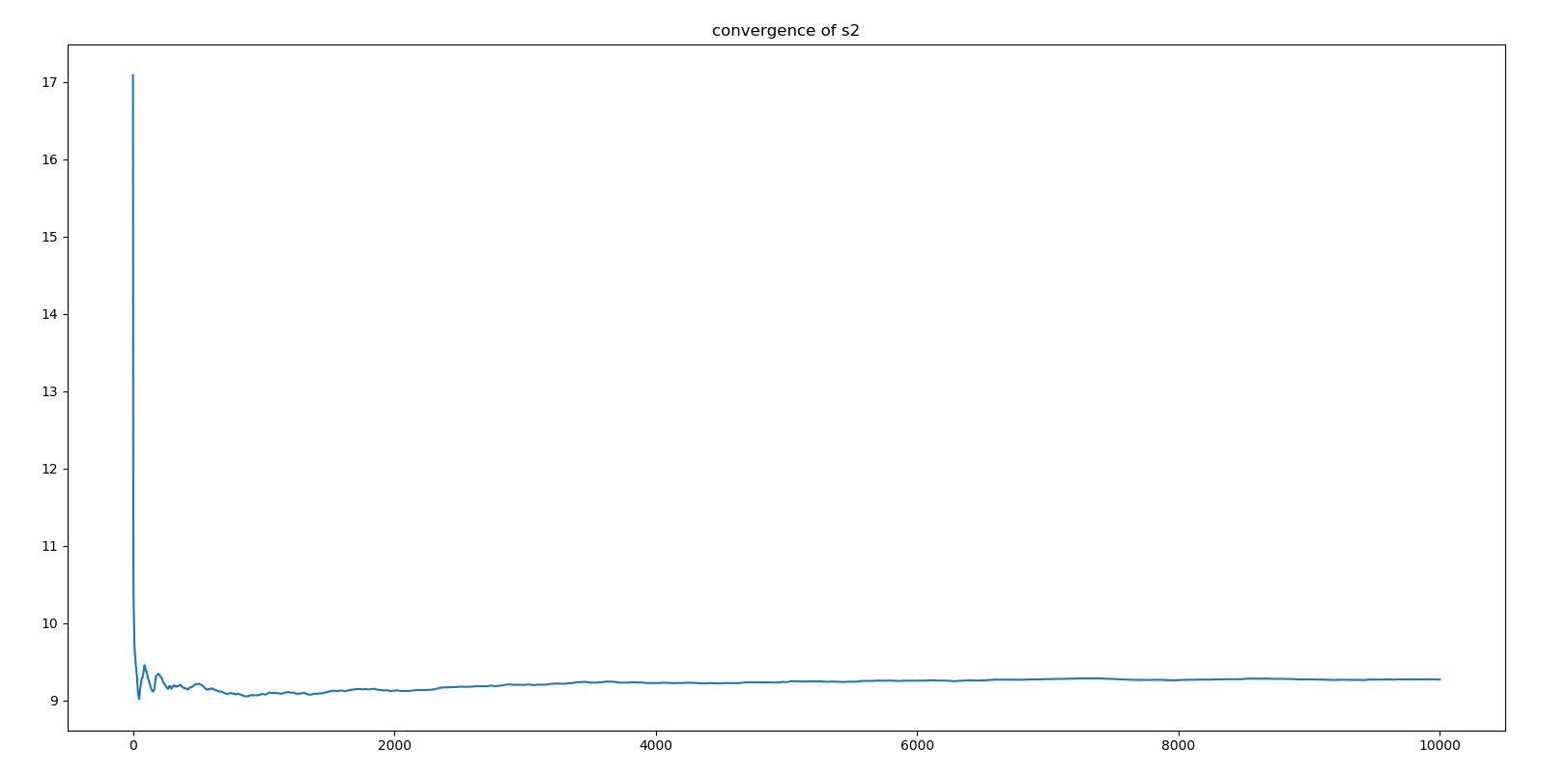
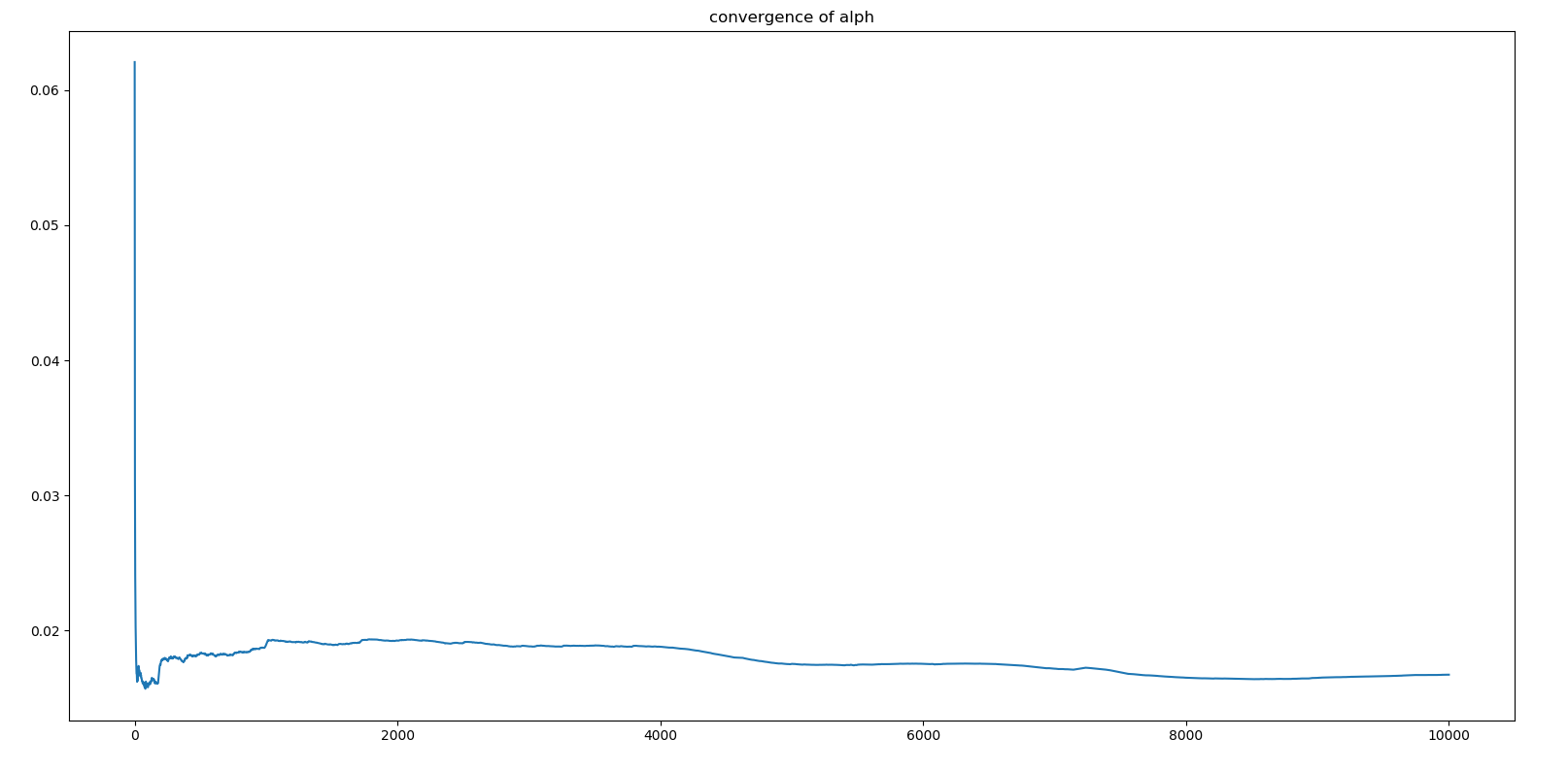
=0

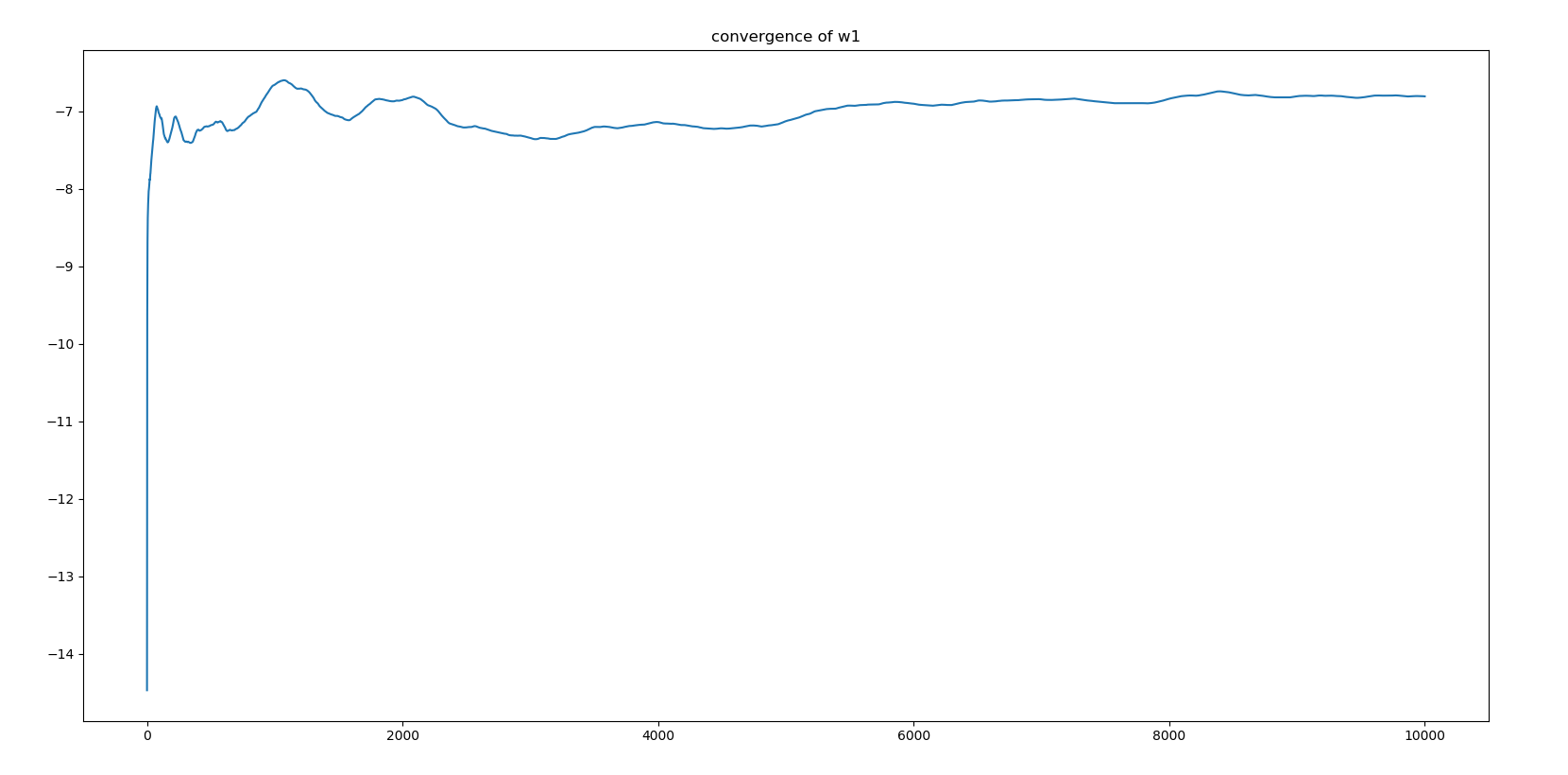
=1

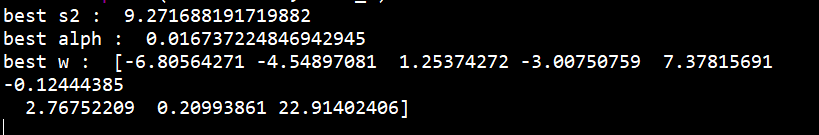
= 0

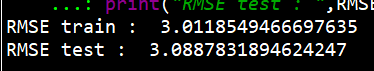
=1

In Figure…. shows the convergence of estimate of parameter







Gap btw train and test error is reduced. Reduced overfitting \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*