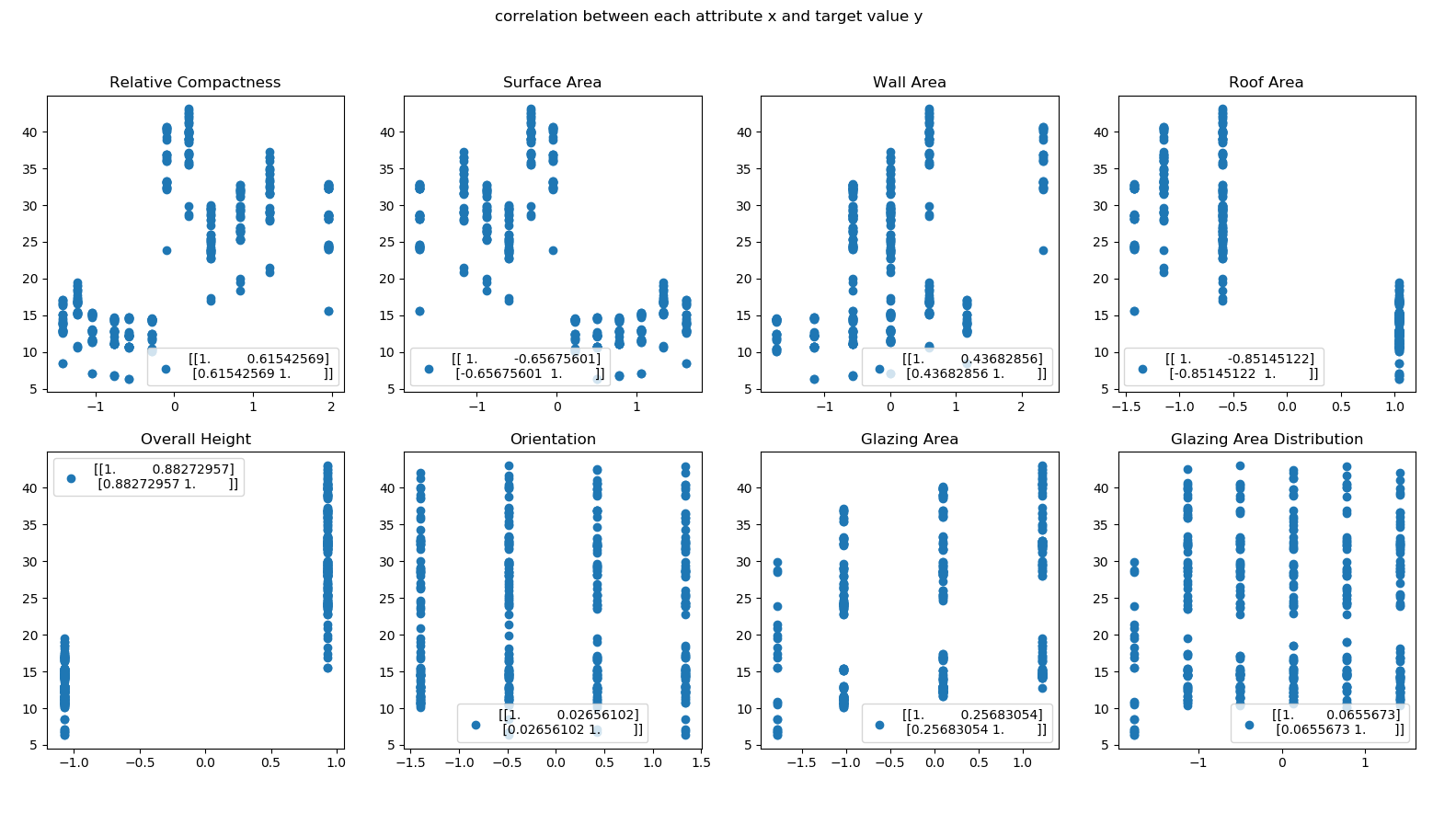
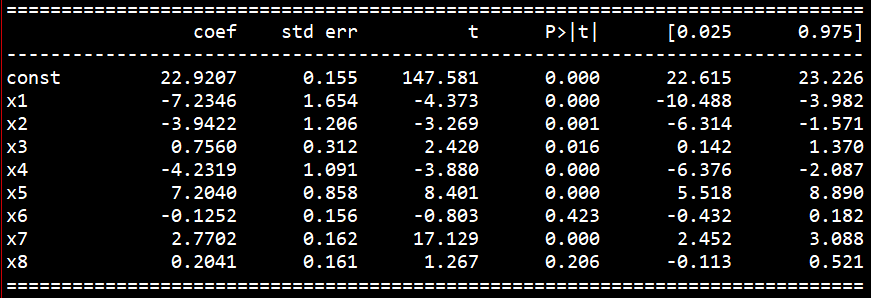
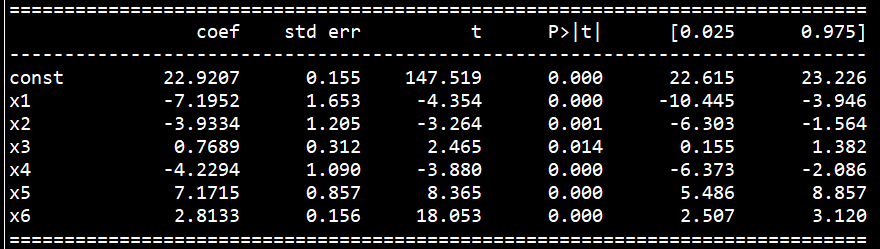
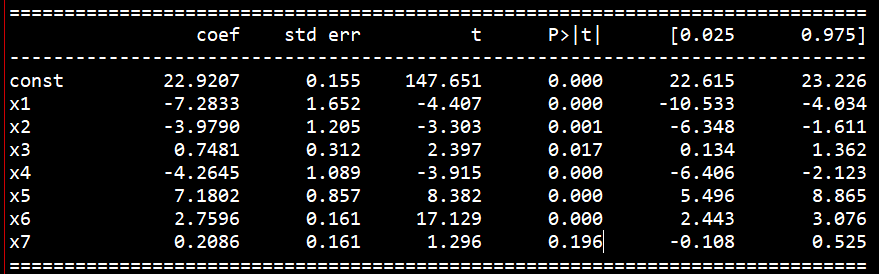
**1.Undertake an initial exploratory analysis of the training data and summarise. [5 marks]**

From correlation plot as shown in Figure…., 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 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

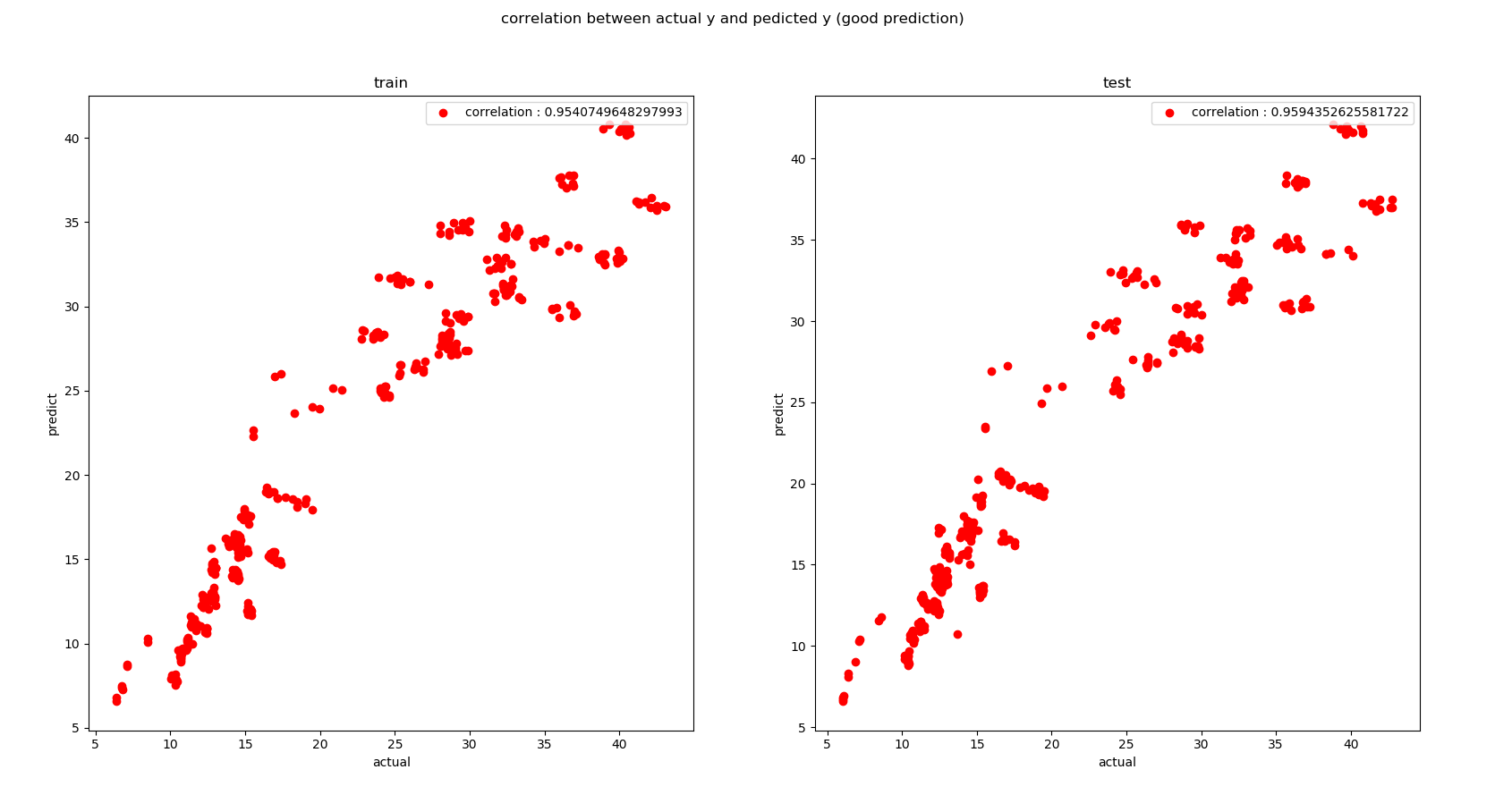
From backward elimination with significant level of 0.05, the first variable to be remove was orientation that have P-value of 0.423 shown in Figure …. which is higher than significant level. The second variable was glazing area orientation with P-value of 0.196 shown in Figure… . After these 2 variables have been remove¸ no variable has p-value more than significant level shown in Figure… . This indicate that orientation and glazing area is irrelevant. The train and test set error after perform least squared linear regression are shown in Table…



|  |  |  |
| --- | --- | --- |
| Case | RMSE train | RMSE test |
| Consider all features | 3.0115517876503612 | 3.0958865845448686 |
| After remove glazing area and orientation | 3.020837151333591 | 3.110716297089577 |

Since the data have more than 2 dimensions, 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) , where α is inverse variance of w and s2 is noise variance, 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|y) = p(w|α,s2,y)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 inverse weight variance α and noise variance 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.

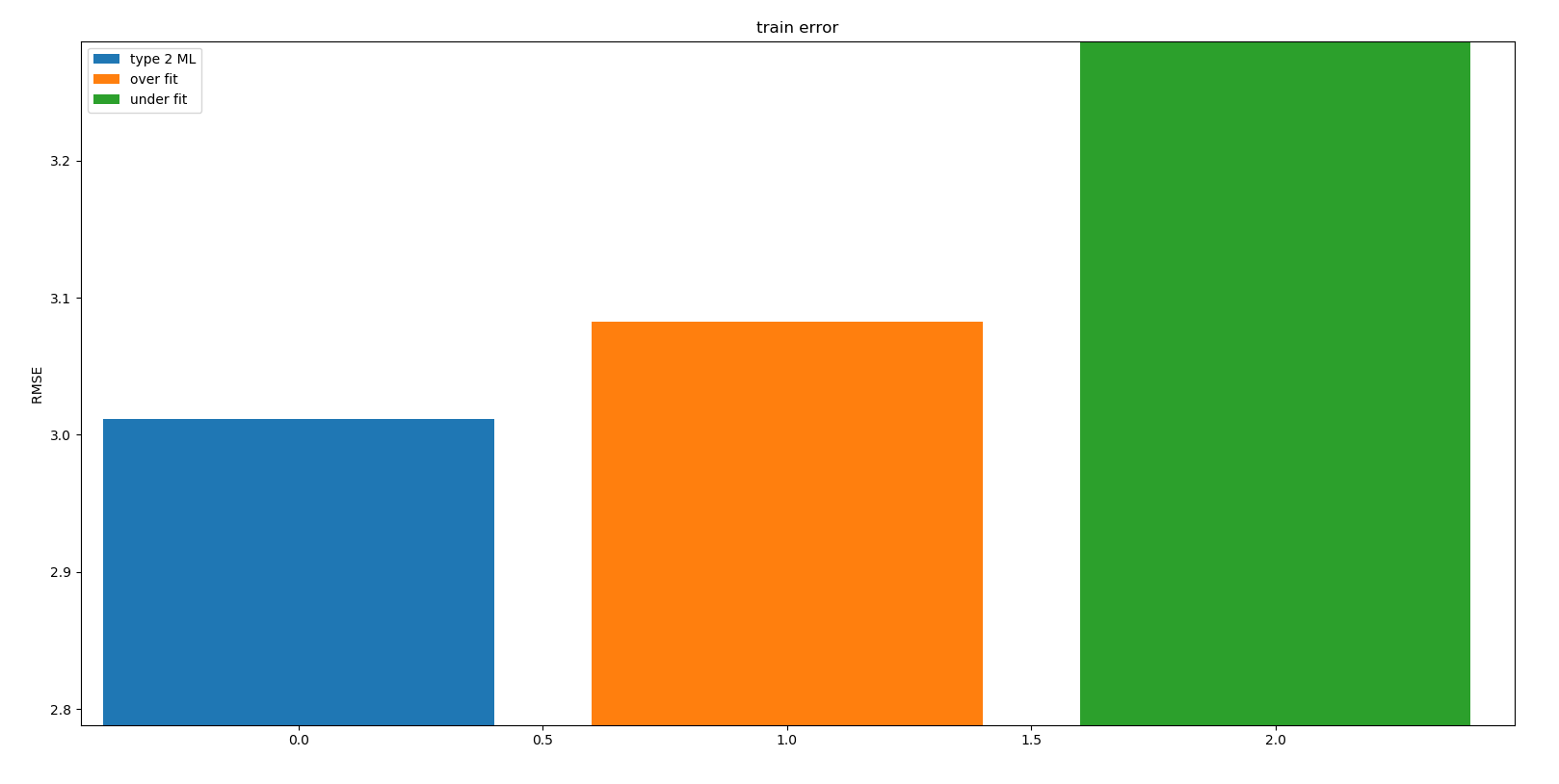
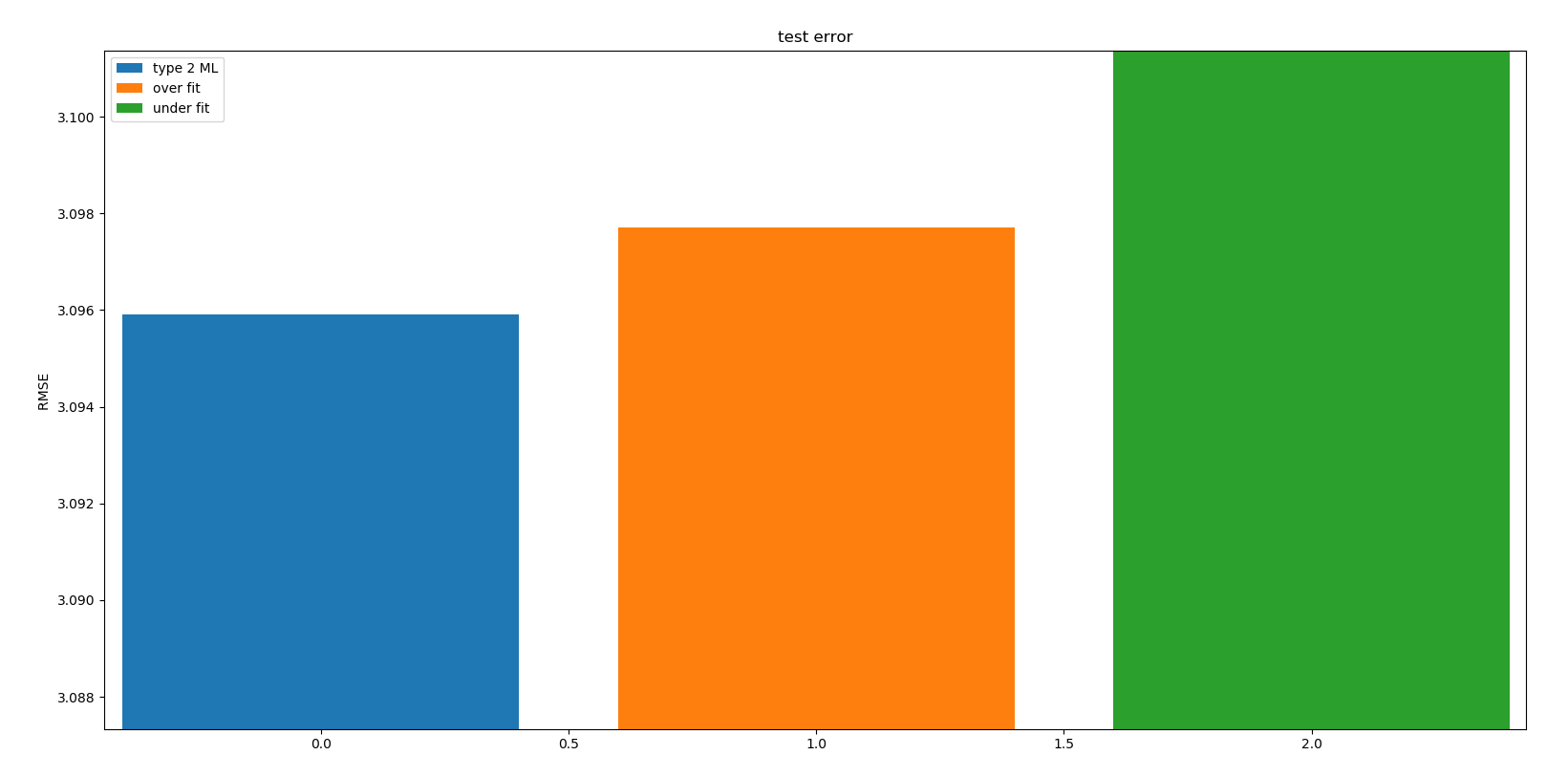
After performing type 2 maximum likelihood¸ the max log likelihood obtained is -49.9791556301691 where best log alpha is-5.0 (alpha is 0.006737946999085467) and best log s2 is -10.0 (s2 is 4.5399929762484854e-05 ) . All parameters are shown in Table…

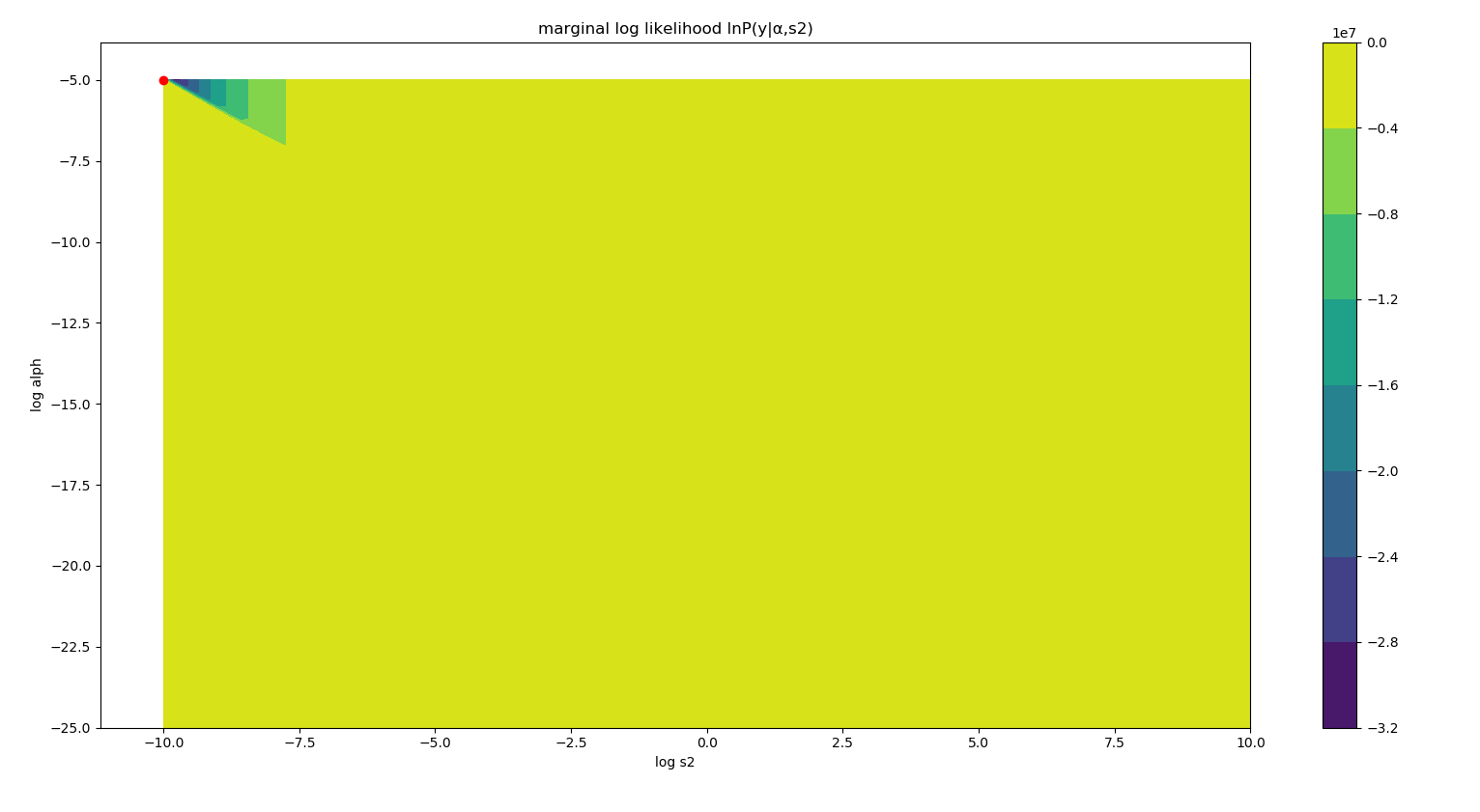
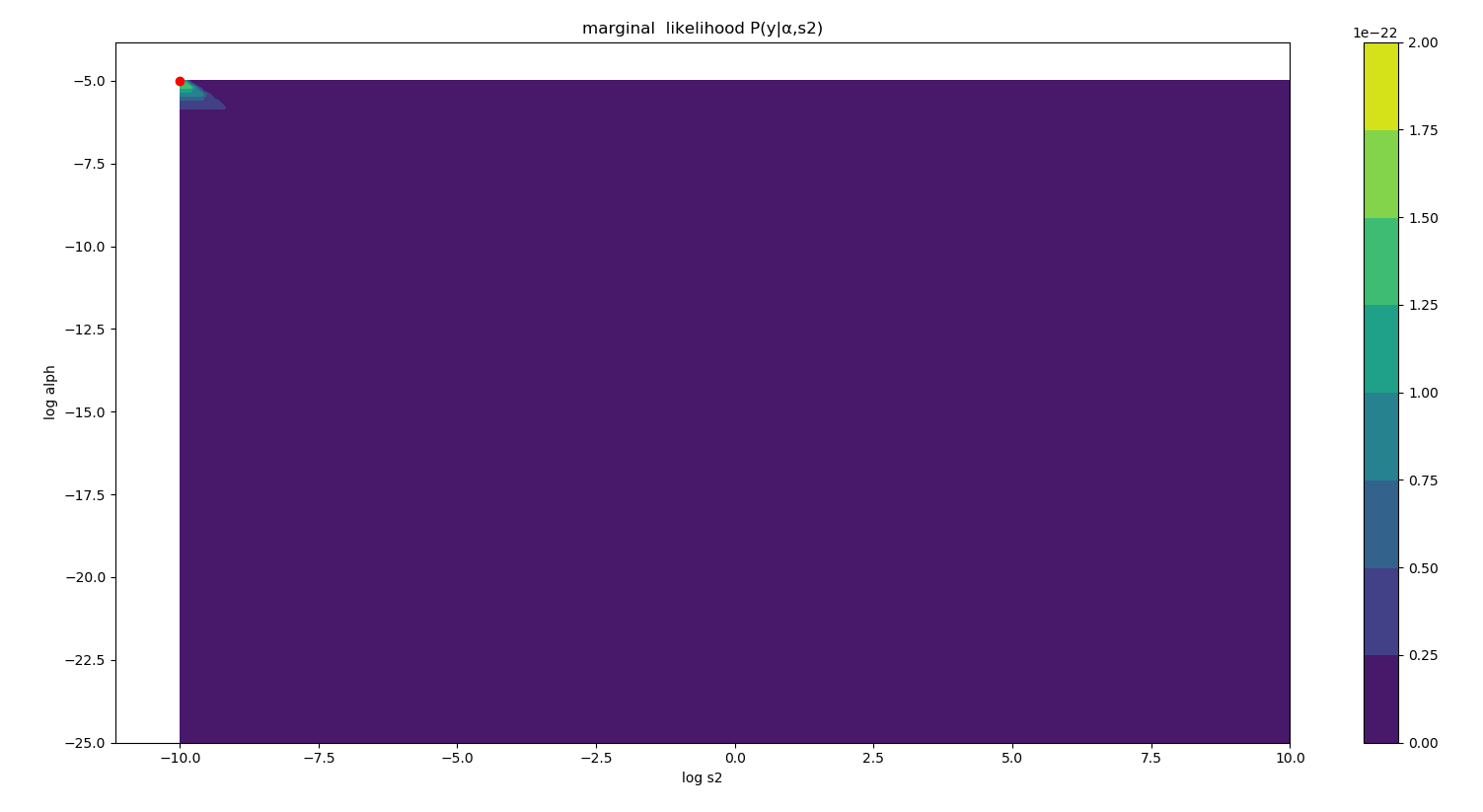
|  |  |
| --- | --- |
| Parameter | Value |
| Best alpha (inverse weight variance) | 0.006737946999085467 |
| Best s2 (inverse noise variance) | 4.5399929762484854e-05 |
| W1 | -7.23462662 |
| W2 | -3.94213963 |
| W3 | 0.75594473 |
| W4 | -4.23192596 |
| W5 | 7.2039515 |
| W6 | -0.12516927 |
| W7 | 2.77021895 |
| W8 | 0.20406264 |
| W9 | 22.92070311 |

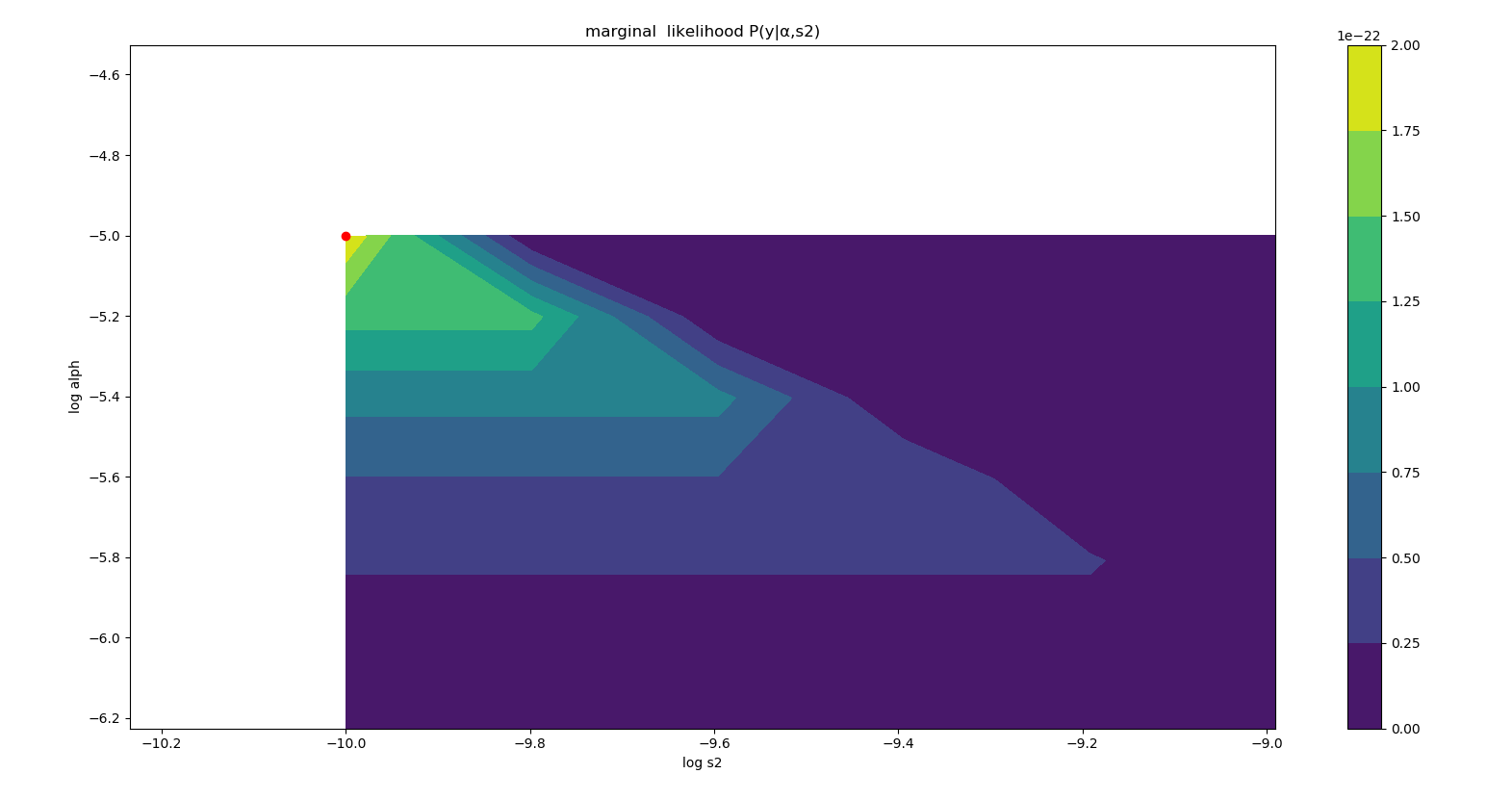
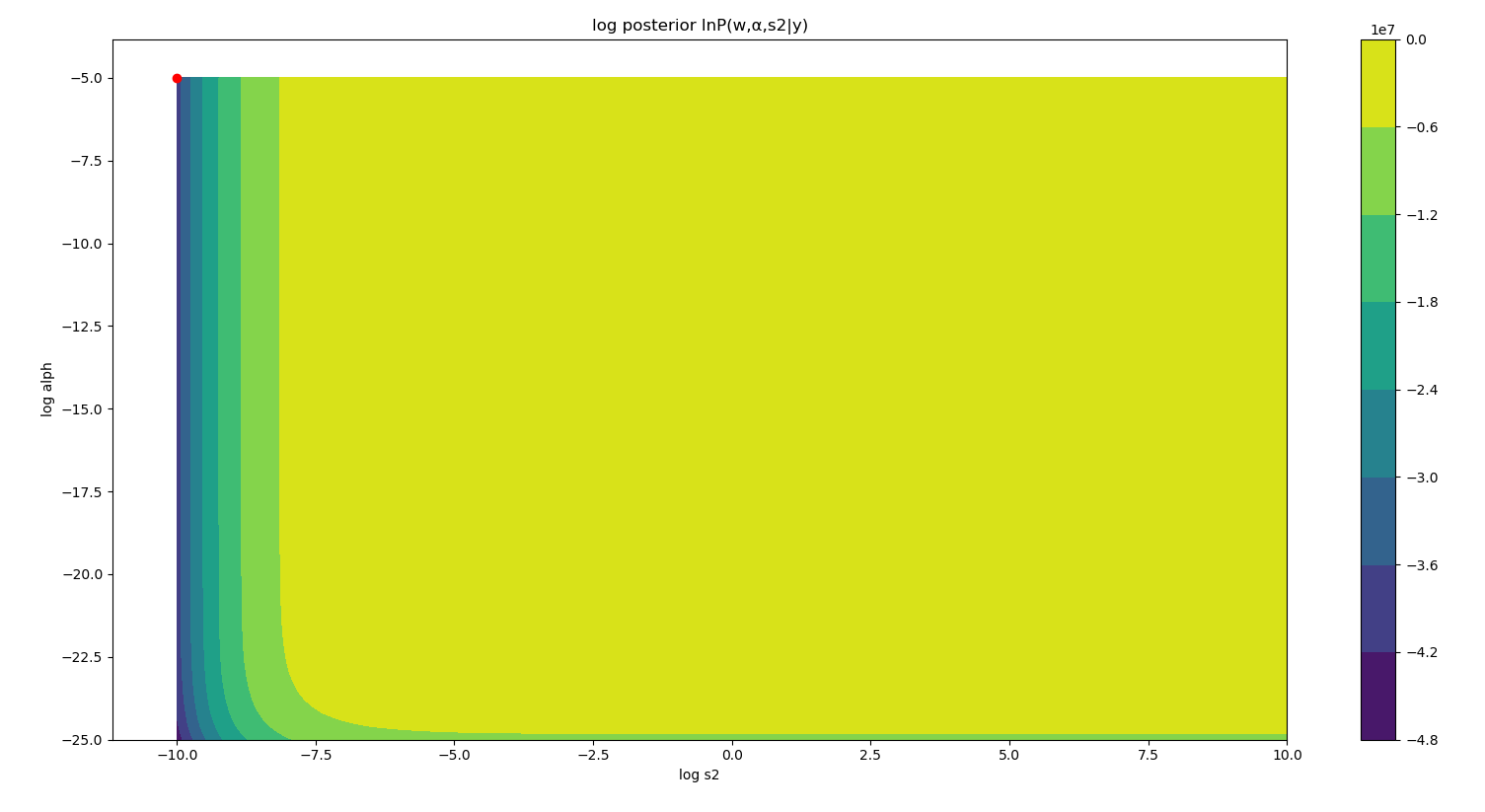
Since regularization parameter (λ) is the product of α and s2, the effect of regularization parameter is shown in Figure…. to distinguish between good and bad hyper-parameter selection. If the value of λ is too large, this will lead to underfitting and too small λ lead to over fit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| case | α | s2 | RMSE train | RMSE test |
| Type 2 ML | 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 |

It is cleared that, by optimally chose the value of alpha and s2 can decrease the RMSE of train and test set as shown in Figure.. and Figure…

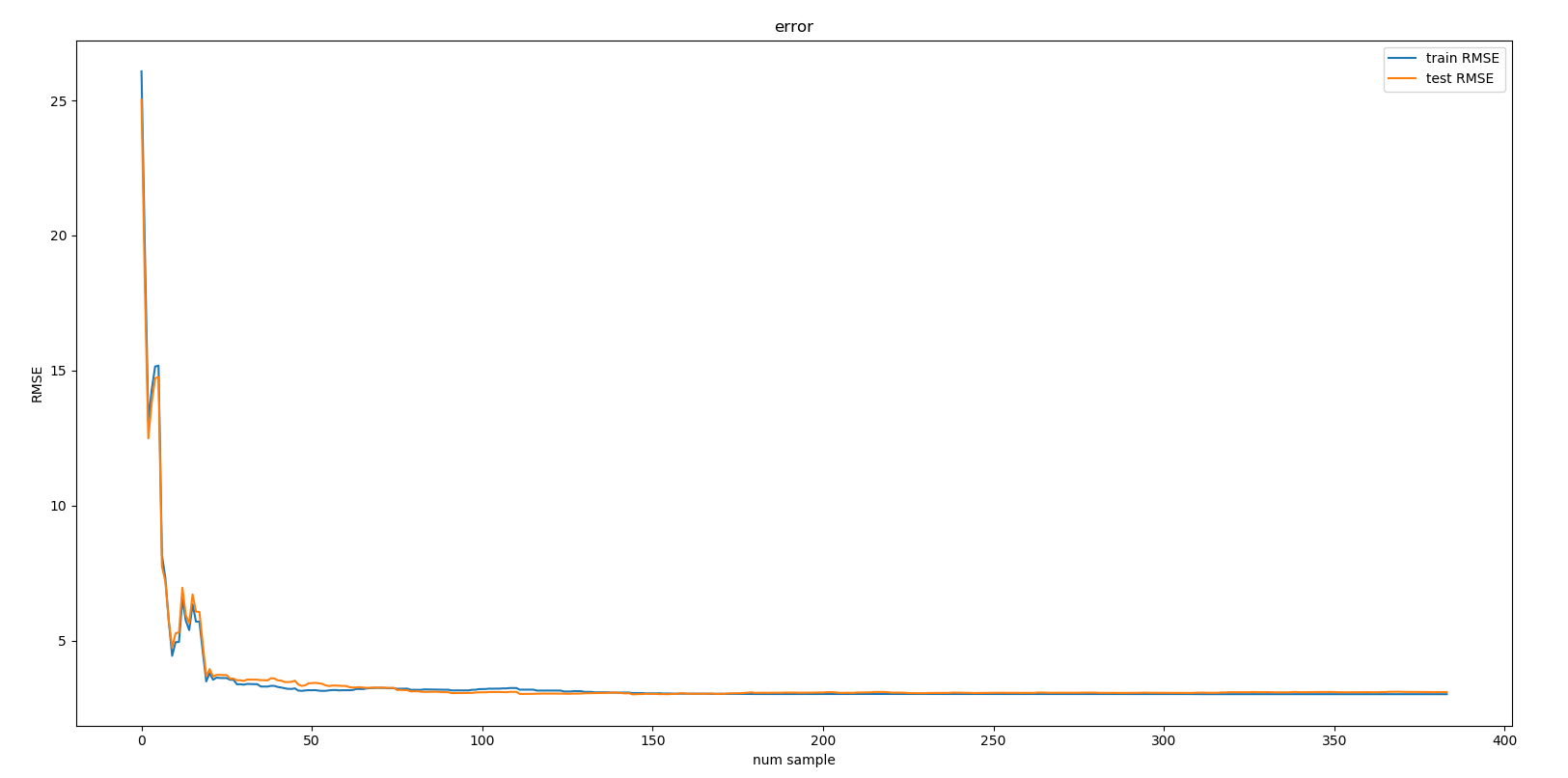
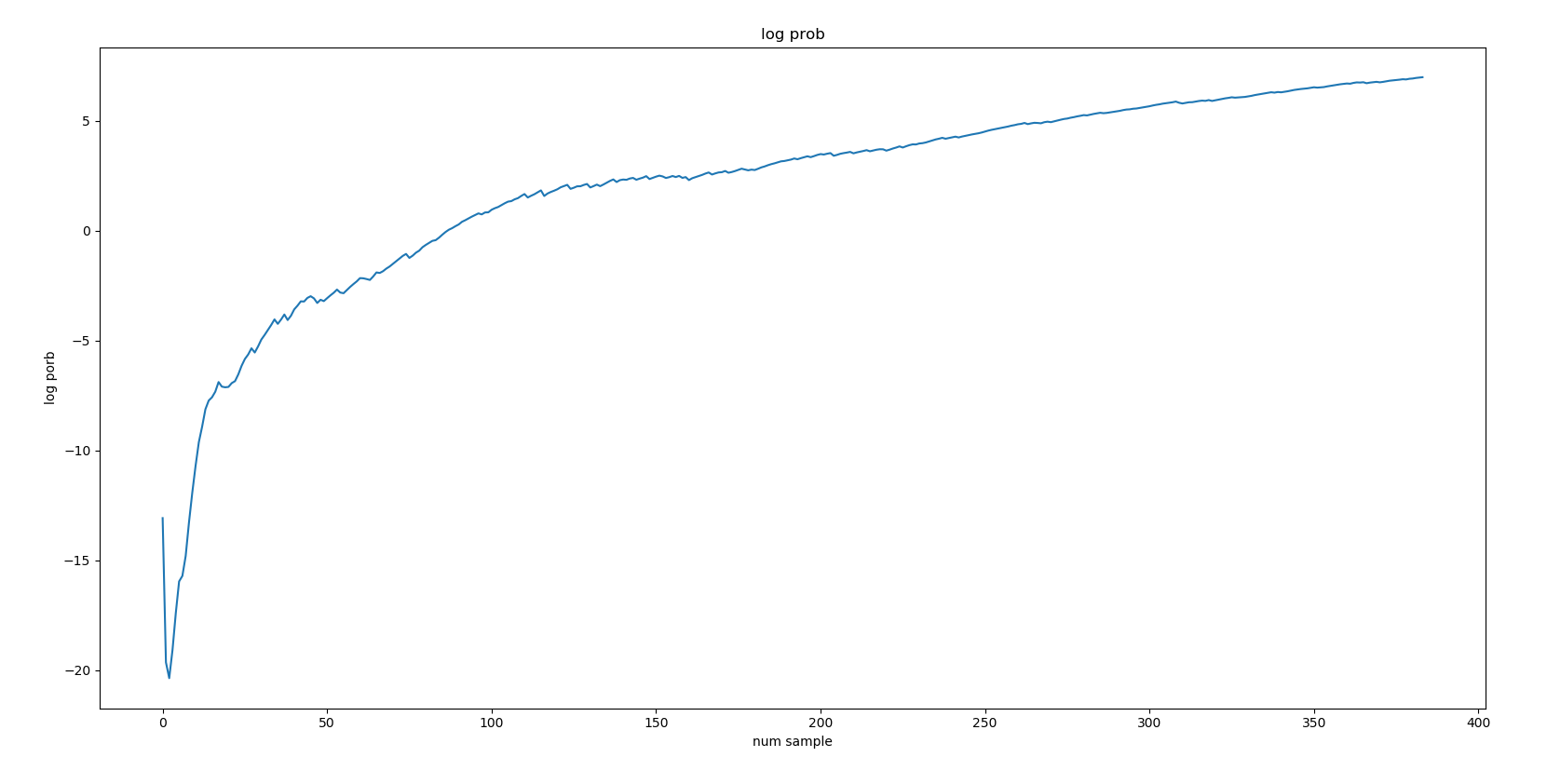
The obtained value of log alpha and log s2 is marked on the contour plot with red color dot. The model performed well to achieve maximum log marginal likelihood as shown in Figure.. and Figure… . However, by visualize the posterior (P(w, α,s2|y) contour plot, the region that the value located is not the best region. The model can do better if the obtained value of log alpha and log s2 fall into yellow region.

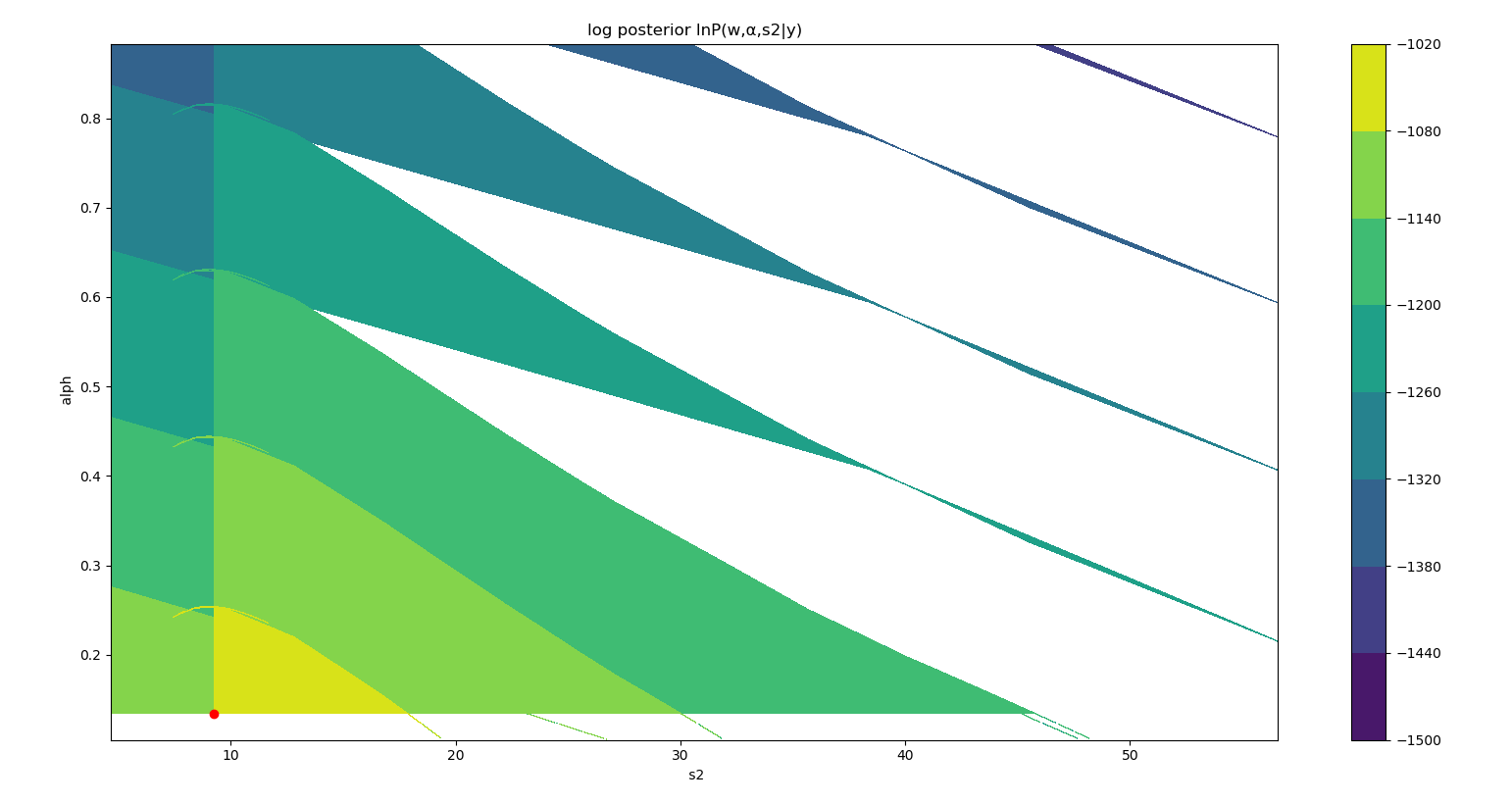
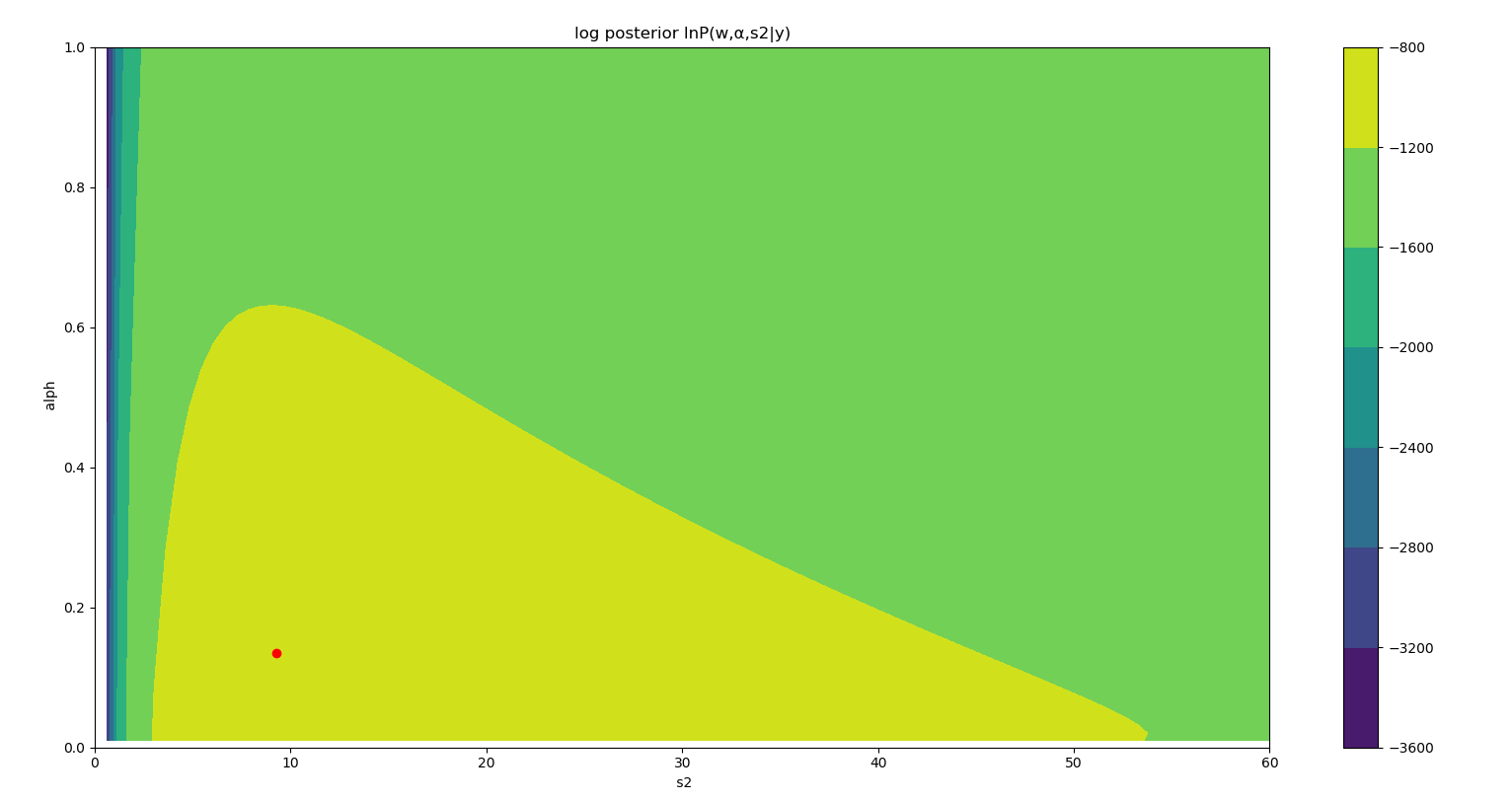
**2(b).using Variational Inference (Lecture 14) with simple ’Mean-Field Theory’ factorisation (Lecture 15) to estimate “most probable” values for the hyper-parameters. [4 marks]**

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. Later, 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 model performed well in terms of error as well as continuously increase the log probability. In Figure…. ,both training and test error decrease rapidly for first 20 samples, and remain stable after 50 sample onwards. Also in Figure.. and Figure… the obtained value of alpha and s2 located in good position on posterior contour plot where the color is yellow which indicate highest value. The differnec between 2 curves is only that the Figure …. Plot in better since some value of s2 and alpha did not get updated in Variational Inference process as there are some white space on the region. Figure… used numpy linspace in order to have smoother contour plot with similar range of values. All values obtained are shown in Table…. and errors are shown in Table…

|  |  |
| --- | --- |
| Parameter | Value |
| Best alpha (inverse weight variance) | 0.13462209254849022 |
| Best s2 (inverse noise variance) | 9.260917427758185 |
| W1 | -6.86459553 |
| W2 | -3.68826174 |
| W3 | 0.81494196 |
| W4 | -4.01007459 |
| W5 | 7.31383394 |
| W6 | -0.12623419 |
| W7 | 2.77138242 |
| W8 | 0.20247917 |
| W9 | 22.91265232 |

|  |  |  |
| --- | --- | --- |
| Case | RMSE train | RMSE test |
| Variational inference | 3.011767383050821 | 3.0918273471103173 |

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) where λ is inverse noise variance. The proposal distribution is factorized in the way as shown in Eq…. .

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))

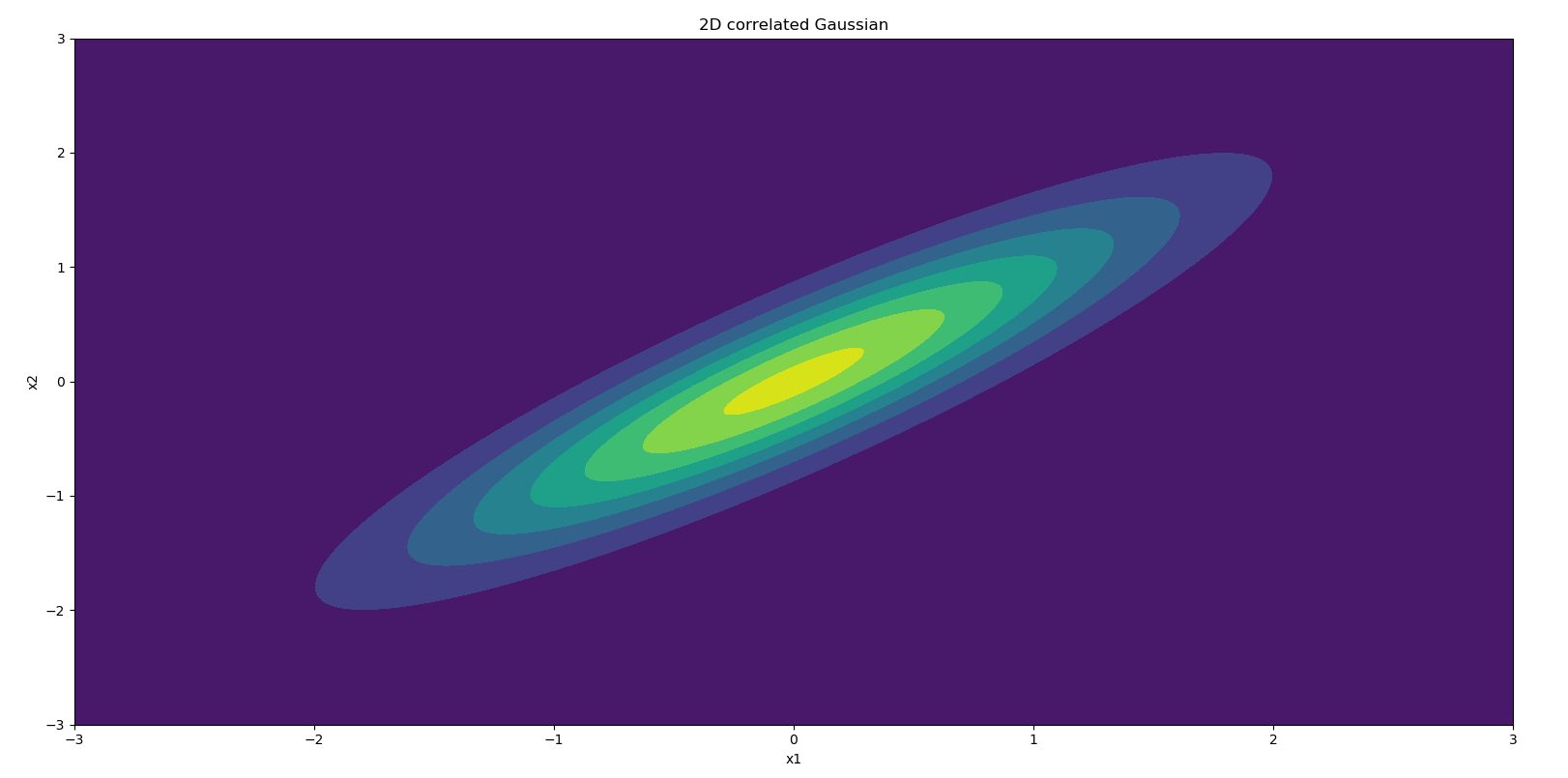
The variable a0 α , b0 α ,a0 λ ,and b0 λ are the hyper-parameter forα and λ

**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 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 probability as shown in Eq… . The energy function is validated by compare with the value obtained from stats library as shown in Table…

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 |

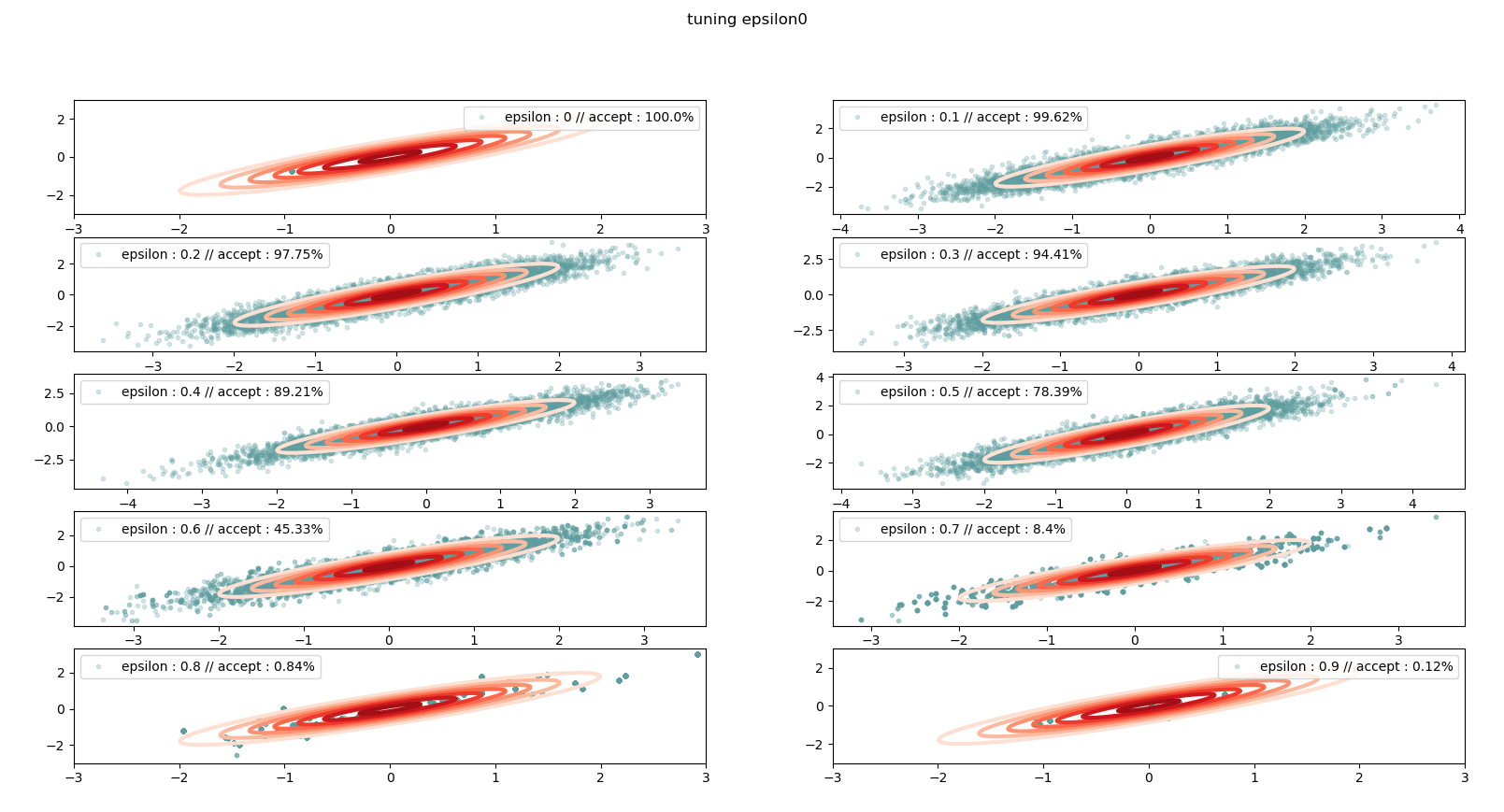
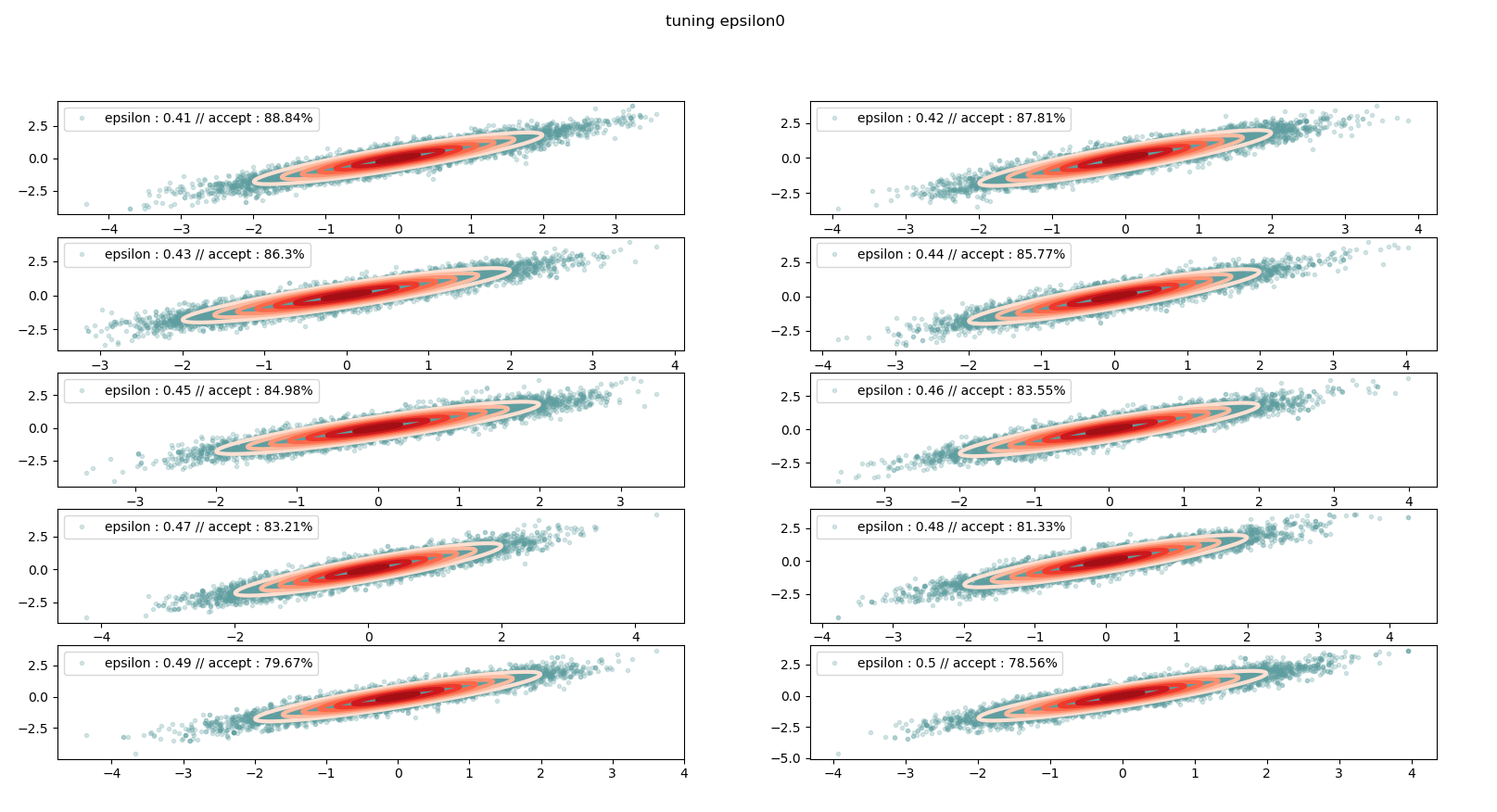
For gradient function, the function takes the input of x1 and x2 and return the corresponding gradient for each variable. Gradient can be determined by taking the derivative of energy function with respect to each variable as shown in Eq… .The gradient function is validated using check gradient function,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

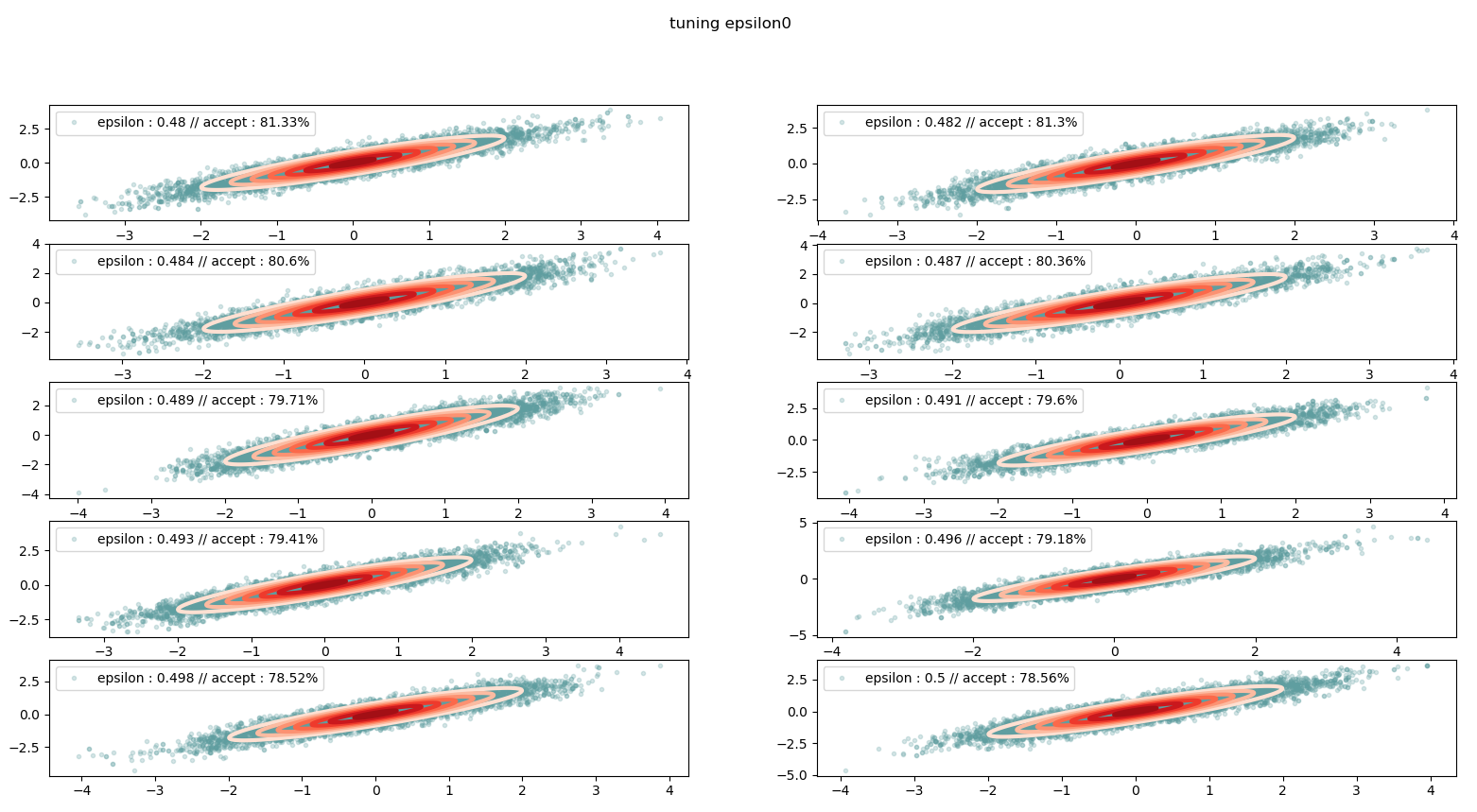
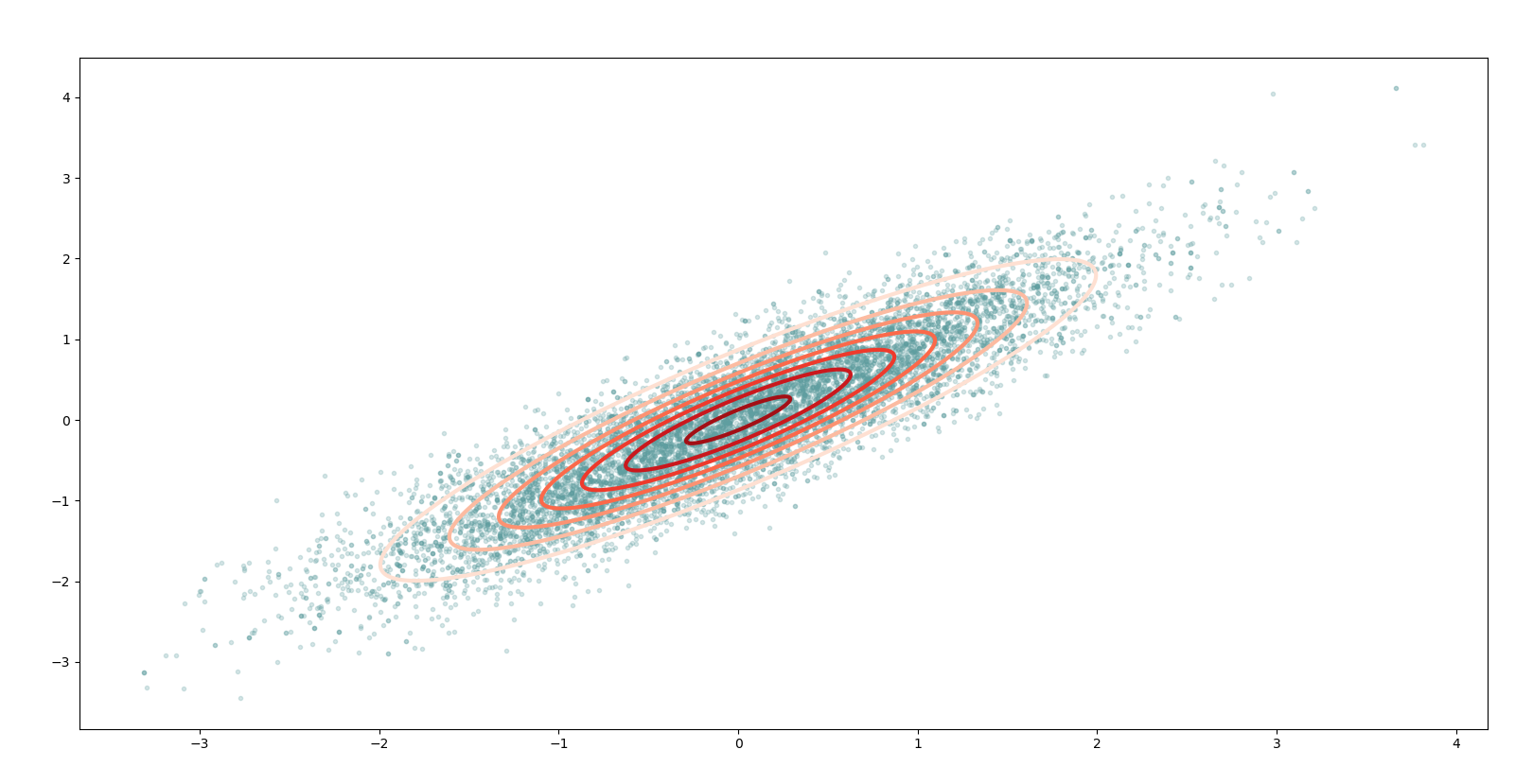
Grad x1 = = (2x1 - 2ρx2)/(2(1- ρ2)

Grad x2 = = (2x2 - 2ρx1)/(2(1- ρ2)

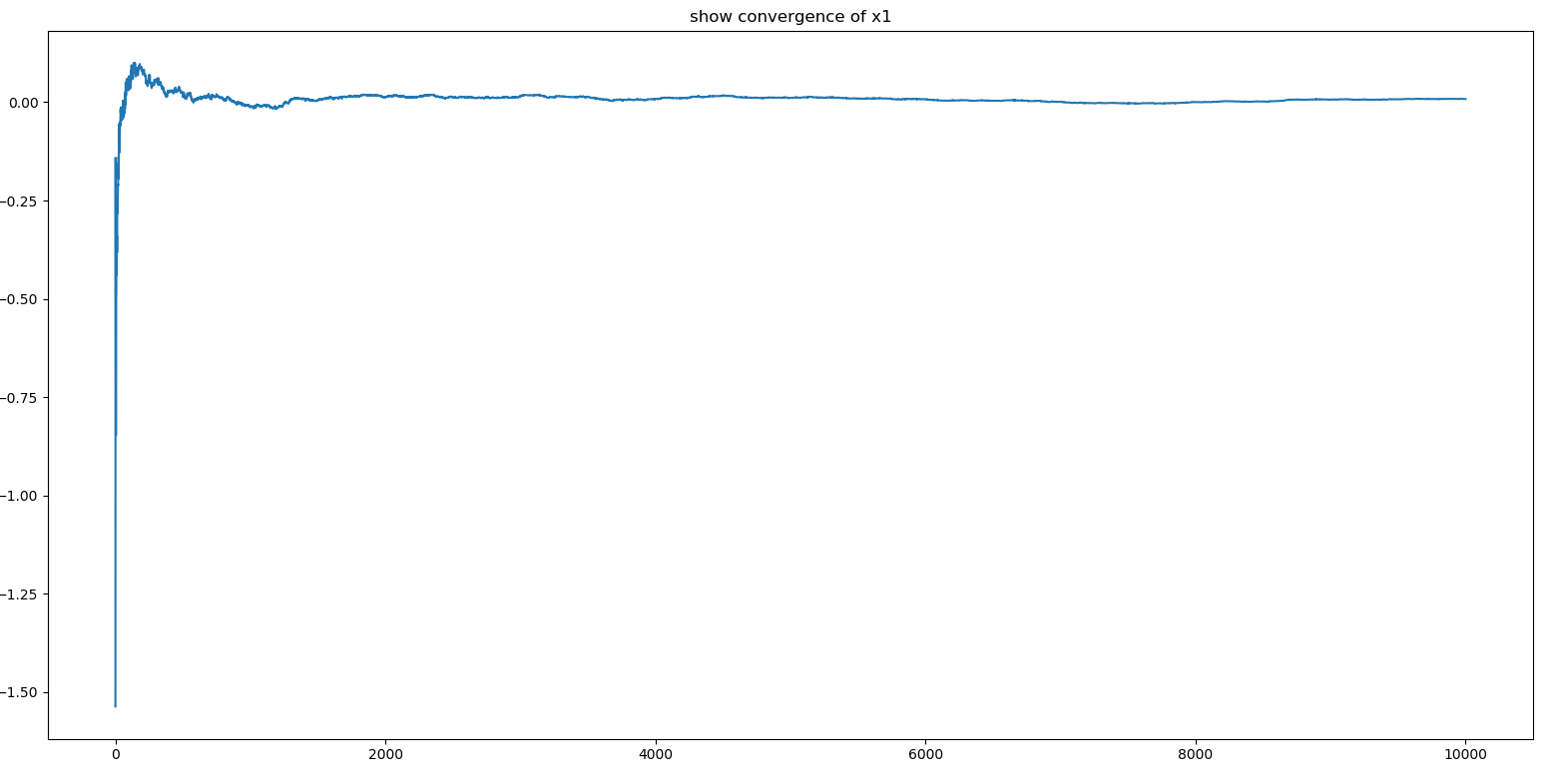
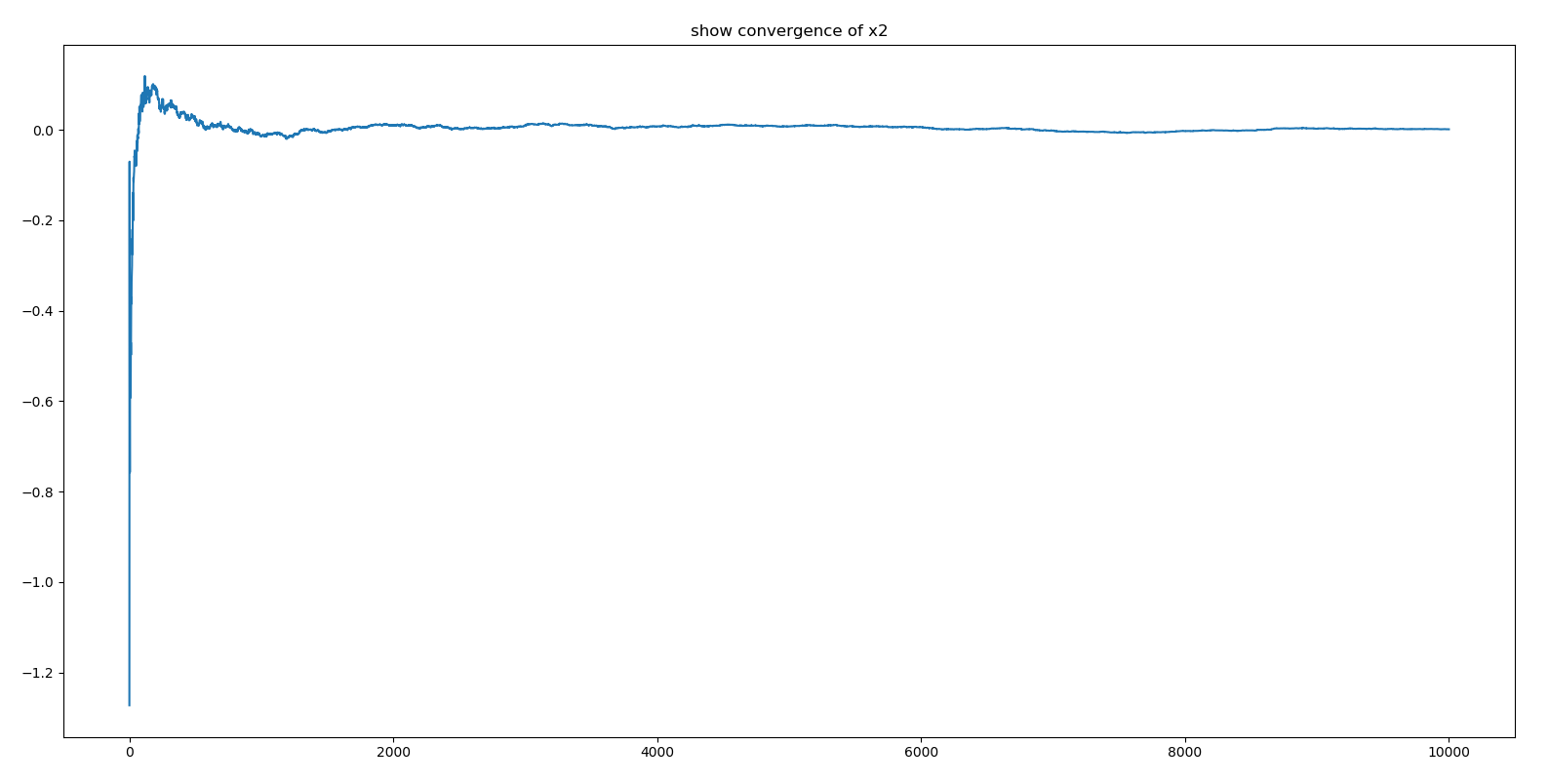
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Calculated | Numeric | Delta | Accuracy |
| X1 | 11.2954 | 11.2954 | -4.116973e-10 | 11 |
| X2 | -11.8463 | -11.8463 | -1.740030e-10 | 11 |

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 numpy cumulative sum.

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

To validate the accuracy of energy function, this is done by comparing the value with output from stats library for 4 terms consists of log likelihood (lnP(y|w,s2)), log weight prior (lnP(w| α)), log hyper prior inverse weight variance (lnP(α)), and log noise variance prior (lnP(s2)). This can be compared by convert energy back to log probability. The results are shown in Table … where the difference are very small and verified the accuracy of the function.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Coding | Stats library | Absolute difference |
| lnP(y|w,s2) | -296357.6249960258 | -296357.6249960263 | 4.656612873077393e-10 |
| lnP(w| α) | -10.09966444922638 | -10.099664449226378 | 1.7763568394002505e-15 |
| lnP(α) | -3.9531513325728818 | -3.953151332572883 | 1.3322676295501878e-15 |
| lnP(s2) | -3.8395808667784257 | -3.8395808667784266 | 8.881784197001252e-16 |

The gradient function is again validated using check gradient function. The results are shown in Table … . All gradients show accuracy of higher than 6 which indicate the absent of error in coding and calculation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Calculated | Numeric | Delta | Accuracy |
| S2 | -31224.3 | -31224.3 | 6.105620e-06 | 10 |
| α | 1.55016 | 1.55016 | -1.411203e-07 | 8 |
| W0 | -1204.22 | -1204.22 | 2.490060e-05 | 8 |
| W1 | 1297.55 | 1297.55 | 6.914611e-06 | 9 |
| W2 | -996.629 | -996.629 | 1.170854e-05 | 8 |
| W3 | 1742.71 | 1742.71 | 2.422461e-05 | 8 |
| W4 | -1820.65 | -1820.65 | -1.494205e-05 | 9 |
| W5 | 302.197 | 302.197 | -2.285516e-05 | 8 |
| W6 | -545.26 | -545.26 | -2.332535e-06 | 9 |
| W7 | -207.931 | -207.931 | 3.188797e-05 | 7 |
| W8 | -4426.91 | -4426.91 | 2.204899e-05 | 9 |

The posterior that we are sample from is P(w, α,s2|y) which is equal to product of P(y|w,s2), P(w| α), P(α),and P(s2). The inverse weight variance hyper-prior and noise variance prior are assume to follow inverse gamma distribution with default parameter of shape and rate of 10^-2 and 10^-4 ;respectively, as recommended by Druguhutsh\*\*\*\*\*\*. The log probability is shown in Eq…

ln P(w, α,s2|y) = a + b + c + d

where:

a = ln P(y|w,s2) = - -

b = ln P(w| α) = - ln ( ) -

c = ln P(α) = ln ( ) – (a0-1)ln(α) -

d = ln P(s2) = ln ( ) + (a0-1)ln(s2) – b0s2

a0=10-2

b0=10-4

Therefore, energy is equal to negative log probability shown in Eq…

Energy = - a – b – c - d

Gradient can be calculated by derivative of Energy function respect to each parameter as shown in Eq…

=

=

= +

= -

= -

= 0

= 0

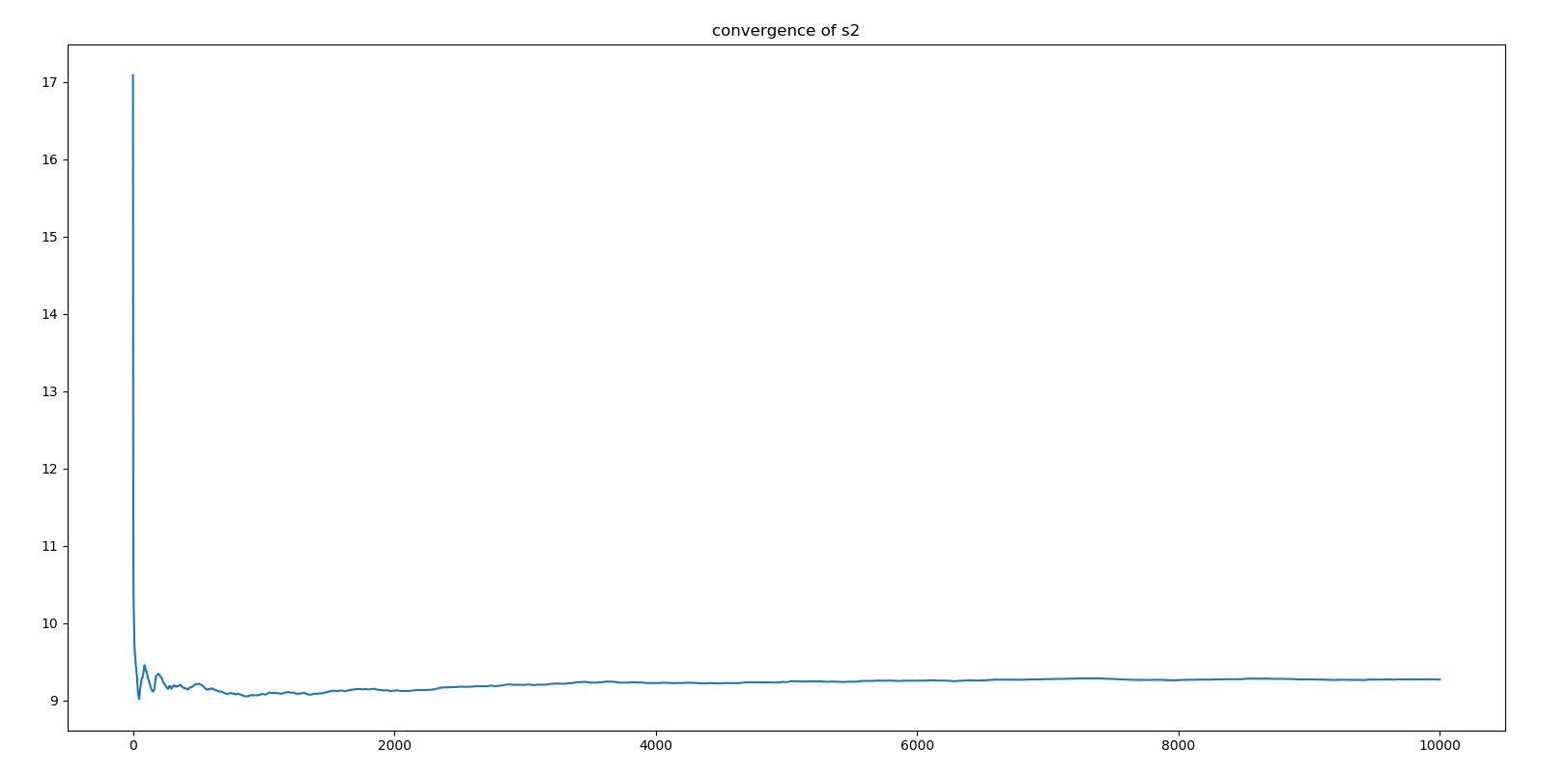
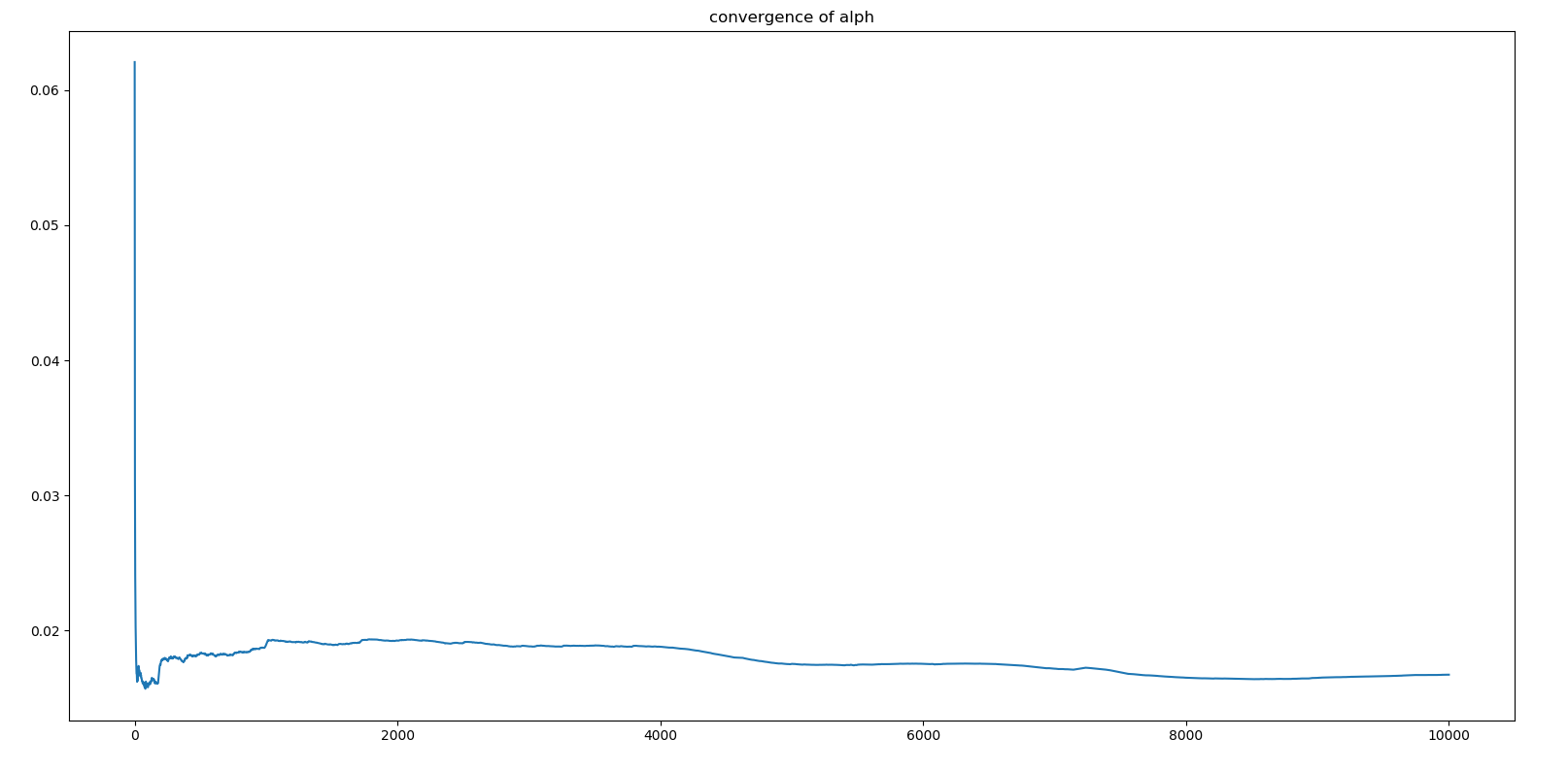
= +

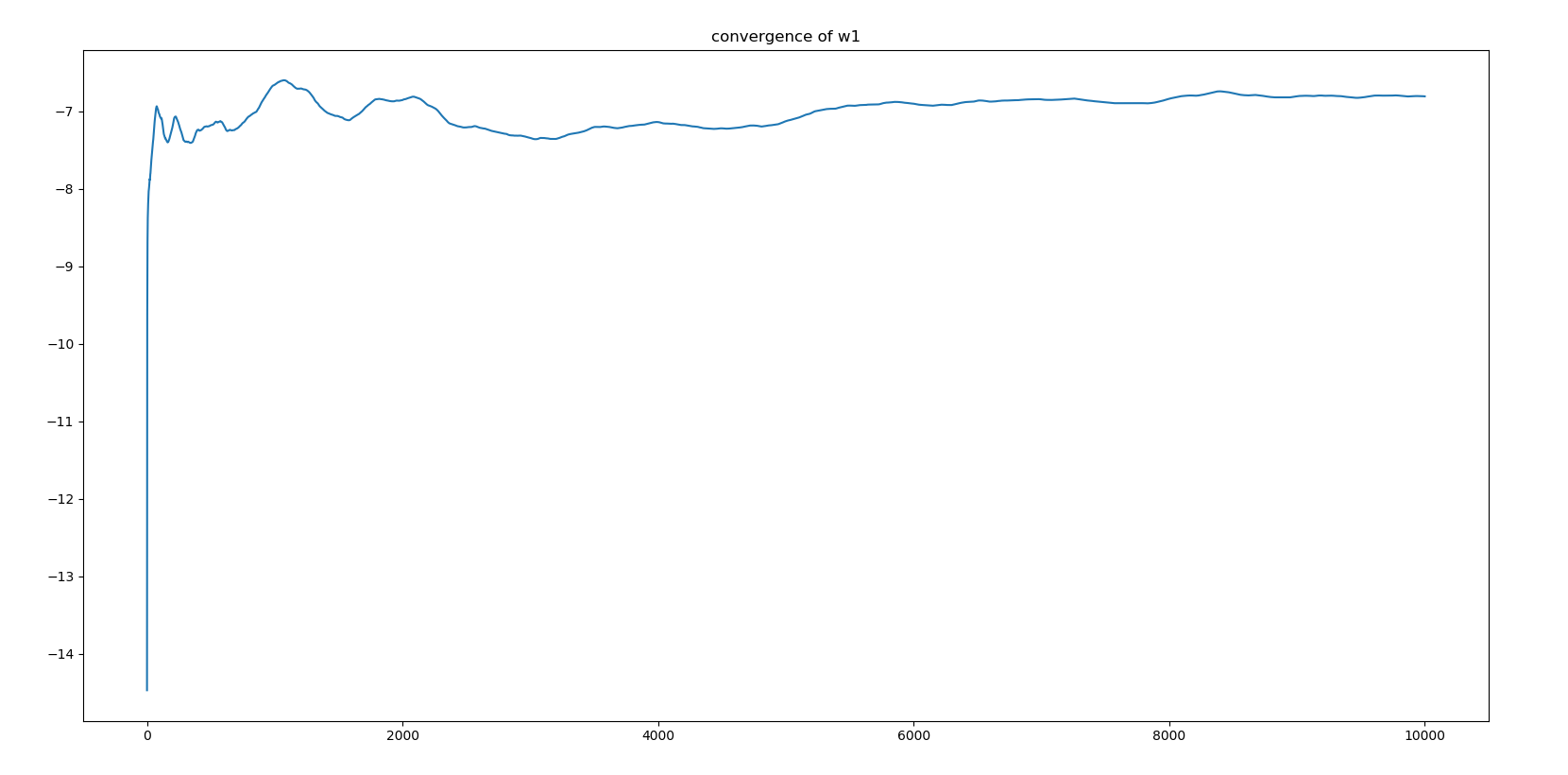
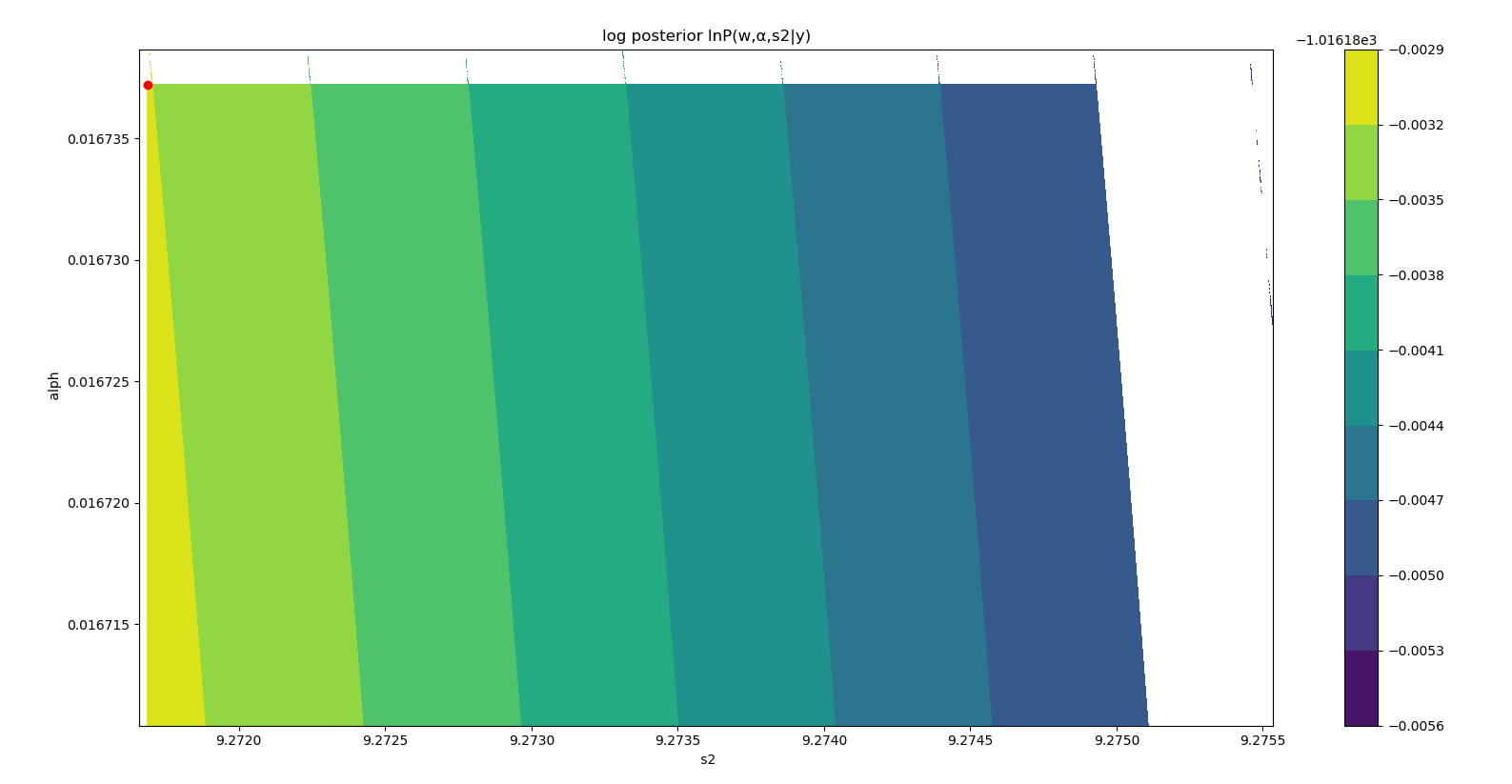
= 0

= 0

= 0

= – b0

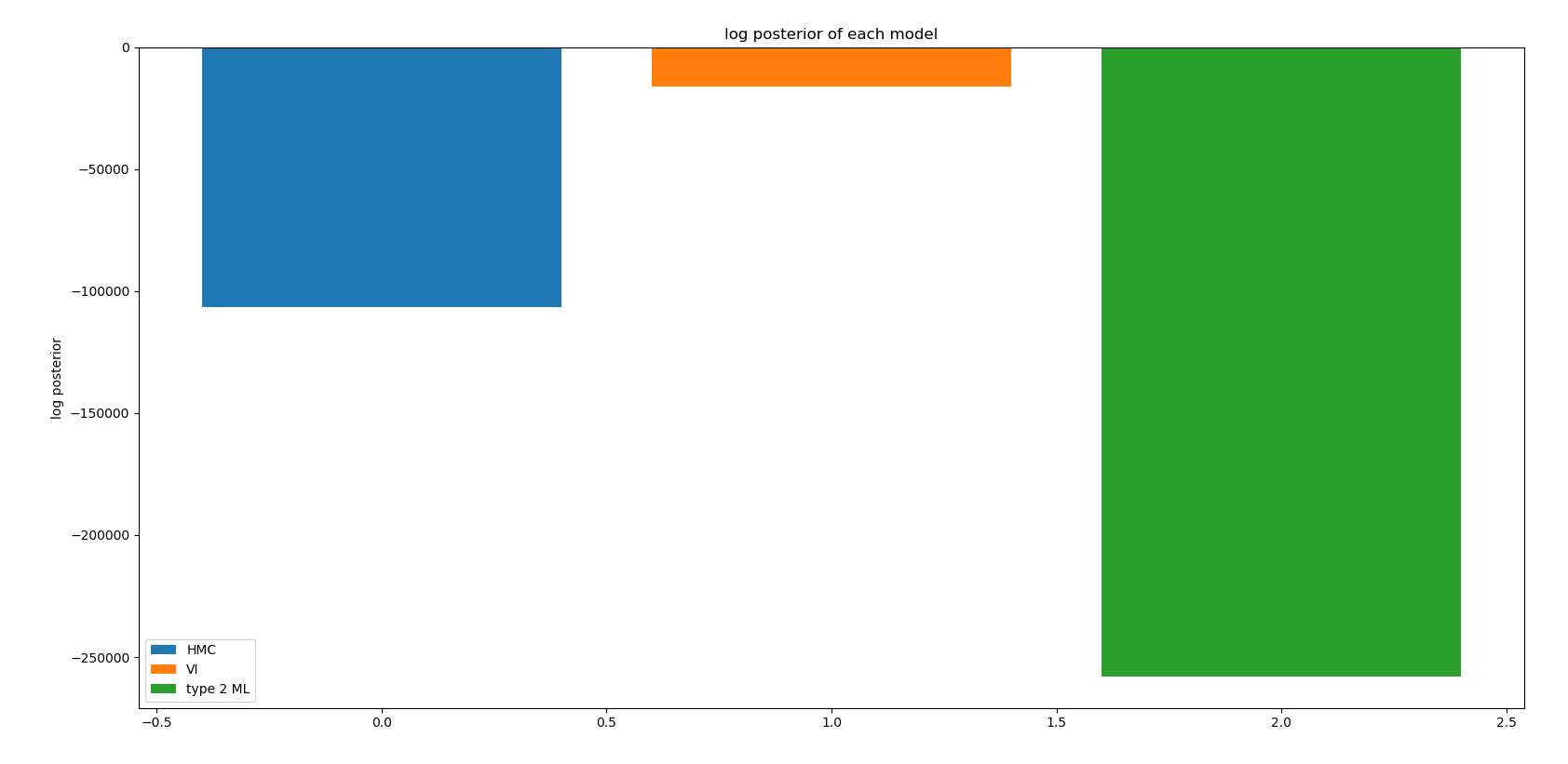
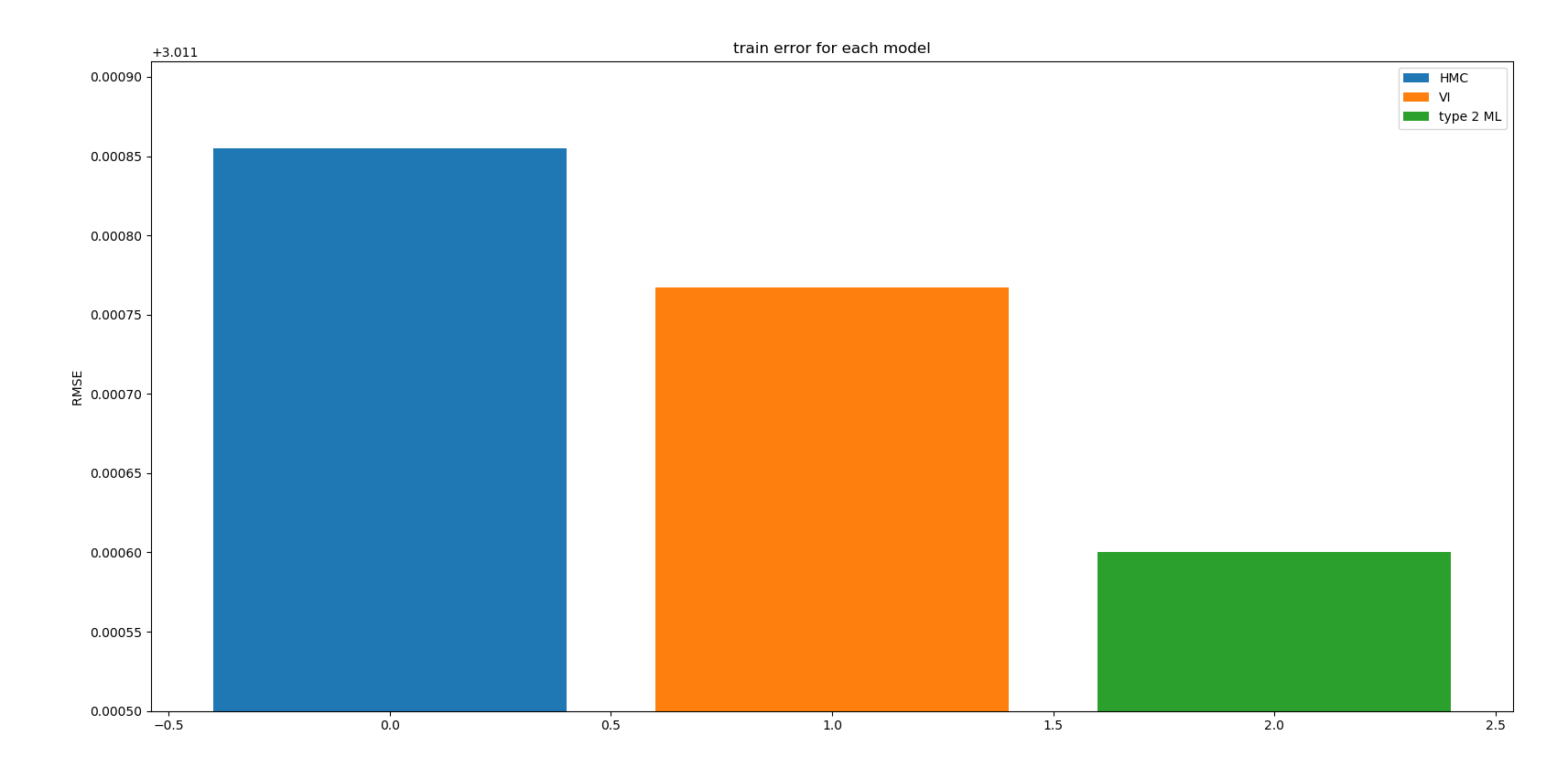
The Figure … , Figure…. , and Figure…. show the convergence of estimate parameter within first 2000 sample. The model also perform well since the location of converged estimate parameter located at highest yellow region of posterior contour plot. The final converged values and errors is shown in Table.. and Table…; respectively.

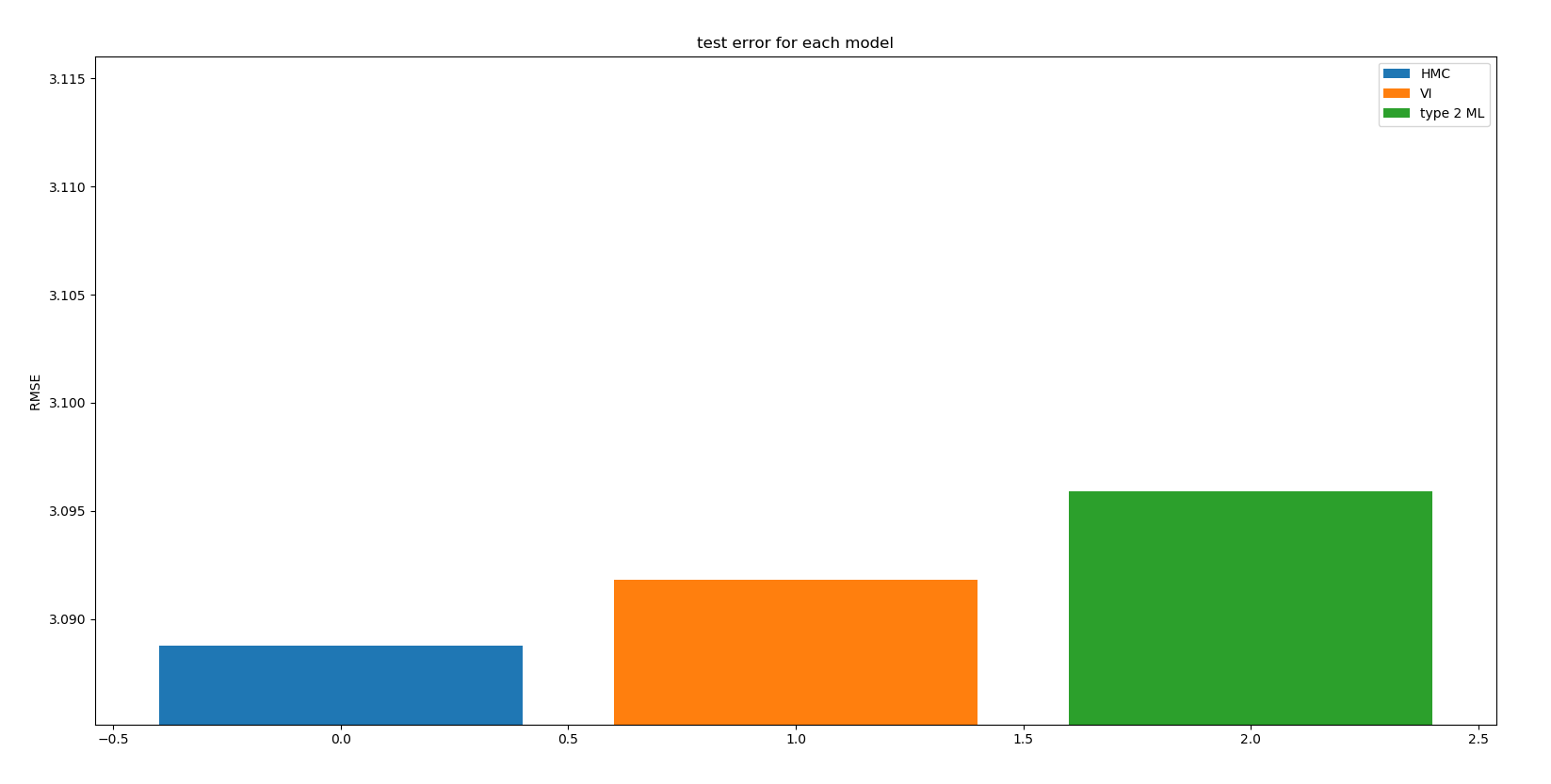
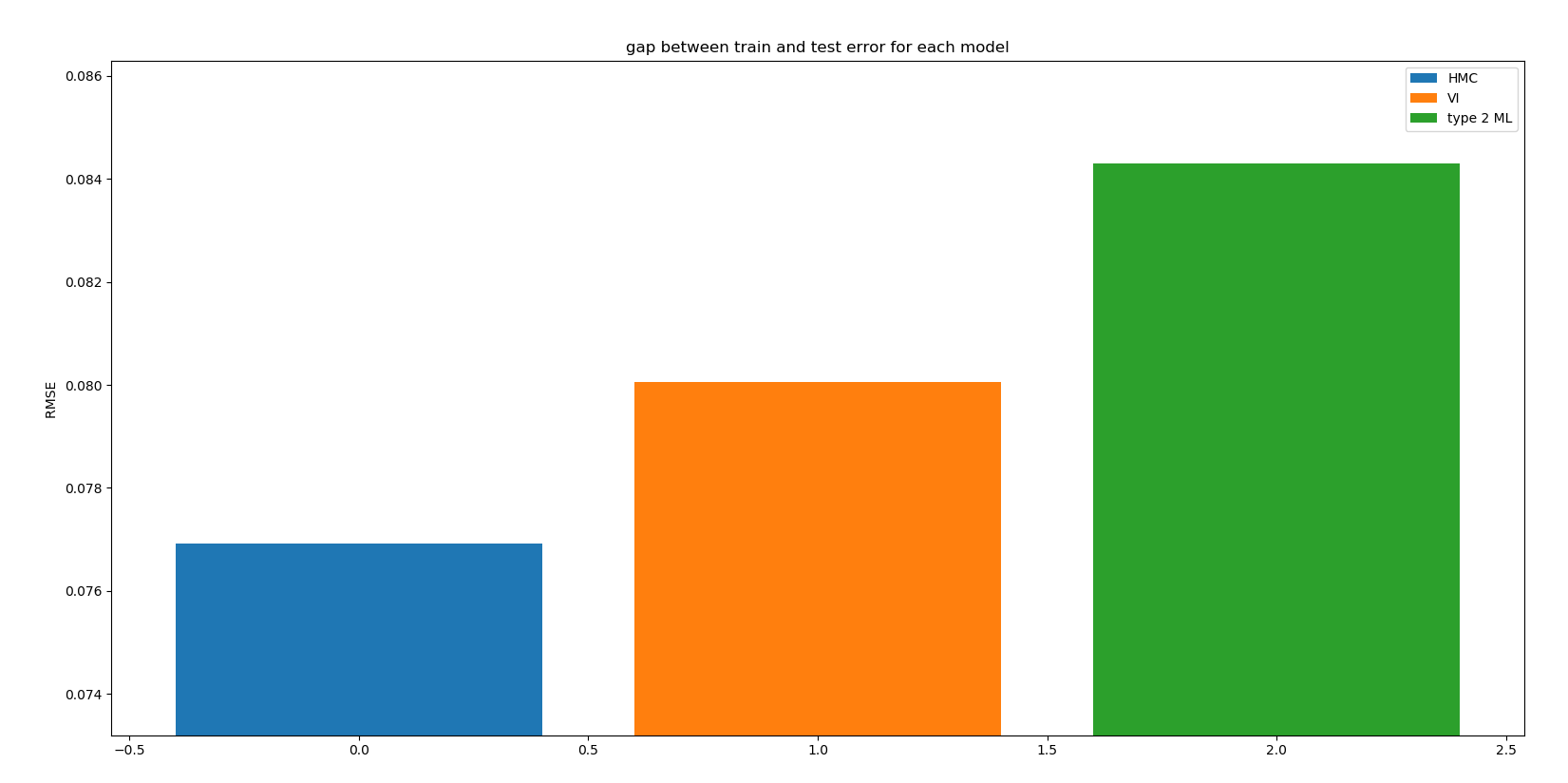
|  |  |
| --- | --- |
| Parameter | Value |
| Best alpha (inverse weight variance) | 0.016737224846942945 |
| Best s2 (inverse noise variance) | 9.271688191719882 |
| W1 | -6.80564271 |
| W2 | -4.54897081 |
| W3 | 1.25374272 |
| W4 | -3.00750759 |
| W5 | 7.37815691 |
| W6 | -0.12444385 |
| W7 | 2.76752209 |
| W8 | 0.20993861 |
| W9 | 22.91402406 |

The model used the value of R for 10000 samples with L and epsilon equal to 100 and 0.0035; respectively. The final percent acceptance is 81.6% where these hyper-parameters are tuned using the same method as described in Task 3.

|  |  |  |
| --- | --- | --- |
| Case | RMSE train | RMSE test |
| HMC | 3.0118549466697635 | 3.0887831894624247 |

Even though HMC model achieved lower value of log posterior as shown in Figure… when compare to Variational inference method, the model shows smallest gap between train and test and error and lowest test set error as shown in Figure.. and Figure … . which indicate that the model does not over fit. While type 2 maximum likelihood method seems to be the worse approach since the model shows highest test set error, largest gap that indicate the most overfit and lowest log posterior.

**5. [Bonus] Modify the HMC sampling framework as a classiﬁer to address a reformulation of the problem. [4 marks]**