DD2434 Machine Learning, Advanced Course Assignment 1

Chia-Hsuan Chou
chchou@kth.se

March 17, 2023

1 The Prior

1.1 Theory

> ***Question 1*** *Explain why Gaussian form of the likelihood is a sensible choice and what assumptions we make by this choice. What assumptions do we make about the data by choosing a spherical covariance matrix for the likelihood?*

Since we have no information about uncertainty here, we can assume that the uncertainty here is a large amount of measurement errors which are independent and identically distributed in the target variable T. According to the Central Limit theorem, the measurement errors here will be normally distributed when the number of observations is sufficiently large. Since the mapping function f and observed variate X are given, the likelihood is only depend on the measurement errors, therefore it is a sensible choice to assume that the likelihood is a gaussian.

A covariance matrix is called spherical if it is proportional to the identity matrix. If the covariance matrix is proportional to the identity matrix, it means that we assume the components of the data are uncorrelated to each other:

$$Cov(t_k^i, t_k^j | f, T) = 0, \forall i, j \subset N, , i \neq j \forall k \leq T$$

> ***Question 2*** *If we do not assume that the data points are independent how would the likelihood look then?*

If the data points are not independent, the likelihood will become :

$$p(\mathbf{T}|f, \mathbf{X}) = p(t_1, t_2, ...t_N|f, \mathbf{X}) = p(t_1|\mathbf{X}, f) \prod_{k=2}^{N} p(t_k|t_1, ..., t_{k-1}, \mathbf{X}, f)$$

1.1.1 Linear Regression

> ***Question 3*** *complete the right-hand side of the expression in*

The likelihood of the data, given the noise in the observations follows additive Gaussian distribution as described in the instruction, will be :

$$p(\mathbf{T}|\mathbf{W}, \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{W}x_n, \sigma^2 I)$$

> ***Question 4*** The prior over each row of W in Eq.8 is a spherical Gaussian: $p(w) = N(w_0, \tau^2 I)$. This means that the preferred model is encoded in terms of L2 distance in the parameter space.
>
> - What would be the effect of encoding the preferred model with L1 norm (for model parameters)?
> - Discuss how these two types of priors affect the posterior from the regularization perspective. Write down the penalization term, i.e. the negative log-prior, and illustrate for a two- dimensional problem (in the two-dimensional parameter space).

In statistics, to encode the preferred model with L2 norm, one can use gaussian distribution to model the prior, at the same time if we want to encode the preferred model with L1 norm, Laplace distribution can be chosen. The following equation shows the Probability density function of Laplace distribution :

$$f(w|w_0, \tau) = \frac{1}{2\tau} e^{-\frac{|w-w_0|}{\tau}}$$

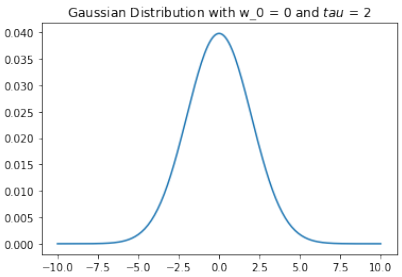

Figure 1: Gaussian Distribution

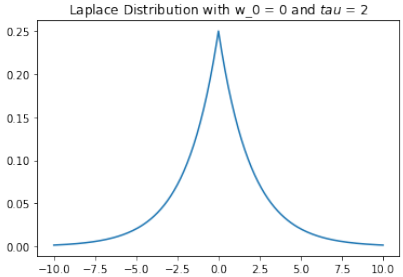

Figure 2: Laplace Distribution

Figures 1 shows the distribution of the Gaussian distribution and figure 2 shows the Laplace distribution with W_0 equals to 0 and τ equals to 2. As we can see from the figures above, Gaussian distribution is much smoother than the Laplace distribution around the mean. This means that the prior in Laplace Distribution, the mean is assigned more weight than other region. Therefore, for those parameters that are far from the mean, their prior will be much less in Laplace distribution than in Gaussian Distribution.

We can estimate the model parameters using maximum a posteriori probability estimate, also known as MAP. To formalize the MAP estimate, we can write it as :[1]

$$\begin{aligned} W_{MAP} &= \underset{W}{\operatorname{argmax}}\, P(T|W, X)P(W) \\ &= \underset{W}{\operatorname{argmax}}\, log(P(T|W, X)P(W)) \\ &= \underset{W}{\operatorname{argmax}}\, logP(T|W, X) + log(P(W)) \end{aligned}$$

The term $\log(P(W))$ can be seen as a regularization. Assuming in the two-dimensional parameter space, For Gaussian Distribution prior (L2 norm distance) W_{MAP} can be written as follow:

$$W_{MAP} = \underset{W}{\operatorname{argmin}}[\sum_{i=1}^{n}(t_i - (W_0 + W_1 x_{i,1} + W_2 x_{i,2}) + \lambda(W_0^2 + W_1^2 + W_2^2)]$$

And for Laplace Distribution prior (L1 norm distance) W_{MAP} can be written as follow:

$$W_{MAP} = \underset{W}{\operatorname{argmin}}[\sum_{i=1}^{n}(t_i - (W_0 + W_1 x_{i,1} + W_2 x_{i,2})) + \lambda(|W_0| + |W_1| + |W_2|)]$$

As λ increases, the parameters w_i will shrink toward zero. There is a slight difference here. Instead of preventing any of the parameters from being too large, L1 norm promotes sparsity. The parameters can be forced to zero with L1 norm while in L2 norm the parameters can only be close to zero, but not exactly 0.[2]

Question 5 Assuming conditional independence of the target variables in t, derive the posterior over the parameters. Please, do these calculations by hand as it is very good practice. To pass the assignment you only need to outline the calculation and highlight the important steps. In summary, please complete the following tasks:

[1] http://bjlkeng.github.io/posts/probabilistic-interpretation-of-regularization/

[2] p33 - p40, Lecture 4 slide, DD2421:Machine Learnig, KTH

- Derive the posterior over the parameters and explain the final form in terms of the mean and covariance.
- How does the posterior form relate to the least square estimator of W (equivalent to the maximum likelihood approach) for this linear regression problem?
- How does the constant Z (Eq.7) affect the solution? Are we interested in it?

From Bayesian theorem, we know that :

$$P(W|X,t) \propto P(W)P(t|W,X)$$

First of all, Let's assume P(W) and P(t|W,X) are Gaussian distribution, so P(W|X,t) will be Gaussian distribution as well.Therefore, we can assume:

$$P(W|X,t) = N(\mu_w, \Sigma_w^{-1})$$

Secondly, We assume Prior over W P(W) has a mean 0 and variance Σ^{-1}, we can write P(W) in the following form:

$$P(W) = N(0, \Sigma I_n) = e^{-\frac{1}{2}W^T\Sigma^{-1}W}$$

Thirdly, the likelihood P(t|W,X). we assume that the noise of the observation is gaussian distribution, and target variables t are conditional independent, the likelihood can we written as :

$$p(t|W,X) = \prod_{i=1}^{N} P(t_i|W,X) = N(XW, \sigma^2 I_N) = e^{-\frac{1}{2\sigma^2}(t-XW)^T(t-XW)}$$

So posterior P(W|t,x) will become

$$P(W|X,t) \propto P(W)P(t|W,X) \propto e^{-\frac{1}{2}W^T\Sigma^{-1}W} e^{-\frac{1}{2\sigma^2}(t-XW)^T(t-XW)}$$

Now, we focus on the exponent:

$$\begin{aligned} -\frac{1}{2}W^T\Sigma^{-1}W - \frac{1}{2\sigma^2}(t-XW)^T(t-XW) &= -\frac{1}{2\sigma^2}t^Tt + \frac{1}{\sigma^2}t^TXW - \frac{1}{2\sigma^2}(XW)^T(XW) - \frac{1}{2}W^T\Sigma^{-1}W \\ &= \frac{-1}{2\sigma^2}t^Tt + \frac{1}{\sigma^2}t^TXW - \frac{1}{2}W^T(\frac{1}{\sigma^2}X^2X + \Sigma^{-1})W \end{aligned}$$

The first term, $\frac{-1}{2}t^Tt$ is a constant for posterior and the second term, $\frac{1}{\sigma^2}t^TXW$ is linear to W, and the third term, $-\frac{1}{2}W^T(\frac{1}{\sigma^2}X^2X + \Sigma^{-1})W$ is quadratic to W. From The third term we can derive Σ_w^{-1}:

$$-\frac{1}{2}W^T(\frac{1}{\sigma^2}X^2X + \Sigma^{-1})W = -\frac{1}{2}W^T\Sigma_w^{-1}W$$

$$\Sigma_w^{-1} = \frac{1}{\sigma^2}X^TX + \Sigma^{-1}$$

Now we know Σ_w^{-1} then we can derive μ_w from the second term:

$$\frac{1}{\sigma^2}t^TXW = \frac{1}{\sigma^2}W^TX^Tt = W^T\Sigma_w^{-1}\mu_w$$

$$\mu_w = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}X^TX + \Sigma^{-1})^{-1}X^Tt$$

So the posterior becomes :

$$P(W|X,t) = N(\frac{1}{\sigma^2}(\frac{1}{\sigma^2}X^TX + \Sigma^{-1})^{-1}X^Tt, \frac{1}{\sigma^2}X^TX + \Sigma^{-1})$$

To find **W** using the least square estimator, We try to find a **W** that can minimize the square error of target variables and observations:

$$W_{MLE} = \underset{W}{\operatorname{argmax}}\, P(t|W,X) = \underset{W}{\operatorname{argmax}}\, log(P(t|W,X))$$

In the equation above we can see that the only different between MLE is that we multiply the a prior of **W** in MAP, which means that we don't do regularization in MLE while in MAP using posterior of **W** regularization is considered.

The constant **Z** here is the normalizing term. It is used to normalize the posterior to make sure the posterior will sum up to 1. We don't care about the constant **Z** here because we only need to compare the linear term and quadratic term to find the mean and variance of posterior here, therefore it is not important in this question but in model selection.

1.1.2 Non-parametric Regression

Question 6 Explain what this prior does. Motivate the choice of this prior and use images to show your reasoning. Clue: use the marginal distribution to explain the prior

From the provided instruction we know that, the relationship between input value **x** and observation **t** is:

$$t_i = f(x_x) + \epsilon \rightarrow f_i + \epsilon$$

where f_i is the output of function at input location x_i.

In Gaussian Process, we define the prior probability distribution over functions directly instead of with parametric model. Since we seldom know that kind of function we should use in the real world, Gaussian process is apparently a good choice for us.

Because we have noise ϵ that have Gaussian distribution which is independent to each data point, and the the values f_1,f_2,.....,f_i are evaluated jointly from x_1,x_2,...,x_i and will be Gaussian distribution, this marginal distribution and the prior can be written as:

$$p(f|X,\theta) = N(0, k(X,X))$$

In most application, we have no prior information about f, so by symmetry we assume the mean of the prior is 0. k(X,X) is the kernel function or we can simply denote it as **K**, as known as Gram matrix. If point X_m and X_n are similar, $y(x_m)$ and $y(x_n)$ will be more strongly correlated than dissimilar points, and k(X,X) here is to represent this property.[3].

A common used Kernel function for Gaussian process is given as follows:

$$k(x_n, x_m) = \theta_0 exp^{-\frac{\theta_1}{2}||x_n - x_m||^2} + \theta_2 + \theta 3 x_n^T x_m$$

In the kernel function, θ_0,θ_1,θ_2 and θ_3 control the shape of the distribution. The following figures are from the book Bishop and shows the experiment on different parameters θ_0,θ_1,θ_2 and θ_3:

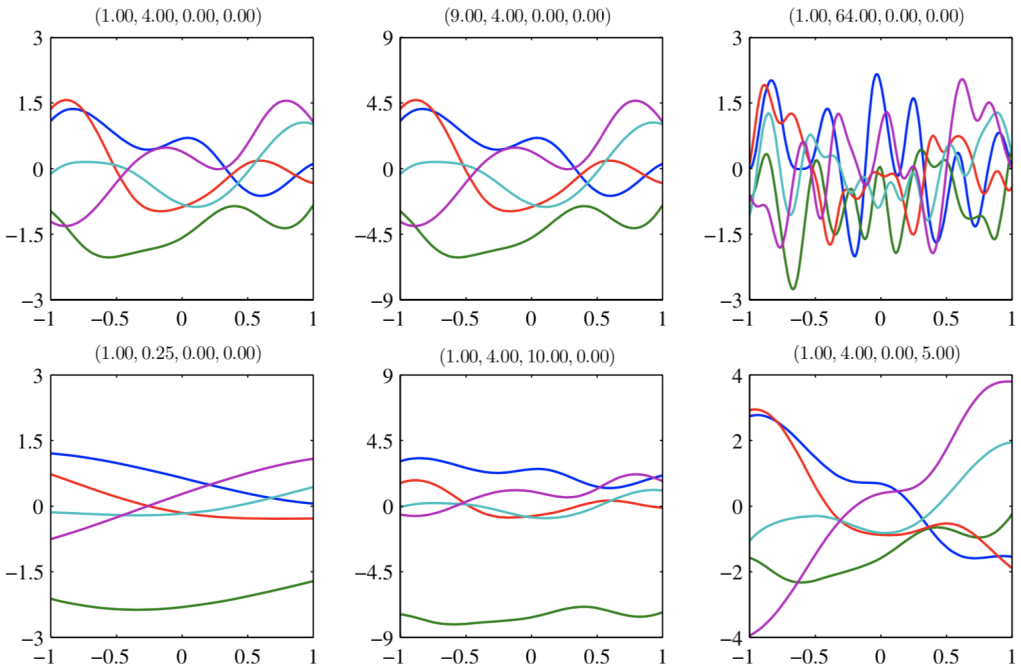

Figure 3: Samples from a Gaussian process prior defined by the covariance function above. The title above each plot denotes (θ_0, θ_1, θ_2, θ_3).(Bishop 308, fig. 6.5)

[3]Bishop p304, p305

Question 7 Formulate the joint likelihood of the full model defined above,

$$p(T, X, f,)$$

and draw a simple graphical model reflecting the assumptions that you have made.

According to the chain rule in probability theory, the joint likelihood of full model can be rewritten as:

$$P(T, X, f, \theta) = P(f|T, X, \theta)P(T|X, \theta)P(X|\theta)P(\theta)$$

First of all, we know that input variables X are independent to the θ. Secondly, we know that f is dependent to X and θ and T is dependent to f, so T is conditional independent to X and θ. Thus, the joint likelihood of full model can be further simplified into following forms:

$$P(T, X, f, \theta) = P(f|T)P(f|X, \theta)P(X)P(\theta)$$

A simple graphical model can reflect my assumption of the joint likelihood of full model :

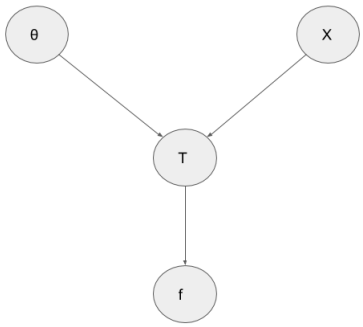

Figure 4: The simple graphical model of joint likelihood of full model

Question 8 Complete the marginalisation formula in Eq.12 (general form) and discuss the following:

- Explain how it connects the prior and the data.
- How does the uncertainty "filter" through the marginalisation? • Why do we still condition on after the marginalisation?

As described in the instruction, we are not interested in f and therefore the variable should be marginalized out. Therefore we get the form:

$$P(T|X, \theta) = \int P(T|f)P(f|X, \theta)df$$

The prior here is P($f|X$,θ). Here we average the likelihood of the data X over all the possible function parameters θ. Here, P($T|f$) represents the uncertainty between T and f, that is, ϵ in the equation (9) in instruction. P($f|X$,θ) represents the uncertainty between X and f, so P($f|X$,θ) and P($T|f$) together we filter the uncertainty between T and X.

The reason why we still condition on θ after the marginalization is that we marginalize f to find P($T|X$,θ), in integral function, we seen θ as constant, so θ remains in the left hand side after marginalization.

1.2 Practical

1.2.1 Linear Regression

Question 9

1. Set the prior distribution over W and visualise it.
2. Pick a single data point (x,t) and visualise the posterior distribution over W.
3. Draw 5 samples from the posterior and plot the resulting functions.
4. Repeat $2-3$ by adding additional data points up to 7.
5. Given the plots explain the effect of adding more data on the posterior as well as the functions. How would you interpret this effect?
6. Finally, test the exercise for different values of 2, e.g. 0.1, 0.4 and 0.8. How does your model account for data with varying noise levels? What is the effect on the posterior?

Figure 5 is the prior distribution over W.

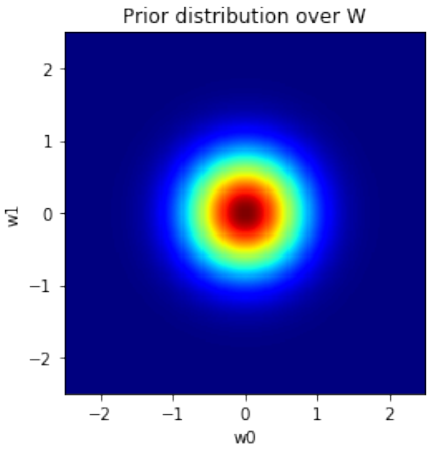

Figure 5: Prior Distribution over Parameter W

We pick n single data point (x,t) randomly and visualize the posterior distribution, then sample 20 parameters W under normal distribution with the mean and variance we found to plot the data space. Figure 6 ,7,8,9,10 and 11 show the results of the posterior distribution and the functions with n = 1,5,6,7,20 and 100.

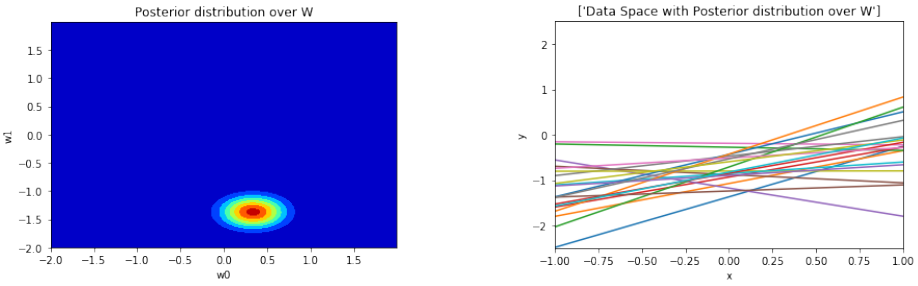

Figure 6: Posterior Distribution and Data Space with 1 data point

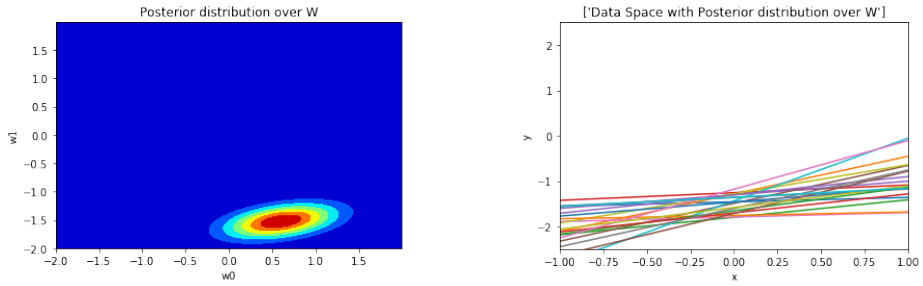

Figure 7: Posterior Distribution and Data Space with 5 data points

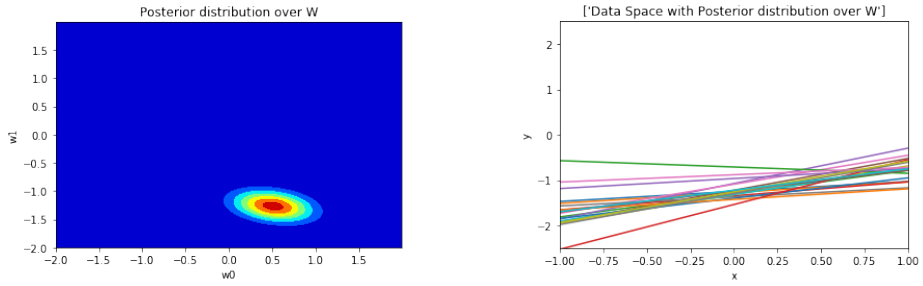

Figure 8: Posterior Distribution and Data Space with 6 data points

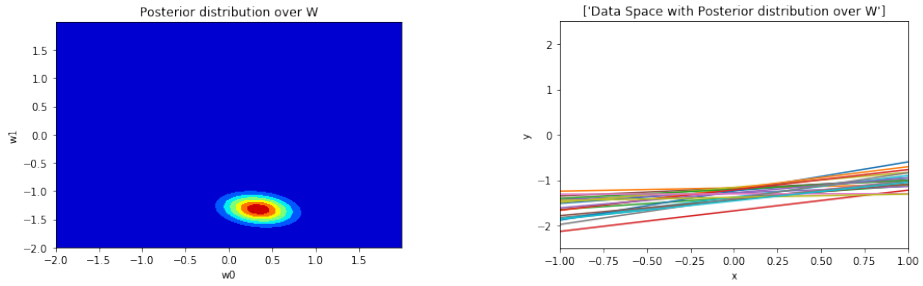

Figure 9: Posterior Distribution and Data Space with 7 data points

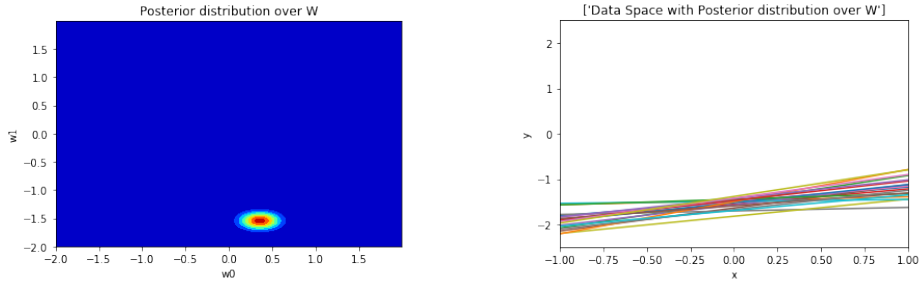

Figure 10: Posterior Distribution and Data Space with 20 data points

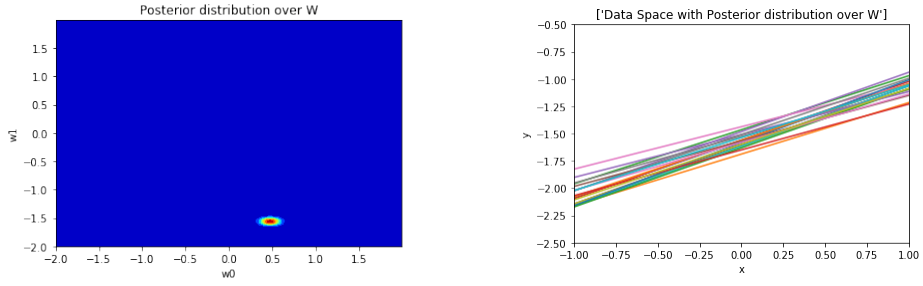

Figure 11: Posterior Distribution and Data Space with 100 data points

From figure 6,7,8,9,10 and 11 we can see that, the more numbers of the data point we pick up, the smaller the variance of the posterior distribution over parameter W will be, and the mean of posterior distribution over parameter w will converge to the actual parameters of function we generate t. At the same time, more numbers of the data points we pick up, the parameters W we sampling from the posterior distribution will give us closer lines in data space, which means the functions we sampling are more likely to be similar because the variance of the posterior distribution over parameter W is smaller.

Now, let's test the exercise for different values of σ, e.g. 0.1, 0.4 and 0.8. The number of data points here is 20. From Figure 12, 13 and 14 we can see that, the smaller σ we test, the smaller the variance of posterior distribution we will get, and the functions we sampling from the posterior distribution will be much similar to each other. This shows that, if we have less noise for the data we observed, we are more 'confident' on the parameters we found under the same number of the data points. If we have noisy observation, we will need more data points to find the parameters with same variance of posterior distribution.

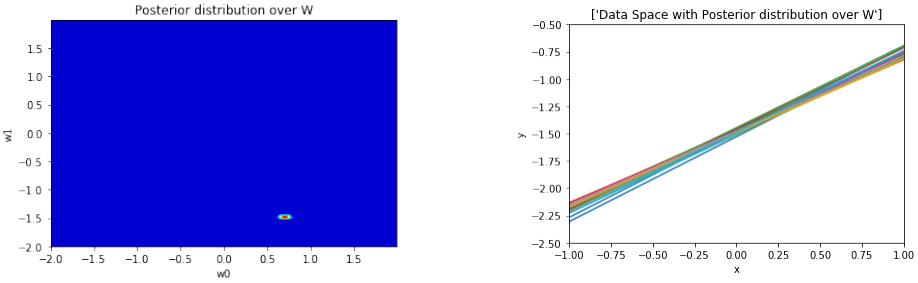

Figure 12: Posterior Distribution and Data Space with $\sigma = 0.01$

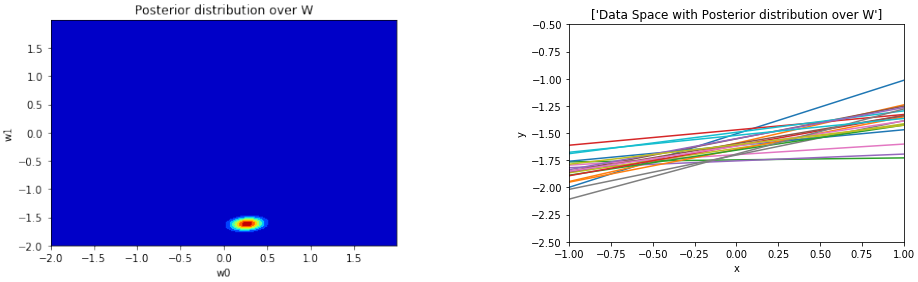

Figure 13: Posterior Distribution and Data Space with $\sigma = 0.1$

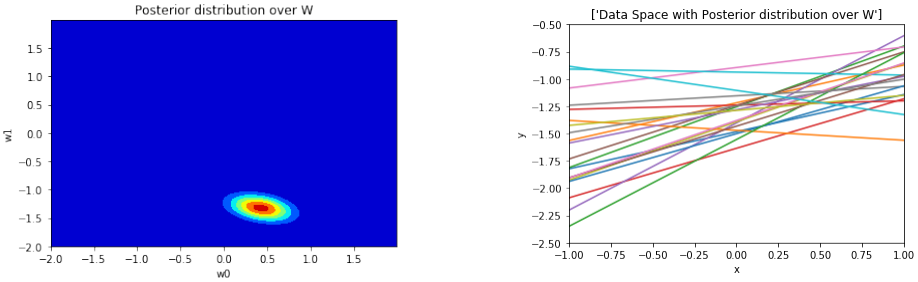

Figure 14: Posterior Distribution and Data Space with $\sigma = 0.4$

1.2.2 Non-parametric Regression

Question 10

- Create a GP-prior with a squared exponential covariance function.
- For each of 4 different length scales, please draw 10 samples from this prior and visualise them. Explain the observed consequences of altering the length-scale of the covariance function.

From question 6 we know that, in gaussian process, the prior can be written as:

$$p(f|X,\theta) = N(0, k(X,X))$$

and $k(X,X)$:

$$k(x_i, x_j) = \sigma_f^2 e^{-\frac{(x_i - x_j)^T (x_i - x_j}{l^2}}$$

$k(X,X)$ here is squared exponential covariance function. To plot the prior under different length scales, the σ_f is set to be 1 and X = -2,-1.999,-1.997,.....1.997,1.999,2. As we can see in figure 15, the larger the length

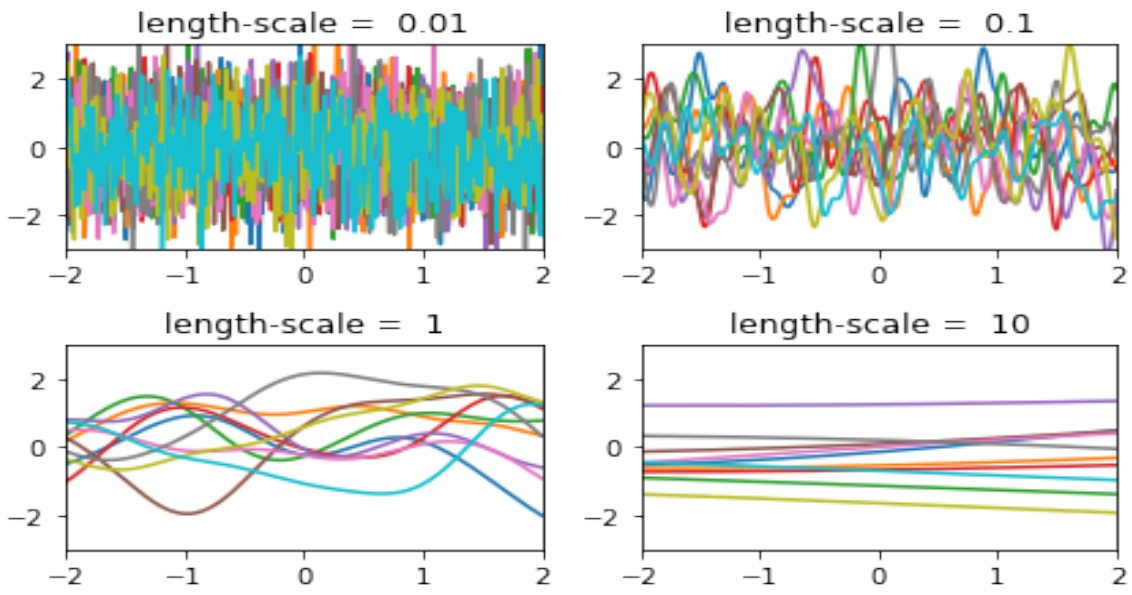

Figure 15: Prior in 4 different length scales

scale is, the smoother the samples are. The larger l in the squared exponential covariance function , the larger $k(x_i, x_j)$ will be, which means f_i and f_j will be more likely correlated. On the contrast, the smaller l we have, the dimensions will be more independent with each other.

Question 11

1. What is the posterior before we observe any data?
2. Compute the predictive posterior distribution of the model.
3. Sample from this posterior with points both close to and far away from the observed data. Explain the observed effects.
4. Plot the data, the predictive mean and the predictive variance of the posterior from the data.
5. Compare the samples of the posterior with the ones from the prior. Is the observed behavior desirable?
6. What would happen if you added a diagonal covariance matrix to the squared exponential?

Before we observe any data, the posterior is the prior.

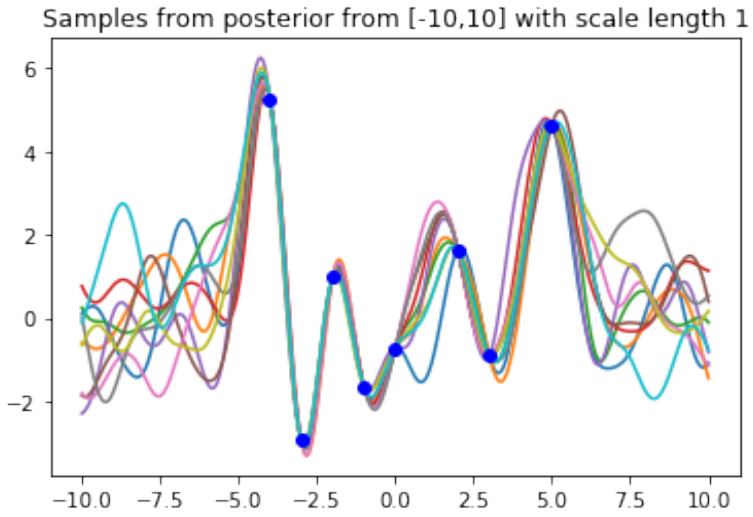

Figure 16: Sampling from posterior over[-10,10] with scale length 1

In figure 16 , blue dots are the data points(x,t) we generate. Here we can see that the sampling are much similar in the range of [-5,5], which is also the range of data points we generate. Outside the range of this, we can see that the sampling have different performance.

In figure 17, the blue curve shows the function provided in the instruction and the red dots are the data points. The green curve are the mean of gaussian predictive distribution, and the green shaded area are plus and minus 2 standard deviation. Since our function diverges when x becomes larger or negative smaller, the region of the standard deviation seems not so significant, the variance with x far away from data points is much larger than x which is close to the data points.

Compare to the prior distribution in figure 15, it is obvious that the behavior in figure 16 is much better than that in prior distribution. We can see that in the region closed to the 8 data points, the green curve(gaussian predictive distribution) can catch up the behavior of those data points. the behavior of sampling in figure 16 is similar to the prior distribution in 15 outside the range of the data points, however, this is what we can accept, since there is not data points provided in that region, the model has no information about that region.

A diagonal covariance matrix is added to the squared exponential covariance function in figure 18. We can see that, the red dots (data points we generate) are no longer on the green curve (the mean of gaussian predictive distribution) anymore. It is equivalent to add independent noise to each value of data point $y_i = f(x_i)$.

2 The Posterior $p(X|Y)$

2.1 Theory

Question 12 What type of "preference" for the latent variable X does this prior encode?

As described in the instruction, we specify the prior over the latent variables as a spherical Gaussian:

$$p(X) = N(0, I)$$

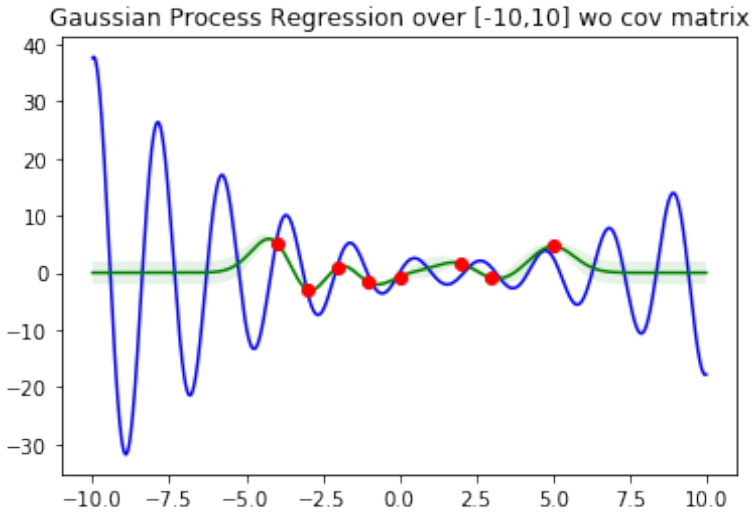

Figure 17: Gaussian Process Regression over [-10,10] without adding cov matrix

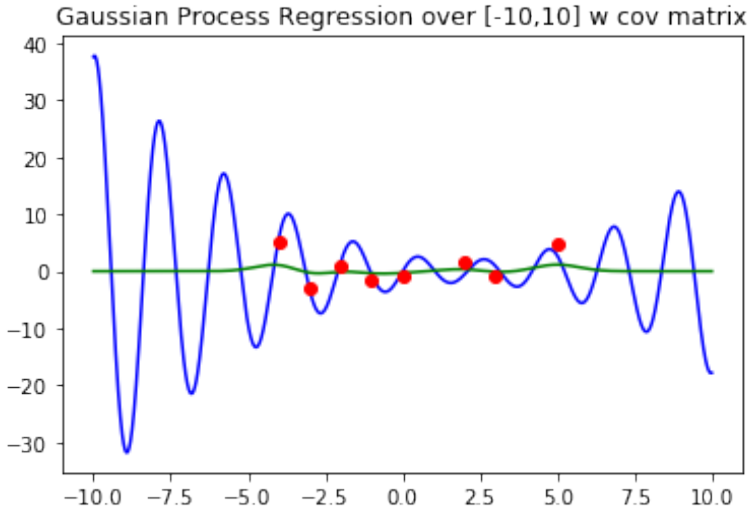

Figure 18: Gaussian Process Regression over [-10,10] with adding cov matrix

This prior encode that we specify the latent variables X has mean 0 and covariance matrix I, which means :

$$Cov(X_k^i, X_k^j) = 0, \forall i, j \subset N, , i \neq j, \forall k \leq T$$

Question 13 Perform the marginalisation in Eq. 23 and write down the expression. As previously, it is recommended that you do this by hand. In the answer outline the calculations and highlight the important steps. Hint: The marginal can be computed by integrating out X with the use of Gaussian algebra we exploited in the exercise derivations and, in particular, by completing the square. However, it is much easier to derive the mean and covariance, knowing that the marginal is Gaussian, from the linear equation of Y(X).

As described in the instruction, we assume that the linear relationship between X and Y is:

$$Y_i = WX_i + \epsilon_i$$

where ϵ is N(0,σ^2I). Because this corresponds to the linear Gaussian model, the marginal distribution will be Gaussian distribution as well and can be written in the following form:

$$P(Y|W) \sim \int P(Y|X,W)P(X)dX = N(Y|\mu, C)$$

μ here is the mean of Gaussian distribution and C here is covariance matrix. We can derive μ and C using expected values of Y and covariance :

$$\mu = E[Y] = E[WX + \sigma] = WE[X] + E[\sigma] = W * 0 + 0 = 0$$

and the covariance matrix is calculated as following:

$$C = E[(WX + \sigma)^T(WX + \sigma)] = E[WXX^TW^T] + E[\sigma\sigma^T] = WW^T + \sigma^2 I$$

P(Y|W) then becomes:

$$P(Y|W) \sim N(Y|\mu, C) = N(Y|0, WW^T + \sigma^2 I)$$

2.1.1 Learning

Question 14 Compare the three different estimation procedures above in log-space.

1. What are their distinctive features and how are they different when we observe more data?
2. Why are the two last expressions of Eq. 25 equal?
3. Explain why Type-II Maximum-Likelihood is a sensible approach to learn the model.

As we have mentioned in previous questions, The maximum-likelihood(ML) in log space is :

$$W_{MLE} = \underset{W}{\operatorname{argmin}} \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - W^T x_i)^2$$

and the maximum-a-posteriori (MAP) estimation in log space is:

$$W_{MAP} = \underset{W}{\operatorname{argmin}} \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - W^T x_i)^2 + (\sum w_i^2)$$

and the Type-II Maximum-likelihood in log-space is derived from the the answer of Question 13:

$$
\begin{aligned}
W &= \underset{W}{\operatorname{argmax}} \int P(Y|W,X)P(X)dX \\
&= \underset{W}{\operatorname{argmax}} P(Y|W) \\
&= \underset{W}{\operatorname{argmax}} \prod_{i=0}^{N} P(y_i|W) \\
&= \underset{W}{\operatorname{argmax}} \prod_{i=0}^{N} N(y_i|0, WW^T + \sigma^2 I) \\
&= \underset{W}{\operatorname{argmax}} \prod_{i=0}^{N} ln(\frac{1}{\sqrt{(2\pi)^D |WW^T + \sigma^2 I|}} e^{y_i^T (WW^T + \sigma^2 I)^{-1} y_i} \\
&= \underset{W}{\operatorname{argmax}} -\frac{ND}{2} ln(2\pi) - \frac{N}{2} ln(|WW^T + \sigma^2 I|) - \frac{1}{2} \sum_{i=0}^{N} (y_i^T (WW^T + \sigma^2 I)^{-1} y_i) \\
&= \underset{W}{\operatorname{argmax}} -\frac{ND}{2} ln(2\pi) - \frac{N}{2} ln(|WW^T + \sigma^2 I|) - \frac{1}{2} tr(Y^T (WW^T + \sigma^2 I)^{-1} Y)
\end{aligned}
$$

If the prior is uniform, then the ML and MAP will be the same. On the other hand, the more data we observe , the weaker the prior belief will be, and our results in MAP will converge more to the results in Maximum-likelihood approach. According to the marginalization rule we know that:

$$\int P(Y|W,X)P(W)dW = P(Y|X)$$

and Based on the Bayesian Rule:

$$P(W|Y,X)P(Y|X) = P(Y|W,X)P(W)$$

we move the term p(Y|X) from left to the right :

$$P(W|Y,X) = \frac{P(Y|W,X)P(W)}{P(Y|X)} = \frac{P(Y|W,X)P(W)}{\int P(Y|W,X)P(W)dW}$$

And since P(Y|X) is independent to W, we can simplify the argmax function :

$$W_{II} = \underset{W}{\operatorname{argmax}} \frac{P(Y|W,X)P(W)}{P(Y|X)} = \frac{P(Y|W,X)P(W)}{\int P(Y|W,X)P(W)dW} = \underset{W}{\operatorname{argmax}} P(Y|W,X)P(W)$$

The reason why the Type-II Maximum-likelihood is a sensible approach to learn the model is that if we have a model with two variables W and X that are dependent to each other, then there will be a problem if we use MAP and ML since we have to use probability of P(Y|X,W). On the other hand, Type-II estimation only take P(Y|W) so that we don't care if X and W are dependent or not, and in representation learning, X is latent variables, meaning that we have no information about X and don't know if X and W are dependent as well. Therefore Type-II ML estimation is a sensible approach to learn the model in representation learning.

Question 15

1. Compute the objective function −log(p(Y|W)) = L(W) for the marginal distribution in Eq. 23.
2. Compute the gradients of the objective with respect to the parameters dLdW

Here we take log in the base of exponential, that is ln, to simplify the equations. From question 14 we already know that:

$$ln(P(Y|W) = -\frac{ND}{2} ln(2\pi) - \frac{N}{2} ln(|WW^T + \sigma^2 I|) - \frac{1}{2} tr(Y^T (WW^T + \sigma^2 I)^{-1} Y)$$

Therefore,

$$-ln(P(Y|W)) = \frac{ND}{2} ln(2\pi) + \frac{N}{2} ln(|WW^T + \sigma^2 I|) + \frac{1}{2} tr(Y^T (WW^T + \sigma^2 I)^{-1} Y) = L(W)$$

Using some general form of gradient [4]:

$$\partial(ln(det|X|)) = tr(X^{-1}\partial X)$$
$$\frac{\partial}{\partial X}Tr(X) = tr(\partial X)$$
$$\frac{\partial Y^{-1}}{\partial x} = -Y^{-1}\frac{\partial Y}{\partial x}Y^{-1}$$
$$\frac{\partial A^T X B)}{\partial X} = AB^T$$

And

$$\frac{\partial (X^T A)_{ij}}{\partial X_{mn}} = (J^{nm}A)_{ij}$$

The gradient of L(W) becomes:

$$\frac{\partial L(W)}{\partial w_{ij}} = \frac{N}{2}tr((WW^T+\sigma^2 I)^{-1}(J_{ij}W^T+WJ_{ij}^T)) - \frac{1}{2}tr(YY^T-(WW^T+\sigma^2 I)^{-1}(J_{ij}W^T+WJ_{ij}^T)(WW^T+\sigma^2 I)^{-1})$$

2.2 Practical

Question 16

1. Plot the representation that you have learned (hint: plot X as a two-dimensional representation).
2. Explain the outcome and discuss key features, elaborate on any invariance you observe. Did you expect this result?
3. How is the effect of representation learning dependent on the number of available samples? Please test lower values of N and discuss the observed implications.

Figure 19 is the real X we generate by $f_{non_lin}(x_i)$ and figure **??** is the learned representation of X,Let's say,X'. The learned representation X is calculated with generated Y and A' we optimized using gradient descent.

X' is calculated from:

$$Y = X'A^T$$
$$YA = X'A^TA$$
$$YA(A^TA)^{-1} = X'$$

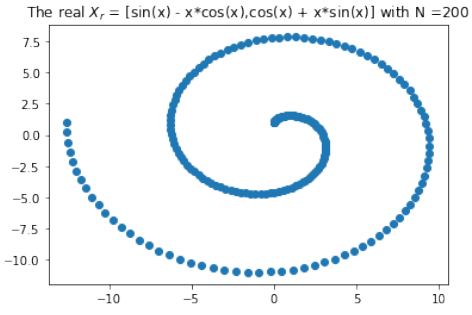

Figure 19: The real X

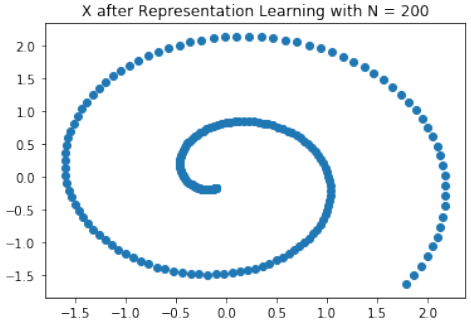

Figure 20: Representation of X

From figure 19 and 20 we notice that the shape of two are similar, but one is the rotation shape of the other. Let's say W' = WR, where R is the rotation matrix. From question 13 we know that, the marginalization likelihood is the following form:

$$P(Y|W) = \prod_{i=0}^{N} N(y_i|0, WW^T + \sigma^2 I)$$

[4]Petersen and Pedersen: Matrix cookbook,2012 edition

For W' we can write the maximum likelihood in the same form:

$$
\begin{aligned}
P(Y|W) &= \prod_{i=0}^{N} N(y_i|0, WR(WR)^T + \sigma^2 I) \\
&= \prod_{i=0}^{N} N(y_i|0, WRR^T W^T + \sigma^2 I) \\
&= \prod_{i=0}^{N} N(y_i|0, WW^T + \sigma^2 I)
\end{aligned}
$$

Therefore we know that the maximum likelihood is invariant of the rotation R, so all possible W' = WR will be the solution and this also explain the reason why the difference of direction between real X and the learned X.

Figure 21,22,23,24,25,26 show the results of difference number of samples with same A matrix we generated Y. From these figures, we found out that the smaller number of sample we test, the smaller the shape of learned representation X is.

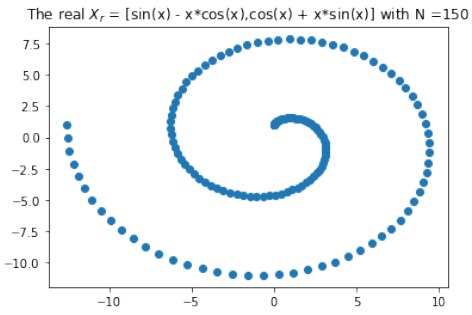

Figure 21: The real X with N=150

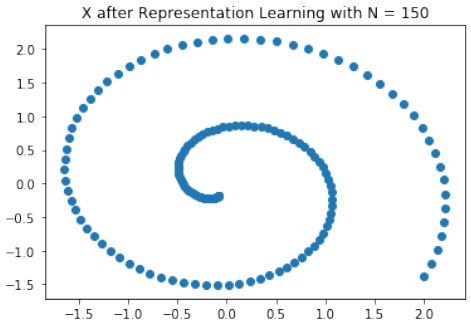

Figure 22: Representation of X N=150

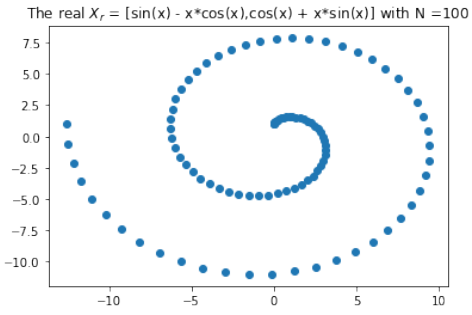

Figure 23: The real X with N=100

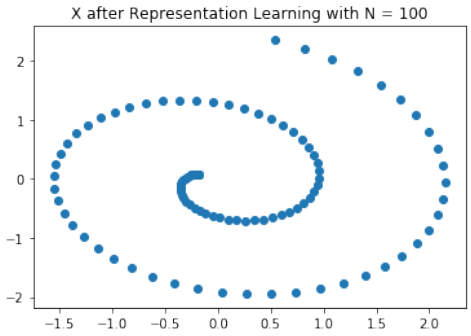

Figure 24: Representation of X N=100

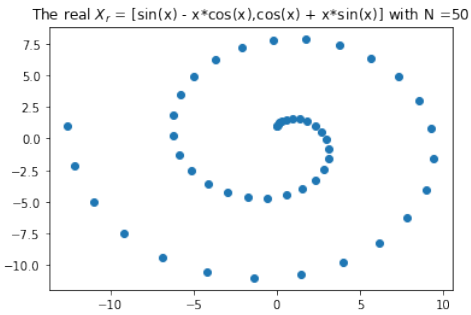

Figure 25: The real X with N=50

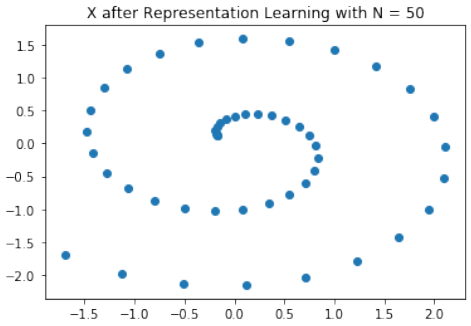

Figure 26: Representation of X N=50

3 The evidence $p(D)$

3.0.1 Models

> ***Question 17*** Why is this the simplest model, and what does it actually imply? What makes it a bad model on the one hand, and a good model on the other hand?

According to Murray and Ghahramani, this simplest model M_0 simply assigns a single distribution over the whole data set without giving any weight on any sample as it has no free parameters. The disadvantage of using this model is that putting a uniform prior yields a different supposedly 'assumption free' model from this simple model, which means that the prior contains no useful information about this dataset.

There are some advantage to use this simplest model. In the paper [5], if the decision boundary is non-linear, then the simplest model is the most likely model among three other models. Also, if the prior is over a large range, then it will have much computational cost. Simpler model can save some computational cost.

> ***Question 18*** Explain how each separate model works. In what way is this model more or less flexible compared to M0? How does this model spread its probability mass over D?

Given θ_1^1, the model try to find a boundary that separate the data t_i in dataset. On side of the boundary has $t_i = 1$ and the other side of the boundary has t_i = -1. If t_i is in the right side with right label, then p($t_i|M_1$,θ_1) will be larger than those t_i who are in the wrong side.

Here, the model only consider the one dimension of x (that is, x_1) but not the other dimension x_2. Compare to Model M_0, this model will give higher probability mass for those sub data set that the boundary is the function of x_1, like dataset (b) in the paper.

> ***Question 19*** Discuss and compare the models. In particular, please address the following questions in your discussion:
>
> - How have the choices we made above restricted the distribution of the model?
> - What datasets is each model suited to model? What does this actually imply in terms of uncertainty?
> - In what way are the different models more flexible and in what way are they more restrictive?

Model M_3 is a standard logistic regression and so is model M_2 but without the bias. The dataset whose boundary does not pass the origin will have high probability in model M_3 because in model M_3 there is a bias term θ_3^3 that allows the boundary to be offset from the origin.

Now we compare model M_1 and model M_2, model M_1 does not take dimension x_2 into account, so model M_1 can not deal with the boundary that is the function of x_2. At the same time, model M_2 can deal with this kind of boundary and if one set of observation is just a rotation of the other, same probability will be given by model M_2 due to the rotation invariance in model M_2. For observations who is not well modeled by a sharp boundary, model M_0 will be a good choice.

Model M_3 is the most complex model because it contains most parameters. It can realize the other three models by setting some parameters equal 0, so model M_3 is the most flexible among the four. However, If we have observation that can be described in a simpler model, then model M_3 will not be a good choice and can be restrictive because it spreads the bulk of its unit probability mass over a wider range of dataset than the important range of this observation should be.

3.0.2 Evidence

> ***Question 20*** Explain the process of marginalisation and briefly discuss its implications in the given context of the model evidence.

Here, we marginalize the model parameters θ by integrating out all parameters. In order to find the evidence of the model, we have to remove the dependency of model parameters. Without marginalizing, We can only get the distribution of probability mass with this model and these specific parameters, but this won't tell us any information about how suitable this 'type' of model is with the datasets but how suitable this model with parameters is. Therefore, we marginalize out the model parameters θ, only the model in the distribution, then we can compare how the probability mass spreads among different 'type' of models.

[5] A note on the evidence and Bayesian Occam's razor, Murray and Ghahramani, 2005

Question 21 What does this choice of prior imply? How does the choice of the parameters of the prior μ and Σ affect the model?

Here, we choose a large variance Σ and 0 as mean in a Gaussian Distribution for the prior. This is due to the reason that we would like to be uncertain and allow for a large range of parameters. First of all, Σ is a diagonal matrix, which means that the parameters in a model will be independent to each other. Secondly, If Σ is small, meaning that we are more confident that the parameters are around the mean, which is not true since we want include more possible parameters of this model so that the model will be capable to find sharp decision boundary to datasets. On the other hand, the mean μ is chosen to be 0 because the parameters are neither necessary to be positive nor necessary to be negative, and we should give equal probability for them, so choosing 0 as μ is reasonable.

3.1 Practical

Question 22 Plot the evidence over the whole dataset for each model (and sum the evidence for the whole of D, explain the numbers you get). The x-axis index the different instances in D and each models evidence is on the y-axis. How do you interpret this? Relate this to the parametrisation of each model.

The sum of the evidence for the whole D is 1 for every model. This is reasonable and the same as what we expected. We sum up the evidence of a model in all possible dataset, and the probability spreads in the whole possible dataset D, that is 512 dataset we generate. By definition, the summation for the probability of D is 1, and here the sum of the evidence for the whole D equals to 1 is expected and we get 1 as well.

Figure 27 is the plot of evidence over the whole dataset for each model, the evidence of dataset are sorted based on the algorithm provided in the paper to order. Figure 28 is the details for the evidence plot.

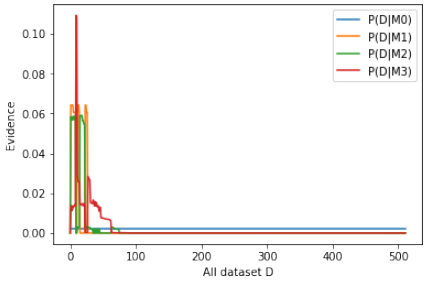

Figure 27: Evidence for all data sets for the models

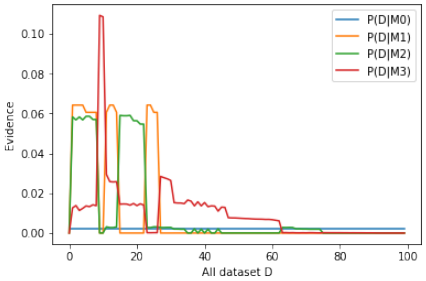

Figure 28: Details for the evidence plot

When D is larger than 100, mostly P(D|M_0) is the highest evidence among the four, this makes sense because in M_0 the probability is equally distributed while in other models, the probability is higher in D $<$ 100. The datasets in these regions should have decision boundaries that are nonlinear, so another models can not well model them.

If a dataset has a very unequal distribution of +1 and -1, for example, dataset(a) in Murray and Ghahramani's paper. Model M_3 is the only model to has a bia to account for it, so the pick of the red curve should be this kind of dataset. On the other hand, the area that both Model M_1 and M_2 has similar distribution will be those dataset that has a sharp linear boundary which is a function of x_1. At the same time, the area that the evidence from M_1 is quite low but M_2 maintains similar evidence will be the dataset that are the rotation of those datasets because model M_1 cannot model decision boundaries with the rotation while model M_2 can. For the region $D \geq 25$ and $D \leq 60$, the evidence of model M_3 is favored over another models, these models' decision boundaries might be offset from the origin because all other models has no bias term.

Question 23 Find using np.argmax and np.argmin which part of the D that is given most and least probability mass by each model. Plot the datasets which are given the highest and lowest evidence for each model. Discuss these results, do the findings make sense?

Figure 29 is the least suitable dataset in model M_1 and figure 30 is the most suitable dataset in model M_1. It makes sense for model M_1 to have a highest evidence of this dataset that the decision boundary is a sharp linear boundary in a function of x_1, as we can see in figure 30. And it also make sense that for the least evidence of dataset, its decision boundary offsets from the origin and model M_3 should have a high evidence in such kind of dataset.

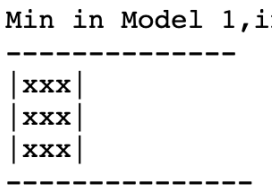

Figure 29: least suitable dataset in M_1

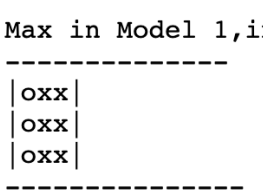

Figure 30: Most suitable dataset in M_1

Figure 31 is the least suitable dataset in model M_2 and figure 32 is the most suitable dataset in model M_2. In figure 31 we can see that the decision boundary is not linear, so it makes sense that the evidence of this dataset with model M_2 is low. From figure 27 we know that this dataset has the highest evidence in model M_0, as our observation described in question 22. In figure 32 we can see that the decision boundary is a function of x_1 and x_2 and cross the origin, so it makes sense as well in this case.

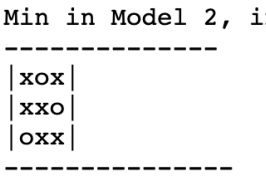

Figure 31: least suitable dataset in M_2

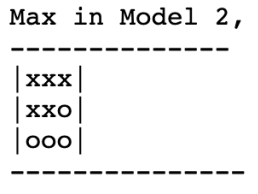

Figure 32: Most suitable dataset in M_2

Figure 33 is the least suitable dataset in model M_3 and figure 34 is the most suitable dataset in model M_3. In figure 33 we can see that the decision boundary is not linear as well, so it makes sense that the evidence of this dataset with model M_3 is low. Model M_3 has a bias term, so the datasets that their decision boundary offsets from the origin should have a high evidence in model M_3, and in figure 34 the decision boundary shifts so it makes sense.

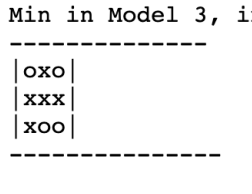

Figure 33: least suitable dataset in M_3

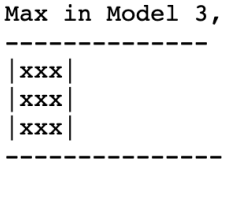

Figure 34: Most suitable dataset in M_3

Question 24 What is the effect of the prior $p(\theta)$

- What happens if we change its parameters?
- What happens if we use a non-diagonal covariance matrix for the prior?
- Alter the prior to have a non-zero mean, such that $\mu = [5, 5]^T$?
- Redo evidence plot for these and explain the changes compared to using zero-mean.

If our variance of the prior is larger, which means that we will get larger setting of parameters θ, then the decision boundary in evidence plot will become sharper. As we can see in figure 35 and figure 36, compare to figure 27 and figure 28, for each dataset, there is no other evidences over different models that have as high probability as the highest evidence over one model, only one evidence in one model is much higher than that of other models in that region. At the same time, if our variance of the prior is smaller, then we will get a much uniform evidence, as we can see in figure 37 and figure 37 where σ^2 is 1.

If we use a non-diagonal covariance matrix for the prior, then the parameters will be dependent, then our evidence will be more or less depend on the value and correlation in the parameters.(figure 39 and figure 40)

If the prior has a non-zero mean, then some specific dataset will have higher evidence while other dataset will have lower evidence. As we can see in figure 41 and figure 42, there are more peaks in the plot, meaning that the probability mass of the models give more weight on some specific datasets.

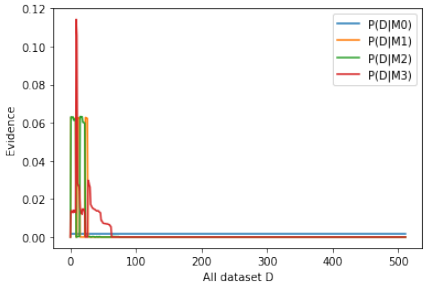

Figure 35: Evidence for datasets with higher σ

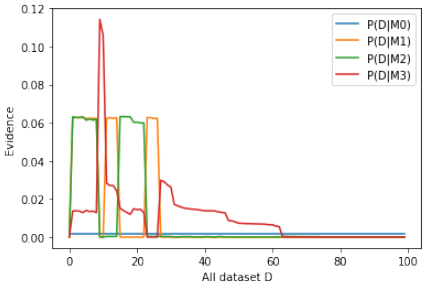

Figure 36: Details for the plot with higher σ

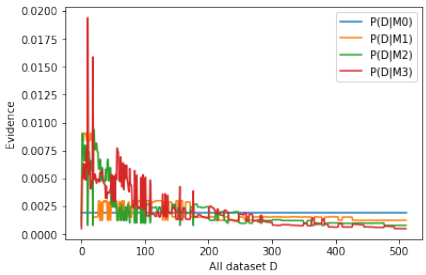

Figure 37: Evidence for datasets with low σ

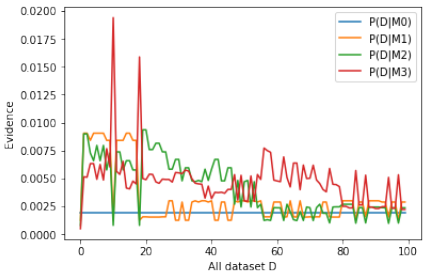

Figure 38: Details for the plot with low σ

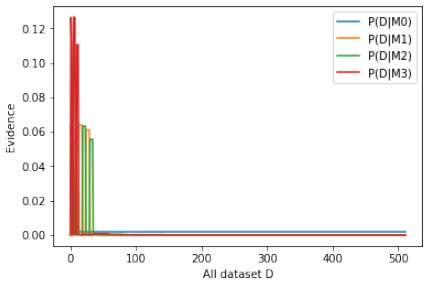

Figure 39: Evidence with non-diagonal cov.

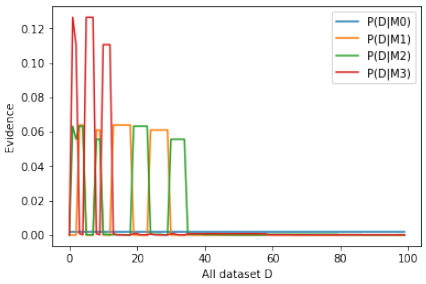

Figure 40: Details with non-diagonal cov.

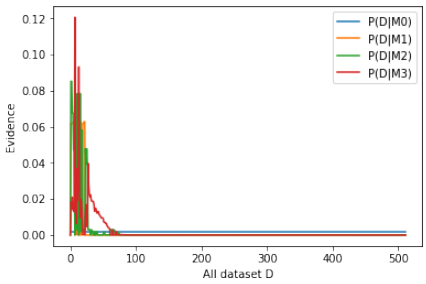

Figure 41: Evidence for datasets with μ=5

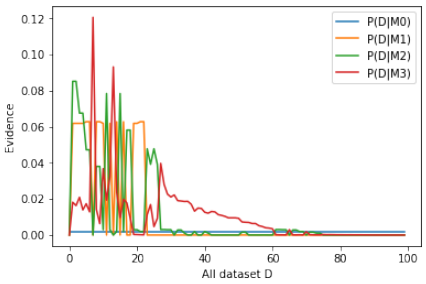

Figure 42: Details for the plot with μ=5