Assignment One: Modelling

Anjie Song	**Xiang He**	**Xing Li**
1705056	1723303	1727134
as17472	xh17500	xl17507

Names not listed in order

Question 1

1)In our first task, using a Gaussian likelihood distribution seems to be an empirical assumption for the regression problem. To explain this, from our opinions, all x_i and y_i are observed in regression problem. As a result, the mean μ and variance σ^2 of observed data are known. In the situation where we actually have no information about what the likelihood should be, we should choose the distribution with a maximum entropy to represent our zero beliefs about it and after counting the μ and σ^2 in, the Gaussian distribution will be our best suitable likelihood distribution[1]. We assume the number of data point tends to infinite.

2)Choosing a spherical covariance matrix for the likelihood means two assumptions. The first is we assume that observed data points are independent of each other(elements not in diagonal of matrix $\sigma^2\mathbf{I}$ are all 0). The second assumption is the distribution over each data point is isotropic or they have the same variance of Gaussian distribution, which is caused by adding the identical Gaussian noise to each independent observed data point.

Question 2

If $\mathbf{Y}$ are dependent with each others, then we have

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{W}) = \mathcal{N}(\mathbf{Y} \mid \boldsymbol{f}, \Sigma)$$

where $\boldsymbol{f}$ is the vector of f and Σ is the covariance of $\mathbf{Y}$.

Question 3

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{W}) = \prod_{i}^{N} \mathcal{N}(\mathbf{y}_i \mid \mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I})$$

where we assume $\mathbf{Y}$ has indenpendent distribution.

Question 4

1)Conjugate could be found in many science and technology areas. Generally, conjugate means two things could be seen as a group in which has a corresponding relationship to a certain extent. Specifically for conjugate distributions in Bayesian statistics, it means the prior distributions and posterior distributions are in the same family, the prior and posterior are then called conjugate distributions.

2)A conjugate prior could be an algebraic convenience, giving a closed-form expression for the posterior; otherwise, numerical integration may be necessary.

Another reason, maybe more important, is conjugate priors may give intuition, by more transparently showing how a likelihood function updates a prior distribution. From question One, we have made our likelihood as a Gaussian distribution, it is natural to choose a conjugate prior for Gaussian likelihood distribution, which is still a Gaussian distribution.

Question 5

Our general regularizer is like the following:

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

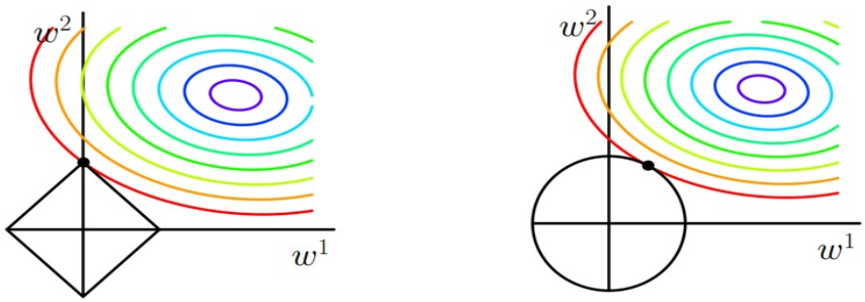
where q denotes our "preference", $q = 2$ corresponds to the quadratic regularizer and $q = 1$ corresponds to the linear regularizer. For convenience of talk, we simplify the regularizer to

$$J_0 + \alpha\sum|\mathbf{w}|^q$$

where J_0 is the original loss function and α is regularization parameter.

The case of q = 1 is known as *lasso* in the statistics literature[3]. Notice that L1 regularization is the sum of absolute value of parameters **W**; therefore, it is incomplete differential. Our task is to find the minimum solution for J_0 restricted by $\alpha\sum|\mathbf{w}|^q$. Considering the two dimensional situation, we have only two weight parameters w1 and w2. Through the process of gradient descent method, we could draw the contour line for J_0 and regularization term in figure 1 (a). It is clear to see that regularization term still give a high penalty when some weight parameter is very close to 0. As a result, in this extreme position and if α is sufficiently large, many coefficients would be zero in high dimensional space and this leads to a sparse model in which the corresponding basis functions play no role.

In contrast, the case of q = 2 gives a gradual penalty so that model would not be too sparse which is sensible in practice. And due to its quadratic form, the absolute value sign disappears, this gives us great mathematical convenience, like applying derivation.



(a) L1 regularization (b) L2 regularization

Figure 1: Comparison between two regularizationss

Question 6

So far, we have the expressions of the likelihood and the prior

$$\begin{aligned} p(\mathbf{Y} \mid \mathbf{X}, \mathbf{W}) &= \prod_i^N \mathcal{N}(\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I}) \\ &= \frac{1}{(2\pi)^{\frac{DN}{2}}}\frac{1}{|\sigma^2\mathbf{I}|^{\frac{N}{2}}}e^{-\frac{1}{2\theta^2}\sum_i^N(\mathbf{y}_i-\mathbf{W}\mathbf{x}_i)^T(\mathbf{y}_i-\mathbf{W}\mathbf{x}_i)} \\ p(\mathbf{W}) &= \mathcal{N}(\mathbf{W}_0, \tau^2\mathbf{I}) \end{aligned} \tag{1}$$

and thus

$$\begin{aligned} p(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}) &= \frac{1}{Z}p(\mathbf{Y} \mid \mathbf{X}, \mathbf{W})p(\mathbf{W}) \\ &= \frac{1}{Z}\frac{1}{(2\pi)^{\frac{D(N+1)}{2}}}\frac{1}{|\sigma^2\mathbf{I}|^{\frac{N}{2}}}\frac{1}{|\tau^2\mathbf{I}|^{\frac{N}{2}}}e^{-\frac{1}{2\theta^2}\sum_i^N(\mathbf{y}_i-\mathbf{W}\mathbf{x}_i)^T(\mathbf{y}_i-\mathbf{W}\mathbf{x}_i)-\frac{1}{2\tau^2}(\mathbf{W}-\mathbf{W}_0)^T(\mathbf{W}-\mathbf{W}_0)} \end{aligned} \tag{2}$$

Question 7

Non-parametric models assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they could often be defined by assuming an infinite dimensional space. Usually, we think of the space as a function space. Non-parametric does not mean there are no parameters in models, but a number of parameters in the models could vary according the amount of data change. Of course, there are finite hyperparameters which help estimate model parameters.

Parametric model could capture all its information about the data within its parameters. These parameters are the only thing it needs to predict a new data. However, the non-parametric model need not only the parameters are hidden in the infinite dimensional space but also the current state of data that has been observed.

Parametric model encodes a stronger assumption in prior. When the assumptions are correct, parametric models will produce more accurate and precise estimates than non-parametric models, However, as more is assumed by parametric models, when the assumptions are not correct they have a greater chance of failing while non-parametric models may be robust and flexible enough to learn more subtle information from data. On the other hand, parametric formulae are often simpler to write down and faster to compute compared with non-parametric model.

Finally, due to the clear parameters set and the explicit assumption, the parametric models may be more interpretable than the non-parametric models.

Question 8

The prior presents the joint distribution of the mapping functions $\boldsymbol{f}$ are Gaussian with zero means and covarience $k(\mathbf{X}, \mathbf{X})$. And the covariance between f_i and f_j has negative correlation with the distance between $\mathbf{x}_i$ and $\mathbf{x}_i$.

Question 9

Yes, we tend to think this prior encode all possible functions because in a Gaussian distribution, any values from $[-\infty, +\infty]$ could be got even though with a tiny probability. In each point of the domain of X, we have a Gaussian distribution over Y from $-\infty$ to $+\infty$, therefore, each point in any function could be got. It means any function is encoded in our prior.

Question 10

According to the product rule,

$$p(\mathbf{Y}, \mathbf{X}, f, \boldsymbol{\theta}) = p(\mathbf{Y} \mid f)p(f \mid \mathbf{X}, \boldsymbol{\theta})p(\mathbf{X})p(\boldsymbol{\theta})$$

where $p(\mathbf{X})$ and $p(\boldsymbol{\theta})$ are given. Figure 2 shows the graphic model.

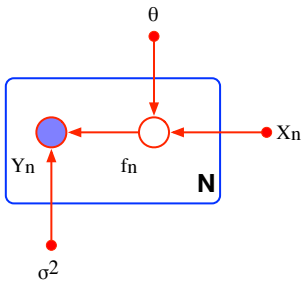


Figure 2: The graphic model of the joint distribution

Question 11

1. Through the marginalisation and multiplying the prior, the prior f and the data $\mathbf{X}$ are encoded into this equation.
2. The prior is our assumption and also the uncertainty. The likelihood is the knowledge which we have some certainty. By multiplying these two probability, the result will be more cercain than the

prior and this filters out some uncertainty from the prior.
3. Because θ is the parameter which controls the kernal function and influences our prior. θ in the left-side means this parameter will also control the distribution of this marginalisation.

Question 12

For the prior of parameters, we chose the alpha which controls the covariance as 2.0. And the mean is 0. Now we can put data points into our prior and compute the posteriors. The process is shown in Figure 3. The first line is the prior and the samples of functions. And then the next three rows represent the posterior of parameters when we put 1, 50 and 150 data points into our model respectively. We did 6 samples from the distributions and drew the functions for every posterior.
When we add more data points into our model, the distribution will be increasingly precise to some

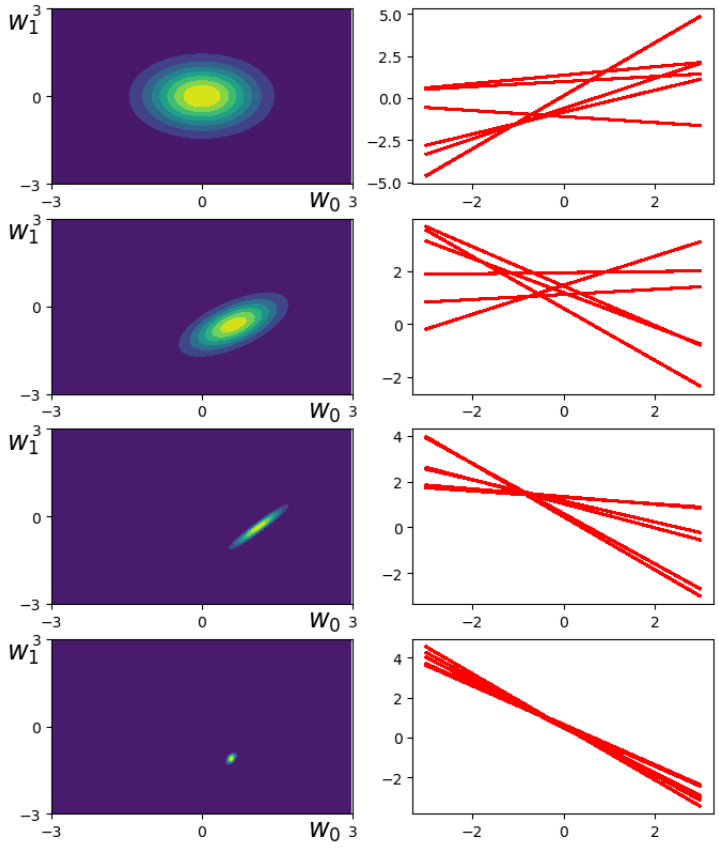


Figure 3: Process of linear regression

point and the samples can be more concentrated. Obviously, the posterior distribution of parameters was closing to the real values due to our train from data.

Question 13

For the prior of GP, we chose the parameters sigma and length-scale as 1.0 and 1.0 respectively. Then the prior distribution can be computed. We sample some functions from the prior, which is shown in Figure 4.

When we change the length-scale to 0.5 and 5.0, the distributions of prior can be very different, which is shown in Figure 5. This behavior can be explained from the squared exponential kernel function. As we alter the length-scale of this function, the uncertainty between data points will change. Therefore, the parameter length-scale just encode the scale of uncertainty into the prior.

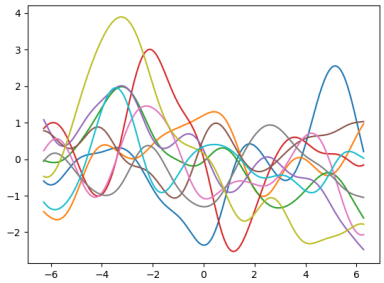
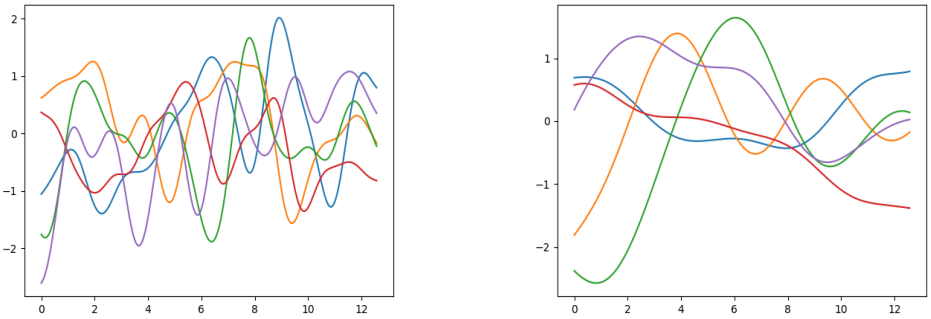


Figure 4: Sample from the prior

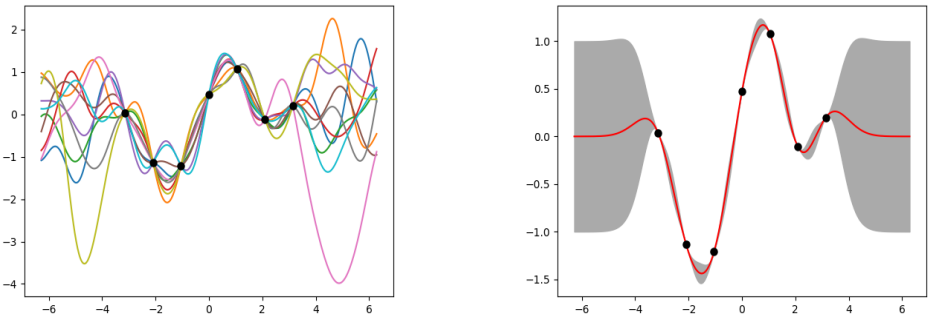


(a) Length-scale = 0.5 (b) Length-scale = 5.0

Figure 5: Change the length-scale of prior distribution

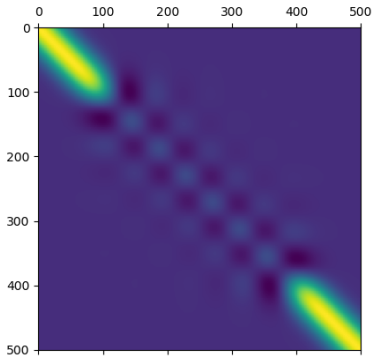
Question 14

The samples of posterior distribution and the predictive mean and predictive variance are shown in Figure 6. We also plot the 2D and 3D graphics for the predictive covariance, which is shown in Figure 7. From these, we can see that the posterior get more and more precise with more data put into the models. And this is exactly the behavior we desire.
When we take the noise into consideration(add the diagonal matrix), the posterior won't be so precise as before. And the uncertainty around the train data is still not small enough, which is shown in Figure 8

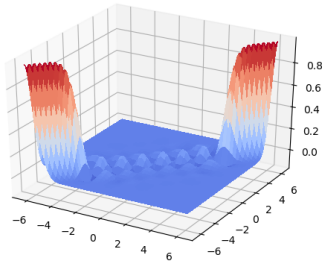


(a) Samples of posterior (b) Predictive mean and variance

Figure 6: Samples of posterior distribution and the predictive mean, variance

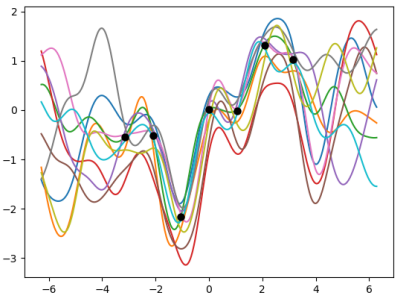


(a) 2D

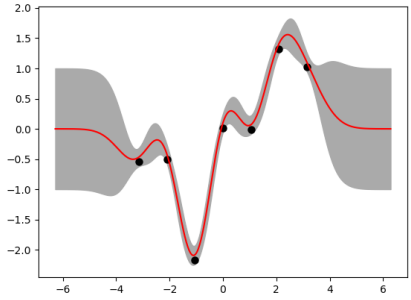


(b) 3D

Figure 7: Samples of posterior distribution and the predictive mean, variance



(a) Samples of posterior



(b) Predictive mean and variance

Figure 8: Consider the noise

Question 15

Assumption is a representation of our beliefs, experiences and suppositions over the model or some specific parameters before seeing any current data sets. In some situations, we may have several possible assumptions all could be acceptable. Now we have to make a decision to choose one assumption due to our interests or preferences. For example, in Probabilistic PCA, we have two waiting estimated value sets $\mathbf{W}$ and $\mathbf{X}$. We can either give assumption over $\mathbf{W}$ or equally $\mathbf{X}$. However, our interest or preference is to estimate $\mathbf{W}$;therefore, we decide firstly give the assumption, Specifically a distribution, over $\mathbf{X}$ so that we could marginalize $\mathbf{X}$ out and get the relationship between observed data and $\mathbf{W}$.

Question 16

In this preference, we encode the distribution of $\mathbf{X}$ as an spherical Gaussian distribution with zero means which means each element in $\mathbf{X}$ is independent.

Question 17

In the beginning, we assume $p(\mathbf{Y})$ is a independent distributed amoung$p(\mathbf{y}_i)$, and the linear relashionship amoung $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{W}$.

$$\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \boldsymbol{\mu}$$

where $\mathbf{y}_i$ is the data in the dataset $\mathbf{Y}$ and $\boldsymbol{\mu}$ is the parameter which controls the predict value of $\mathbf{y}_i$. Then we have

$$\begin{aligned} p(\mathbf{Y} \mid \mathbf{W}) &= \prod_i^N p(\mathbf{y}_i \mid \mathbf{W}) \\ &= \prod_i^N \int p(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{W}) p(\mathbf{x}_i) \mathrm{d}x \end{aligned} \tag{3}$$

where

$$p(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i \mid \mathbf{0}, \mathbf{I})$$
$$p(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{W}) = \mathcal{N}(\mathbf{y}_i \mid \mathbf{W}\mathbf{x}_i + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

According to the Eq.2.115 [1], we can state the marginalisation as

$$\begin{aligned} p(\mathbf{Y} \mid \mathbf{W}) &= \prod_i^N \mathcal{N}(\mathbf{y}_i \mid \mathbf{W} \times 0 + \boldsymbol{\mu}, \sigma^2\mathbf{I} + \mathbf{W}\mathbf{I}\mathbf{W}^T) \\ &= \prod_i^N \mathcal{N}(\mathbf{y}_i \mid \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \end{aligned} \tag{4}$$

Question 18

$$MAP : \hat{\mathbf{W}} = argmax_{\mathbf{W}} \frac{p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})}{\int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})d\mathbf{W}} = argmax_{\mathbf{W}} p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})$$

$$ML : \hat{\mathbf{W}} = argmax_{\mathbf{W}} \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{X})d\mathbf{X}$$

1) In MAP, we encode our prior with the likelihood and generate the posterior. The maximum of posterior is effected by both likelihood and prior. The ML only focus on the likelihood by marginalize data X out.

2) When the data is few, the MAP may by effected by our prior a lot when the ML is fully dependent on data set X, hence it cause over-fitting problem. With the growth of the amount of observed data, the effect from prior in MAP and over-fitting problem in ML could be weakened due to the growth of power of data.

3) The denominator marginalizes the W out, so the $argmax_{\mathbf{W}}$ only relates to the numerator, this is just the right side term. They are equal.

Question 19

(i) Now, we have derived the marginalisation $p(\mathbf{Y} \mid \mathbf{W})$ which we know the means and the covarience. The Eq.2.118 gives the general solution of the $log|\cdot|$ for a joint Gaussian distribution. Then, we can easily state that

$$-\log p(\mathbf{Y} \mid \mathbf{W}) = \frac{ND}{2}\log(2\pi) + \frac{N}{2}\log|\boldsymbol{C}(\mathbf{W})| + \frac{1}{2}\sum_{n=1}^{N} \mathbf{y}_n^T(\boldsymbol{C}(\mathbf{W}))^{-1}\mathbf{y}_n \tag{5}$$

where D is the dimension of $\mathbf{y}_i$

$$\boldsymbol{C}(\mathbf{W}) = \mathbf{W}\mathbf{W}^T + \sigma^2\boldsymbol{I}$$

(ii) In order to obtain the $\mathbf{W}$ which minimizes the $\mathcal{L}(\mathbf{W})$ ($\mathrm{argmin}_{\mathbf{W}}\, \mathcal{L}(\mathbf{W})$), we need to derive the gradient of the $\mathcal{L}(\mathbf{W})$ over $\mathbf{W}$ and make it be zero.

$$\frac{\delta\mathcal{L}(\mathbf{W})}{\delta\mathbf{W}_{ij}} = \frac{N}{2}\frac{\delta}{\delta\mathbf{W}_{ij}}\log|\boldsymbol{C}(\mathbf{W})| + \frac{1}{2}\frac{\delta}{\delta\mathbf{W}_{ij}}\sum_{n=1}^{N} \mathbf{y}_n^T(\boldsymbol{C}(\mathbf{W}))^{-1}\mathbf{y}_n \tag{6}$$

Firstly, we focus on the $\frac{\delta}{\delta \mathbf{W}_{ij}} \log|\boldsymbol{C}(\mathbf{W})|$. According to the equation Eq.43 in *The Matrix Cookbook* [2], we have

$$\frac{\delta}{\delta \mathbf{W}_{ij}} \log|\boldsymbol{C}(\mathbf{W})| = \text{Tr}\left(\boldsymbol{C}(\mathbf{W})^{-1}\frac{\delta \boldsymbol{C}(\mathbf{W})}{\delta \mathbf{W}_{ij}}\right)$$

Now we derive the part from the previous form to the form which calculates the difference of $\boldsymbol{C}(\mathbf{W})$ over $\mathbf{W}$. We will derive this later and move to the second part of Eq.6.
The

$$\sum_{n=1}^{N} \mathbf{y}_n^T(\boldsymbol{C}(\mathbf{W}))^{-1}\mathbf{y}_n$$

can be rewriten as

$$\text{Tr}(\mathbf{Y}^T(\boldsymbol{C}(\mathbf{W}))^{-1}\mathbf{Y})$$

where $\mathbf{Y}$ is the matrix of $\mathbf{y}_n$. Using the Eq.36[2], Eq.40[2], Eq.71[2] and the chain rule, we have

$$\begin{aligned}
\frac{\delta}{\delta \mathbf{W}_{ij}}\text{Tr}(\mathbf{Y}^T(\boldsymbol{C}(\mathbf{W}))^{-1}\mathbf{Y}) &= \text{Tr}\left(\frac{\delta}{\delta \mathbf{W}_{ij}}(\mathbf{Y}^T(\boldsymbol{C}(\mathbf{W}))^{-1}\mathbf{Y})\right) \\
&= \text{Tr}\left(\frac{\delta}{\delta \boldsymbol{C}(\mathbf{W})^{-1}}(\mathbf{Y}^T(\boldsymbol{C}(\mathbf{W}))^{-1}\mathbf{Y})\frac{\delta(\boldsymbol{C}(\mathbf{W}))^{-1}}{\delta \mathbf{W}_{ij}}\right) \\
&= \text{Tr}\left(\mathbf{Y}^T\mathbf{Y}\frac{\delta(\boldsymbol{C}(\mathbf{W}))^{-1}}{\delta \mathbf{W}_{ij}}\right) \\
&= \text{Tr}\left(\mathbf{Y}^T\mathbf{Y}\left(-\boldsymbol{C}(\mathbf{W})^{-1}\frac{\delta \boldsymbol{C}(\mathbf{W})}{\delta \mathbf{W}_{ij}}\boldsymbol{C}(\mathbf{W})^{-1}\right)\right)
\end{aligned} \tag{7}$$

So far, we have derived the two parts to the form of $\frac{\delta \boldsymbol{C}(\mathbf{W})}{\delta \mathbf{W}_{ij}}$, and now we start to derive this difference.

$$\frac{\delta \boldsymbol{C}(\mathbf{W})}{\delta \mathbf{W}_{ij}} = \frac{\delta \mathbf{W}\mathbf{W}^T}{\delta \mathbf{W}_{ij}} = \mathbf{W}\frac{\delta \mathbf{W}^T}{\delta \mathbf{W}_{ij}} + \frac{\delta \mathbf{W}}{\delta \mathbf{W}_{ij}}\mathbf{W}^T = \mathbf{W}\mathbf{J}_{ji} + \mathbf{J}_{ij}\mathbf{W}^T$$

where $\mathbf{J}_{ij}$ is a matrix whose all entries are zero except for $(\mathbf{J}_{ij})_{ij} = 1$ and $\mathbf{J}_{ji} = \mathbf{J}_{ij}^T$.
Combine the whole things, we can state

$$\frac{\delta \mathcal{L}(\mathbf{W})}{\delta \mathbf{W}_{ij}} = \frac{N}{2}\text{Tr}\left(\boldsymbol{C}(\mathbf{W})^{-1}(\mathbf{W}\mathbf{J}_{ji} + \mathbf{J}_{ij}\mathbf{W}^T)\right) + \frac{1}{2}\text{Tr}\left(\mathbf{Y}^T\mathbf{Y}\left(-\boldsymbol{C}(\mathbf{W})^{-1}(\mathbf{W}\mathbf{J}_{ji} + \mathbf{J}_{ij}\mathbf{W}^T)\boldsymbol{C}(\mathbf{W})^{-1}\right)\right)$$

Question 20

The non-parametric graphic model is shown in Figure 2. And the parametric one is shown in 9. In non-parametric model, Y is formed by f and f is formed by $\mathbf{X}$ and $\boldsymbol{\theta}$. We can integrate to f easily. In parametric model, there are two seperate parts that need to consider: W and X. When we do marginalization to this kind of likelihood, we need to fix one of them. Normally, it is more difficult to finish it.

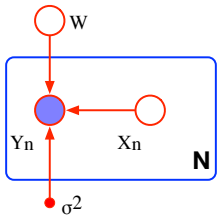
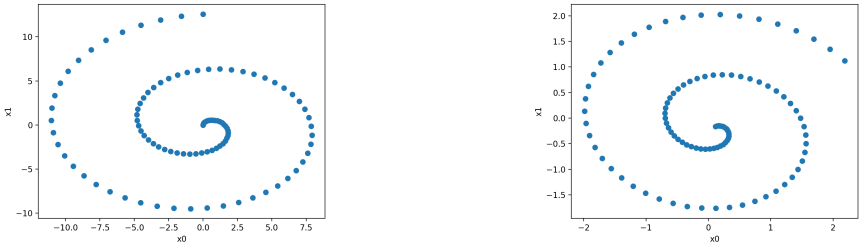


Figure 9: Graphic model of parametric method

Question 21

The result of the recovered data $\mathbf{X}$ is shown in Figure 10b. Comparing with the original data point in Figure 10a, the scale of the recovered data is smaller than the original. This is because we assume the exception value of $\mathbf{Y}$, $\boldsymbol{\mu}$, is zero which leads the result.



(a) The original data points

(b) The recovered data points

Figure 10: The result

Question 22

This model is the most complicated one among these four models. The reason is that simple models always choose to concentrate their probability mass on a limited number of data sets. And complex models predict that data will be drawn from a large range of possibilities. Since this model doesn't concentrate on any data set and give all range of data the same probabilities, it is the most complex one. Parameter counting has no meaning since we can set them to zero.

Question 23

The choice restricted the distribution by the parameters. Model 3 is a standard logistic regression and it is the only model which has the bias parameter, it has a lot more flexibility, therefore it is more suitable for data set which has an unequal distribution or the decision boundaries are offset from the origin. Model 2 is the same as model 3 but without bias parameter. When it comes to the distribution that the decision boundary is near the origin or there is a data point at the origin, model 2 is a better choice since it has no bias. Compared to model 2, model 1 ignores the second dimension. Hence, it is more suitable for the distribution which the decision boundary is only or nearly only related to the first dimension. Model 0 gives every data set the same probability. As a result, while there is no sharp linear boundary can model the data set, uniform model 0 is a reasonable choice. When it comes to uncertainty, every model actually has an assumption to the data. And to some extent, they 'filter' uncertainty of the data sets.

Question 24

As the parameters can affect the modelling process directly, the distribution of parameters is of great importance. However, we have no idea with the parameters. Therefore, a zero-mean and huge-covariance distribution is a simple but reasonable choice to cover more values of parameters. We think that means won't affect the model significantly if the covariance is big enough. But the structure of covariance will change things. For example, it can affect the flexibility of models since the parameters are not dependent on each other.

Question 25

If we sum the evidence for the whole data sets, we can get 1 for every model. It means that these evidence are all regularized.

Question 26

The plot is as shown in Figure 11. It can be explained that model 3 is more focus on the very special part of the data sets since it is the only model with bias parameter. For these data sets, model 3 can get a good performance. Model 2 and Model 1 are very familiar with each other, but somehow model 2 is more flexible due to its two-dimensional parameters. For every data set, model 0 gives the same evidence. Hence it can only be considered when other models are all failed to represent the data.

Question 27

Since every data set in model 0 has the same evidence, we do not talk about model 0 in this part. For model 1, 2 and 3, the data sets which get the maximum and minimum evidence are shown in Figure

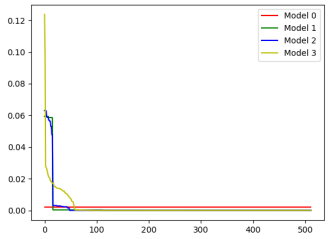


Figure 11: Plot of evidence for all data sets for all models

12. (a) and (e) cannot be solved by a linear model, so model 1 and model 3 failed on it. Actually, model 0 is more suitable for them; (b) is only related to the first dimension, hence model 1 did a very good job on it; (c) and (f) are same(very unequal), so it needs a bias parameter to model it. That's why model 2 failed but model 3 had the highest evidence. (d) can be modelled by a linear through the origin, and (d) could finish it well. Everything does make sense.

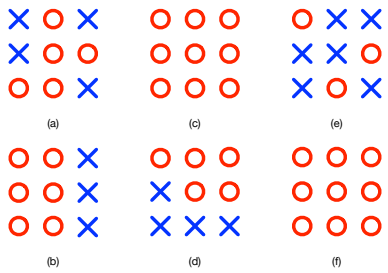
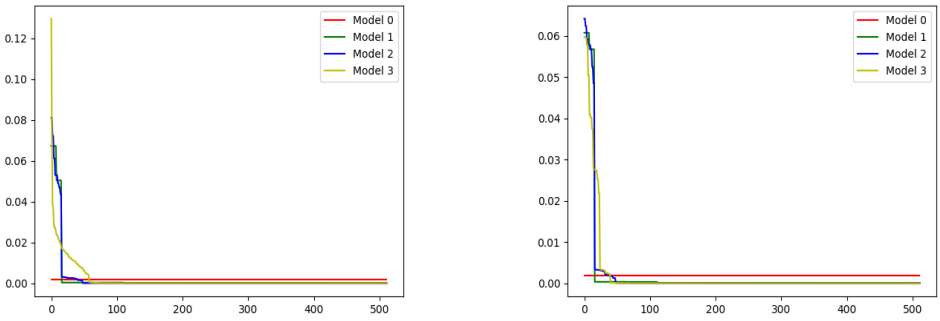


Figure 12: data-sets which are given the highest and lowest evidence for each model. (a), (b) are the data sets of highest and lowest evidence for model 1 respectively. (c), (d) and (e), (f) represent for model 2 and model 3 respectively. Red circle means label -1, Blue cross means label 1.

Question 28

From the plot shown in Figure 13, we can conclude that if the covariance is very big, mean has a limited effect on the models. The highest evidence of model 2 and model 3 increased. However, when the covariance is not big enough, mean changing can affect the models a lot. When it comes to the structure of covariance, things are different. identity matrix means that the parameters are independent. And when the parameters are dependent, flexibility will be affected. That's why model 3 cannot get higher evidence on certain data sets.



(a) Change the mean

(b) Change the covariance

Figure 13: Change the mean and covariance of the parameters distribution

Question 30

Through this project, we take our first step into the Bayesian theory and applications of Gaussian distribution. We have explored many basic concepts in probabilistic machine learning, including assumption, model, prior, likelihood, posterior, Gaussian process and so on. Some interpretation questions lead our group into a lively discuss, some mathematic problems help us build a solid foundation and the rest practical questions provide us a good chance to gain an insight into machine learning. However, there is still a long way to go.

Advanced Topic Question 1

$\mathbf{X}$ is in the kernel $\mathbf{K}$. Therefore, we will compute the gradients with respect to the kernel first and then use the chain rule to get the gradients of X. Then let $\frac{\partial \mathcal{L}}{\partial \mathbf{X}}$ equals 0, we could finally get the solution.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{K}} = \mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T\mathbf{K}^{-1} - \mathbf{K}^{-1}$$

since $\mathbf{K}$ is the linear kernel, then

$$\mathbf{K} = \mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}$$

finallly,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{X} - \mathbf{K}^{-1}\mathbf{X}$$

$$\mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{V}^T$$

where $\mathbf{U}$ is a an $N * q$ matrix, L is a $q * q$ diagonal matrix whose jth element is $(\lambda_j - \frac{1}{\beta})^{-\frac{1}{2}}$ and $\mathbf{V}$ is an arbitrary $q * q$ rotation matrix. 4.

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Version 20121115. Nov. 2012. URL: http://www2.imm.dtu.dk/pubdb/p.php?3274.

[3] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.