Models

Justin Salmon[1] and George Lancaster[2]

[1] *wr18313*
[2] *qv18258*

October 31, 2018

1 The Prior

1.1 Theory

Question 1.1

Choosing a Gaussian likelihood encodes the assumption that the observations of the variates are going to behave like most probabilistic natural processes and are going to contain some inherent noise. The central limit theorem states that in most cases, when independent random variables are sampled the result is a Gaussian distribution. In other words, most probabilistic processes in nature tend to be noisy, and that noise tends to follow a Gaussian distribution. Hence this is generally a good first assumption to make about unknown data.

Question 1.2

Choosing a spherical covariance matrix means that we are assuming that the distribution is equally likely to deviate from the mean in all directions. Additionally, we assume that all dimensions of y are independent, and therefore do not covary with one another. Again, this is a good place to start.

Choosing a non-spherical covariance would imply that we know something in advance about the relationship between the different dimensions of y, which is not true in this case.

Question 2

If we did not assume independence of the data, the covariance matrix would not be in terms of the identity matrix. We would have non-zero values in the offset diagonals which correspond to the correlations between different variables.

1.1.1 Linear Regression

Question 3

The specific form of the likelihood can be written as

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \beta) = \prod_{i=0}^{N} \mathcal{N}(y_i|\mathbf{W}^T\phi(x_i), \beta^{-1}). \qquad (1)$$

Question 4

A distribution is conjugate to another if they both take the same algebraic form, meaning that they are in the same probability distribution family. For example, Gaussians are conjugate to each other, and the conjugate to a Bernoulli distribution is a Beta distribution. Conjugates are used as a convenience to avoid calculating the denominator in Baye's rule (the evidence) which can often be an integral. If the prior and likelihood are conjugate, then their product will be proportional to the posterior.

Question 5

The distance function of a Gaussian distribution represents the dissimilarity of a given value of a random variable (in this case a parameter choice $\mathbf{W}$) from the expected value (the mean). The further away from the mean the parameter choice is, the less likely we think it will be true.

In the 2 dimensional case, it is easy to think about this geometrically as the distance between two points on a Euclidean plane. For a spherical covariance matrix, the distance is going to be the squared Euclidean distance from the mean $||x - \mu||$. This is because the variables are independent and we have picked the distribution such that it has unit variance. If those two conditions were not true, the Gaussian would not be spherical and the distribution would be deformed. The spherical covariance case can be seen as a special case where the distance function is Euclidean. In general the distance between two points in a distribution is called the Mahalanobis distance.

Question 6

We have data generated from a linear model, with added Gaussian noise using the following mapping.

$$y_i = \mathbf{W}x_i + \epsilon$$

The added noise is in the form of a gaussian distribution.

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

The likelihood has been specified using,

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{X}) = \mathcal{N}(\mathbf{WX}, \sigma^2 I)$$

and we have specified a prior distribution.

$$p(\mathbf{W}) = \mathcal{N}(0, \Sigma)$$

To avoid calculating the evidence, we use a posterior distribution that is conjugate to the likelihood. Gaussian distributions are conjugate to themselves, so we use a gaussian for the posterior. The posterior distribution is proportional to the likelihood times the prior.

$$p(\mathbf{W}|\mathbf{Y}, \mathbf{X}) \propto p(\mathbf{Y}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

By writing the exponent of the likelihood times the prior, we can split the exponent into three pieces.

$$= \underbrace{\frac{1}{2\sigma^2}\mathbf{Y}^T\mathbf{Y}}_{\text{A}} + \underbrace{\frac{1}{\sigma^2}\mathbf{Y}^T(\mathbf{XW})}_{\text{B}} - \underbrace{\frac{1}{2\sigma^2}(\mathbf{XW})^T(\mathbf{XW}) - \frac{1}{2}\mathbf{W}^T\Sigma^{-1}\mathbf{W}}_{\text{C}} \tag{2}$$

- A is the constant. As it does not contain $\mathbf{W}$, it is used to find the posterior covariance;
- B is linear, and we can use it to find the mean;
- C is quadratic in $\mathbf{W}$.

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

Which allows us to define the posterior covariance matrix,

$$S^{-1} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \Sigma^{-1}$$

and use B to find the mean.

$$B = \frac{1}{\sigma^2}\mathbf{Y}^T(\mathbf{XW}) = \frac{1}{\sigma^2}\mathbf{W}^T\mathbf{X}^T\mathbf{Y}^T$$

We can use the linear term (B) to solve for μ.

$$\mathbf{W}^T S^{-1}\mu = \mathbf{W}^T(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \Sigma^{-1})\mu = \frac{1}{\sigma^2}\mathbf{W}^T\mathbf{X}^T\mathbf{Y}$$

$$\mu = \frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \Sigma^{-1})^{-1}\mathbf{X}^T\mathbf{Y}$$

Finally, we can write the posterior distribution for linear regression.

$$p(\mathbf{W}|\mathbf{Y}, \mathbf{X}) = \mathcal{N}(\mathbf{W}|\frac{1}{\sigma^2}(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \Sigma^{-1})^{-1}\mathbf{X}^T\mathbf{Y}, (\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \Sigma^{-1})^{-1}) \tag{3}$$

The posterior uses $\frac{1}{\mathbf{Z}}$ as a normalisation constant to ensure that the posterior is a probability distribution.

$$p(\mathbf{W}|\mathbf{Y}, \mathbf{X}) = \frac{1}{\mathbf{Z}}p(\mathbf{Y}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

1.1.2 Non-parametric Regression

Question 7

Parametric models make an assumption about the form of the relationship between variates i.e. they specify a finite set of basis functions in advance. For example, we can assume linear basis functions, nonlinear polynomial basis functions of varying degrees, or other types of nonlinear basis function. The "parameters" of these models are the coefficients of the basis functions. Conversely, non-parametric models do not specify explicit basis functions but rather seek to allow an infinite set of functions. Gaussian Processes are a type of non-parametric model that wrap this infinite set of functions in a prior distribution.

Non-parametric models focused on using what we know about the current data to classify new unseen data points. The data can be seen as analogous to the parameters. A good example of this is the K-nearest-neighbour model, which classifies new data points based on its surrounding classes.

Non-parametric models may be more difficult to interpret as they do not have direct parameters to describe the model. It is not always clear how the relationship between variates has been learnt.

Question 8

The GP prior represents a distribution of the space of all possible functions on the 2D plane. By giving the prior a zero mean and uniform variance, we are essentially saying that we think the most likely intersection of a function passing through a vertical line at x will have a y coordinate of 0. Extending out to along the entire x axis, the most likely function will be a flat horizontal line (the x axis itself). However in reality we know that the functions are not likely to be linear - the prior allows for this in that every vertical line through the x axis has a Gaussian likelihood associated with it.

Question 9

Since the prior mean and covariance is the same for all x coordinates, as mentioned above some functions are more likely than others. For example, smoother and flatter functions are considered more likely by this prior, since the difference between the y coordinates of two adjacent intersection points will be smaller and hence closer to the mean. The covariance $k(\mathbf{X},\mathbf{X})$ allows us to assume that any two values x_i and x_j covary, therefore f_i and f_j can also be expected to covary.

This means that we think that smooth functions are more likely, however the probability of more rapidly changing functions is non-zero. In fact, all functions have a non-zero probability.

Question 10

The joint distribution of the GP model can be written as

$$p(\mathbf{Y},\mathbf{X},f,\theta) = p(\mathbf{Y}|f)p(f|\mathbf{X},\theta)p(\mathbf{X})p(\theta). \qquad (4)$$

The corresponding graphical model is shown in Figure 1.

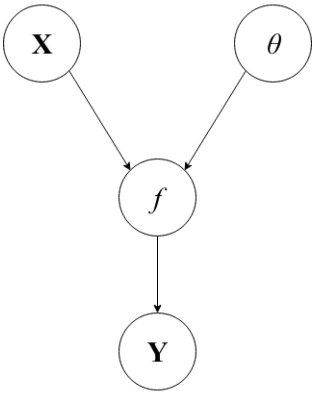


Figure 1: *Graphical model of the joint distribution.*

The assumptions that have been made are:

- $\mathbf{X}$ and θ do not depend on anything;
- f depends on θ and $\mathbf{X}$;
- $\mathbf{Y}$ depends on f, but is is conditionally independent of $\mathbf{X}$ and θ.

Question 11

The marginalisation $p(\mathbf{Y}|\mathbf{X},\theta)$ connects the prior and the data because we now have a way to directly generate $\mathbf{Y}$ values given $\mathbf{X}$ and θ values without knowing the actual form of f.

Because we are uncertain about f, when we marginalise it out, the uncertainty gets pushed onto $\mathbf{Y}$.

The fact that θ is left on the left hand side of the expression after marginalisation means that it is needed, along with $\mathbf{x}$, to calculate $\mathbf{Y}$. This implies that $\mathbf{Y}$ is dependent on theta.

The graphical model of the marginal $p(\mathbf{Y}|\mathbf{X},\theta)$ is shown in Figure 2 and shows how $\mathbf{Y}$ is now dependent on only $\mathbf{X}$ and θ. The function f (and its uncertainty) has been "baked in" to $\mathbf{Y}$.

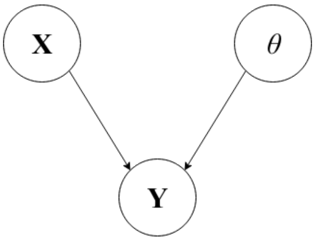


Figure 2: *Graphical model of the marginalised distribution.*

1.2 Practical

1.2.1 Linear Regression

Question 12.1

The prior distribution over $\mathbf{W}$ can be seen in the middle column of the top row of Figure 3. The red cross represents the true value of $\mathbf{W}$. The spherical covariance gives rise to a symmetrical distribution.

The rightmost column of the top row of Figure 3 shows six random samples drawn from the prior which clearly do not have any well defined direction, which is exactly as expected since we have no data to make predictions with.

Question 12.2

The second row of Figure 3 show the result of picking a single data point. The leftmost column plots the likelihood distribution in $\mathbf{W}$-space of the data point. The middle column shows the posterior distribution after the likelihood has been combined with the prior. It can be seen visually that the prior has been squashed along the dimension of the likelihood.

Question 12.3

The rightmost column shows that the sampled functions now all pass through the observed data point, but do not yet agree completely in direction.

Question 12.4

The lower two columns of Figure 3 show the likelihood, posterior and data space samples after adding two and twenty points respectively.

Question 12.5

After a single data point is added, we can see that the posterior distribution begins to squash in one direction to converge onto the parameters $\mathbf{W}$. When we sample from this distribution, there are many lines passing through the point at different gradients, which is because we cannot define two parameters from a single data point. As we begin to reveal more data, the posterior distribution centres onto $\mathbf{W}$ and the sample functions fit more closely to the data points. This is exactly what we want to happen as it shows that we

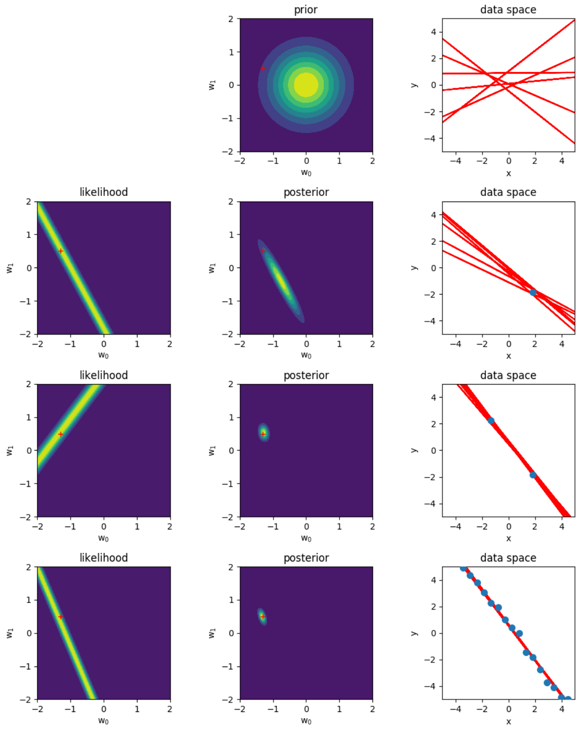


Figure 3: *Implementation of linear regression. The left hand column plots the likelihood. The middle column plots the prior/posterior, and the right hand column shows six random sample functions. Plots are drawn after one, two and twenty data points are revealed.*

have relearned the parameters of the model used to generate the data $\mathbf{X}$.

Question 12.6

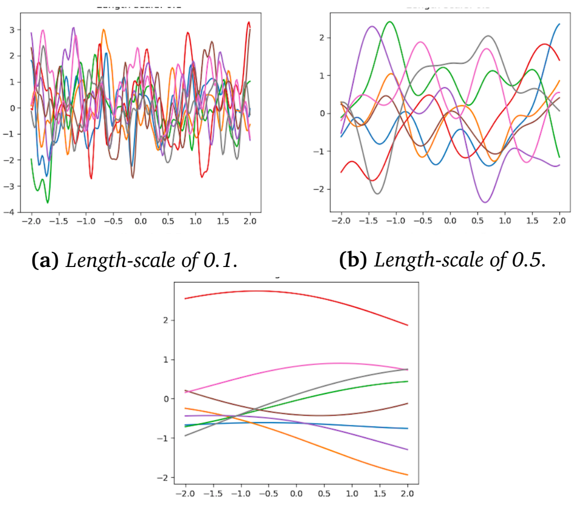
The posterior converges on the position of $\mathbf{W}$ because after a new data point is added, we update the parameters using the newly calculated mean and covariance. As we see more data, we become more and more confident about the true value of $\mathbf{W}$.

1.2.2 Non-parametric Regression

Question 13

Figure 4 shows three examples of samples taken from a GP prior with a squared exponential covariance function that has a single hyper-parameter l which encodes the length scale of the sample functions. The three values of l are 0.1, 0.5 and 5 respectively.

Increasing the length-scale of the covariance function allows us to constrain the smoothness of the sample functions. A smaller length-scale creates functions with more rapid changes than those with a larger length-scale. This is because a higher length-scale means that, for two random variables x_i and x_j where the covariance is $k(x_i, x_j)$, their instantiations as f_i and f_j are similar.



(a) *Length-scale of 0.1.* **(b)** *Length-scale of 0.5.*

(c) *Length-scale of 5*

Figure 4: *Squared exponential with varying length scales.*

Choosing a larger length-scale encodes the assumption that smoother function with lower frequency is preferable to high frequency, more oscillatory functions. In other words, we prefer functions that change less rapidly.

1.2.3 Question 14

The predictive posterior distribution is a Gaussian given by

$$p(\mathbf{Y}_{N+1}|\mathbf{Y}) = \mathcal{N}(\mathbf{M}_{N+1}, \mathbf{C}_{N+1}) \tag{5}$$

where

$$\mathbf{M}_{N+1} = \mathbf{k}^{\mathbf{T}}\mathbf{C}_N^{-1}\mathbf{Y} \tag{6}$$
$$\mathbf{C}_{N+1} = c - \mathbf{k}^{\mathbf{T}}\mathbf{C}_N^{-1}\mathbf{k} \tag{7}$$

and $\mathbf{C_N}$ is the previous covariance matrix, $\mathbf{k}$ has elements $k(\mathbf{x}_n, \mathbf{x}_{N+1}$ (i.e. the kernel function invoked with all the previous data points and the new data point) and c is a scalar that equals $k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}$ (i.e. the kernel function over just the new data point).

Figure 5 shows three rows of the process of computing the predictive posterior and sampling from it for the three different length scales. The first column is the random samples from the prior. The middle column shows the noisy data points and the generated predictive mean, with the true function plotted underneath. It also shows the predictive variance, shaded in blue. The last column shows samples from the predictive posterior.

It is clear from the plots that the samples from the posterior are more or less following the observed data. Using a longer length scale results in functions that more closely resemble the true function. If the length scale is too short, it is almost as if the functions have too many degrees of freedom and oscillate up and down quite a lot in order to pass as close as possible to the observed data points. This is not necessarily desirable as it seems as though the data has been overfitted.

The functions do not pass through the data points exactly, due to the additive noise. Setting σ to zero gets rid of this residual noise and causes the functions to go exactly through the data. However, this causes the functions to oscillate more rapidly as if they are being "forced" to pass through exact points.

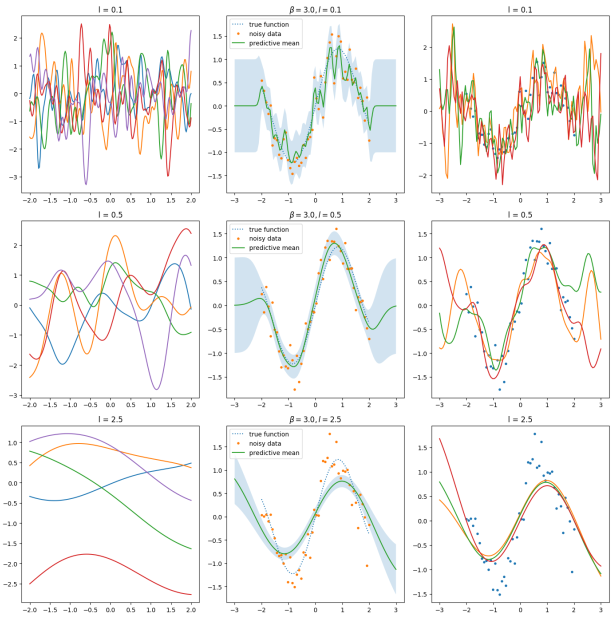


Figure 5: *Implementation of non-parametric regression with changing length-scales.*

It is interesting to note that the predictive variance "falls off" more sharply in regions where there is no data when using a shorter length scale. This is because a shorter length scale produces functions that change more rapidly, so we can be less certain about where adjacent functions will intersect. With a longer length scale, the opposite is true; since we can be fairly sure that the function isn't going to change very rapidly we can make a decent estimate as to where it will intersect in adjacent regions.

2 Posterior

Question 15

We use our beliefs in order to find a starting point (or "handle") for a particular problem. These beliefs allow us to make assumptions about either the input data itself or the relationship between the inputs and the outputs (i.e. the parameters). These assumptions are used to formulate our prior distributions.

In other words, assumptions are what we assume to be true about something. If we had all the data we required for a problem, there would be no need to assume anything. Making assumptions based on beliefs gives us a way to reason about a problem when we have very little information.

A preference is an assumption that we want to use because it is easier to do so. A preference does not necessarily represent some truth about a variable, but is rather used for convenience. For example, using a Gaussian likelihood so that a posterior distribution is also Gaussian, or "preferring" a certain class of functions over another because it simplifies calculations in some way.

Question 16

Since the covariance matrix is spherical, we have assumed that the latent input variables are independent and that they are centred around the origin.

Question 17

One way to obtain the marginal distribution $p(\mathbf{Y}|\mathbf{W})$ would be to write out the exponents of the likelihood and the prior and do a bunch of algebra to work out the actual integral. But there is a simpler way.

Firstly, we know that $\mathbf{Y} = \mathbf{WX} + \epsilon$ and we know that $\mathbf{X}$ is Gaussian since we put a prior over it. We know that we can do any linear transformation to a Gaussian and it is still a Gaussian because Gaussians are closed under linear transformation. So we know that $p(\mathbf{Y}|\mathbf{W})$ is going to be a gaussian. All we need to do is work out the mean and the covariance of the distribution we are looking for.

We know that the mean is the expectation of $\mathbf{Y}$, so we can therefore write

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{WX} + \epsilon]. \tag{8}$$

We know the expected value of $\mathbf{X}$ since we picked a prior over $\mathbf{X}$ and gave it a zero mean, so ti will be zero. We can also ignore ϵ in the expectation since it is just a noise parameter. Therefore we can deduce that the mean μ is just going to be zero.

Now to work out the covariance, we know that it is going to be the expected value of $\mathbf{Y}$ multiplied by its transpose, which we can write similarly to (8) as

$$\begin{aligned} \mathbb{E}[\mathbf{Y}\mathbf{Y}^\mathbf{T}] &= \mathbb{E}[(\mathbf{WX} + \epsilon)(\mathbf{WX} + \epsilon)^\mathbf{T}] \qquad (9) \\ &= \mathbb{E}[\mathbf{WX}\mathbf{X}^\mathbf{T}\mathbf{W}^\mathbf{T}] + \mathbb{E}[\epsilon\epsilon^\mathbf{T}] \qquad (10) \end{aligned}$$

The expectation of the noise parameter ϵ we have already defined as being $\sigma^2\mathbf{I}$. Since our prior over $\mathbf{X}$ was zero-mean, we know that $\mathbf{X}\mathbf{X}^\mathbf{T}$ is going to be the identity matrix, therefore can simply write the covariance as

$$\mathbb{E}[\mathbf{Y}] = \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^\mathbf{2}\mathbf{I} \tag{11}$$

which gives us the resulting marginal distribution

$$p(\mathbf{Y}|\mathbf{W}) = \mathcal{N}(0, \mathbf{W}\mathbf{W}^{\mathbf{T}} + \sigma^{\mathbf{2}}\mathbf{I}) \tag{12}$$

and we are done.

Question 18.1

All three estimation techniques attempt to find some parameters that maximise a distribution. ML simply maximises the likelihood, and does not take any prior information prior into account, which can often be seen as a limitation. MAP attempts to do better by taking the prior into account and maximising on the posterior distribution instead, however that makes it more computationally expensive. ML can be seen as a special case of MAP which assumes a uniform prior of the parameters. Type-II ML maximises the marginal likelihood after integrating out some parameters and is kind of an in-between stage of ML and MAP.

Question 18.2

With only one data point, ML will be less accurate than MAP since it does not take the prior into account. However as more data is seen, the two will begin to converge on one another.

Question 18.3

The denominator of the posterior distribution has no bearing on the result as it will always be positive, and depends on the parameters $\mathbf{W}$. Therefore the two expressions are equivalent.

Question 19

The objective function can be written as:

$$\begin{aligned}\mathcal{L}(\mathbf{W}) &= -log(p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)) \\ &= -\frac{ND}{2}log(2\pi) - \frac{N}{2}log|\mathbf{C}| - tr((\mathbf{YC})^{-\mathbf{1}}\mathbf{Y}^{\mathbf{T}})\end{aligned} \tag{13}$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathbf{T}} + \boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I}$. The gradients of the objective function with respect to the parameters can be written as:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= tr(\mathbf{C}^{-\mathbf{1}}(\mathbf{W}\mathbf{J}_{\mathbf{ij}} + \mathbf{J}_{\mathbf{ji}}\mathbf{W}^{\mathbf{T}})) \\ &+ tr(\mathbf{Y}^{\mathbf{T}}\mathbf{Y}(-\mathbf{C}^{-\mathbf{1}}(\mathbf{W}\mathbf{J}_{\mathbf{ij}} + \mathbf{J}_{\mathbf{ji}}\mathbf{W}^{\mathbf{T}})\mathbf{C}^{-\mathbf{1}})\end{aligned} \tag{14}$$

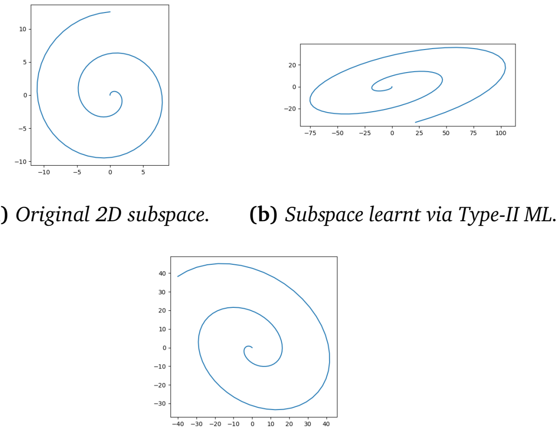
where $\mathbf{W}\mathbf{J}_{\mathbf{ij}} + \mathbf{J}_{\mathbf{ji}}\mathbf{W}^{\mathbf{T}} = \frac{\partial \mathbf{C}}{\partial \mathbf{W}_{\mathbf{ij}}}$.

Question 20

The uncertainty in $\mathbf{X}$ and θ is pushed into f. Therefore, by marginalising out $\mathbf{X}$, we lose some of the information required to calculate an accurate $\mathbf{Y}$.

Question 21

We first mapped the input variables $\mathbf{x}$ non-linearly to a 2D space and plotted the result. This gives a uniform spiral shape as seen in Figure 4a. We then mapped this linearly to a 10D space. Given the objective function and gradients from above, we then minimised the log likelihood to "learn" the linear mapping parameters W. Using the new parameters, we performed the reverse of the linear mapping back to 2D space and obtained another (slightly different) spiral, a particular example of which can be seen in Figure 6b.



(a) *Original 2D subspace.* **(b)** *Subspace learnt via Type-II ML.*

(c) *Random subspace.*

Figure 6: *Subspaces representing a 2D non-linear mapping from a 1D input vector.*

We expected the recovered shape to be more similar to the original shape. We considered that there may be an error in the minimisation process (probably somewhere in computing the gradients). We obtain a different shape every time we run the code, due to the randomised starting point for the gradient descent. We assume this is because there are many local minima in the 10D space, and the gradient descent is not guaranteed to find the global minima and is likely to get stuck in a local minima.

Since we have only learned the parameters for the linear part of the mapping and not the non-linear part, it does not immediately appear possible to recover the original latent variables themselves.

Question 22

Choosing any randomly initialised W matrix and performing the reverse mapping also results in a spiral shape, as can be seen in Figure 6c. However, the shape is slightly different in dimension. In fact, this shape is often more similar to the original than the one obtained via gradient descent. It is not understood why this is the case.

3 The Evidence

Question 23

When ewe maerginalise out everything,that is normally not possible.

Tic tac toe, only 512 possible boards. Can compute integral as sum. Trying to build different models for them. Given model, try to evalate probability.

Question 24

Question 25

4 Final Thoughts

Question 30

The authors feel that the purpose of this assignment was to introduce the basics of machine learning using a statistical, low level, brutal, sadistic, trauma-inducing and slightly evil approach that is terrifying for those without a strong background in mathematics. Students are thrown in to the deep end of a swimming pool and forced to learn the theory behind the techniques in gory detail before being able to truly understand or implement them. However, in the end, the differences between the various techniques we have explored have started to become clearer (although knowing which technique to pick in which situation based on real data still seems a long way off). One main learning point that is more or less clear is that all the techniques involve making some assumptions about some data, formulating those assumptions into prior distributions and then churning the Bayesian handle to produce some intelligible posterior result.