Machine Learning Coursework

Eddie Jones *36147* **and Kheeran Naidu** *35838*

1 The Prior

Question 1

1) The likelihood function represents the probability of getting the data sampled from a given model. Models can often be represented by a specific set of parameters where we can assume a functional form of the distribution in advance. Notated by $p(\mathbf{D}|\mathbf{W})$ where D is our observed data and $\mathbf{W}$ the parameter vector. A common likelihood function comes from the Gaussian distribution. A Gaussian likelihood function is used because of the central limit theorem [CN09] which states that if we repeatedly sample the same random variable (where each observation is independent), then the average of these samples will converge to a Gaussian distribution. This applies to our scenario as when we take samples of $\mathbf{y}_i$, we are effectively taking samples of the error $\mathbf{y}_i - \mathbf{W}\mathbf{x}_i$. Each measurement of the error we shall assume to come from the same distribution and to be independent and hence the average of these errors will converge to a Gaussian distribution. The main assumption that the Gaussian therefore encodes is that the additive errors of each measurement are independent.

2) For a Gaussian probability distribution $p(\mathbf{y}|f,\mathbf{x}) \sim \mathcal{N}(f(\mathbf{x}),\Sigma)$, with mean $f(\mathbf{x})$ and co-variance matrix Σ. $(\Sigma)_{mn} = (\Sigma)_{nm}$ is the co-variance between y_m and y_n and for the diagonal (Σ_{mm}) is the variance of y_m with mean $f(x_m)$. In other words, how dependent y_m and y_n are on each other, and the spread of each y_m around its mean respectively. In this case where we have chosen a spherical co-variance matrix which is a multiple of the identity matrix of the form $\sigma^2\mathbf{I}$. Since it is a diagonal matrix, each dimension, y_m and y_n, $m \neq n$, are not co-variant and have the same uni-variate normal distribution from their respective means. As they are specifically normally distributed we can say that they are also independent, but in general not being co-variant doesn't imply independence. It is called a spherical co-variance matrix because if we apply some rotation to the coordinate space there will be no change in the distribution.

Question 2

If we can't assume that the outputs are independent of each other, then we have to assume some chronological causal chain, in other words the first output y_1 can only be dependent on our given function f and set $\mathbf{X}$; $p(y_1|f,\mathbf{X})$. Any subsequent output, however, can then dependent on the previous outputs as well as the function f and set $\mathbf{X}$; for the case $i = 2$, $p(y_2|f,\mathbf{X},y_1)$, and in general $\mathbf{X}$; $p(y_i|f,\mathbf{X},y_1,y_2,...,y_{i-1})$, $\forall 2 \leq i \leq N$. Now we use the product rule to string them together;

$$p(\mathbf{Y}|f,\mathbf{X}) = p(y_1|f,\mathbf{X})\prod_{i=2}^{N} p(y_i|f,\mathbf{X},y_1,y_2,...,y_{i-1}).$$

Question 3

We have assumed the distribution of the noise ϵ and therefore we can derive the distribution of $\mathbf{y}_i$:
$\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i + \boldsymbol{\epsilon} \implies \boldsymbol{\epsilon} = \mathbf{y}_i - \mathbf{W}^T\mathbf{x}_i$ and since $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$, we have that $p(\mathbf{y}_i|\mathbf{W}^T,\mathbf{x}_i) \sim \mathcal{N}(\mathbf{W}^T\mathbf{x}_i,\sigma^2\mathbf{I})$. Using the probability density function of a Gaussian distribution and assuming that the data point are independent, we have that;

$$p(\mathbf{Y}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{N} \frac{e^{-\frac{1}{2}(\mathbf{y}_i-\mathbf{W}^T\mathbf{x}_i)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y}_i-\mathbf{W}^T\mathbf{x}_i)}}{\sqrt{(2\pi)^D\sigma^{2D}}}$$
$$= \left((2\pi)^D\,\sigma^{2D}\right)^{-\frac{N}{2}} e^{\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(\mathbf{y}_i-\mathbf{W}^T\mathbf{x}_i\right)^T\left(\mathbf{y}_i-\mathbf{W}^T\mathbf{x}_i\right)}$$

where $\mathbf{x}_i \in \mathbb{R}^D$

Question 4

Bayes' Theorem states that:

$$(Posterior) = \frac{(Likelihood)(Prior)}{(Evidence)}$$

Here *Likelihood* refers to the probability of observing our data given a model, the *Prior* is a random variable encoding our beliefs/assumptions about the model, the *Evidence* is the marginal likelihood of our data, and the resulting *Posterior* is our updated beliefs.

In order to apply this theorem and update our system iteratively (i.e. learn), we can formulate the *Posterior*, and use it as the new *Prior*.

A conjugate prior is one where the functional form of the posterior distribution will be the same as the prior's. This is particularly useful as the formulation of the evidence in general, will be intractable.

$$(Evidence) = \int (Likelihood)(Prior_X)dX$$

However, if we use a conjugate prior then we already have the functional form of the posterior, and from this we can easily calculate specific parameters which gives us the evidence/normalising factor for free. Without the ability to compute the specific normalising factor, *Evidence*, then our posterior doesn't form a proper probability distribution, and consequently will be difficult to interpret; defeating the purpose of the posterior.

Another advantage of using a conjuagte prior is that the family of distribution which the prior and posterior belong to will be invariant through the learning process and hence we can iteratively update our beliefs. This property also has the advantage that, in general, we dont want the distribution of our beliefs to change with more data, only the parameters to be updated. This helps counter over-fitting, for example we could end up considering a multimodal distribution with peaks at all data points.

A very useful property of a Gaussian distribution is that it is conjugate to itself. So in our example referring to the application of Bayes' Theorem in the question, since both our *Likelihood* and *Prior* have a Gaussian distribution, the *Posterior* is also a Gaussian distribution.

A practical drawback of this is making false assumptions about the distribution of our beliefs in order to simplify computation.

Question 5

The distance function of a Gaussian distribution in D-dimensions between two points $\mathbf{x}$ and $\mathbf{y}$ is encoded by the squared Mahalanobis distance; $(\mathbf{x}-\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{y})$. In the context of $p(\mathbf{W}) \sim \mathcal{N}(\mathbf{W}_0, \tau^2\mathbf{I})$, with a spherical co-variance matrix, the squared Mahalanobis Distance between $\mathbf{W}$ and the mean $\mathbf{W}_0$ becomes the scaled squared Euclidean distance; $\frac{(\mathbf{W}-\mathbf{W}_0)^T(\mathbf{W}-\mathbf{W0})}{\tau^2}$.

This metric is useful as it tells us how the probability density of the Gaussian exponentially decays with the squared Mahalanobis distance from the mean.

Question 6

To calculate the posterior distribution over the parameters $\mathbf{W}$ we need to consider the likelihood multiplied by the prior as this will be proportional to the posterior probability. We don't need to consider any normalisation factors as we deal with conjugate distribution (see question 4). For the same reason we are only going to consider the exponential term in the probability density function.

Here the likelihood, $p(\mathbf{Y}|\mathbf{X},\mathbf{W})$ follows a Gaussian distribution with mean $\mathbf{W}^T\mathbf{X}$ and covariance $\sigma^2\mathbf{I}$. The prior is also Gaussian $p(\mathbf{W}) \sim N(\mathbf{W}_0, \sigma^2\mathbf{I})$, where we are assuming that $\mathbf{W}_0 = \mathbf{0}$.

$$
\begin{aligned}
p(\mathbf{W}|\mathbf{X},\mathbf{Y}) &\propto p(\mathbf{Y}|\mathbf{X},\mathbf{W})\, p(\mathbf{W}) \\
&\propto -\frac{1}{2\sigma^2}\left(\mathbf{Y}-\mathbf{W}^T\mathbf{X}\right)^T\left(\mathbf{Y}-\mathbf{W}^T\mathbf{X}\right) \\
&\quad -\frac{1}{2\tau^2}\mathbf{W}^T\mathbf{W} \\
&= -\frac{1}{2\sigma^2}\mathbf{Y}^T\mathbf{Y} + \frac{1}{\sigma^2}\mathbf{Y}^T\mathbf{W}^T\mathbf{X} \\
&\quad -\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X} - \frac{1}{2\tau^2}\mathbf{W}^T\mathbf{W}
\end{aligned}
$$

[Bis06]

The term has now been split into three parts: a constant (with respect to $\mathbf{W}$), a linear term and a quadratic. Hence we can deduce the new precision matrix (inverse of the co-variance) from the quadratic coefficient:

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

By examining the linear term in conjunction with our co-variance matrix we can deduce that the mean is:

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

The posterior distribution of $\mathbf{W}$ therefore will be $\mathcal{N}(\mu, \mathbf{S})$, with parameters defined above.

Question 7

Parametric representation:
In a parametric representation of our model, we pre-specify our belief of the form of the generating model (the type of our regression); for example if it is linear ($y_i = w_0 + w_1x_i$), quadratic ($y_i = w_0 + w_1x_i + w_2x_i^2$) or represented by a trigonometric function ($y_i = w_0 + w_1 sin(x_i)$), etc. This category of regressions is equivalent to optimising the parameters with regards to how well the model would fit the data (as we have done in the previous questions).

Non-parametric Representation:
In contrast, in the non-parametric representation we don't make any assumptions about the functional instead we describe each output point of the generating function as a distribution. For example with a Gaussian process instead of attempting to parameterise the space of functions, we represent the likelihood that our data fits a specific function at a certain point (a margin) as its own Gaussian distribution. This is much more representative of the data as it is not limited to well defined and understood functions (linear, quadratic, trig, etc), it also means the uncertainty varies at different points in the data domain, so in a region where we have lots of samples we can be more certain and less certain in regions where we have extrapolated, this helps prevent over-fitting. The model still has parameters of sorts, but now they are called hyper-parameters as they don't need to be changed as we learn. In the case of the Gaussian process the defining hyper-paremeters determine the kernel, a form of inner product, ($k : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$) between two points in the input space, through which we build the co-variance matrix $\Sigma = \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & ... \\ k(x_2,x_1) & k(x_2,x_2) & ... \\ \vdots & ... & \ddots \end{bmatrix}$. The kernel function usually contains some distance term and so we make the assumption that, in general, when two points $\mathbf{x}_i$ and $\mathbf{x}_j$ are closer together, they are more dependent on each other and will have similar output values $\mathbf{y}_i and \mathbf{y}_j$. This property encodes the preference for smoothness in our function.

Parametric Interpretation:
We can easily interpret the posterior model in the parametric case as we have belief about the parameters and can therefore say how well the data fits our assumption, which makes presenting the findings very easy. On the other hand if there is a large number of parameters, how each of them affects the shape of our model is not-intuitive to reason about. For example if we wanted to consider the entire space of functions we'd need a infinite amount of parameters, which is impossible and even approximations with a very large number wouldn't be interpretable.

Non-Parametric Interpretation:
The shape and form of the data is now represented by the function of the inner product combined with the co-variance matrix. Just by looking at these two structures is much harder to interpret and generalise. We can plot the shape of the most likely function and the error space around it to give a good representation of our findings, however the function

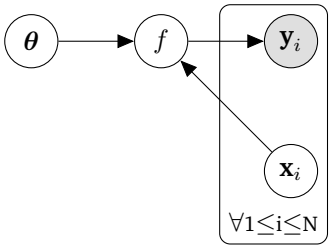

Figure 1: *Joint distribution of the full model*

computed could be unrepresented by well known functions such as the quadratic/trig functions, instead it can learn to fit any function, getting round the problem with parametric representations.

Question 8

The prior $p\left(f|\mathbf{X},\boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{0}, k\left(\mathbf{X},\mathbf{X}\right)\right)$ is describing the probability of a particular output value at a marginal slice as a Gaussian distribution. The distribution has mean zero, which we can assume without loss of generality by zero-meaning our data space. The distribution has co-variance matrix determined by the kernel function k where $(\Sigma)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The model also contains some hyper-parameters, θ which determine kernel function.

This places structure on the space of functions by assigning a distribution to each marginal output value, which enables us to compare the likelihood of any function's output value. We can then consider a functions overall likelihood by integrating this over the entire space. At this stage we have no information about the kernel function and so no additional structure has upon it.

Question 9

Each marginal slice will follow a Gaussian distribution, and therefore every output value has a non-zero probability. Which means there is a likelihood, however small, of a function from this Gaussian Process passing through any set of output values which means encodes the entire space of all possible functions.

Question 10

Using the laws of conditional probability, we can rewrite the joint distribution as $p(\mathbf{Y}, f, \mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{Y}|f, \mathbf{X}, \boldsymbol{\theta})p(f, \mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{X}, f, \boldsymbol{\theta})p(f|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{X}, \boldsymbol{\theta})$. If we make the assumption that $\mathbf{X}$ and $\boldsymbol{\theta}$ are independent we get the following;

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

We illustrate this relationship, along with our assumption, in figure 1.

- $\mathbf{X}$ and θ are independent, in other words $p\left(X,\theta\right) = p\left(X\right)p\left(\theta\right)$.

Question 11

The marginalisation in Eq.2 is the process of summing over all possibilities of f so that it is no longer a dependent in marginal probability but it's distribution is accounted for. In this case $\mathbf{Y}$ is dependent on f, and f depends on both $\mathbf{X}$ and $\boldsymbol{\theta}$. We are effectively taking the expected value of the probability of $\mathbf{Y}$ given $\mathbf{X}$, $\boldsymbol{\theta}$ and f over the distribution of f. Since we have accounted for all instances of f, the uncertainty in f has been fully passed/filtered through to uncertainty in $\mathbf{Y}$ (as shown in figure 2).

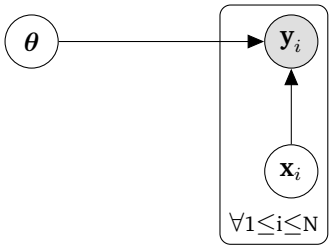

Figure 2: *The marginalisation of f*

The marginal probability is still condition and dependent on $\boldsymbol{\theta}$. This because the marginalisation process doesn't omit any information about the dependencies of the variable being marginalised, their effects are passed through to $\mathbf{Y}$ when considering all the values of f, which as with the uncertainty is dependent on these hyper-parameters $\boldsymbol{\theta}$.

Question 12

In figure 3 we visualise the prior distribution over $\mathbf{W}$, which represents our encoded beliefs, as seen in figure 3. After generating a particular data point, say, (-0.5, 1.15) with some error in the y value coming from a normal distribution, we can consider the posterior distribution as in figure 4. Here we can see the mean has been shifted towards the observed data and the co-variance is smaller as we can now be more certain about the distribution's spread. It is also no longer spherical as the different dimensions have different variances where our observed data isn't diagonal from the mean. We can now sample from the posterior distribution, these are particular linear functions which have parameters w_0 and w_1, as seen in figure 5

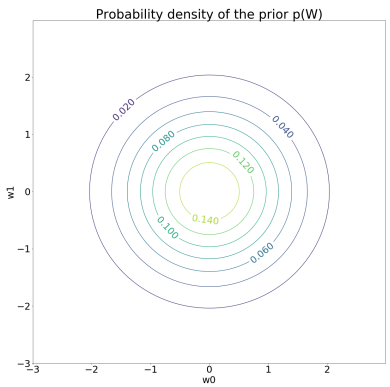

Figure 3: *The Probability Distribution of the Prior $p\left(W\right)$ where w_0 is the X-axis and w_1 the Y-axis*

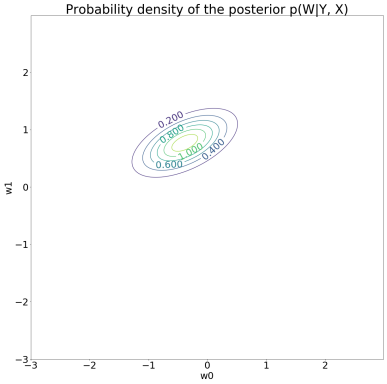

Figure 4: *The Probability Distribution of the Posterior $p\left(W|X\right)$ where w_0 is the X-axis and w_1 the Y-axis*

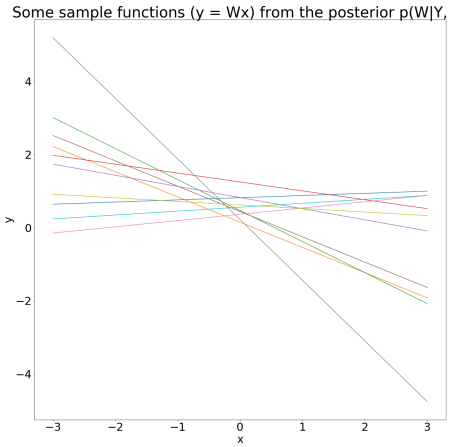

Figure 5: *Sample Functions from this Posterior Distribution*

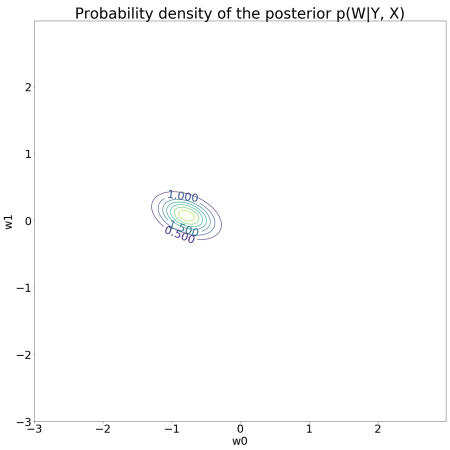

Figure 6: *The new Posterior Distribution p(W|X)*

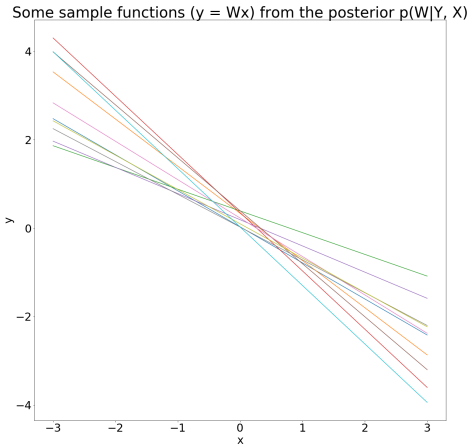

Figure 7: *Sample Functions from the new Posterior Distribution*

If we continue to add data this behaviour is repeated as seen in figure 6. When we add two more points (0.5, -0.15) and (0.9, -0.67), again the distribution gets closer to the underlying mean and the co-variance gets smaller. This behaviour is desired as it tells us the method for calculating the posterior is working (the mean of the posterior is getting closer to the true mean) and the model gets more and more sure as we add more data.

As we can see the functions sampled from the posterior distribution more closely approximate the underlying distribution, this is because the more data you add the larger the value of X^TX becomes and as we can see the precision, S^{-1}, increases with X^TX, and thus the co-variance decreases, i.e. we can be more confident about the values of W.

As the number of observations increase the precision approaches $\frac{1}{\sigma^2}X^TX$ so the mean approaches $\frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^TX\right)^{-1}X^TY = X^{-1}Y$, in other words $\mu^TX = Y$ and hence we can say μ is approaching the underlying mean.

Question 13

A Gaussian process is an example of a non-parametric model, for which we consider the likelihood of each output value, of the function we are trying to learn, to be a particular marginal distribution. In this case a Gaussian distribution. The prior will have mean zero (as we can adjust the data to be zero mean-ed) and the co-variance will be constructed from our kernel function to quantify how closely related output values of a certain distance are.

The kernel function we used here is of the form $k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma_f^2 e^{-\frac{\left(\mathbf{x}_i - \mathbf{x}_j\right)^T\left(\mathbf{x}_i - \mathbf{x}_j\right)}{l^2}}$, this particular kernel function depends on two hyper-parameters, l and $sigma$.

The prior for our Gaussian process will be $p\left(f\left(\mathbf{x}_j\right)\right) \sim N\left(0, k\left(\mathbf{X}, \mathbf{X}\right)\right)$. Figure 8 is a sample of 10 functions from this prior. The blue region shows $\mu \pm \sigma$, 1 standard deviation from the mean.

The lengthscale l is one of the hyper-parameters for our model, it encodes the assumption of the degree at which the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ affects the certainty, $p(f(\mathbf{x}_j)|f(\mathbf{x}_i)) \sim N(f(\mathbf{x}_i), \tau^2)$. A larger lengthscale, there-

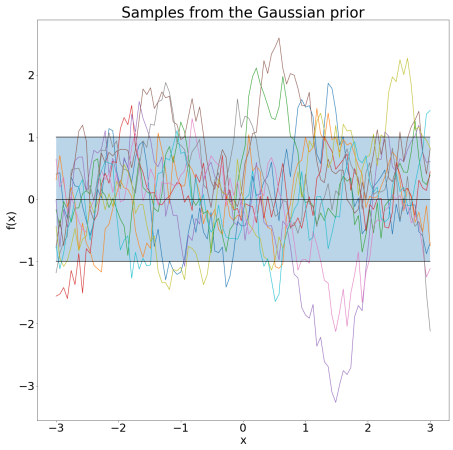

Figure 8: *Samples from the Gaussian Process Prior with Standard Deviation and Mean Indicated*

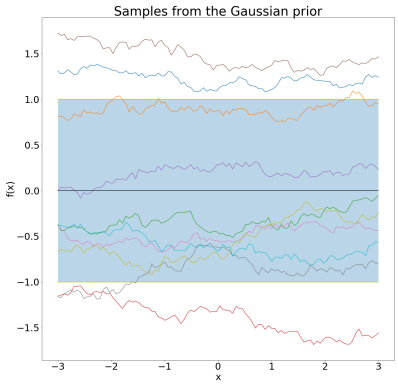

Figure 9: *Samples from the Gaussian Process Prior with a Larger Lengthscale*

fore, effectively flattens out the curve since we get more and more certain about what the output $f(\mathbf{x}_i)$ will be.

Question 14

The posterior of a Gaussian Process has the same role as with parametric models. After observing some data $\mathbf{X}, \mathbf{Y}$ we can use the self-conjugate nature of the normal distribution to determine the new mean and covariance for the posterior over a certain range $\mathbf{x}^*$, using standard Bayesian inference, this we will also include a bias term to encode our distrust of the data, or equivalently our belief about the error in our samples: $\sigma^2\mathbf{I}$.

$$\mu = K\left(\mathbf{X}, \mathbf{x}^*\right)^T \left(K\left(\mathbf{X}, \mathbf{X}\right) - \sigma^2\mathbf{I}\right)^{-1} Y,$$
$$\Sigma = K\left(\mathbf{x}^*, \mathbf{x}^*\right) - K\left(\mathbf{X}, \mathbf{x}^*\right)^T \left(K\left(\mathbf{X}, \mathbf{X}\right) - \sigma^2\mathbf{I}\right)^{-1} K\left(\mathbf{X}, \mathbf{x}^*\right)$$

The following plot shows some samples from the posterior. Here the space which is plotted extends beyond the region where we have concrete data and the effect of this can be seen in the more sporadic nature of the samples at these extrema

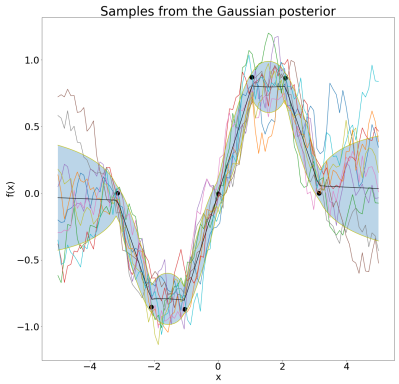

Figure 10: *Samples from the Gaussian Process Prior with Standard Deviation and Mean Indicated*

and the corresponding increase in width of the standard deviation lines. In general though the samples pass near the sample points (not exactly though as we take into account error in their value), the function clearly behaves similarly to a sine function however at the extrema the prior and uncertainty in our observed data is making pulling the function back towards zero (the mean of the prior). The figure also shows that the process is less certain about positions further from the mean as we can see from the standard deviation plots.

2 Posterior

Question 15

The first thing is to think about what kind of beliefs we have about a system (or a problem that needs solving). The beliefs represent our ideas and opinions, potentially a source of bias.

We then would make assumptions about the data in order to represent our beliefs. In other words our assumptions encode our beliefs. The more data we are provided with the less assumptions we are making as we can be more certain about properties of the data and the model.

In the context of Bayesian learning, our beliefs and assumptions are encoded by the distribution of the prior. Our beliefs now can be updated by the new observations in the formulation of the posterior where the assumptions are our the initial basis and intelligence of the entire learning process. This means that our assumptions are dependent on our beliefs, and our updated beliefs are dependent on our assumptions and so on.

A preference is an idealised form of something we are attempting to model. This will lead to a bias in the results produced but it is not rigid and given enough data the preference will effectively be ignored. A prior usually encodes a preference/assumption for example, in the case of a coin flip, we may presume or have a preference for fair coins and build a prior as such, or in the case of a Gaussian process we are in some sense preferring a smooth function that explains the data.

We can then go on to say that our preference is also the depth at which we want our assumptions and belief to go, and so the preference is the point in which we stop going deeper in our assumptions.

Question 16

This prior is a spherical Gaussian, which means that each dimension is univariate and the co-variance between dimensions is 0. This encodes assumptions are that each dimension of **X** is independent from each other and no dimension has precedence is terms of how much it varies and consequently its impact on our likelihood.

Question 17 [Bis06]

From the question we have that $p(\mathbf{y}_i|\mathbf{W}) \sim N(\boldsymbol{\mu}, \mathbf{C})$ because

1. $p(\mathbf{x}_i)$, the conjugate prior of $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W})$, has a Gaussian distribution $N(\mathbf{0}, \mathbf{I})$ and so $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W})p(\mathbf{x}_i)$ is Gaussian
2. The sum of Gaussian distributions is Gaussian (including an infinite sum, which is integration)

Knowing this we can then perform the marginalisation and determine the mean and covariance of $p(\mathbf{Y}|\mathbf{W})$ as follows;

Mean:

$$\mathbb{E}(\mathbf{y}_i) = \mathbb{E}(\mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon}) = \mathbf{W}\mathbb{E}(\mathbf{x}_i) + \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{W}\mathbb{E}(\mathbf{x}_i) = \mathbf{0}$$

Since **X** is zero meaned.

Co-variance: It follows that

$$\begin{aligned}\mathbf{C} = cov(\mathbf{Y}) &= \mathbb{E}((\mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon})^T) \\ &= \mathbb{E}(\mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T) + \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \\ &= \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}\end{aligned}$$

Question 18

With a maximum likelihood estimate, we dont make any assumptions about the distribution of W simply pick a W that will maximise the likelihood of the values of Y and X occurring. A MAP estimation (maximum a-posteriori) uses a prior as well as the data. This prior encodes some belief about the distribution of W and hence will produce a result not just based on the data. Going back to the coin tossing example, even if wed only seen heads so far the prior belief that the coin is close to fair will prevent us from concluding the coin will always heads. A Type-II Maximum Likelihood, on the other hand, is a hybrid method where the prior of one parameter is marginalised, as with a MAP estimate, but the other parameter is simply optimised over as with an ML.

The more data observed the less significant the prior becomes and so ML and MAP will tend to the same thing.

We say that ML is a frequentist approach as we make no prior assumptions about the data, which, in the case of a coin toss, leaves us the ability to predict that a coin will always give us heads. MAP and Type-II ML, however, are Bayesian approaches as we encode our beliefs in the prior. ML and MAP can only be performed if there is one latent variable. This means that all other variables must be marginalised out before we can perform ML or MAP, and so this will always be a case of Type-II ML.

The two expressions in Eq.10 are equal as the denominator of the fraction on the right-hand side is an integral over W and so actually positive constant in terms of W (as W has been marginalised out). An argument to maximise a term will also maximise the same term multiplied by a positive constant and hence the left and right-hand side are equal.

Question 19 [Bis06] [Ped12]

Given some probability function $p(\mathbf{Y}|\mathbf{W})$, we want to find the unknown **W** that maximises the likelihood of **Y** (which is known). To do so, we need it's objective function and corresponding gradient to use in gradient decent method.

Objective function:

$$\begin{aligned}\mathcal{L}(\mathbf{W}) &= -\log p(\mathbf{Y}|\mathbf{W}) \\ &= \frac{ND}{2}\log(2\pi) + \frac{N}{2}\log(\mathbf{C}) + \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i)^T\mathbf{C}^{-1}(\mathbf{y}_i) \\ &= \frac{N}{2}(D\log(2\pi) + \log|\mathbf{C}| + Tr(\mathbf{Y}\mathbf{C}^{-1}\mathbf{Y}^T)\end{aligned}$$

where $\mathbf{Y}^T = \{\mathbf{y}_i \in \mathbb{R}^D : \forall 1 \leq i \leq N\}$,
$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$

Gradient: The gradient matrix is a matrix **M** with entries $(\mathbf{M})_{ij}$ such that

$$\begin{aligned}(\mathbf{M})_{ij} &= \frac{\partial\mathcal{L}(\mathbf{W})}{\partial\mathbf{W}_{ij}} \\ &= \frac{N}{2}(Tr(\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\mathbf{W}_{ij}}) + Tr(\mathbf{Y}\mathbf{Y}^T(-\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\mathbf{W}_{ij}}\mathbf{C}^{-1})))\end{aligned}$$

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

$$(\mathbf{J}_{ij})_{xy} = \begin{cases}1, & \text{if } x = i,\ y = j \\ 0 & \text{otherwise}\end{cases}$$

Question 20

Marginalising out **X** is more difficult because there are more random variables dependent on **X**. There are two key reasons why that is for the marginalisation of variables further down the chain of dependence. First of all, even if all the distribution involved are although Gaussian, their probability mass function may not have simple exponents due to complex interdependent co-variances and hence the integral we have to compute can quickly become analytically intractable.

The second reason, is that further down the chain we simply have more probabilities to consider; $p(\mathbf{X}_N|\mathbf{X}_{N-1}...\mathbf{X}_1)\,p(\mathbf{X}_{N-1}|\mathbf{X}_{N-2}...\mathbf{X}_1)\,...p(\mathbf{X}_1)$ which again results in a more complicated function to integrate as in figure 11.

On the contrary, f is further up the dependency chain and so there are fewer terms dependent on it allowing us to marginalise it out with far less computation.

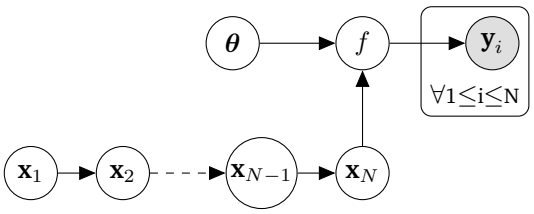

Figure 11: *Joint distribution of the model where $\mathbf{x}_i$'s for a chain of dependency*

Question 21

In theory, when using linear representation learning, we want to be able to represent our data in a lower dimension than it actually is in. This makes interpreting and understanding the data much easier (especially in the case of very high dimensional data such as photos).

For this part of the practical, we generate some 10D data from a predefined 2D space (given in the left of figure 12). We use this latent space, which the data was generated from, to verify the results of our process.

From the data we are using linear representation learning to find our 2D representation of the 10D data. So we assume that the representation will look like the generating space.

We clearly see from the comparison in figure 12 that the representation learned is rotated. This is because in our objective function, the calculation of the co-variance involves $\mathbf{x}_*\mathbf{x}_*^T$, which is unique only up to rotation since the transpose of a rotation matrix is its inverse. This is described as follows, $\mathbf{x}_*\mathbf{x}_*^T = \mathbf{x}_*\mathbf{R}(\mathbf{x}_*\mathbf{R})^T = \mathbf{x}_*\mathbf{R}\mathbf{R}^T\mathbf{x}_*^T = \mathbf{x}_*\mathbf{x}_*^T$

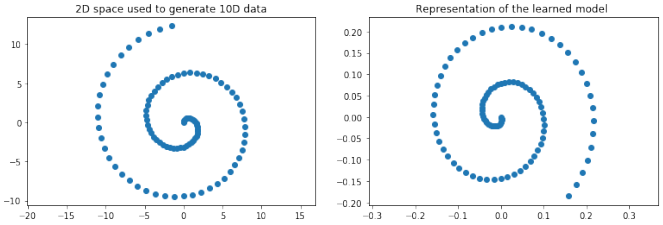

Figure 12: *2D space used to generate the 10D data vs the representation of the model, using an optimised projection matrix*

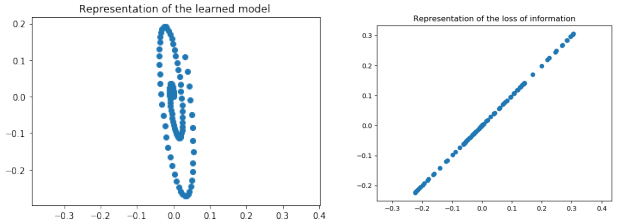

Figure 13: *The representation of the model, using random projection matrices of 2 different cases*

Question 22

By using a linear transformation of the subspace in Q21, we change the basis vector of the subspace from [1,0] [0,1] to [4,0] [3,1] by applying a change of base matrix $\begin{bmatrix} 4 & 3 \\ 0 & 1 \end{bmatrix}$ to each point and then plotting our points in the new subspace (seen in the left of figure 13). This is because a change of base is equivalent to applying a random matrix $\mathbf{x}_*$ which reduces the dimensionality of the 10D data to 2D (since it is a change of base, the dimensionality isn't reduced any furthur from 2D), and so the structure of the data would still be maintained, however, it would be scaled and rotated. In the case where dimensionality is reduced further from 2D, then information about the structure is lost, as seen in the right of fig 13. There is one final case where the structure is squashed onto 1 point, [0,0], losing all information about the structure. The likelihood of randomly getting these 2 latter cases are much lower than the first.

3 Evidence

For the evidence section we are considering a very simple data domain of grid points $\boldsymbol{X} = \{\mathbf{x}\}_{i=1}^{9}$ where $\mathbf{x}^i = (\{-1, 0, 1\}, \{-1, 0, 1\})$ as the underlying parameterisation of a collection of classifications $D = \{y^i\}_{i=1}^{9}$ where $y^i \in \{-1, 1\}$.

An example grid may have the following form:

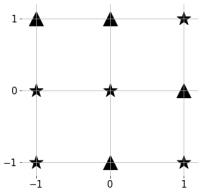

Figure 14: *A Representation of a Sample Grid, with Triangles for +1 Points and Stars for -1 Points*

Question 23

This assumption is stating that all values in the data domain are equally likely. This is the simplest model as it has the most entropy, all possible outcomes are equally weighted in the uniform distribution. So it has effectively made the smallest possible assumption. It can be considered the most complex model as it has no flexibility around the observed values of D as well allocating probability to a much wider space, it doesnt depend on the value of M_0 or θ_0 and hence it cant learn. It is therefore assuming that the model is already correct.

Question 24

These models all work by assuming each value of y is independent and then assigning a probability to each using a variation of the logistic function, so the limit towards positive infinity is 1 and towards negative infinity is 0. The limit we are considering is a linear combinations of the dimensions of $\mathbf{x}$ with potentially an extra linear term in the case of M_3. This is effectively a probability in terms of $\boldsymbol{\theta} \cdot \boldsymbol{x}^*$ where $\mathbf{x}^* = (x_1, x_2, 1)$. The models M_3 and M_3 have made more assumptions about the shape of the distribution than M_0 however it is more flexible as the likelihood now depends on a certain parameter $\boldsymbol{\theta}$ and therefore Bayesian style learning

can be applied. The choices made for the models M_1, M_2 and M_3 have locked the distribution into a particular shape. Even with lots of data which suggests a different shape the distribution of the likelihood doesnt change. In model M_1 we are assuming that only a linear variation of x_1 (the first dimension of **x**) is sufficient to determine y, in M_2 we are considering both dimensions of **x** and for M_3 we also considering an additional linear offset term θ_3 as part of our exponent. These different models are suited to data sets which classify **x** values based on only a reduced number of dimensions. The more dimensions considered the less potential for uncertainty as we are considering more data in our classification, if for example we used M_1 it may be ignoring significant data which determines the class of a particular **x** value.

M_3 is the most general model and therefore the most flexible as it allows all components of the vector x to be considered as well as a linear term not dependant on it. This gives it the potential to describe the underlying distribution more accurately. Although, the more complexity here may result in over-fitting if the data is noisy, or there isnt a large enough sample. This will especially be an issue if the second dimension of **x** doesnt actually determine its y value. M_2 is slightly less complex disregarding the linear additional term, hence it will be less susceptible to over fitting but may not be able to accurately describe the distribution or classify the points. A similar decrease in complexity occurs for M_1 and M_0.

Question 25

Marginalisation is the process of calculating the expected value of a probability over the entire range of the dependent variable. In this case we are considering all values of θ, and their respective probability, to determine a term for the probability of the evidence given by a certain model without considering the parameters that model may take. This enables us to determine which model would fit the data best before attempting to learn the specific parameters.

The choice of prior is assuming the possible parameters could vary massively, this is an assumption wed like to make as we have no reason to assume the parameters have any particular value. The choice of mu and sigma will determine which model is more or less likely in the marginalisation process, for example if we assumed θ_3 has a very low spread around 0 then there will be less evidence for the model M_3 which may be most accurate with higher values of θ_3.

Question 26

We can not compute the marginalisation in this case as the integral is too complex, instead we used Monte Carlo integration which uses samples of random variables to produce a numerical approximation of the integral, by considering the integral as an evidence formulation. In our model the variable S represents the number of samples taken in this process. For an S value of 50 the sum of evidence over the entire data set for each model can be seen in the table.

For the first model (M_0) the evidence is exactly 1 as expected, clearly any given model should have all of its probability mass in the data domain and therefore should a summation over the entire domain should be 1. The other models however (M_1, M_2, M_3) are only around 1, this is due to the error in our approximation of the marginalisation integral where we are using the stochastic method Monte Carlo integration. The error, $|1 - \Sigma p|$, in the evidence for each model is also in correlation with its complexity. This intuitively make sense as each model is considering effectively a higher dimension for the parameter θ and therefore would require more samples to span the same proportion of the space. The higher the value for S the more accurate the approximation of the integral, but this becomes slow to compute showing the intractability of overly complex models.

Table 1: *Comparison of Total Evidence over Domain Space*

Model	Sum Σ	Error
M_0	1.000	0.000
M_1	0.997	0.003
M_2	1.073	0.073
M_3	0.902	0.098

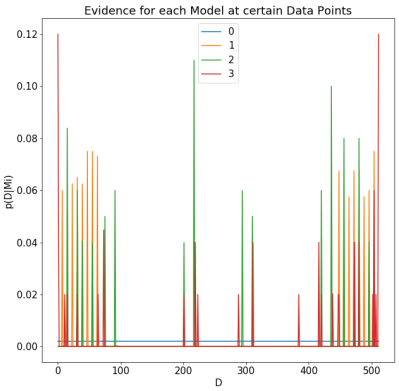

Figure 15: *The Evidence for each Model as we Enumerate the Domain Space*

Question 27

We can consider how the evidence for various models changes over the domain space in the next plot. Here the domain is iterated through changing one dimension at time. Although this is a perfectly valid logical way to enumerate the space, it isn't necessarily representative of the same structure the models will place upon.

The plot in figure 15 doesnt initially show any clear interpretation, i.e. the probability mass of each model doesnt appear to be distributed over the data domain in any particular way. However we can now reorder the domain for one model, say, the most complex M_3, where the x-axis goes from the most likely element in the domain to the least likely, we can then compare how the other models agree with M_3 within certain regions of the domain. This visualisation will also better mirror the inherent structure placed on the domain by the models.

From the plot in figure 16 we can see the region where the model M_3 attributes probability is quite small in comparison to the other models, it is much more specific about which data points are likely to occur. Although the evidence for the other (non-trivial) models drop off quite quickly too,

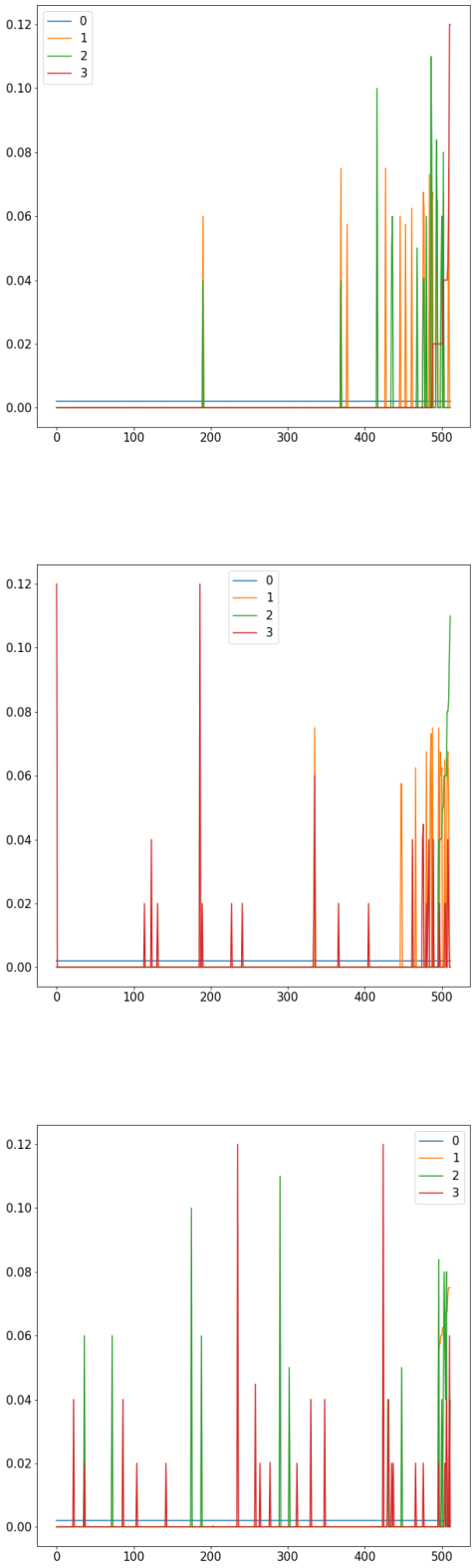

Figure 16: *The Evidence for each Model Ordered by the Model M_3, M_2 and M_1*

with most of their probability mass in-agreement by visual inspection. We can also see that M_3 has a relatively higher certainty of its most likely configuration. If, however, we order the domain by the probability of other models, say M_2, we can see the region where m2 attributes probability mass is correlated with m1 but isnt strongly correlated with the m3 at all. It also has a lower certainty of its most likely configuration. Again for the last (non-trivial) model M_1, there is little to no correlation between it and the other models, with even less certainty of its most likely point. From this we can deduce as the (parametric) complexity of our models decrease they become less precise in terms of the spread of the probability mass, but there does seem to be an overlap in the core region where they agree.

Question 28 [MG05]

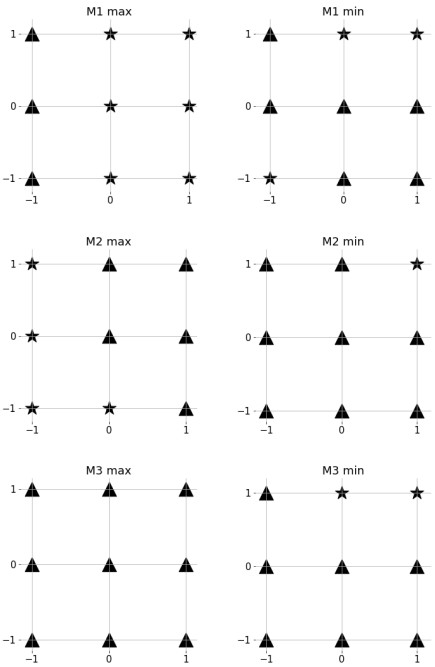

Figure 17: *The Extremum for each Model, with Positive Points as Triangles and Negative Points as Stars*

The elements of D which maximise or minimise the first model M_0 are trivial, as they all have the same probability. The other models may also potentially have multiple minimum or maximum, although the results aren't trivial.

The extremum for the other models can be seen in figure 17. As the M_1 model only classifies based on the first axis (the x-axis on our plots) it intuitively makes sense that the classification is consistent for each column. The minimum can be noted to have no columns of one value and similarly no rows of one value so requires quite a complex model. In other words this data would only be generated by a model which can distinguish different rows. Effectively the maximum will have a vertical decision boundary and the minimum some line which is not vertical at all, in this case a parabola can be seen between the positive points (△) and negative (★). For M_2 we now have the ability to consider diagonal decision lines within the grid as we take into account both dimensions of the domain space, this can

be seen where the maximum has such a diagonal boundary and is therefore likely to be generated by M_2. The minimum has a positive point separated from the other by negative points and therefore no such line could be said to exist from visual inspection, and therefore intuitively M_2 is unlikely to generate this data. Finally the most likely outcome for M_3 is where all points are classified positive consistently. This is different from the previous model as the decision boundary is outside the grid, this effect comes from the third dimension of θ we has the ability to shift the line not just rotate it. The minimum is the same as with M_2 and again we can again there is no straight line that would split the points as such.

Question 29

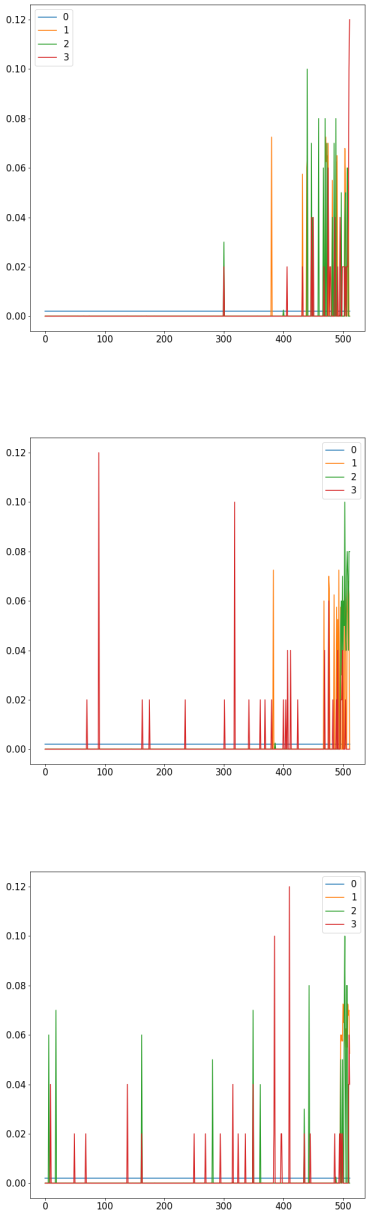

Figure 18: *The Evidence for each Model Ordered by the Model M_3, M_2 and M_1 with altered prior*

The prior dictates the distribution of θ which tells us the probability of seeing a particular likelihood $p(\mathcal{D}|M, \theta)$ and consequently, when marginalised it will affect the evidence for each model. For example if one model has the potential to accurately predict the data with a certain value of θ but that value of θ is considered unlikely by the prior then the model will be considered less likely to have generated that particular data set. A non-diagonal covariance matrix means that there can be correlation between the dimensions of θ.

If we change the mean to [5, 5, 5] but keep the same orderings as before we can see fairly similar graphs in figure 18 are produced with the overlapping regions following the same pattern. For any particular ordering, the points which were likely with the original prior remain likely, this is intuitive as we are assigning very little certainty to the distribution of θ and therefore moving the mean has little effect on specific evidence values.

4 Final thoughts

Question 30

This assignment has provided a range of examples and contexts where machine learning applies, each section requires different approaches and tools however they all have the same underlying theme of making assumptions and modelling. The only way to make progress in learning, broadly, is to assign some preference otherwise the problem is often ill-defined. Regardless of which variables are latent or whether the regression is parametric or not the key aspect is applying Bayesian statistics (where ever possible) opposed to a frequentist perspective.

Bibliography

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.

[CN09] Christopher Clapham and James Nicholson. *The Concise Oxford Dictionary of Mathematics (4 ed.)* Oxford University Press, 2009. ISBN: 9780199235940.

[Law05] Neil Lawrence. *Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models*. 2005. URL: `http://jmlr.csail.mit.edu/papers/volume6/lawrence05a/lawrence05a.pdf`.

[MG05] Iain Murray and Zoubin Ghahramani. "A note on the evidence and Bayesian Occams razor". In: (2005).

[Ped12] Kaare Brandt Petersen Michael Syskind Pedersen. *The Matrix Cookbook*. 2012. URL: `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`.

Attempt on GP-LVM

[Law05]PLVM2005

This question is very similar to the Linear representation learning task however the co variance of the marginalisation of our likelihood in Q17 is now a GP process so **C** = **K** = k(**X**,**X**)

Objective function:

$$\begin{aligned}\mathcal{L}(\mathbf{X}) &= -\log p(\mathbf{Y}|\mathbf{X}) \\ &= \frac{ND}{2}\log(2\pi) + \frac{N}{2}\log(\mathbf{K}) + \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i)^T\mathbf{K}^{-1}(\mathbf{y}_i) \\ &= \frac{N}{2}(D\log(2\pi) + \log|\mathbf{K}| + Tr(\mathbf{Y}\mathbf{K}^{-1}\mathbf{Y}^T)\end{aligned}$$

where $\mathbf{Y}^T = \{\mathbf{y}_i \in \mathbb{R}^D : \forall 1 \leq i \leq N\}$,
$\mathbf{K} = k(\mathbf{X}, \mathbf{X})$

In order to calculate the gradient we need to find the derivative with respect to **X**. We do this using the chain rule with respect to the kernel **K**.

Gradient: The gradient matrix is a matrix **M** with entries $(\mathbf{M})_{ij}$ such that

$(\mathbf{M}_{ij}) = \frac{\partial\mathcal{L}(\mathbf{X})}{\partial\mathbf{X}_{ij}} = \frac{\partial\mathcal{L}(\mathbf{X})}{\partial\mathbf{K}}\frac{\mathbf{K}(\mathbf{X})}{\partial\mathbf{X}_{ij}}$

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