Modelling
Assignment in COMS30007 Machine Learning

Luke Storry LS14172

Louis Wyborn LW15771

Ahmer Butt AB15015

1 The Prior

Q1.1 It is sensible to assume the value of Y follows a Gaussian distribution due to the Central Limit Theorem. This states that if a large enough number of independent and identically distributed samples are drawn, their sum will tend towards a Gaussian distribution, regardless of what their original distribution is.

Q1.2 A spherical covariance matrix is proportionate to the identity matrix: the only non-zero elements are on the diagonal. This implies that there is no correlation between any of the random variables in x_i, ie. each variable is independent of every other. We are also stating that the variance of each x_i is the same, thus encoding the assumption that each element of x_i is identically distributed.

Q2 We wouldn't be able to assume the likelihood follows a Gaussian distribution. If we assume the Markov property is satisfied, such that the probability of each point depends only on the single point before it and not the entire sequence of points before, the likelihood would be:

$$p(Y|f,X) \ \propto p(X|f,Y) = p(y_1|f,x_1)\prod_{i=1}^{N-1} p(y_{i+1}|f,y_i,x_{i+1})$$

Q3

$$p(Y|X,W) = \prod_{i=1}^{N} \mathcal{N}(y_i|W^T x_i, \sigma^2)$$

Variables:

- X is a column vector of the observed variables.
- Y is a column vector of the target variables.
- W is a column vector of the coefficients of X.
- σ is a known constant.

Assumptions:

- The data is independent and identically distributed.
- The mapping between the variates X and Y is linear.
- The observations have been corrupted by additive Gaussian noise.

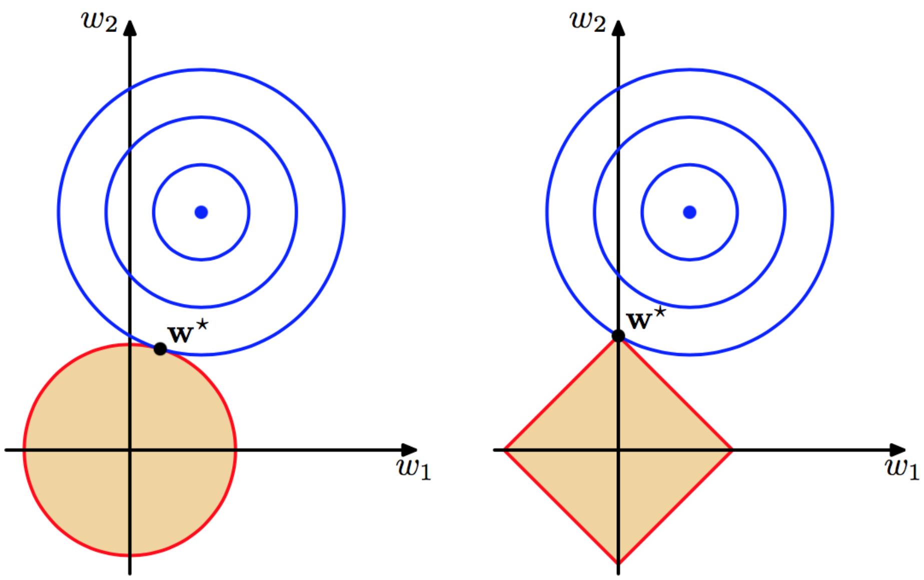
Q4 A prior is "conjugate" if it is chosen such that the posterior becomes a function in the same family as the prior.

As our likelihood $p(\mathrm{Y}|\mathrm{X},\mathrm{W})$ is Gaussian, we choose the prior $p(\mathrm{W})$ to be Gaussian, and also spherical, as we continue the same assumption that samples are i.i.d.

A conjugate prior also means that we do not have to compute the denominator (normalising factor) in Bayes Rule, but can instead just multiply the likelihood and prior, then easily calculate the parameters of the posterior distribution by identification as we already know which family it is part of. This ease of calculation is another motivating reason for choosing conjugate priors.

Q5 The goal of a preference in the space of the parameters is to allow complex models to be trained on limited data without severe over-fitting, by limiting the effective model complexity. This is called regularisation and can be carried out with different Lagrange multipliers to achieve different results. The solution found for X when using an L1 distance as a regularisation term will be sparser, meaning it will have fewer parameters than the solution found when using an L2 distance. This is because some of the coefficients in W will be driven to zero, such as in Figure 1 from Bishop [3], where the L2 regulariser would find a small, but non-zero w_1 value, but the L1 regulariser finds the value of w_1 to be zero.

Figure 1: Diagram from Bishop [3] comparing the solution for 2 parameters in W, with an L2 distance on the left and an L1 distance on the right.



Q6 To derive the posterior, we first note that:

$$\text{Posterior} \propto \text{Prior} \cdot \text{Likelihood}$$

Also as we know it is a conjugate, it must be a Gaussian distribution.

Likelihood $\sim \mathcal{N}(y_i|\mathrm{W}^T x_i, \sigma^2)$

Prior $\sim \mathcal{N}(\mathrm{W}_0, \tau^2\mathrm{I})$

Posterior $\sim \mathcal{N}(W^T|y_i, \mu, \sigma^2)$

We can then rewrite these in exponential form:

$$\text{Posterior} \propto \frac{1}{\sqrt{2\pi\tau^2\mathrm{I}}} e^{-\frac{(x-\mathrm{W}_0)^2}{2(\tau^2\mathrm{I})}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-(y_i|\mathrm{W}^T x_i))^2}{2\sigma^2}}$$

And due to proportionality:

$$e^{-\frac{(x-\mu)^2}{2(\sigma^2\mathrm{I})}} \propto e^{-\frac{(x-\mathrm{W}_0)^2}{2(\tau^2\mathrm{I})}} \cdot e^{-\frac{(x-(y_i|\mathrm{W}^T x_i))^2}{2\sigma^2}}$$

Because the exponential coefficients are normalising factors in the proportionality, they can be combined.

Then, after re-writing and rearranging the exponents, it is possible to identify the inverse covariance matrix and mean for the posterior distribution.

Q7 When we are dealing with a regression problem, we are aiming to estimate the function f which maps input variables in X to target variables in Y. When we choose a parametric approach, we make an explicit assumption about the functional form of f. For example, in the last section, we assumed that f is linear. The regression problem was thus reduced to estimating the parameters W, instead of the entire function f.

In a non-parametric approach, we do not make explicit assumptions about the functional form of f. We seek an estimate of the entire function f (as opposed to just scalar parameters W). By not assuming a particular functional form, our estimation may now give us a model that is of any shape: linear, polynomial, sinusoidal et cetera. The non-parametric approach completely avoids the danger of choosing a model that is very different to the true f. If we have no prior knowledge about the shape or form of family of the data, then the non-parametric approach avoids having to make a potentially incorrect assumption.

When we are interested in inference, i.e. understanding the way in which Y is affected by the input variables X_i, then using a simpler, less flexible, models makes this easier. Such a model is said to be more interpretable. When we use a parametric approach, we can force our estimate for f to be interpretable by selecting a simple model shape, e.g. linear. However, when we use a non-parametric method, this restriction is not there so the model we create is far more flexible and may result in an extremely complicated estimate of f. This will be difficult to understand and hence a more uninterpretable model.

Q8 In a Gaussian Process (GP), we have a potentially uncountably infinite number of input variables. Hence, we model the input space as the space of all functions, where a function can take any number of the input variables to produce the output variable. This space of functions is extremely large. The prior is an assumption we make that encodes our 'guess' of the probabilities of the possible functions that f may be. The prior is thus a probability distribution over f. This prior states that, given the hyper-parameter θ, then for every finite subset of the domain X, the probability of the definition of the function f is given by a multivariate normal distribution with mean zero and covariance matrix $k(X, X)$.

The function $k(X, X)$ is the kernel function. It is any function such that $k(x, y) = k(y, x)$ and the output covariance matrix K is a positive semidefinite matrix. The covariance matrix that the kernel function outputs has a significant affect on the functions the prior deems as more probable. The values not on the diagonal represent the correlation between the ith and jth elements of the random variables in X_1 and X_2. If the value is zero, then the two values for the random variables at the ith and jth indices are independent. However, if the value is nonzero, then the value of one influences the other. So, if the value if strongly positive, then the value of the function f will be very similar at i and j.

Q9 By definition, this prior is a Gaussian distribution. A Gaussian distribution never assigns a probability of zero to any input value(s). Therefore, this prior states that every function is possible, but the probability may be negligible. In fact, 95% of probability density lies within two standard deviations of the mean.

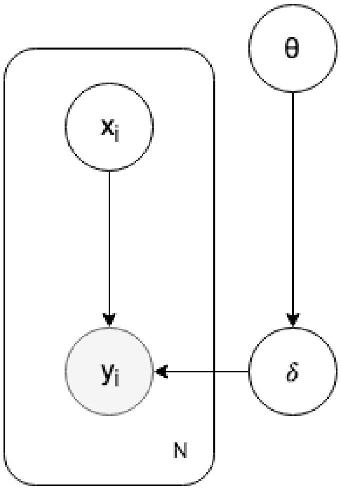
Q10 We can formulate the joint distribution using the chain rule.

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

Assumptions

- Our additive noise ϵ are independent and identically distributed.
- We have a Gaussian prior on f as given above in the question text.

Figure 2: Graphical Model



Q11 We have marginalised over the space of all functions f that connect the input data to the output data. By doing this, we get the probability of the output variable Y conditioned on only the input variable X and hyper-parameters θ. Therefore, we are removing a condition between the input and output random variables. The uncertainty we encode via our selected prior distribution can now be updated with data via applying Bayes' Rule, $p(\mathrm{X}|\mathrm{Y},\theta) = p(\mathrm{Y}|\mathrm{X},\theta)p(\mathrm{X}|\theta)$. The presence of θ on the left hand side of the equation after marginalisation implies that it was not marginalised out. It is in fact a hyper-parameter to the prior distribution over the input variable, X. Therefore, we do not necessarily wish to marginalise it out when we want to formulate $p(\mathrm{Y}|\mathrm{X})$.

Q12 See Figure 4.

Figure 3: Prior Distribution

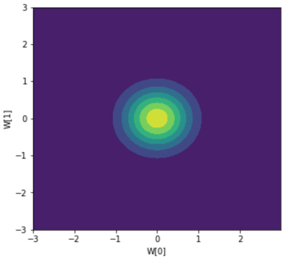
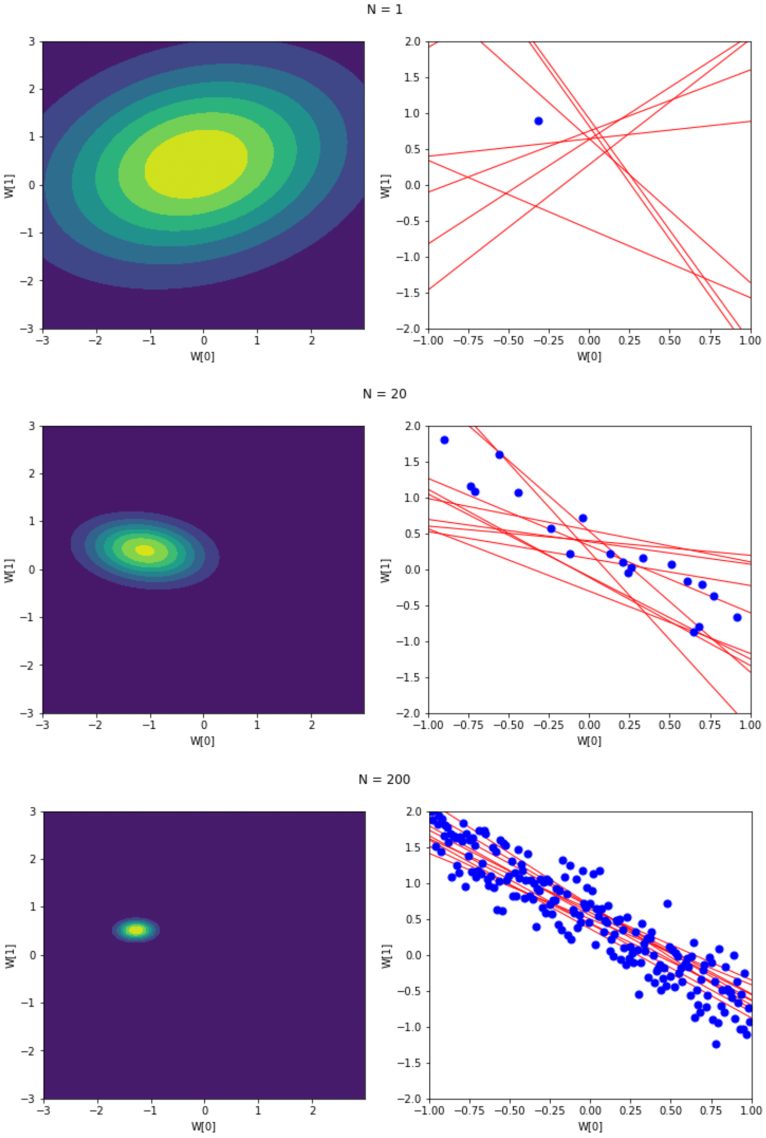


Figure 4: Visualisations of a series of posterior distributions, with corresponding plots of N samples and 10 functions



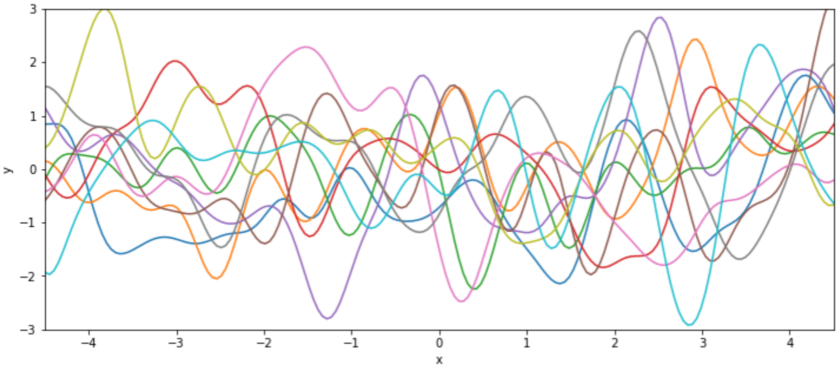
Q12.3 Analysis of Figure 3:

With more data, the contours representing the Gaussian distribution become smaller and more precise. This means that the predicted gradients and intercepts are closer together. As more data points are added, the Gaussian distribution converges around the point on the graph that represents the solution for the slope and y intercept.

The lines that are created from these values vary less and less, moving closer together and pointing more in the same direction. This is desirable behavior as we gain better and better estimates for the original values of W.

Q13

Figure 5: 10 random samples from the GP prior with a length-scale of 0.5



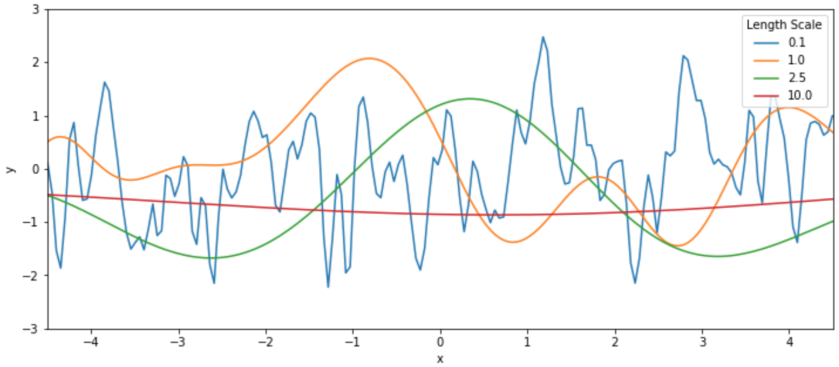
The kernel function maps an input space, i.e. a selection of random variables x_i and x_j, to a real-valued covariance matrix. The kernel function we have chosen to use is the squared exponential. The output covariance matrix is defined such that K_{ij} is the result $k(x_i, x_j)$. In other words, take the function k, take all the pairs of points from x_i and x_j, evaluate for each and put in a matrix.

The squared exponential kernel produces a positive covariance matrix such that the value at K_{ij} is proportional to the difference between indices i and j. This is determined in the exponent in the definition above. Therefore, the resulting matrix will have highest values on the diagonal. As you move away from the diagonal in any direction, the values will decrease. So, the probability of two random variables x_i and x_j being close together is greater when the difference between i and j is smaller, as the correlation between the two (dependent) variables increases.

The kernel function output is also subject to the hyper-parameters σ and l. The length-scale, l, divides the exponent of the exponential. Therefore, w.l.o.g., as the value of l decreases, then the values in the covariance matrix decrease faster. Therefore, the correlation between the two random variables and index i and j decreases with their distance faster. Therefore, each point on the output function f is less likely to be close to its surrounding points. Hence, the line drawn by the function f appears more 'wiggly'.

There are some implications of this that one should note. We can see that decreasing l means the model f we generate becomes more complex. This risks over-fitting to the training data. Equally, as the line becomes more 'wiggly', it becomes less accurate to extrapolate. In general, one should not extrapolate more than l units away from the training data.

Figure 6: Random samples from GP priors with different length-scales



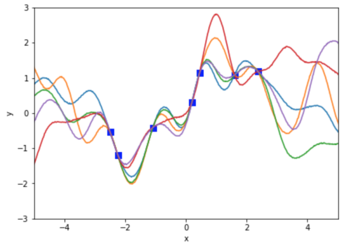
The length-scale encodes our assumption about how smooth the original function used to generate the data is. A higher length-scale encodes a belief in a smoother original function. This can also be interpreted as encoding our assumption about the correlation between each random variable and its neighbouring random variables.

Q14

Compared to the samples from the prior, samples drawn from the posterior distribution are much more likely to pass closer to the observed data points (See Figures 5 and 7).

Therefore the sample functions drawn are more likely to closer represent the original $sin()$ curve used to generate the data. This is desirable because as we add more data points we can update our beliefs such that the posterior distribution comes closer to representing the original function.

Figure 7: N=5 samples from GP posterior



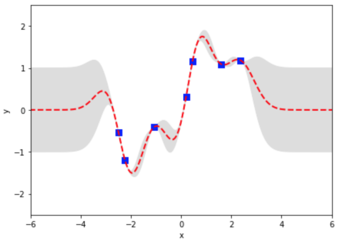
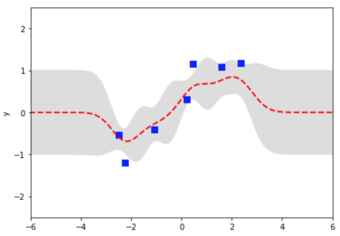


Figure 8: Predictive mean and variance

Figure 9: Added noise variance to kernel



Adding a diagonal covariance matrix to the squared exponential increases the variance on the data points without increasing the covariance. As can be seen from Figure 9, this results in a larger variance from the mean in the Gaussian Process posterior distribution, so lines are “allowed” to be slightly further from the mean, which can be very useful with noisy data. Some definitions[1] of squared exponential kernels even include this added diagonal matrix as a “noise variance” parameter to specify how much noise is expected to be present in the training data.

2 The Posterior

Q15 Assumptions are what you already know or believe to be true, whereas a preference is what you want to be true. It can also be seen that both of these are ways of encoding how you wish the model to look, by giving it a prior shape to form. A Bayesian machine learner should be careful not to too easily choose priors to be of a certain family of distributions purely in order to make the calculations easier.

Q16 By choosing a spherical gaussian as the prior, we have encoded our assumption that the value of the random variable x is normally distributed, with a mean vector of **0**. By choosing the normal distribution with a variance of ***I***, we have encoded our assumption that no components of the vector x are correlated (independence) and that the variance of each component is equal.

Q17 The *MxM* latent space , which we may refer to as the principal subspace, given by X is a lower-dimensional subspace of the *DxD, (M $<$ D)* data space of Y. We will assume that the space given by Y is a linear transformation of the space given by X. Therefore, the conditional distribution of the observed variable Y on latent variable X is given by,

$$p(\mathrm{Y}|\mathrm{X}) = \mathcal{N}(\mathrm{Y}|\mathrm{WX} + \mu, \sigma^2\mathrm{I})$$

where W is a *DxM* matrix such that it's columns span a linear subspace within the data space that corresponds to the principal subspace. μ is the mean of the observed data. σ^2 defines the variance of the distribution. There are no correlations. So, we interpret the observations as being generated from a linear transformation of the latent variable plus additive Gaussian noise.

$$\mathrm{Y} = \mathrm{WX} + \mu + \epsilon, \quad where \quad \mathcal{N}(\epsilon|0, \sigma^2\mathrm{I})$$

Now, because we are using a linear-Gaussian model, the marginal distribution is also Gaussian,

$$p(\mathrm{Y}|\mathrm{W}) = \mathcal{N}(\mathrm{Y}|\mu, \mathrm{C})$$

where C is a *DxD* covariance matrix.

Therefore, rather than computing an integral over a large data space, we can derive the parameters for this Gaussian. Then, we have the marginal distribution.

$$\mathbb{E}[\mathrm{Y}] = \mathbb{E}[\mathrm{WX} + \mu + \epsilon] = \mu$$

Now derive the covariance.

$$cov[\mathrm{Y}] = \mathbb{E}[(\mathrm{WX} + \epsilon)(\mathrm{WX} + \epsilon)^T]$$

Now, expanding out and using the fact that X and ϵ are independent,

$$\mathbb{E}[\mathrm{WXX^TW^T}] + \mathbb{E}[\epsilon\epsilon^T]$$

As W is constant and X is a zero mean Gaussian with covariance of I,

$$\mathrm{WW^T} + \sigma^2\mathrm{I} = \mathrm{C}$$

Therefore, we can now write the marginal distribution as

$$p(\mathrm{Y}|\mathrm{X}) = \mathcal{N}(\mu, \mathrm{WW^T} + \sigma^2\mathrm{I})$$

Q18 Maximum Likelihood Estimation can have a tendency to over-fit data, so the regularization offered by the prior in Maximum A Posteriori can alleviate that, lessening the impact of observed data on the model.

Type-II ML chooses a function that balances both matching the assumption (encoded by marginalizing out certain variables) and being close to the observed data, which is a good middle ground for occasions where the complete marginalisation of all variables is analytically intractable.

As the amount of data tends towards infinity, the results from ML and MAP converge to being equal. This is due to the prior in MAP becoming relatively weaker in relation to the larger amount of observed data.

Because the example is simple (only using X, Y and W), the effect of integrating out X is the same as using MAP:

$$argmax_w p(Y|X,W)p(W) = argmax_W \int p(Y|X,W)p(X)dX$$

Q19

We want to maximise the negative log likelihood of the observed data given the parameters

$$-\log(p((\mathbf{Y}|\mathbf{X})) = \text{constant} + \log|\mathbf{C}| + tr(\mathbf{Y}\mathbf{C}^{-1}\mathbf{Y}^T) = \mathcal{L}(\mathbf{W})$$

where **C** is the covariance matrix discussed in **Q17**. First, we will find the derivative of the matrix **C** with respect to W_{ij}.

For any i, j,

$$\frac{\partial C}{\partial W_{ij}} = \frac{\partial W W^T}{\partial W_{ij}} = W\frac{\partial W^T}{W_{ij}} + \frac{\partial W}{\partial W_{ij}}W^T = WJ_{ij} + J_{ji}W^T$$

where J_{ij} is the matrix which has all zero entries except $(J_{ij})_{ij} = 1$. Now, we will differentiate $\mathcal{L}(W)$ with respect to W.

$$\frac{\partial \mathcal{L}(W)}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}} log(|C|) + \frac{\partial}{\partial W_{ij}} tr(YC^{-1}Y^T)$$

We will differentiate these two terms separately. First, the left hand term,

$$\frac{\partial}{\partial W_{ij}} log(|C|) = tr(C^{-1}\frac{\partial C}{\partial W_{ij}}) = tr(C^{-1}(WJ_{ij} + J_{ji}W^T))$$

Now, differentiate the second term.

$$\frac{\partial}{\partial W_{ij}} tr(YC^{-1}Y^T) = tr(\frac{\partial}{\partial W_{ij}} YC^(-1)Y^T) = tr(\frac{\partial}{\partial W_{ij}} YCY^T) = tr(\frac{\partial}{\partial C}(YC^(-1)Y^T)\frac{\partial^{-1}}{\partial W_{ij}}$$

$$= tr(YY^T\frac{\partial C^{-1}}{\partial W_{ij}}) = tr(YY^T(-C^{-1}\frac{\partial C^(- 1)}{\partial W_{ij}}C^{-1})) = tr(YY^T(-C^{-1}(C^{-1}(WJ_{ij}+J_{ji}W^T))C^{-1}))$$

Now, putting these together we get,

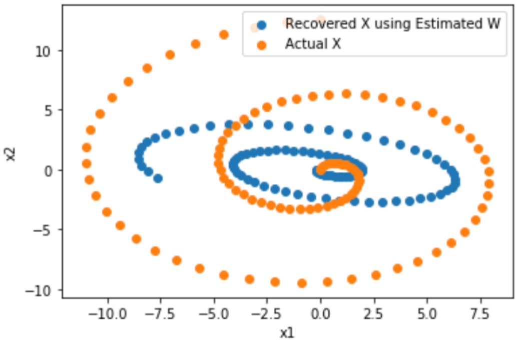
$$\frac{\partial \mathcal{L}(W)}{\partial W_{ij}} = tr(C^{-1}(WJ_{ij} + J_{ji}W^T)) + tr(YY^T(-C^{-1}(C^{-1}(WJ_{ij} + J_{ji}W^T))C^{-1}))$$

Q20 Y is parametised by f, which is parametised by X and θ. Therefore, to marginalise out the latent variable (or more generally with a Gaussian process, the random variables indexed by the input indices), we would have to marginalise out f anyway. The size of the latent variable space grows combinatorially as the data space grows, which means that the solution will easily become intractable to compute.

Q21 In Figure 10, we have plotted both the actual X and our recovered X as two-dimensional representations.

Although both spirals, there are some discrepancies, for a variety of reasons. The observed data variable Y is a linear mapping of the latent variable x, plus some additive Gaussian noise. Gradient descent will not necessarily find the perfect minimum of the objective function (negative log likelihood). Therefore, the linear mapping A that is generated by the process will not perfectly generate Y from X. The resulting output matrix W also relies on the initial guess, which we randomly choose by sampling a zero-mean Gaussian variable.

Figure 10: Result of performing PCA using Gradient Descent to recover generating parameters X from Y



3 The Evidence

Q22 This assumption implies that all possible datasets are equally likely.

This could be considered the simplest possible model because the model has no fixed parameters and thus a very simple shape: a straight horizontal line. This means that performing prediction or inference is a simple task.

Alternatively one could consider this the most complex model because it assigns many different types of behaviours similar probability.

Q23 Choosing logistic functions for the models restricts the distribution to following a logistic regression, even if the data could be better suited to a different distribution.

Each successive model can realise the previous models by simply setting some of its parameters to 0. Thus Model 3 is the most flexible (and arguably the most complex), as it can emulate all the other models. This flexibility has a downside: because it has a higher evidence for a wider range of more complex datasets, there is less probability mass available for the more simple datasets. Model 1 is much more simple, so although it cannot model the more complex datasets, it can concentrate its probability mass on the simple datasets that it can model.

Model 0 is most suited to model a data set with no clear linear boundaries. Model 1 is a simple model that can capture vertical but not horizontal decision boundaries, so is the best model for data with vertical decision boundaries. Model 2's added parameter over Model 1 gives it rotational invariance - the ability to model datasets with horizontal as well as vertical boundaries. Model 3 is the only model to have a bias term (θ_3^3), which allows it to model very unequal distributions of data in a dataset, and detect decision boundaries that are offset from the origin.

Q24 The choice of a Gaussian distribution for the prior allows infinitely many potential values for the parameters. The large variance encodes our large uncertainty about the parameters. Also, due to the covariance matrix Σ being diagonal, we are assuming that all of the parameters are independent. Moreover, the covariance matrix is also isotropic, resulting in all the parameters of the model being both independent and identically distributed.

Setting $\mu = 0$ implies that the parameters are equally likely to be positive or negative, which further implies that our model encodes no preference between each binary observation being -1 or 1 in the data.

Q30 In this assignment we have learnt how to formulate and compute parametric and non-parametric Bayesian Learning techniques for regression and feature recovery. We learned how to exactly encode our beliefs by selecting appropriate prior distributions and paramatising them over hyperparameters. We also learnt how to update these beliefs by observing data in order to form a posterior distribution from which we can predict values for data we haven't yet observed. We looked at out how different model parameters affect the evidence produced for a given dataset. Furthermore this showed us how the evidence allows us to automatically select the appropriate model complexity for a given dataset.

We have gained an appreciation for how powerful probabilistic models can be, as they allow us to encode the uncertainty of our beliefs and update these beliefs with data. This is a very intuitive approach, and is how we as humans naturally understand the world.

References

[1] Ariadne. *Covariance functions*. URL: http://evelinag.com/Ariadne/covarianceFunctions.html.

[2] Katherine Bailey. *Gaussian Processes for Dummies*. URL: http://katbailey.github.io/post/gaussian-processes-for-dummies/.

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

[4] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN: 1461471370, 9781461471370.

[5] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020, 9780262018029.

[6] Iain Murray and Zoubin Ghahramani. *A note on the evidence and Bayesian Occam's razor*. Tech. rep. GCNU-TR 2005-003. Gatsby Computational Neuroscience Unit, University College London, 2005.

[7] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN: 026218253X.