Coursework 1 - Modelling

Jonas Osborn jo14944 | **Tristan Saunders** ts16802 | **Corin Varney** cv14985

1 The prior

1.1 Theory

Question 1 *1.* Choosing a gaussian is a sensible thing to do because it makes calculation much easier, Gaussians are quite easily multiplied and integrated which are things you often have to do in machine learning contexts e.g. multiplying to compute a joint distribution or integrating to marginalise out a variable. Choosing a Gaussian likelihood is also sensible as it fits with our model because we often assume observation errors to be independent and identically distributed and therefore Gaussian due to the Central Limit Theorem.

2. Choosing a spherical covariance matrix for the likelihood means that we are assuming the different dimensions of y_i to be independent and identically distributed. As they are independent they do not covary with each other and so the covariance matrix is diagonal. As they are identically distributed they all have the same variance. Therefore the covariance matrix is spherical.

Question 2

$$\begin{aligned} p(\boldsymbol{Y}|f,\boldsymbol{X}) &= p(y_1,\ldots,y_{N-1},y_N|f,\boldsymbol{X}) \\ &= p(y_N|y_{N-1},\ldots,y_1,f,\boldsymbol{X})p(y_{N-1}|y_{N-2},\ldots,y_1,f,\boldsymbol{X})\ldots p(y_1|f,\boldsymbol{X}) \end{aligned}$$

Question 3 We assume that the data is independent and so the joint probability is a product.

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W}) &= \prod_i^N p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{W}) \\ &= \prod_i^N \mathcal{N}(\boldsymbol{W}\boldsymbol{x}_i,\sigma^2\boldsymbol{I}) \end{aligned}$$

Question 4 A conjugate prior is one that is conjugate to the posterior, meaning they are in the same family of distributions. A conjugate prior is useful as it gives a closed-form solution for the posterior. If we didn't choose a conjugate prior then numerical integration may be necessary to calculate the posterior, which may mean the solution is potentially intractable. The conjugate prior for a Gaussian posterior is a Gaussian.

Question 5 Just as encoding the prefrence in a L_2 norm is equivalent to having a Gaussian prior, encoding the preference using a L_1 norm is equivalent to having a Laplace prior. This is because the Laplace distribution estimates median rather than the mean estimated by the Guassian and median minimises the L_1 norm and mean the L_2.

The shape of the laplace distribution's probability density function, with it's higher peak around zero compared to the probability density function of a Gaussian means that more co-efficients are likely to be equal to zero and this leads to a sparser model than those produced by a Gaussian prior.

Question 6

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) &= \frac{1}{Z} p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W}) p(\boldsymbol{W}) \\ &= \frac{1}{Z} \mathcal{N}(\boldsymbol{W}\boldsymbol{X}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{W}_0, \tau^2 \boldsymbol{I}) \end{aligned}$$

$\frac{1}{Z}$ is the normalising constant to ensure that the posterior is a probability density function by making the area under the graph equal to 1, this constant is called the evidence. We will ignore it for now.

By the probability density function of the multivariate normal distribution, we have:

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) \propto \frac{1}{\sqrt{(2\pi)^N \sigma^2}} \exp\Big(-\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}) \Big) \\ \cdot \frac{1}{\sqrt{(2\pi)^N \tau^2}} \exp\Big(-\frac{1}{2\tau^2} (\boldsymbol{W} - \boldsymbol{W}_0)^T (\boldsymbol{W} - \boldsymbol{W}_0) \Big) \end{aligned}$$

We ignore the normalising constants as we re-normalise with Z and then combine and multiply out the exponents.

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) &\propto \exp\Big(-\frac{1}{2\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}) - \frac{1}{2\tau^2} (\boldsymbol{W} - \boldsymbol{W}_0)^T (\boldsymbol{W} - \boldsymbol{W}_0) \Big) \\ &\propto \exp\Big(-\frac{1}{2\sigma^2} (\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{W} + \boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}) \\ &\quad - \frac{1}{2\tau^2} (\boldsymbol{W}_0^T\boldsymbol{W}_0 - 2\boldsymbol{W}^T\boldsymbol{W}_0 + \boldsymbol{W}^T\boldsymbol{W}) \Big) \end{aligned}$$

We know the posterior will be Gaussian as both the likelihood and prior are, so we can assume it will take the form: $\exp\big((\boldsymbol{W} - \mu)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{W} - \mu)\big)$. If we multiply the exponent out we get a quadratic so we try and make the posterior we have look like this quadratic.

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) \propto \exp\Big(&-\frac{1}{2\sigma^2}\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W} - \frac{1}{2\tau^2}\boldsymbol{W}^T\boldsymbol{W} && \text{quadratic term} \\ &+ \frac{1}{\sigma^2}\boldsymbol{W}^T\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}^T\boldsymbol{W}_0 && \text{mixed term} \\ &- \frac{1}{2\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{2\tau^2}\boldsymbol{W}_0^T\boldsymbol{W}_0 \Big) && \text{constant term} \end{aligned}$$

By re-arranging and completing the square we can find both $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}^{-1}\mu$ in the quadratic and mixed terms respectively.

$$\begin{aligned} p(\boldsymbol{W}|\boldsymbol{X},\boldsymbol{Y}) \propto \exp\Big(&-\frac{1}{2}\boldsymbol{W}^T \overbrace{\Big(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\Big)}^{\boldsymbol{\Sigma}^{-1}} \boldsymbol{W} && \text{quadratic term} \\ &+ \boldsymbol{W}^T \underbrace{\Big(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0\Big)}_{\boldsymbol{\Sigma}^{-1}\mu} && \text{mixed term} \\ &- \frac{1}{2\sigma^2}\boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{2\tau^2}\boldsymbol{W}_0^T\boldsymbol{W}_0 \Big) && \text{constant term} \end{aligned}$$

From these we can easily find $\boldsymbol{\Sigma}$ and μ.

$$\begin{aligned}\boldsymbol{\Sigma} &= \left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1} \\ \boldsymbol{\Sigma}^{-1}\mu &= \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0 \\ \mu &= \boldsymbol{\Sigma}\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0\right)\end{aligned}$$

Thus we have our posterior as a Guassian with our values for $\boldsymbol{\Sigma}$ and μ:

$$\begin{aligned}p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{W}-\mu)^T\Sigma^{-1}(\boldsymbol{W}-\mu)\right) \\ &\propto \mathcal{N}\left[\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{W}_0\right), \left(\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1}\right]\end{aligned}$$

Question 7 Parametric models assume that the distribution the data comes from is based on a finite, fixed set of parameters and models future predictions based off these parameters, they capture everything there is to know about the data. Non-parametric models do not make such assumptions about the model structure and instead infer structure from the data, they have parameters but these are not fixed in advance and there can be an infinite set of parameters.

Non-parametric models are more flexible and can represent a wider variety of data and will represent the data better if the assumptions made in the parametric model are incorrect but are less precise and accurate than parametric methods if the right assumptions are made.

Parametric models are often easier interpreted as they are simpler to transcribe and are also often faster to compute due to lacking the complexity and flexibility of the non-parametric models.

Question 8 As we use a Gaussian process we define this prior over functions and want our prior to put some constraints on the space of functions. The fact that its a Gaussian process means that for an arbitary set of points $x_i, \ldots, x_j$ we assume that $p(f_i), \ldots, f(x_j))$ is jointly Gaussian with mean 0 and covariance function k, from this we have the equation for our prior:

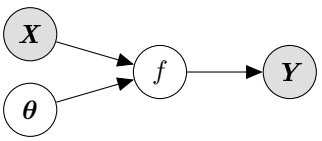
$$p(f|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(0, k(\boldsymbol{X}, \boldsymbol{X}))$$

The covariance function k allows us to set constraints from our assumptions about the mapping f. For example, our assumption about the functions smoothness that if x_i and x_j are similar then we expect f_i and f_j to be similar too, with certain kernel functions for f we can only sample functions that have sufficient smoothness by ensuring that constraint.

Question 9 Gaussian processes encode all possible functions for which inferring a posterior distribution from the data is tractable.

Question 10

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



- $\boldsymbol{X}$ and $\boldsymbol{\theta}$ are independent
- f is conditionally dependent on both $\boldsymbol{X}$ and $\boldsymbol{\theta}$
- $\boldsymbol{Y}$ is conditionally dependent on f and conditionally independent of $\boldsymbol{X}$ and $\boldsymbol{\theta}$

Question 11 This marginalization shows the likelihood of the data we have observed over the function we are testing which is constrained by the prior.

There are two sources of uncertainty here, that associated with f in the prior and that associated with ϵ in the likelihood, these are independent and as such are merged by simply adding to form the covariance of the marginal likelihood Gaussian.

Leaving the θ on the left-hand side of the expression implies that we still have specific hyperparameters rather than undefined ones. It remains throughout the integral.

1.2 Practical

Question 12 Figure 1 represents our prior assumption over $\boldsymbol{W}$. Our prior assumption states that the distribution over $\boldsymbol{W}$ is a Gaussian with a mean of 0 and the identity matrix as its covariance. As discussed in question 3 the likelihood is a Gaussian, therefore choosing a Gaussian prior allows us to obtain a Gaussian posterior.

In Figures 2, 3 and 4 we can see how our assumptions about $\boldsymbol{W}$ change as we observe more data. Making the distribution over $\boldsymbol{W}$ a posterior or our updated belief. These are paired with samples from our posterior that allow us to plot some sample functions. It is clear that the more data we observe, the more the uncertainty in our posterior reduces, giving us a posterior that approaches the exact values of the actual parameters $\boldsymbol{W}$ (displayed as a white cross) along with increasingly accurate sample functions.

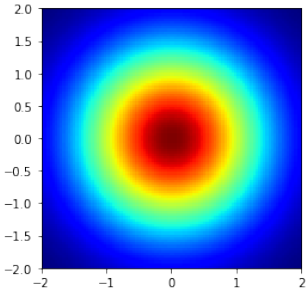
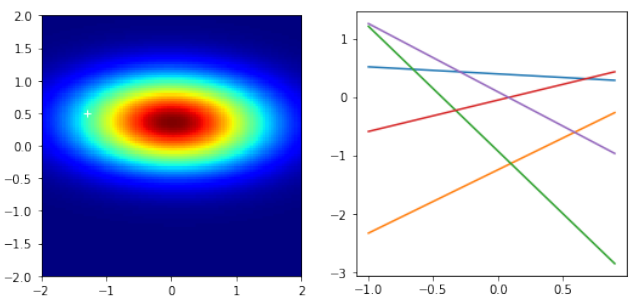
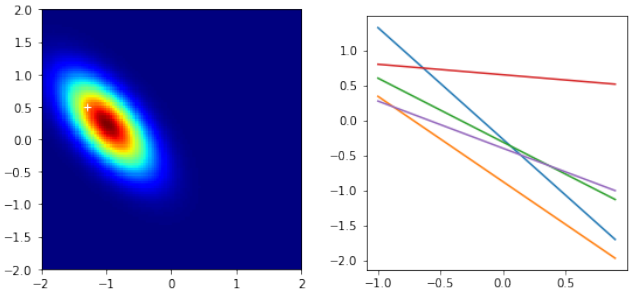


Figure 1: Prior for linear regression



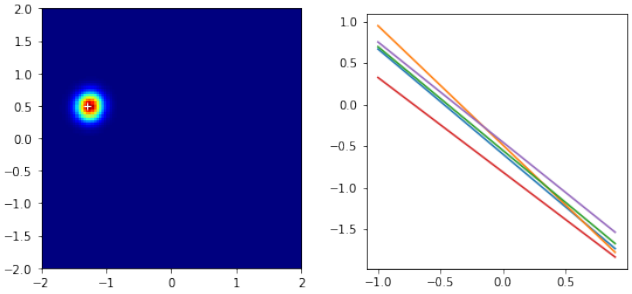
(a) Posterior (b) Samples of y

Figure 2: Linear regression - 1 observation



(a) Posterior (b) Samples of y

Figure 3: Linear regression - 2 observations



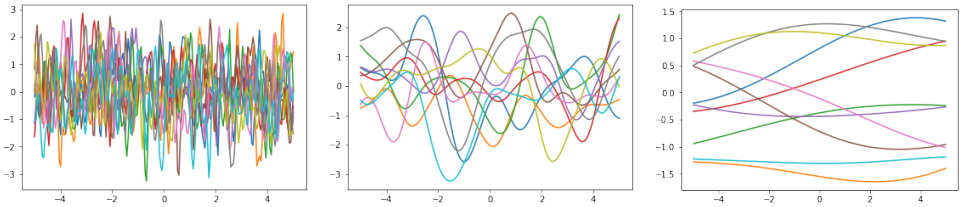
(a) Posterior (b) Samples of y

Figure 4: Linear regression - 15 observations

Question 13 The covariance function, or kernel, encodes all the assumptions about the form of functions we are modelling, and often represents some form of "distance" or similarity between the data.

Figure 5 shows samples from our GP-Prior created with a squared exponential covariance function. The length scale allows us to put constraints on how smooth the functions are. Small lengthscale values such as the one in Figure 5a characterize functions that change quickly, whereas larger values such as the one in 5c characterize functions that change slowly.

The lengthscale therefore encodes our assumption of how smooth we think the underlying function will be.



(a) length scale = 0.1 (b) length scale = 1.0 (c) length scale = 10

Figure 5: GP-prior samples

Question 14 Figure 6 shows the predictive posterior samples, our updated prior after having observed data. While our prior contained functions that corresponded to our general prior assumptions, the posterior shows only the set of constrained functions that pass by the observed data points.

We can observe from the samples in Figure 6 and the variance in Figure 7a (displayed as red shading at one standard deviation from the mean) that the uncertainty drops as we

observe data but rises again as we move away from the observed data in the centre. This is due to our squared exponential covariance matrix that encodes our assumption that if two points x_1 and x_2 are close together, we expect their corresponding y values to be close too. Therefore the further away a new x value is from the x values of observed data points the more uncertain we are about its y value.

Adding a diagonal covariance matrix to the covariance function encodes a general uncertainty in our observations. This leads to a smoother function that approximates the data points rather than strictly interpolating between them. As we can see in Figure 7b this leads to a much larger standard deviation but a line that much closer approximates the underlying sine function.

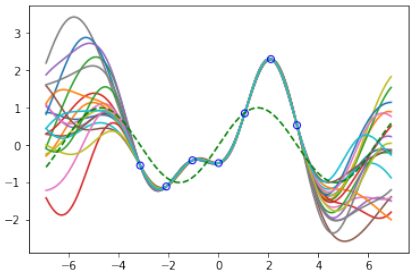
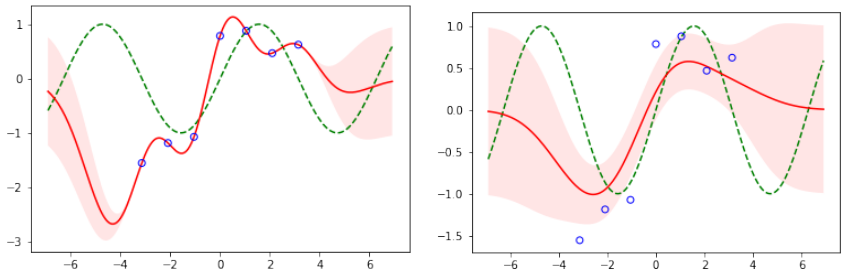


Figure 6: Predictive posterior samples



(a) Just kernel co-variance matrix (b) With diagonal co-variance matrix

Figure 7: Predictive mean and variances

2 The posterior

2.1 Theory

Question 15 Specifiying a prior assumption over $\boldsymbol{X}$ allows us to encode our preference about the nature of the properties that $\boldsymbol{X}$ should have. This assumption also constrains the variable $\boldsymbol{W}$ as $\boldsymbol{W}$ and $\boldsymbol{X}$ have a simple relationship.

Question 16 We have encoded our preference that the Gaussian have zero mean and our assumption that all the dimensions in each variable of $\boldsymbol{X}$ are independent and identically distributed as we have used a Gaussian with an identity matrix as its covariance matrix.

Question 17 We assume additive gaussian noise for a linear non-parametric Gaussian process and so our likelihood is as follows:

$$p(y_i|x_i,\, \boldsymbol{W}) = \mathcal{N}(y_i|\, \boldsymbol{W}\, x_i, \sigma^2 \boldsymbol{I})$$

Our marginalisation is the product of two Gaussians which is itself another Gaussian.

$$\begin{aligned} p(y_i|\, \boldsymbol{W}) &= \int p(y_i|x_i,\, \boldsymbol{W}) p(x_i)\, dx \\ &= \int \mathcal{N}(y_i|\, \boldsymbol{W}\, x_i, \sigma^2 \boldsymbol{I})\, \mathcal{N}(0, \boldsymbol{I})\, dx \end{aligned}$$

By replacing these Gaussians with their probability density function and combining we can form a single exponential, by then completing the square with regard to x we can integrate x out to form a new exponent for the marginal which by re-arranging as we did in Question 6 we can find the mean 0 and covariance $\boldsymbol{W}\,\boldsymbol{W}^T+\sigma^2\boldsymbol{I}$ thus making our marginal distribution:

$$p(y_i|\,\boldsymbol{W}) = \mathcal{N}(y_i|0,\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})$$

Question 18 *1.* The MAP estimate equals the mode of the posterior distribution while ML simply finds the parameters that maxmimise the likelihood. ML can be seen as a MAP that simply assumes a uniform prior distribution, or MAP can be seen as ML that employs a modified optimization objective by regularizing with the prior.

2. As we observe more data the prior becomes less important relatively and the MAP estimate tends towards the ML estimate.

3. We can ignore the denominator in Eq. 8 because the integral is constant with respect to $\boldsymbol{W}$ and so has no bearing on the $\arg\max_{\boldsymbol{W}}$.

Question 19 *1.* We have from Question 17 that:

$$p(\,\boldsymbol{Y}|\,\boldsymbol{W}) = \prod_{i=1}^{N}\mathcal{N}(y_i|0,\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})$$

then the objective function is derived as follows using log rules:

$$\begin{aligned}
\mathcal{L}(\,\boldsymbol{W}) &= -\log\Big(\prod_{i=1}^{N}\mathcal{N}(y_i|0,\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})\Big)\\
&= -\sum_{i=1}^{N}\log\big(\mathcal{N}(y_i|0,\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})\big)\\
&= -\sum_{i=1}^{N}\log\Big(\frac{1}{\sqrt{2\pi|\Sigma|}}\Big) + \log\Big(\exp\Big(-\frac{1}{2}y_i^T\Sigma^{-1}y_i\Big)\Big)\\
&= -\sum_{i=1}^{N}\log\frac{1}{(2\pi)^{\frac{D}{2}}|\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I}|^{\frac{1}{2}}} + \log\Big(\exp\Big(-\frac{1}{2}y_i^T(\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}y_i\Big)\Big)\\
&= -\log\big((2\pi)^D|\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I}|\big)^{-\frac{N}{2}} - \sum_{i=1}^{N}-\frac{1}{2}y_i^T(\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}y_i\\
&= \frac{N}{2}\Big(D\log 2\pi + \log(|\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I}|)\Big) + \frac{1}{2}\sum_{i=1}^{N}y_i^T(\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}y_i
\end{aligned}$$

From the fact that $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{B}\,\boldsymbol{A})$ and $\mathrm{Tr}(\boldsymbol{C} = \boldsymbol{c})$ for the dimensions in our sum we can evaluate it as follows:

$$\mathcal{L}(\,\boldsymbol{W}) = \frac{N}{2}\Big(D\log 2\pi + \log(|\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I}|) + \mathrm{Tr}\big((\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}\,\boldsymbol{Y}\,\boldsymbol{Y}^T\big)\Big)$$

2. To find the gradient of $\mathcal{L}$ we will look at each term in turn:

$$\mathcal{L}(\,\boldsymbol{W}) = \frac{N}{2}\Bigg(\overbrace{\mathrm{Tr}\big((\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}\,\boldsymbol{Y}\,\boldsymbol{Y}^T\big)}^{A} + \overbrace{\log(|\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I}|)}^{B} + \overbrace{D\log 2\pi}^{C}\Bigg)$$

C is a constant term so we shall discard that. To find the derivative of B we use the rules that $\partial(\log(\det(\boldsymbol{X}))) = \mathrm{Tr}(\boldsymbol{X}^{-1}\partial\boldsymbol{X})$ and that $\sigma^2\boldsymbol{I}$ is constant with respect to $\boldsymbol{W}$ and so evaluate B to:

$$\frac{\partial B}{\partial\,\boldsymbol{W}} = \mathrm{Tr}\Big((\,\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}\partial(\,\boldsymbol{W}\,\boldsymbol{W}^T)\Big)$$

We also have the rules that $\partial(\boldsymbol{X}\,\boldsymbol{Y}) = (\partial\boldsymbol{X})\,\boldsymbol{Y} + \boldsymbol{X}(\partial\boldsymbol{Y})$ and that $\frac{\partial \boldsymbol{X}}{\partial \boldsymbol{X}_{ij}} = \boldsymbol{J}^{ij}$ where $\boldsymbol{J}$ is the single-entry matrix, having 1 at (i,j) and 0 elsewhere and so can reduce B to:

$$\frac{\partial B}{\partial \boldsymbol{W}_{ij}} = \mathrm{Tr}\left((\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{J}^{ij}\,\boldsymbol{W}^T + \boldsymbol{W}\,\boldsymbol{J}^{ijT})\right)$$

Using the earlier rules along with the rule that $\partial(\mathrm{Tr}(\boldsymbol{X})) = \mathrm{Tr}(\partial\boldsymbol{X})$ we can find the derivative of A:

$$\begin{aligned}\frac{\partial A}{\partial \boldsymbol{W}} &= \mathrm{Tr}\left(\partial\Big(\boldsymbol{Y}(\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}\,\boldsymbol{Y}^T\Big)\right)\\ &= \mathrm{Tr}\left(\boldsymbol{Y}\,\boldsymbol{Y}^T\partial\Big((\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}\Big) + (\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}\Big(\partial(\boldsymbol{Y}^T\,\boldsymbol{Y})\Big)\right)\end{aligned}$$

As $\partial(\boldsymbol{X}^{-1}) = -\boldsymbol{X}^{-1}(\partial\boldsymbol{X})\boldsymbol{X}^{-1}$ we can reduce A to:

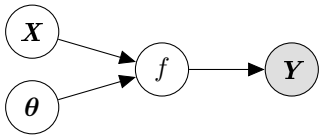
$$\frac{\partial A}{\partial \boldsymbol{W}} = \mathrm{Tr}\left(\boldsymbol{Y}\,\boldsymbol{Y}^T\partial\Big((\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}\Big)\right)$$

$$\frac{\partial B}{\partial \boldsymbol{W}_{ij}} = \mathrm{Tr}\left(\boldsymbol{Y}\,\boldsymbol{Y}^T\Big(-(\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{J}^{ij}\,\boldsymbol{W}^T + \boldsymbol{W}\,\boldsymbol{J}^{ijT})(\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})\Big)\right)$$

And so by combining our terms A and B we have the gradient for our log likelihood $\mathcal{L}$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{ij}} = \frac{N}{2}\,\mathrm{Tr}&\left(\boldsymbol{Y}\Big(-(\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{J}^{ij}\,\boldsymbol{W}^T + \boldsymbol{W}\,\boldsymbol{J}^{ijT})(\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})\Big)\,\boldsymbol{Y}^T\right)\\ &+ \frac{N}{2}\,\mathrm{Tr}\left((\boldsymbol{W}\,\boldsymbol{W}^T + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{J}^{ij}\,\boldsymbol{W}^T + \boldsymbol{W}\,\boldsymbol{J}^{ijT})\right)\end{aligned}$$

Question 20 Our marginalisation over f rather than $\boldsymbol{X}$ is clearly a step shorter and captures uncertainty in both θ and $\boldsymbol{X}$.



2.2 Practical

Question 21 We minimised the objective function to get our maximum likelihood estimate for $\boldsymbol{W}$

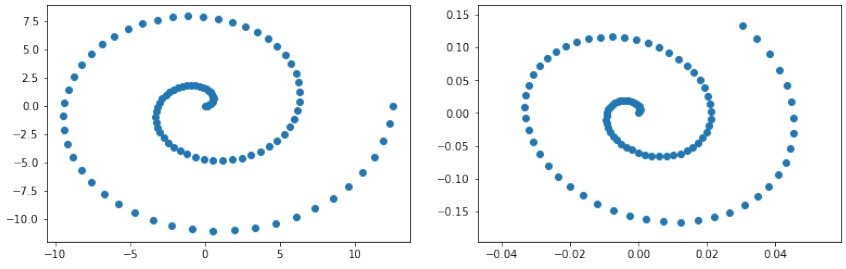
$$\boldsymbol{W}' = \arg\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W})$$

To find the representation for $\boldsymbol{X}'$ we must know the reverse of the linear mapping.

$$\boldsymbol{Y} = \boldsymbol{X}'\,\boldsymbol{W}'^T$$

$$\boldsymbol{X}' = \boldsymbol{Y}\,\boldsymbol{W}'(\boldsymbol{W}'^T\,\boldsymbol{W}')^{-1}$$

And then use minimization of our objective function using the non-linear conjugate gradient method to find $\boldsymbol{X}'$.



(a) Original $\boldsymbol{X}'$ (b) Learned $\boldsymbol{X}'$

Figure 8: Representation Learning of $\boldsymbol{X}'$

Our learned representation of $\boldsymbol{X}'$ is very close to the correct shape only rotated slightly, this is because the marginal likelihood we are maximising is invariant to any matrix transformation whose inverse is its own transposition i.e. any matrix for which $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}$ (rotations along with a few other transformations fall into this category). Therefore any of these transformations can be applied to $\boldsymbol{W}$ without the marginal likelihood changing. We can see this by setting $\boldsymbol{W}$ in the likelihood to $\boldsymbol{W}'\boldsymbol{R}$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}) &= -\sum_{i=1}^{N} \log\left(\mathcal{N}(y_i|0, (\boldsymbol{W}'\boldsymbol{R})(\boldsymbol{W}'\boldsymbol{R})^T + \sigma^2\boldsymbol{I})\right) \\ &= -\sum_{i=1}^{N} \log\left(\mathcal{N}(y_i|0, \boldsymbol{W}'\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{W}'^T + \sigma^2\boldsymbol{I})\right) \\ &= -\sum_{i=1}^{N} \log\left(\mathcal{N}(y_i|0, \boldsymbol{W}'\boldsymbol{W}'^T + \sigma^2\boldsymbol{I})\right)\end{aligned}$$

3 The Evidence

3.1 Theory

Question 22 This assumption implies that all data sets are equally likely assigning them uniform probability $\frac{1}{512}$. This is the most simple model in the sense that it has no free parameters but is the most complex model in the sense that it assigns lots of different models the same probability and therefore can't assign much probability mass to simple data-sets.

Question 23 M_3 is the most complex and most flexible model, it is analogous to full logistic regression having a bias parameter θ^3 and parameters for both dimensions of x. M_3 can be made to realize the other models by setting θ^3 to zero for M_2 or both θ^3 and θ^2 to zero for M_1. Because of this complexity and flexibility it assigns less probability mass to a larger set of data sets and so if the data sets to align with the simpler models will lose out to them.

M_2 is the next most complex model and will assign more probability mass than M_3 to all data sets that aren't very unequal or whose decision boundary separating $y = 1$ and $y = -1$ is offset from the origin as both these cases would be better modeled with the bias parameter in M_3.

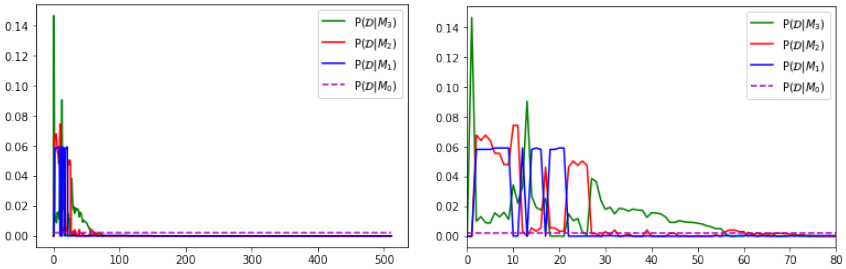
M_1 is the simplest of these models bar the uniform one and so assigns more probability mass to the set of data sets it can describe but can describe less data sets, as without a parameter for x_2 it can only describe data sets for which the decision boundary is a function of x_1.

Question 24 The prior $p(\theta|M_i) = \mathcal{N}(0, 10^3\boldsymbol{I})$ implies that all the dimensions are independent as it's a diagonal covariance matrix and the high standard deviation affects the model in that it leads to sharp linear boundaries in the data space as the parameters can be very wide.

3.2 Practical

Question 25 For all the models the evidence for the whole of $\mathcal{D}$ sums to 1 this is because they are all probability density functions.

Question 26 Our plots of the evidence in Figure 9 show that for the vast majority of the data set the model with the highest evidence is M_0 as most datasets cannot by decided by a linear decision boundary. Within the subset of data sets pictured in Figure 9b however, the other models outperform M_0 considerably, these are the datasets do follow a linear boundary. We can see that M_3 has a few datasets it is extremely adept at identifying, but overall just has a general trend of low evidence over a large period and is generally beaten by models M_2 and M_1 where $\mathcal{D} < 25$ as these datasets are probably bounded by lines crossing the origin which don't require a third bias parameter to formulate.



(a) Whole data set (b) Subset data set

Figure 9: Plot of evidence for each model

Question 27 The maximum and minimum data sets for M_0 are exactly the same as M_0 has uniform probability distribution so every data set is just as probable as the other ones. The minimum probability data sets for M_1, M_2 and M_3 are all datasets with boundaries that are not possible to express linearly as all our models are linear.

The maximum probability dataset for M_1 has a horizontal decision boundary as that is the only boundary M_1 can express with its single parameter. The maximum probablity dataset for M_2 has a sloped decision boundary but still centred around the origin as that is the boundary M_2 is most adept at expressing. The dataset which is most probable for M_3 is all one particular value as with its third bias parameter it can push the boundary all the way off the grid.

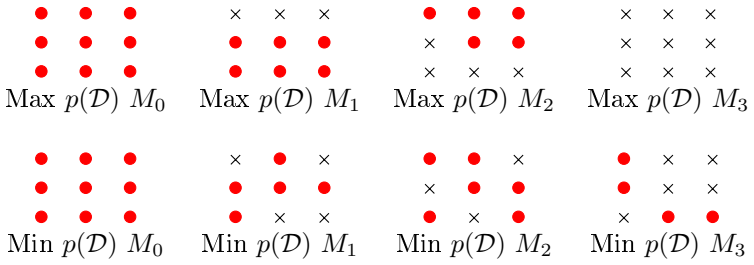


Table 1: Elements of D that maximise and minimise probability mass for each model

Question 30 The assignment was made of three parts containing both theoretical and practical questions: supervised learning or parametric regression (linear regression), unsupervised learning or non-parametric regression, and model selection; with an emphasis on the role of priors, assumptions, preferences and uncertainty. The assignment was presented in a way such that we were given the necessary theoretical knowledge throughout the assignment, only allowing us to complete it step by step. The lectures material mainly being theoretical forced us to grasp a general idea of machine learning concepts in order to implement these with actual data in the practical exercises. The two first parts of the assignment were covered in lectures, however the third section was not, forcing us to take a to apply what we learned in order to solve a new problem, model selection.

It appears the aim of this assignment has been to lead us a step further into the practice of machine learning, taking us from the material covered in lectures, generally theoretical, to a mix of theoretical questions and applied practical exercises. This enabled us to extend our understanding of the importance of integrating our beliefs in our observations in order to extract meaning from data. We were able to do this hands-on by actually encoding assumptions and uncertainty in our models, allowing us to learn from these models.