I. The prior $p(\mathbf{X}), p(\boldsymbol{W}), p(f)$

Question 1: *Why Gaussian form of the likelihood is a sensible choice? What does it mean that we have chosen a spherical covariance matrix for the likelihood?*

Answer: Since we are assuming additive noise in our probabilistic reasoning, the samples from the noise or error distribution are independent and identical, i.i.d. Hence, by the central theorem, the sum of error variables will become a Gaussian as the number of the variables in the sum increases, with same mean and variance.

A typical two-dimensional Gaussian likelihood function can be represented with a zero-mean $\boldsymbol{\mu} = [0\ 0]^T$ and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ which is called a spherical Gaussian, because of the spherical symmetry it describes. The covariance matrix above is also called an isotropic covariance has no other elements than just on the diagonal. Thus, shows that the variables are uncorrelated and independent. The problem is that using this type of covariance matrix limits the ability of capturing interesting correlations in the data. But mathematically is easier to work with, because conjugate Gaussians is a Gaussian.

Question 2: *If we do not assume that the data points are independent how would the likelihood look then? Remember that* $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_N]$

Answer: Assuming that the data points are not independent gives mathematically that the joint probability from a set of data points cannot factorises, hence result in joint probability distribution. With given input data, mapping functions and with the help of the chaining rule in probability result with non- dependant conditional expression:

$$p(\mathbf{T}|f,\mathbf{X}) = p(\cap_{i=1}^{N} \mathbf{t}_i \,|f,\mathbf{X}) = \prod_{i=1}^{N} p\left(\mathbf{t}_i \middle| \cap_{j=1}^{i-1} \mathbf{t}_j, f, \mathbf{X}\right) = \prod_{i=1}^{N} p(\mathbf{t}_i|\mathbf{t}_{i-1}, \mathbf{t}_{i-2},,, \mathbf{t}_1, f, \mathbf{X})$$

Q3: *What is the specific form of the likelihood* $p(\mathbf{T}|\mathbf{X}, \mathbf{W})$:

Answer: The output data are mapped with a function expressed as $f(\mathbf{x}_i) = \mathbf{W}\mathbf{x}_i$, the output is expressed with an added Gaussian noise $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$: $\mathbf{t}_i = \mathbf{W}\mathbf{x}_i + \boldsymbol{\epsilon}$ where $i = 1,,,N$. Assuming that the output is conditionally independent given, the input and parameters, the data points are i.i.d:

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} p(\boldsymbol{t}_i|\mathbf{x}_i, \mathbf{W}) = \prod_{i=1}^{N} N(\mathbf{W}\mathbf{x}_i + \mathbf{0}, \sigma^2\mathbf{I}) \quad (1)$$

Where the mean of the Gaussian function is the function f, distributed by every data point, and the variance corresponding to the noise distribution.

Question 4: *The prior in Eq.2 is a spherical Gaussian. This means that the "preference" is encoded in terms of a L_2 distance in the space of the parameters. With this view, how would the preference change if the preference was rather encoded using a L_1 norm? Compare and discuss the different type of solutions these two priors would encode*.

$$p(\mathbf{W}) = N(\mathbf{W}_0, \tau^2 \boldsymbol{I}) \qquad (2)$$

Answer: Equation 1 is a Gaussian prior with mean $\mathbf{W}_0$, its exponential function is expressed in terms of the Euclidian distance L_2, for example the distance between two input vectors **x** and **y** with N elements, is expressed by the notation: $\left|\left|\mathbf{x}-\mathbf{y}\right|\right|_2 = \sqrt{\sum_i^N (x_i - y_i)^2}$. And the L_1 norm or taxicab distance is $\left|\left|\mathbf{x}-\mathbf{y}\right|\right|_1 = \sum_i^N |x_i - y_i|$. Hence, the prior will be different with L_1 space: $P(\mathbf{W}) \propto e^{-\frac{|\mathbf{W}-\mathbf{W}_0|}{2\tau^2}}$ which is familiar with Laplacian distribution. This type of distribution is part of the exponential family, can be implemented with Gaussian likelihood but will give a Laplacian posterior (conjugate priors). The Laplace prior is more centred around zero if $\mathbf{W}_0$ is zero, for the distribution of its parameters. The distribution characterizes with a peak and wider tails, most estimated parameters ends up at zero. Instead with a Gaussian distribution varies a lot more around zero. Therefore, using Laplace prior can give zero valued parameters (any moderate size, any large size). While Gaussian prior tends to give moderate coefficients that cannot exactly be zero. Something else that could be considered with solution of prior with L_2 are the outliers. It squares the error, hence shows higher values than with L_1. In the perspective of minimizing the error. Using L_2 gives a better adjustment for data containing outliers than with L_1 in regression problems. According to the error function with L_1 term also known as Lasso, and with L_2 called Ridge regression.

Question 5: *Derive the posterior over the parameters. Please, do these calculations by hand as it is very good practice. However, in order to pass the assignment you only need to outline the calculation and highlight the important steps. You can make derivations for individual samples* $(\mathbf{x_i}, \mathbf{t_i})$ *and then generalize to the dataset or operate on matrices keeping the concept of vectorization in mind.*

- Briefly comment/discuss the form (mean and covariance).
- What is the effect of the constant Z, are we interested in this?

The expression of the posterior over the parameters, **W**:

$$\mathrm{p}(\mathbf{W}|\mathbf{X},\mathbf{T}) = \frac{1}{\mathrm{Z}}\mathrm{p}(\mathbf{T}|\mathbf{X},\mathbf{W})\mathrm{p}(\mathbf{W}) \qquad (3)$$

Answer: Where the posterior can be described with data points being independent and following the in general form.

$$p(\boldsymbol{W}|\mathbf{X},\mathbf{T}) \sim N(\boldsymbol{\mu_w}|\boldsymbol{\Sigma_w}) \propto e^{-\frac{1}{2}(\mathbf{W}-\boldsymbol{\mu_w})^T\boldsymbol{\Sigma_w^{-1}}(\mathbf{W}-\boldsymbol{\mu_w})} =$$

$$= \underbrace{e^{-\frac{1}{2}\mathbf{W}^\mathbf{T}\boldsymbol{\Sigma_w^{-1}}\mathbf{W}}}_{(4)} * \underbrace{\mathrm{e}^{\mathbf{W}^\mathbf{T}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{\mu_w}}}_{(5)} * \underbrace{e^{-\frac{1}{2}\boldsymbol{\mu_w^T}\boldsymbol{\Sigma_w^{-1}}\boldsymbol{\mu_w}}}_{(6)}$$

$\boldsymbol{\mu_w}$ is a mean vector and $\boldsymbol{\Sigma_w}$ the covariance of the parameter conditional probability distribution.

The prior from equation 2:

$$P(\mathbf{W}) = N(\mathbf{W_o}, \tau^2\mathbf{I}) \propto e^{-\frac{1}{2}[(\mathbf{W}-\mathbf{W_o})^T\frac{1}{\tau^2}\mathbf{I}(\mathbf{W}-\mathbf{W_o})]}$$

The likelihood from equation 1:

$$p(\boldsymbol{T}|\mathbf{X},\mathbf{W}) = N(\mathbf{WX}, \sigma^2 I) \propto e^{-\frac{1}{2\sigma^2}[(\mathbf{T}-\mathbf{WX})^\mathrm{T}(\mathbf{T}-\mathbf{WX})]}$$

The posterior: $\mathrm{p}(\mathbf{W}|\mathbf{X},\mathbf{T}) \propto \mathrm{p}(\mathbf{T}|\mathbf{X},\mathbf{W})\mathrm{P}(\mathbf{W}) = \mathrm{e}^{-\frac{1}{2\sigma^2}\left[(\mathbf{T}-\mathbf{WX})^\mathrm{T}(\mathbf{t}-\mathbf{WX})\right]} * \mathrm{e}^{-\frac{1}{2}\left[(\mathbf{W}-\mathbf{W_o})^\mathrm{T}\frac{1}{\tau^2}\mathbf{I}(\mathbf{W}-\mathbf{W_o})\right]} =$

$$= \underbrace{\mathrm{e}^{-\frac{1}{2\sigma^2}\mathbf{W}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{W}-\frac{1}{2}\mathbf{W}^\mathbf{T}\frac{1}{\tau^2}\mathbf{W}}}_{(4')} * \underbrace{\mathrm{e}^{\frac{1}{\sigma^2}\mathbf{T}^\mathbf{T}\mathbf{X}\mathbf{W}+\frac{1}{\tau^2}\mathbf{W}^\mathbf{T}\mathbf{W_0}}}_{(5')} * \underbrace{\mathrm{e}^{-\frac{1}{2\sigma^2}\mathbf{T}^\mathbf{T}\mathbf{T}-\frac{1}{2\tau^2}\mathbf{W_0^T}\mathbf{W_0}}}_{(6')}$$

We want the result to be in the form described first, not taking the constants term into consideration, eq $(6')$ and (6):

$(4')$ = (4) the quadratic form: $e^{-\frac{1}{2\sigma^2}\mathbf{W^T X^T X W}-\frac{1}{2\tau^2}\mathbf{W^T W}} = e^{-\frac{1}{2}(\frac{1}{\sigma^2}\mathbf{W^T X^T X W}+\mathbf{W^T}\frac{\mathbf{1}}{\boldsymbol{\tau^2}}\mathbf{W})} = e^{-\frac{1}{2}\mathbf{W^T}(\frac{1}{\sigma^2}\mathbf{X^T X}+\frac{1}{\tau^2}\mathbf{I})\mathbf{W}} = e^{-\frac{1}{2}\mathbf{W^T}\,\boldsymbol{\Sigma_w^{-1}}\,\mathbf{W}}$, hence the precision matrix of **W is**:

$$\boldsymbol{\Sigma_w^{-1}} = \frac{1}{\sigma^2}\mathbf{X^T X} + \frac{1}{\tau^2}\mathbf{I} \qquad (7)$$

$(5')$ = (5) the linear term: $e^{\frac{1}{\sigma^2}\mathbf{T^T X W}+\frac{1}{\tau^2}\mathbf{W^T W_o}} = e^{\mathbf{W^T}(\frac{1}{\sigma^2}\mathbf{X^T T}+\frac{1}{\tau^2}\mathbf{W_o})} = e^{\mathbf{W^T}\,\boldsymbol{\Sigma_w^{-1}}\,\boldsymbol{\mu_w}}$

In the exponent the equation with respect to the mean (7):

$$\mathbf{W^T}\,\boldsymbol{\Sigma_w^{-1}}\,\boldsymbol{\mu_w} = \mathbf{W^T}\left(\frac{1}{\sigma^2}\mathbf{X^T X} + \frac{1}{\tau^2}\mathbf{I}\right)\boldsymbol{\mu_w} = \mathbf{W}^{\mathrm{T}}\left(\frac{1}{\sigma^2}\mathbf{X^T T} + \frac{1}{\tau^2}\mathbf{W_o}\right) \rightarrow \left(\frac{1}{\sigma^2}\mathbf{X^T X} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X^T T} + \frac{1}{\tau^2}\mathbf{W_o}\right) = \boldsymbol{\mu_w} \qquad (8)$$

which is the mean of the distribution of the parameter **W.**

Because it is conditional independent, the expression of the multidimensional gaussian is:

$$\mathrm{p}(\mathbf{W}|\mathbf{X},\mathbf{T}) \sim N\left(\left(\frac{1}{\sigma^2}\mathbf{X^T X} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X^T T} + \frac{1}{\tau^2}\mathbf{W_o}\right), \left(\frac{1}{\sigma^2}\mathbf{X^T X} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\right)$$

The mean corresponds to the input and output data, but the variance only encounters the input data.

In equation 6, Z is the normalization constant. Deriving the value provides the probability of the posterior distribution having the maximum value equal to one, which is proper for probability distribution. For model selection, Z gives the property of evidence.

Question 6: *Explain what this prior does? Why is it a sensible choice? Use images to show your reasoning. Clue: use the marginal distribution to explain the prior*

$$p(f|\mathbf{X},\boldsymbol{\theta}) = N(\mathbf{0}|\mathrm{k}(\mathbf{X},\mathbf{X})) \qquad (9)$$

Answer: Gaussian process is a non-parametric approach, to find a distribution over possible functions based on the observed or given data, see equation 9. Instead of relying on a linear relationship between points, the regression function can take up too many kinds of functions. To derive the prior, you want to look at a certain domain of data points with a mean equal to zero in order to get proper functions that is not too wiggly. And to get a smoother function you have to use a covariance matrix. And the covariance matrix represents a kernel type of function, the points are expected by the kernel to be similar to the output of the function. Thus, the prior shows the first probability of the distribution of functions. With a mean of being zeroes also shows that the covariance function completely defines the process behaviour. Fig 1 represents an examples of the sampling functions.

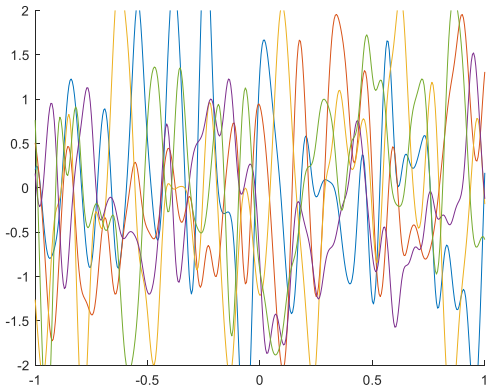

Fig 1. Samples f from prior using Gaussian process, x- axis being x, y- axis being t

Question 7: *Formulate the joint likelihood of the full model that you have defined above,*

$$p(\mathbf{T}, \mathbf{X}, f, \boldsymbol{\theta})$$

(Try to draw a very simple graphical model to clearly show the assumptions that you have made.)

Answer: Our hyper parameter and data points are independent $p(\mathbf{X}), p(\boldsymbol{\theta})$. For the prior using Bayer theorem, we know that f is conditionally independent with given $\mathbf{X}$ and $\boldsymbol{\theta}$ hence $p(f|X, \theta)$.

Last we know that with given f after having observed $\mathbf{X}$ and f will give the output probability: $p(\mathbf{T}|f)$, factorisation of the joint distribution of the model can be expressed as:

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

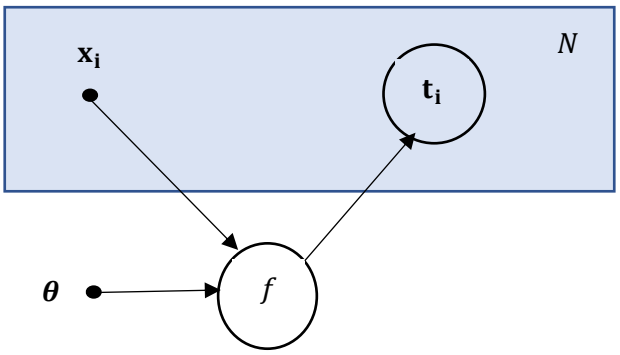

Fig 2. Graphical model of the joint likelihood of the full model

Where $i = [1,2,,,N]$ and $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, , , , \mathbf{x_N}]$, $\mathbf{T} = [\mathbf{t_1}, \mathbf{t_2}, , , , \mathbf{t_N}]$

Question 8: *Complete the marginalisation formula in Eq.10 (general form) and discuss,*

- *Explain how this connects the prior and the data?*
- *How does the uncertainty "filter" through this?*
- *What does it imply that* $\boldsymbol{\theta}$ *is left on the left-hand side of the expression after marginalisation?*

$$p(\mathbf{T}|\mathbf{X}, \boldsymbol{\theta}) \qquad (10)$$

Answer: $p(T|X,\theta) = \int p(T|f)p(f|X,\theta)df$, we want to marginalize out the mapping function f. The last term is the Gaussian process or prior which shows our first belief of the mapping functions over the parameters and data input. The Gaussian process is finite depending on the amount of data observed thus have a condition sent in the data subset. The first term is the likelihood, the output data with given mapping function. And get weighted by the prior.

We got two types of uncertainties from two Gaussians, one of them id from the Gaussian process and the other from the possible functions over the data. From the likelihood term with the output derived as $\mathbf{t}_i = f(\mathbf{x}_i) + \mathbf{e}$, where e is the noise. Since each of the terms are Gaussian, their covariance adds up in the integral, because of independence.

That $\boldsymbol{\theta}$ is left of the expression implies that the parameter is a constant. Being Bayesian, we still need to take account of uncertainties. To deal with the parameter can be done by a Bayesian approach to identify the prior, posterior and average it out. It is still dependent on the kernel, hence stays constant. And the expression consequently say that our output primarily depends on the parameter and data points.

Question 9:

1. Set the prior distribution over **W** *and visualise it.*

2. Pick a single data-point from the data and visualise the posterior distribution over **W**.

3. Sample from the posterior and show a couple of functions.

4. Repeat 2 - 3 by adding additional data points.

Describe the plots and the behavior when adding more data? Is this a desirable behavior? Provide an intuitive explanation.

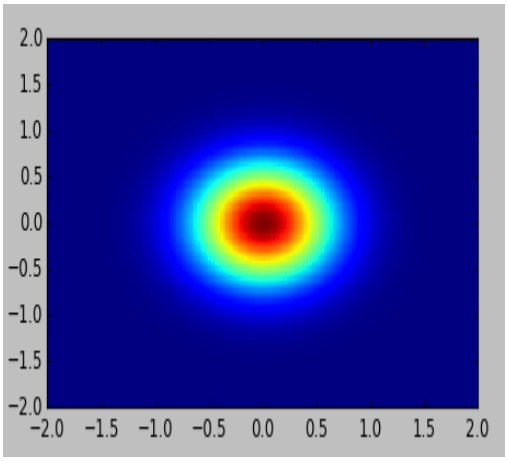

Fig 3. The prior over **W**. x-axis: w_0, y-axis: w_1

Answer: We can clearly see in fig 3, the prior probability distribution with mean zero and a uniform covariance matrix in two-dimensional perspective with the highest probability of function $p(f)$ being in the middle with the parameters $w_0 = 0$, $w_1 = 0$.

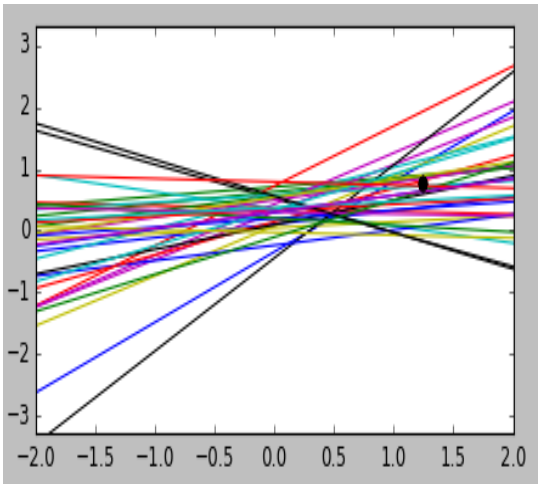

Fig 4: Data space with one observation point. x-axis: w_0, y-axis: w_1

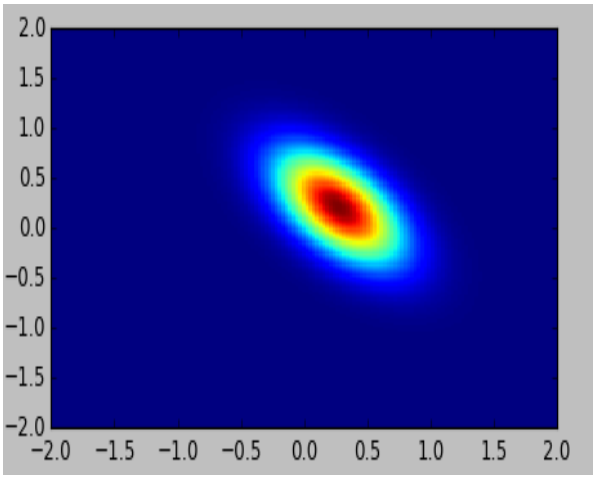

Fig 5: the posterior over **W**, x-axis: w_0, y-axis: w_1

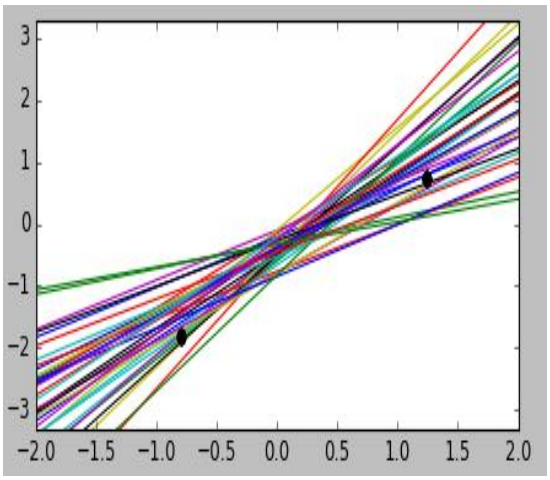

Fig 6: Data space with two observation points. x-axis: w_0, y-axis: w_1

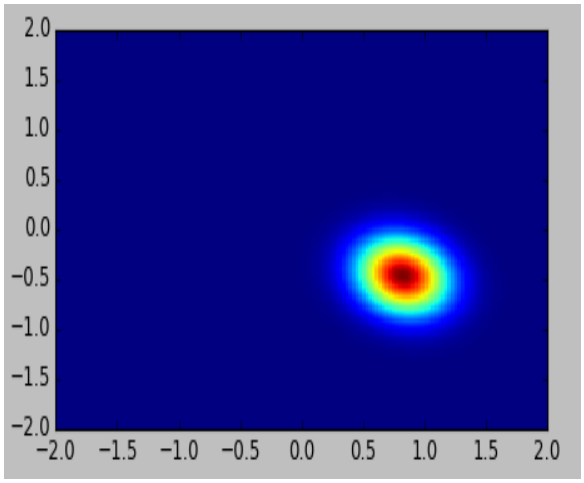

Fig 7: The posterior over **W**, x-axis: w_0, y-axis: w_1

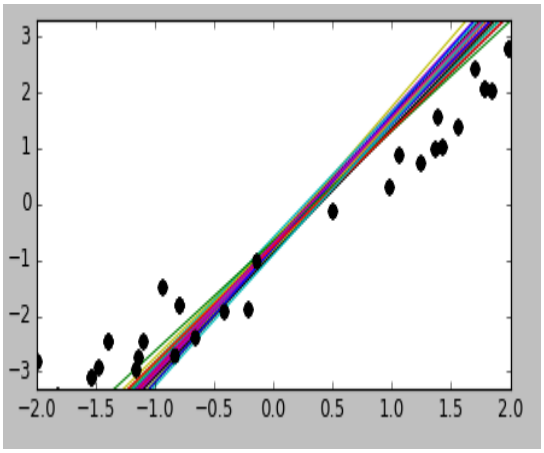

Fig 8: Data space with 30 observation points. x-axis: w_0, y-axis: w_1

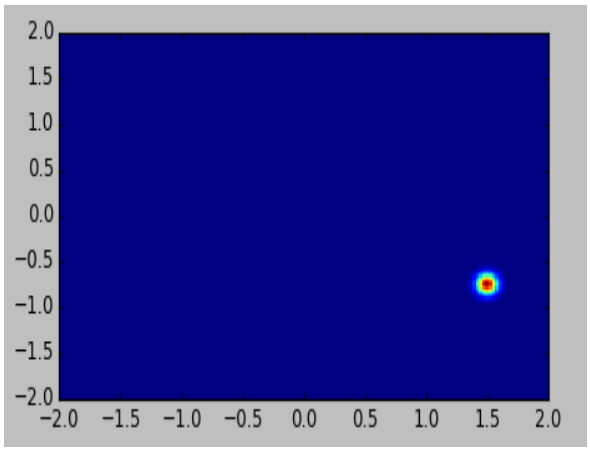

Fig 9: The posterior over **W**, x-axis: w_0, y-axis: w_1

Having observed data points (black dot in data space plots), result that the samples of Gaussian process posterior to be constrained near the observed data point is later used for interference. The observed data puts a conditional joint distribution of the other unknown data point. In fig 9 one can see that the estimated parameters were achieved with $w_0 \approx 1.5$ and $w_1 \approx -0.8$. Clearly for more data points being observed and labelled, the more the estimation will converge to the sought-after parameters and the variance decreased.

Question 10:

1. Create a GP-prior with a squared exponential co-variance function.

2. Sample from this prior and visualise the samples.

3. Show samples using different length-scale for the squared exponential.

Explain the behavior of altering the length-scale of the covariance function.

Answer: The GP prior is $p(f|\mathbf{X},\boldsymbol{\theta}) = N\big(\mathbf{0}, k(\mathbf{X},\mathbf{X})\big)$, where $k(\mathbf{X},\mathbf{X}) = \sigma_f^2 e^{-\frac{(x_i-x_j)^T(x_i-x_j)}{l^2}}$, σ_f^2 is the gain, l is the length-scale, indicates how close two data points x_i and x_j are, to be able to influence each other. The plots below are shown with five samples and the gain equal to one and the x- axis: x, y- axis: t.

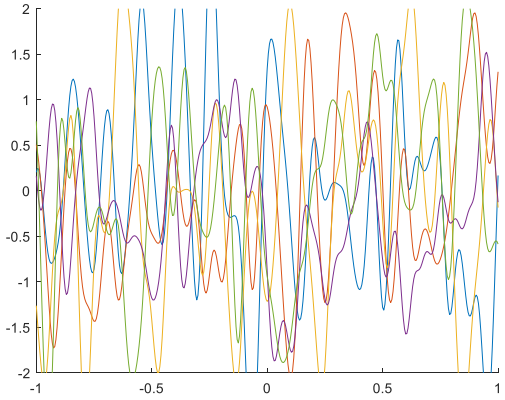

Fig 10. Samples, l = 0.05.

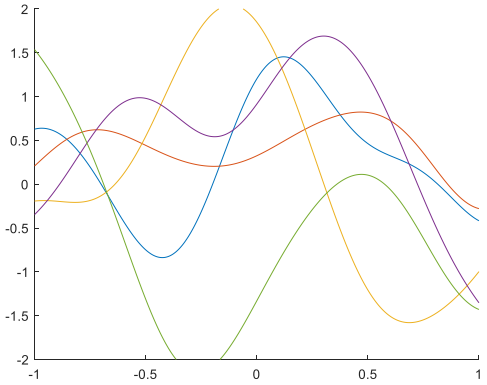

Fig 11. Samples, l = 0.5.

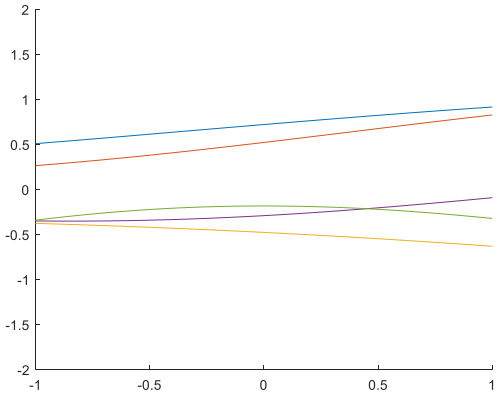

Fig 12. l = 5

The conclusion that can be drawn from the plots above is that for a distance d between the data points gives for larger l, more covariation as can be seen in fig 12. For smaller l, weaker covariation, similar to white noise, see fig 10.

Question 11:

1. How do we interpret the posterior before we observe any data?

2. Compute the predictive posterior distribution of the model.

3. Sample from this posterior with points both close to the data and far away from the observed
data.

4. Plot the data, the predictive mean and the predictive variance of the posterior from the data.

Explain the behavior of the samples and compare the samples of the posterior with the ones from the prior. Is this behavior desirable? What would happen if you would add a diagonal covariance matrix to the squared exponential?

Answer: To create gaussian process using the prior definition $p(f|\mathbf{X},\boldsymbol{\theta}) = N(\boldsymbol{\mu}, k(\mathbf{X},\mathbf{X}))$ where $\boldsymbol{\mu} = \mathbf{0}$. To predict new samples f^* for selected new data $\mathbf{X}^*$, is used by the predictive posterior $p(f^*|\mathbf{X}^*,\mathbf{X},f,\boldsymbol{\theta})$. The joint distrubion of the new and observed samples gives by:

$$p(f^*,f|\mathbf{X}^*,\mathbf{X},\boldsymbol{\theta}) = \begin{bmatrix} f \\ f^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(\mathbf{X},\mathbf{X}) & k(\mathbf{X},\mathbf{X}^*) \\ k(\mathbf{X}^*,\mathbf{X}) & k(\mathbf{X}^*,\mathbf{X}^*) \end{bmatrix}\right),$$

having observed samples, for new samples it is gives by

$$p(f^*,f|\mathbf{X}^*,\mathbf{X},\boldsymbol{\theta}) = N\left(k(\mathbf{X}^*,\mathbf{X})^T k(\mathbf{X},\mathbf{X}) f, k(\mathbf{X}^*,\mathbf{X}^*) - k(\mathbf{X}^*,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},\mathbf{X}^*)\right),$$

and we can sample from this distribution.

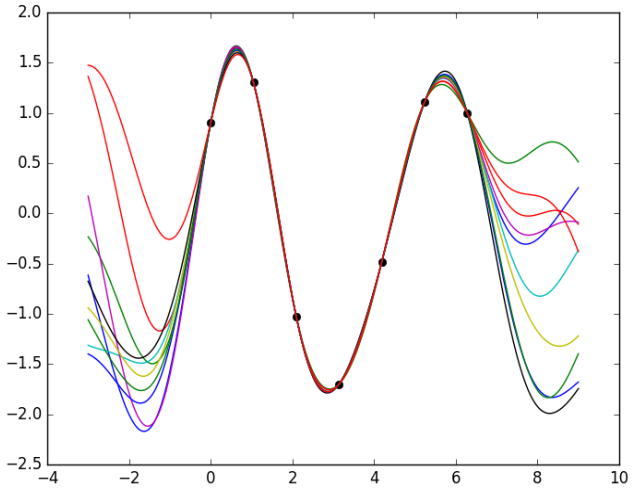

Fig 13. 10 samples and the given x values represents as the black dots. x-axis: x, y-axis t

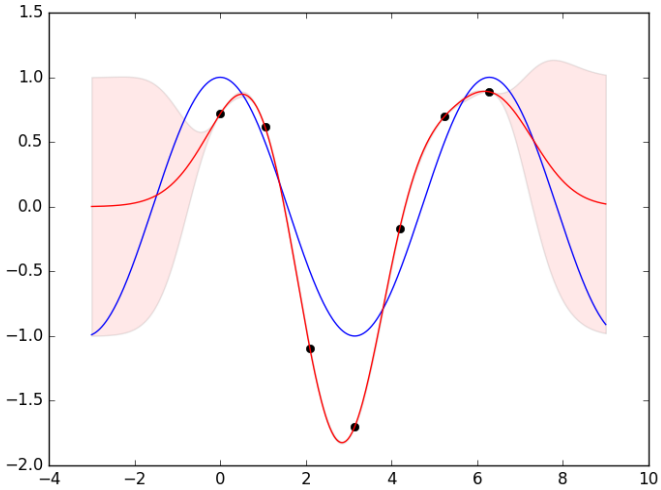

Fig 14: The blue line is a cosine function, the red line is the predictive mean and the red shadow is the predictive variance of the posterior samples. x-axis: x, y-axis t

The variance of the samples changes a lot before and after the observed x values which can be seen in fig 13. In the other plot, fig 14, the predictive mean curve follows relatively a cosine function. Between the behaviour of the samples in the prior (question 10) and the samples of the posterior which are the two plots above. The difference is the observation of data points by the posterior. It makes the samples pass near the data point, hence for a larger number data points will result in a more accurate posterior.

Adding a diagonal covariance matrix, would definitely affect the output as the variance changes. The new model will make room to take account for the noise, thus will make the predictive mean for example more robust with a bias. Resulting that it won't go through the given data points because of the noise it inherits.

II The posterior $p(\mathbf{X}|\mathbf{Y})$

Question 12: *What type of preference does this prior encode?*

$$p(\mathbf{X}) = N(\mathbf{0}, \mathbf{I}) \quad (11)$$

Answer: The latent data points $\mathbf{x_i}$ comes with a distributed Gaussian function within a certain interval, and because the variance being an identity matrix thus implies that the data points are independent. Also recall because of the relation between **X** and **W** mentioned in the theory, the preference of the prior over parameter **W** is affected by the encoded preference of $p(\mathbf{X})$.

Question 13: *Perform the marginalisation in Eq. 12 and write down the expression. As previously, it is recommended that you do this by hand even though you only need to outline the calculations and show the approach that you would take to pass the assignment.*
Hint*: The marginal can be computed by integrating out* **X** *with the use of Gaussian algebra we exploited in the exercise derivations and, in particular, by completing the square. However it is much easier to derive the mean and covariance, knowing that the marginal is Gaussian, from the linear equation of* **Y(X).**

$$p(\mathbf{Y}|\mathbf{W}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) p(\mathbf{X}) d\mathbf{X} \quad (12)$$

Answer: $\mathbf{x_i}$ is considered a latent variable $x_i \rightarrow y_i$ by a mapping f. The likelihood $p(\mathbf{Y}|\mathbf{W}, \mathbf{X}) = \prod_{n=i}^{N} p(\mathbf{y_i}|\mathbf{W}, \mathbf{x_i})$ for the total independent conditional distribution. There is a linear transformation of **y** and **W** with the output $\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\epsilon}$, the noise has a zero mean Gaussian distribution with a covariance $\sigma^2\mathbf{I}$. Hence $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = N(\mathbf{W}\mathbf{x}, \sigma^2\mathbf{I})$. And the prior is equation 11.

The marginalisation distribution $p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} p(\mathbf{y_n}|\mathbf{W})$ is expressed by:

$$p(\mathbf{y}|\mathbf{W}) = \int p\,(\mathbf{y}|\mathbf{x}, \mathbf{W}) p(\mathbf{x}) d\mathbf{x}.$$

Because this correspond to a linear-gaussian model it can be expressed by $p(\mathbf{y}|\mathbf{W}) = N(\mathbf{y}|\,\mathbf{C})$

Where **C** is the covariance matrix and can be derived by using the predictive distribution that also will be an Gaussian. Then calculating the mean and covariance with respect to $\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\epsilon}$. **x** and $\boldsymbol{\epsilon}$ is a distribution of independent random variables that are uncorrelated and have a zero mean.

$$E[\mathbf{y}|\mathbf{W}] = E[\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon}] = \mathbf{0}$$

$$\begin{aligned} cov[\mathbf{y}|\mathbf{W}] &= E[(\mathbf{y} - E[\mathbf{y}])(\mathbf{y} - E[\mathbf{y}])^T] = E[\mathbf{y}\mathbf{y}^\mathbf{T}] = E[(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})^T] \\ &= E[\mathbf{W}\mathbf{x}\mathbf{x}^\mathbf{T}\mathbf{W}^\mathbf{T} + 2\mathbf{W}\mathbf{x}\boldsymbol{\epsilon}^\mathbf{T} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathbf{T}] = \mathbf{W}E[\mathbf{x}\mathbf{x}^\mathbf{T}]\mathbf{W}^T + E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathbf{T}] = \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I} \\ &= \mathbf{C} \end{aligned}$$

Resulting that:

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

Question 14. *Compare these three estimation procedures above in log-space.*

- *How are they different?*
- *How are MAP and ML different when we observe more data?*
- *Why are the two last expressions of Eq. 13 equal?*

$$argmax_{\mathbf{W}} \frac{p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W})}{\int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W})d\mathbf{W}} = argmax_{\mathbf{W}} p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W}) \quad (13)$$

Answer: Learning the probabilistic models regarding the likelihood can be done maximizing its parameters. The Maximum-likelihood (ML) of $p(\mathbf{Y}|\mathbf{W},\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{y_n}|\mathbf{W},\mathbf{x_n})$ in log space derives as, taken from Bishop [1]:

$$\ln[p(\mathbf{Y}|\mathbf{W},\mathbf{X})] = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\{\mathbf{y_n} - \mathbf{W}\mathbf{x_n}\}^2 - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\sigma^2, \quad (14)$$

where σ^2 is the variance from the likelihood and N as the total number of data points.

Maximum-a-posteriori (MAP) is derived by maximizing the posterior distribution, taken from Bishop [1]:

$$\ln\big(p(\mathbf{Y}|\mathbf{W},\mathbf{X})p(\mathbf{W})\big) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}\{\mathbf{y_n} - \mathbf{W}\mathbf{x_n}\}^2 + \frac{1}{2}\sum_{n=1}^{N}\mathbf{w_n^T}\mathbf{w_n}, \quad (15)$$

where $\mathbf{w} \in \mathbf{W}$**,**

Maximum -likelihood type-2 is like MAP except it marginalize over **X**:

The marginalisation distribution is $p(\mathbf{Y}|\mathbf{W}) = \int p\,(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})d\mathbf{X}$ and $p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{\mathrm{N}} N(\mathbf{y_i}|\mathbf{0},\mathbf{W}\mathbf{W^T} + \boldsymbol{\sigma^2}\mathbf{I})$. In log-space it will look like, taken from Bishop [1]:

$$\ln\big(p(\mathbf{y}|\mathbf{W})\big) = -\frac{1}{2}\mathbf{y}^T(\mathbf{W}\mathbf{W^T} + \sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{ND}{2}\ln|\mathbf{W}\mathbf{W^T} + \sigma^2\mathbf{I}|, \quad (16)$$

Where output $\mathbf{y} \in \mathbb{R}^{\mathrm{D\,x\,1}}$**.**

We can see that MAP got 2 quadratic terms, one more than ML which estimates **W.** Thus implies for higher amount of data, hence requires heavier computations, especially in the MAP where the polynomial term computes faster over the least square term. MAP is expressed by the prior which includes the regularization term, unlike ML.

For fixed data points, the likelihood function tells the likelihood, not about the probability because it is not normalised. But with MAP sets a belief that estimates the likelihood. If the

prior belief is strong than the data has less impact on the estimation of the parameters, for weak belief, the outcome will show more of a standard ML.

The equation 13 stays equal because we marginalize over the distribution of **X** to get the posterior $p(\mathbf{Y}|\mathbf{W})$, hence the denominator will not depend on **W**. The maximization of the expression with regard of the parameter **W** will not consider the denominator, instead it is a constant term that is irrelevant, and will be removed by the derivative being equal to zero. Thus, only the numerator stays equivalent.

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

Question 15.

1. Write down the objective function $-log(p(\mathbf{Y}|\mathbf{W})) = L(\mathbf{W})$ *for the marginal distribution in Eq.* 17

2. Write down the gradients of the objective with respect to the parameters $\frac{\delta L}{\delta W}$

$$p(\mathbf{Y}|\mathbf{W}) = \int p\,(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})d\mathbf{X} \quad (17)$$

Answer: Some useful formulas from bishop [1] and matrix cookbook [2], example for a matrix $\mathbf{C}(\mathbf{A})$:

$$\frac{\partial}{\partial \mathbf{A}}\boldsymbol{C}^{-1} = \boldsymbol{C}^{-1} * \frac{\partial}{\partial \mathbf{A}}\boldsymbol{C} * \boldsymbol{C}^{-1}$$
$$\frac{\partial}{\partial \mathbf{A}}\ln|\boldsymbol{C}| = (\boldsymbol{C}^{-1})^{T} * \frac{\partial}{\partial \mathbf{A}}\boldsymbol{C}$$

The objective function (16), where $\boldsymbol{\Lambda} = \mathbf{W}\mathbf{W}^{\mathbf{T}} + \sigma^2\mathbf{I}$ is a quadratic matrix and $p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} p(\mathbf{y_n}|\mathbf{W})$ is expressed as following:

$$L(\mathbf{W}) = \frac{1}{2}\mathbf{y}^{T}(\boldsymbol{\Lambda})^{-1}\mathbf{y} + \frac{N}{2}\ln|\boldsymbol{\Lambda}| + constant$$

The gradient of the objective with respect to the parameters:

$$\frac{\delta L}{\delta \mathbf{W}} = \frac{1}{2}\mathbf{y}^{T}\frac{d}{d\mathbf{W}}(\boldsymbol{\Lambda})^{-1}\mathbf{y} + \frac{N}{2}\frac{d}{d\mathbf{W}}\ln|\boldsymbol{\Lambda}| = \{\text{with formulas above}\} =$$
$$-\frac{1}{2}\mathbf{y}^{T}(\boldsymbol{\Lambda})^{-1}\frac{d}{d\mathbf{W}}(\mathbf{W}\mathbf{W}^{T})(\boldsymbol{\Lambda})^{-1}\mathbf{y} + \frac{N}{2}(\boldsymbol{\Lambda})^{-1}\frac{d}{d\mathbf{W}}(\mathbf{W}\mathbf{W}^{T})$$

Where: $\frac{d}{d\mathbf{W}}(\mathbf{W}\mathbf{W}^{T}) = \left(\frac{d}{d\mathbf{W}}\mathbf{W}^{T}\mathbf{W}\right)^{T} = (\mathbf{W}^{T}\boldsymbol{J}^{ij} + \boldsymbol{J}^{ji}\mathbf{W})^{T} = \boldsymbol{J}^{ji}\mathbf{W} + \mathbf{W}^{T}\boldsymbol{J}^{ij}$, $\frac{d\mathbf{W}}{dW_{ij}} = \boldsymbol{J}^{ij}$

Question 16: *Plot the representation that you have learned (hint: plot* $\mathbf{X}$ *as a two-dimensional representation). Explain the outcome and discuss key features, elaborate on any invariance you observe. Did you expect this result?*

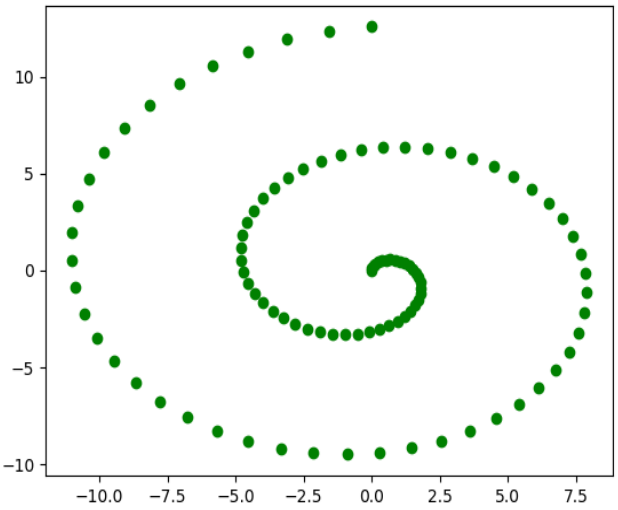

Fig 15. Original data points, x-axis: x_1, y-axis: x_2

Answer: In fig 15, inserted the 100 data input points from 0 to 4π. The output is a linear function taking the input data as nonlinear: $f_{lin}(x_i) = [x_i \sin(x_i), x_i \cos(x_i)] * A^T$, where A is a matrix that maps the input to the output containing random values from the distribution $N(0,1)$.

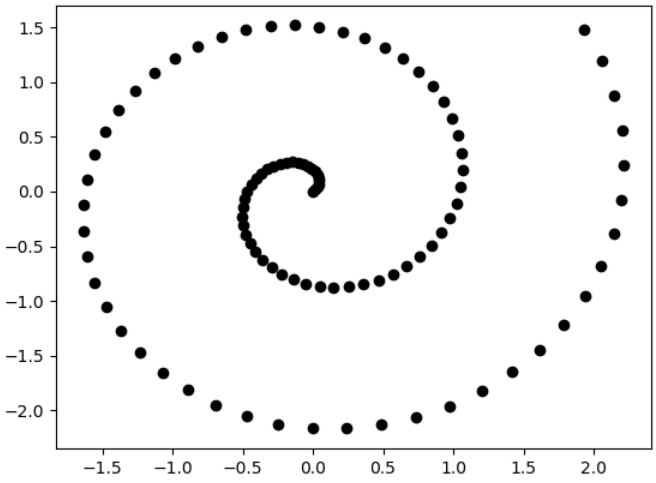

Fig 16. Representation learning with optimizing input data $\mathbf{x_{star}}$, x-axis: x_1, y-axis: x_2

Used a minimizing function with a nonlinear conjugate algorithm to get a linear term. Had to subtract the represented data which is represented as $\mathbf{Y} = \mathbf{x_{star}}\widehat{\mathbf{W}}$ where $\widehat{\mathbf{W}} = argmin_{\mathbf{w}}(-\log(p(\mathbf{Y}|\mathbf{W}))$ with respect to the gradient of the conditional probability. Hence, by some algebra calculation the learned input data were derived from:

$\mathbf{x_{star}} = \mathbf{Y} * \widehat{\mathbf{W}}\left(\widehat{\mathbf{W}}^{\mathbf{T}}\widehat{\mathbf{W}}\right)^{-1}$ which resulted in fig 16.

The result exceeded the expectations, both plots shows a similar pattern, but not the same. The magnitude of the scaling differs, because of the generated parameters from the prior. The reason for the different rotations can be explained by using a rotation of the weight matrix $\widetilde{\mathbf{W}} = \mathbf{WR}$, where $\mathbf{R}$ is an orthogonal matrix. The conditional probability, then states:

$$P\left(\mathbf{Y}\middle|\widetilde{\mathbf{W}}\right) = N(\mathbf{0}, \mathbf{WRR^T W^T}) = N(\mathbf{0}, \mathbf{WIW^T}),$$

hence the original marginal likelihood is invariant to the rotation. What can also be discussed is how the patterns still have its spiral shape, that is because it occurs a dominant correlation.

III The evidence $p(D)$

Question 17: *Why is this the simplest model, and what does it actually imply? Discuss its implications, why is this a bad model and why is it a good model?*

Answer: The model $p(D|M_0, \boldsymbol{\theta_0}) = \frac{1}{512}$ takes account for all the data sets and placing the same probability mass over them and has no free parameters. Occam's Razor states we should choose the simplest model that explains all the data and this model corresponds with the statement. The conditional probability expression tells that there is an equal probability of all data, a uniform distribution. The disadvantage of using this model is the complexity. Due to that it fits more to the data, but assigns relatively small probability to each one of them, hence for data that should be graded as more important or simple data sets, are neglected or less assigned probability mass. And for data sets that are not well modelled by any sharp linear boundary, hence not complexed, this uniform model is the most suitable one.

Question 18: *Explain how each separate model works. In what way is this model more or less flexible compared to M_0? How does this model spread its probability mass over D?*

Answer: The model $p(D|M_1, \theta_1)$ spread its probability, taking account of independent data points. With help of logistic regression. For each location $\mathbf{x}^{\mathrm{i}}$ corresponding to parameters θ_1 follows up with a decision boundary $\theta_1^1 x_1^n$. For each given data point, the probability function which is represented with a sigmoid function, gives a value from 0 to 1. But M_1 considers only the input x_1^n. Therefore, will give high probability for the given data points t^n and x_1^n. Comparing with the uniform probability M_o, the model M_1 will be less simple but more flexible. This is better for well modelled data sets that's has sharp linear boundaries.

Question 19: *How have the choices we made above restricted the distribution of the model? What datasets are each model suited to model? What does this actually imply in terms of uncertainty? In what way are the different models more flexible and in what way are they more restrictive? Discuss and compare the models to each other.*

Answer: The models have different decision boundaries influencing the regression. Especially with the dimension of the decision boundary. M_3 is a full logistic regression, and the only model having a bias term w_0 that considers the unequal distribution of the data set, because w_0 allows the decision boundary to be offset from the origin. What can also be noticed is that

the model M_3 is similar to the other models, if some of the parameters are set to zero, hence indicates it is far more flexible than the other models, spreading the probability mass over a wider range.
But the flexibility comes with a cost of sometimes losing out to the simpler models, when it comes to assigning probability of simpler datasets. Model M_1 and M_2, their decision boundaries are linearly that goes through the origin. M_1 only give evidence of dataset belonging to x_1 and not for x_2 like M_2 and M_3, hence these two works with the all given data set. But it is more likely that M_1 would give a high probability of x_1 as mentioned before. But M_3 should give spread out probability for all data sets.

Question 20: *Explain the process of marginalisation and briefly discuss its implications.*

Answer: $p(D|M_i) = \int p(D|M_i, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ for all $\boldsymbol{\theta}$. The evidence of models is given by the sum and product of probability rules, with models that are governed by the parameters $\boldsymbol{\theta}$. By being Bayesian, we have to take uncertainty into account on each step. We accommodate the parameters to eliminate its dependency on the evidence.

Question 21: *What does this choice of prior imply? How does the choice of the parameters of the prior* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *effect the model?*

Answer: The chosen covariance matrix tells us that the distribution of the parameters is independent which is optimal because we don't know much about the parameters. The mean being set to zero helps us to working with simpler datasets, for example, enhance model 3 to take on the other models with its parameters getting the value zero. And with large sigma helps the models getting different parameters to draw sharp linear boundaries on different complex data sets.

Question 22: *Plot the evidence over the whole dataset for each model (and sum the evidence for the whole of D, explain the numbers you get). The x-axis index the diffeerent instances in D and each models evidence is on the y-axis. How do you interpret this? Relate this to the parametrisation of each model.*

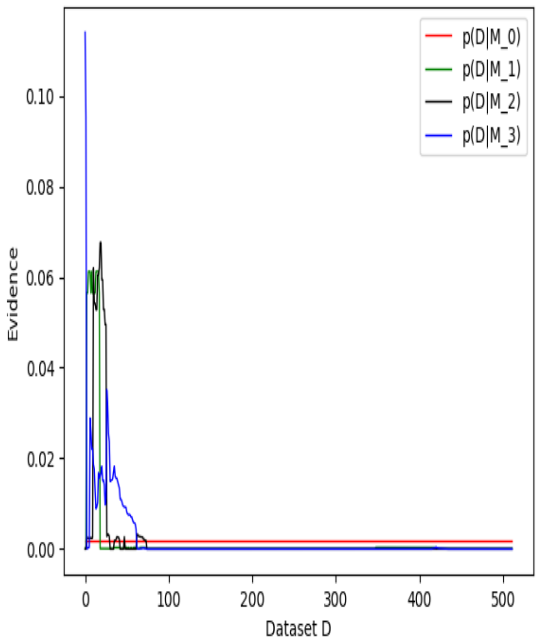

Fig 17. Evidence plot on the whole data domain

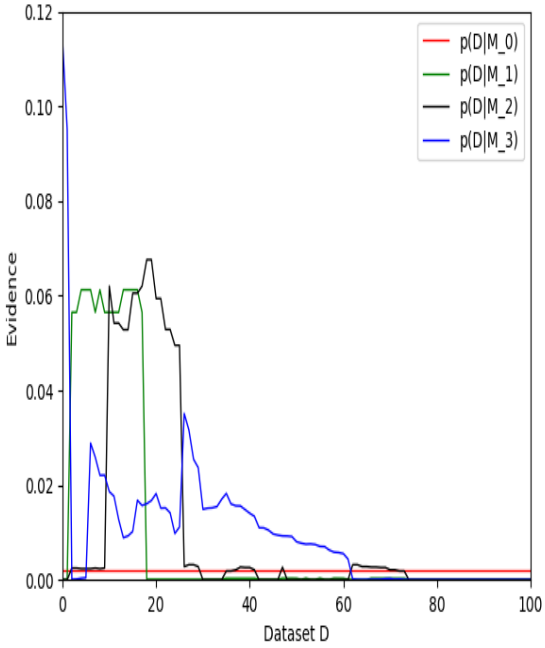

Fig 18. Sub set of the data domain

Answer: The sum of the evidence for the whole dataset D is equal to one corresponding to the total sum of probability distribution for each model. The interoperation worked as following

Fig 17 shows the evidence plot for all four models with respect to the datasets. As mentioned before in the theoretical part. M_0 with a uniform probability covers the whole data domain and thus is the simplest model. M_3 which favours the other models spreads more of its probability mass on a wider range of the D - axis, which can be seen in fig 17. Especially, for D above approximately 70 one can clearly see that for more complex datasets (on far left), model 3 covers the data sets because of its bias term. Consequently, model 1 and model 2 outperforms regarding the evidence probability of simple data sets, which you can draw a simple linear decision boundary on. This can be seen with the larger evidence on the datasets D less than approximately 30 on the D- axis in fig 18.

Question 21. *Find using* **np.argmax** *and* **np.argmin** *which part of the D that is given most and least probability mass by each model. Plot the data-sets which are given the highest and lowest evidence for each model. Discuss these results, does it make sense?*

Answer: Part of the data domain that was given most and least probability mass is shown below, the label corresponding to -1 is characterized by X and 1 with O. On the horizontal line, far down of each table is the x_1 axis $\{-1, 0, 1\}$ and the vertical line to far left is the x_2 axis $\{-1, 0, 1\}$.

Most probability mass on data sets:

model 0 model 1 model 2 model 3

X	X	X
X	X	X
X	X	X

X	X	O
X	X	O
X	X	O

O	O	O
X	O	O
X	X	X

X	X	X
X	X	X
X	X	X

Under model 0, data sets have same probability mass. Model 1 consists of x_1, thus can for a simple data set separate the labels on the x_1 axis. Model 2 takes x_1 and x_2 on consideration, therefore, can simply separate the data set above, linearly through the origin. Last model 3 favours other models and has a bias term which also allows sharp linear boundaries.

Least probability mass on data sets:

Model 0 model 1 model 2 model 3

X	X	X
X	X	X
X	X	X

O	X	X
X	X	X
X	X	O

X	X	O
X	X	X
X	O	X

X	O	O
X	O	X
O	O	X

Under model 0, data sets have same probability mass. For model 1, model 2 and model 3 the data sets have the lowest probability because of its complexity regarding separating the labels linearly, thus the decision boundaries are non-linear.

Q24: What is the effect of the prior $p(\boldsymbol{\theta})$.

- *What happens if we change its parameters?*
- *What happens if we use a non-diagonal covariance matrix for the prior?*
- *Alter the prior to have a non-zero mean, such that $\boldsymbol{\mu} = [5, 5]^T$?*
- *Redo evidence plot for these and explain the changes compared to using zero-mean.*

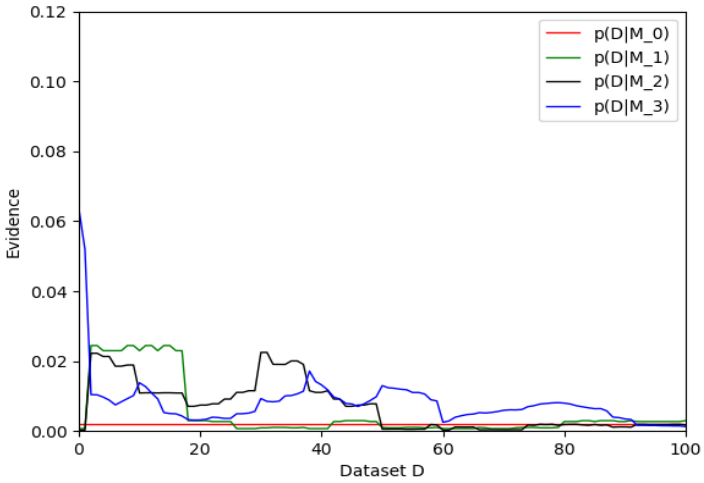

Fig 19. Evidence plot with $\sigma^2 = 5$

The first prior had a large variance $\sigma^2 = 1000$. For large sigma resulted in a sharper boundary which can be seen in fig 17. And for smaller variance, the result can be seen in fig 19, shows that the probability mass is wider distributed over the data domain. For even smaller sigma will result in much more uniformly evidence similar to the model M_0.

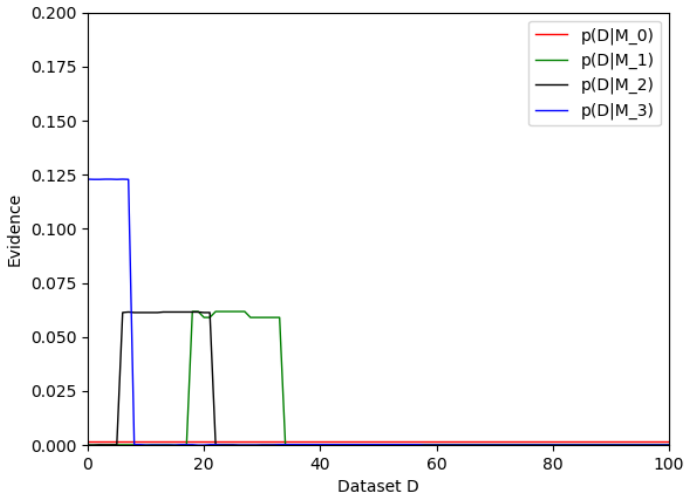

Fig 22. Evidence plot with non diagonal covariance matrix

The first prior had a covariance matrix $\boldsymbol{\Sigma}$ that showed that the parameters are independent. Because of the uncertainty regarding the parameters, we wanted to work with them independently. Inserting a non-diagonal covariance matrix in the prior will result in dependant parameters. Special case with figure 22. The used covariance matrix had the same value in the elements. Thus, tells us that it is seemingly perfect covariance between the parameters.

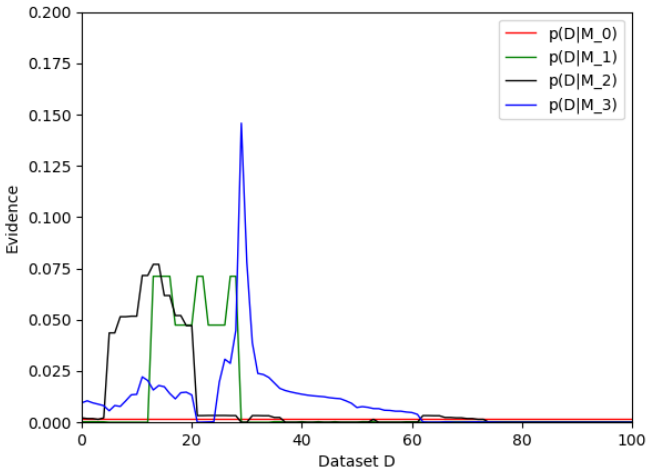

Fig 21. Evidence plot with $\mu = [5, 5]^T$

Prior with zero mean tells that the parameters are distributed with centre of $\boldsymbol{\theta} = \mathbf{0}$. By changing the mean to 5 results that the centre of the distribution is on that value. Hence our parameters are more likely getting higher values compared to with zero mean. Looking at the the mathematical expression of the evidence probabilities from the models. The model with larger number of parameters will show higher probability if the data points $\mathbf{t_i}$ and $\mathbf{x_i}$ are greater than zero for example. Depending on the data set, the evidence shows larger probability value. This can be seen in fig 20. Because M_3 is the most complexed one, with more parameters compared to the other models.

References:

[1] Pattern Recognition and Machine Learning, Christopher Bishop, 2006

[2] Matrix cookbook, Kaare Brandt Petersen and Michael Syskind Pedersen, 2012

[3Lectures slides

[4] “A note on the evidence and Bayesian Occam’s razor”, Iain Murray and Zoubin Ghahramani, 2005