DD2434 Machine Learning, Advanced Course Assignment 1

Alice Karnsund

November 2017

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

1.1 Theory

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?*

In a Bayesian approach one is interested in the uncertainty of the observations (Bishop 22). Since this is not known one can assume that the uncertainty over the data points constitute a great set of i.i.d. errors or noise. The *central limit theorem* states that the sum of a set of random variables has a distribution that approaches the Gaussian distribution as the number of terms in the sum increases (Bishop 79). Therefore one can assume that the noise constitutes a Gaussian distribution, and since the likelihood depends on the noise it is sensible to have a Gaussian form of the likelihood (Bishop 140).

A spherical covariance matrix means that the covariance matrix is proportional to the identity matrix, $\Sigma = \sigma^2\mathbf{I}$, known as an *isotropic* covariance, giving D+1 independent parameters in the model and spherical surfaces of constant density (Bishop 84). In such a case the dimensions of the data points will be independent of each other since the off diagonal elements in Σ are zero.

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

In a case where we assume that the data points are independent the likelihood takes the form:

$$p(\mathbf{T}|f,\mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{t}_i|f,\mathbf{x}_i) \tag{1}$$

But in the case where the data points are not independent we can, based on this and with the product rule of probability, get the following expression for

the likelihood:

$$p(\mathbf{T}|f,\mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{t}_i|\mathbf{t}_1,...,\mathbf{t}_{i-1},f,\mathbf{X}) \tag{2}$$

Here each part depends conditionally on all the inputs through $\mathbf{X}$ and all previous outputs through $\mathbf{t}_1$,...,$\mathbf{t}_{i-1}$.

1.1.1 Linear Regression

Question 3: *What is the specific form of the likelihood above, complete the right-hand side of side of the expression in (6).*

Given that the target variables $\mathbf{t}_i$ are given by a linear mapping $\mathbf{t}_i$=$\mathbf{W}\mathbf{x}_i$+ϵ where $\mathbf{W}\mathbf{x}_i$ is the deterministic function $\mathbf{y(x}_i\mathbf{,W)}$ and that the additive noise is Gaussian distributed $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$, we can write:

$$p(\mathbf{t}_i|\mathbf{x}_i,\mathbf{W},\sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{t}_i|\mathbf{y}(\mathbf{x}_i\mathbf{,W}),\sigma^2\mathbf{I}) \tag{3}$$

Now considering the whole data set and assuming that the data points are independent we get the following result (Bishop 141):

$$p(\mathbf{T}|\mathbf{X},\mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{t}_i|\mathbf{W}\mathbf{x}_i,\sigma^2\mathbf{I}) \tag{4}$$

Where $\mathbf{W}\mathbf{x}_i$ is the mean and $\sigma^2\mathbf{I}$ is the covariance, adapted from the Gaussian noise distribution $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$.

Question 4: *The prior in Eq.8 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.*

The spherical Gaussian prior encodes the preference as an L_2 norm which means that $p(\mathbf{W}) \propto exp(\mathbf{W}^2)$. Whereas if the prior is encoded in L_1 norm the prior takes the form $p(\mathbf{W}) \propto exp(|\mathbf{W}|)$. It can be shown that the median minimizes the L_1 norm while the mean minimizes the L_2 norm, which is the case in Gaussian distributions. The median is in practical less sensitive to outliers than the mean, thus if modelling using an L_1 norm one do not have to worry about outliers.

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 ($\boldsymbol{x}_i$,$\boldsymbol{t}_i$) 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 posterior is given by,

$$p(\mathbf{W}|\mathbf{X},\mathbf{T}) = \frac{1}{Z}p(\mathbf{T}|\mathbf{X},\mathbf{W})p(\mathbf{W}) \tag{5}$$

where the likelihood p($\mathbf{T}|\mathbf{X}$, $\mathbf{W}$) is as in equation 4,

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

If $\mathbf{W} \sim (D \times q)$ and $\mathbf{X} \sim (q \times N)$, giving $\mathbf{Y},\mathbf{T} \sim (D \times N)$. And because the product of two Gaussians is a Gaussian and the covariance is spherical we can write the likelihood in a simpler form,

$$p(\mathbf{T}|\mathbf{X},\mathbf{W}) = \mathcal{N}(\mathbf{T}|\mathbf{W}\mathbf{X},\sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{W}\mathbf{X},\sigma^2\mathbf{I}) \tag{6}$$

The prior p($\mathbf{W}$) is given by,

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

In a general case a multivariate Gaussian distribution takes the form,

$$\mathcal{N}(\mathbf{x}|\mu,S) \propto e^{\frac{-1}{2}(\mathbf{x}-\mu)^T S^{-1}(\mathbf{x}-\mu)}$$

where μ is a D-dimensional mean vector, Σ is a $D \times D$ covariance matrix. We can write the exponent as follows,

$$-\frac{1}{2}(\mathbf{x}-\mu)^T S^{-1}(\mathbf{x}-\mu) = -\frac{1}{2}\mathbf{x}^T S^{-1}\mathbf{x} + \mathbf{x}^T S^{-1}\mu + const \tag{7}$$

The crucial trick now is to complete the square in the exponent and from that identify the mean and covariance for the posterior distribution $\mathcal{N}(\mu,S)$. Following the steps in (Exercise 1) and making use of Bayes' theorem we proceed as follows (note, for simplicity I'll drop the bold notation from now on),

$$p(W|T,X) \propto e^{\frac{-1}{2\sigma^2}(T-WX)^T(T-WX)}\ e^{\frac{-1}{2\tau^2}(W-W_0)^T(W-W_0)}$$

We rewrite the exponent,

$$\frac{-1}{2\sigma^2}(T-WX)^T(T-WX)\ -\frac{1}{2\tau^2}(W-W_0)^T(W-W_0) =$$

$$= \frac{-1}{2\sigma^2}(WX)^T(WX) - \frac{1}{2\tau^2}W^TW + \frac{1}{\sigma^2}T^TWX + \frac{1}{\tau^2}W^TW_0 - \tag{8}$$

$$- \frac{1}{2\sigma^2}T^TT - \frac{1}{2\tau^2}W_0^TW_0 \tag{9}$$

The first two terms in 8 formulates a squared-term in W, the third and the forth term constitutes a linear term and the two terms in 9 are constants (independent of W).

Squared-term:

$$\frac{-1}{2\sigma^2}(WX)^T(WX) - \frac{1}{2\tau^2}W^TW = \frac{-1}{2}W^T(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)W$$

From this we can identify the posterior covariance S from 7 as

$$S^{-1} = \frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I \Rightarrow S = (\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1} \tag{10}$$

which is a combination of the input set X and the precision matrix (inverse of covariance matrix) of the prior distribution $\frac{1}{\tau^2}I$.

Linear-term:

$$\frac{1}{\sigma^2}T^TWX + \frac{1}{\tau^2}W^TW_0 = W^TS^{-1}\mu = W^T(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)\mu$$

Where I again made use of 7, for the linear term, and of 10. Rearranging we finally get the mean for the posterior,

$$\mu = (\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1}(\frac{1}{\sigma^2}X^TT + \frac{1}{\tau^2}W_0) \tag{11}$$

which is a product of the precision matrix S^{-1} of the posterior, the target matrix T, the input values X and the mean vector of the prior W_0. The whole posterior distribution $p(W|T,X)$ can now be written as

$$\mathcal{N}(W|\mu,S) = \mathcal{N}((\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1}(\frac{1}{\sigma^2}X^TT + \frac{1}{\tau^2}W_0), (\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I)^{-1})$$

This expression also match the one you get by simply matching your expressions for the likelihood an prior with those in Bishop eq. 2.113-2.117.

The constant Z is a normalizing constant, this means that the posterior distribution represents a true probability distribution, $p(W|T,X) \leq 1$. The Z term also represents the evidence function (Bishop 162), which we are interested in when considering model selection.

1.1.2 Non-parametric Regression

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.*

In a Gaussian process (GP) viewpoint, the prior is a probability distribution over the functions directly, instead of over the parameters. This is desirable since it is seldom obvious what kind of functions one should use. The relationship between the target value and the input value is

$$\mathbf{t}_i = f(\mathbf{x}_i) + \epsilon \rightarrow \mathbf{t}_i = f_i + \epsilon$$

where f_i is the output of the function at input location $\mathbf{x}_i$. The set of values $f(\mathbf{x}_i)$ evaluated at an arbitrary set of points $\mathbf{x}_1$,...,$\mathbf{x}_N$ jointly have a Gaussian distribution. Thus the marginal distribution and prior can be written,

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

where $\mathbf{K}$ is the Gram matrix determined by a kernel function. The kernel function is typically chosen to express the property that, for points $\mathbf{x}_n$ and $\mathbf{x}_m$ that are similar, the corresponding values $f(\mathbf{x}_n)$ and $f(\mathbf{x}_m)$ will be more strongly correlated than for dissimilar points (Bishop 306). This correlation is encoded in parameters of the kernel, which makes it possible to control the smoothness etc of the function.
For example, a common kernel function in GP regression is given by,

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 e^{\frac{\theta_1}{2}||\mathbf{x}_n - \mathbf{x}_m||^2} + \theta_2 + \theta_3 \mathbf{x}_n{}^T \mathbf{x}_m \tag{12}$$

Samples from this prior is shown in Figure 1, (taken from Bishop 308), where the parameters θ gives an idea how they can control the shape of the function.

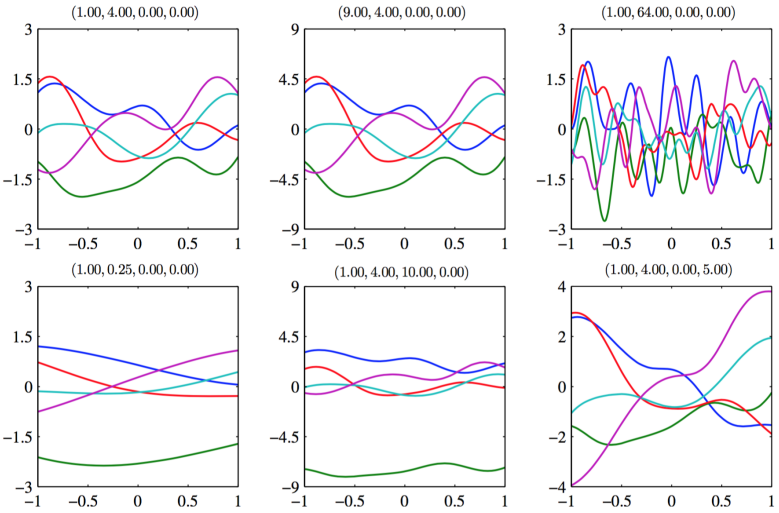


Figure 1: Samples from a Gaussian process prior defined by the covariance function (12). The title above each plot denotes (θ_0, θ_1, θ_2, θ_3), figure taken from Bishop p.308

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

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

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

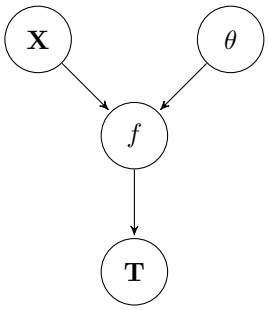
Given the prior and the relationship between the target values $\mathbf{T}$ and the input values,

$$\mathbf{t}_i = f_i + \epsilon$$
$$p(\mathbf{f}|\mathbf{X}, \theta) = \mathcal{N}(\mathbf{f}|\mathbf{0}, k(\mathbf{X},\mathbf{X}))$$

it is clear that $\mathbf{T}$ is dependent on f and that f in turn is dependent on $\mathbf{X}$ and θ, that is to say that $\mathbf{T}$ is conditionally dependent on $\mathbf{X}$ and θ. Also θ and $\mathbf{X}$ are independent. Using this information and the rules of probability theory we can formulate the joint likelihood as,

$$p(\mathbf{T}, \mathbf{X}, f, \theta) = p(\mathbf{T}|f)p(f|\theta, \mathbf{X})p(\mathbf{X})p(\theta) \tag{13}$$

These connections can also be put forward in a simple graphical model,



Question 8: *Complete the marginalization formula in Eq.12 (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 θ is left on the left-hand side of the expression after marginalization?*

Since we are not really interested in f itself, we have to marginalize it out. And from the previous question we know that $\mathbf{T}$ is conditionally dependent on θ and $\mathbf{X}$. Thus we get the general form (Bishop 306),

$$p(\mathbf{T}|\mathbf{X}, \theta) = \int p(\mathbf{T}|f)p(f|\mathbf{X}, \theta)df \tag{14}$$

Where the prior is $p(f|\mathbf{X}, \theta)$. Here we consider the average of all data $\mathbf{X}$ over all functions f.

The uncertainty between the function value and the true target value is stated in $p(\mathbf{T}|f)$, and $p(f|\mathbf{X}, \theta)$ states the uncertainty between $\mathbf{X}$ and f. And so these two terms together filter the uncertainty between the inputs $\mathbf{X}$ and the target values $\mathbf{T}$.
Since we were not interested in f we marginalized out f, and in the operation kept θ constant, therefore θ appears on the left-hand side as well.

1.2 Practical

1.2.1 Linear Regression

Question 9:
1. Set the prior distribution over W and visualize it.
2. Pick a single data-point from the data and visualize 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.

1.) In Figure 2 the prior distribution is shown for the parameters w_0 and w_1. At this stage we have no information available yet, thus this is a prior "guess" for the parameters.

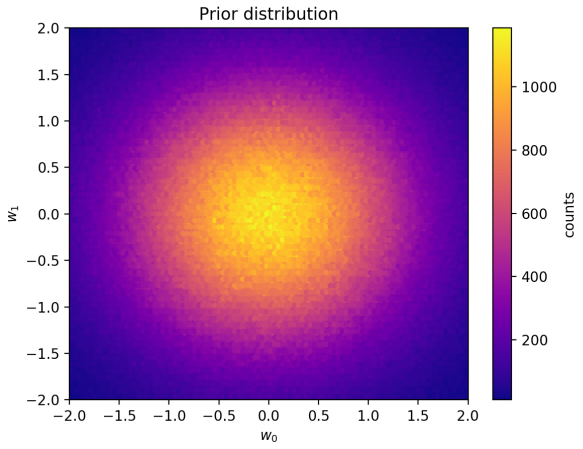


Figure 2: The prior distribution over $\mathbf{W}$=(w_0,w_1)

2.) After observing a single data point we get a slightly better intuition about the parameters. This posterior distribution is shown in Figure 3,

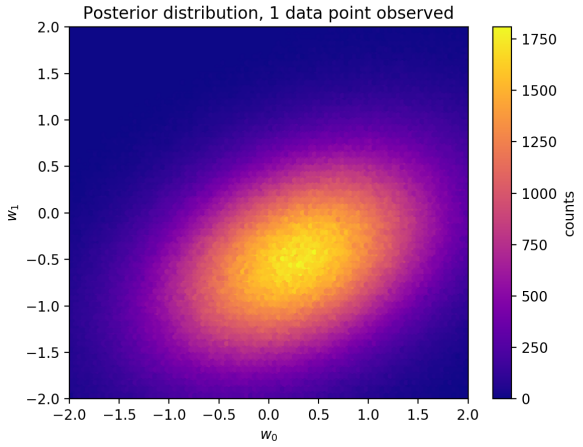


Figure 3: The posterior distribution over $\mathbf{W}$=(w_0,w_1) after observing a single data point

3.) Now we draw a couple of $\mathbf{w}$ values from the posterior in Figure 3 and show a couple of functions, Figure 4,

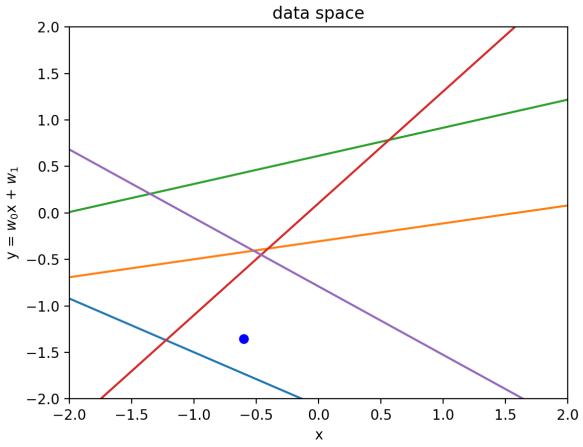


Figure 4: Functions y, where $\mathbf{W}$=(w_0,w_1) are drawn from the posterior (Figure 3) after observing a single data point, blue.

4.) Below, in Figure 5, I have computed and plotted the posterior after observing 2 and 15 new data points. To the right of these are function samples from each corresponding posterior. Note, not all data points are shown in the forth picture due to the fixed axes, but the relevant results are clearly shown.

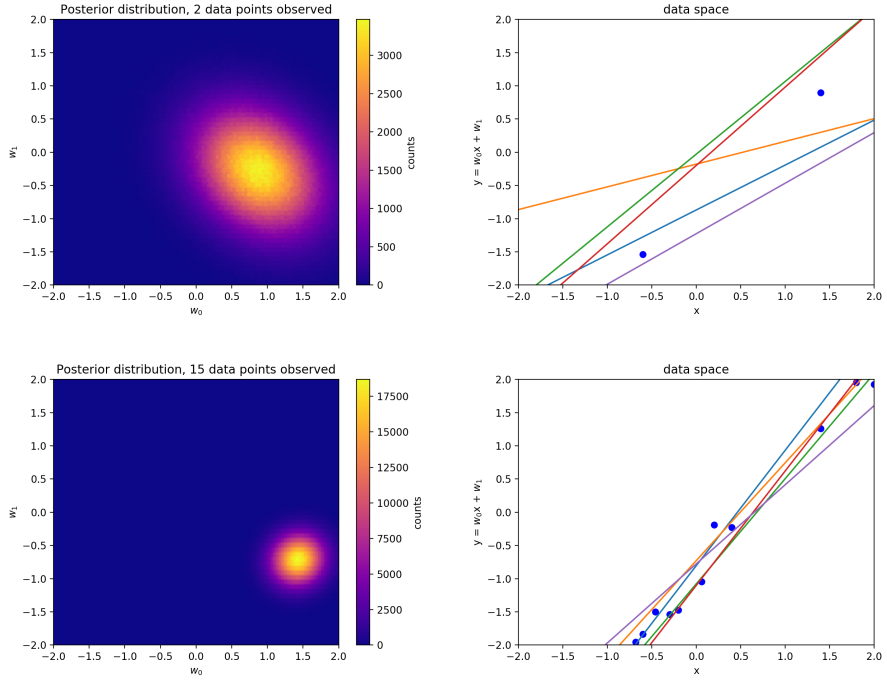


Figure 5: Posterior and function samples. The first row is after observing two points and the second row is after observing 15 points

Now I make use of all the 201 data points to show that we end up with the true w-values, $\mathbf{W} = (1.5, -0.8)$.

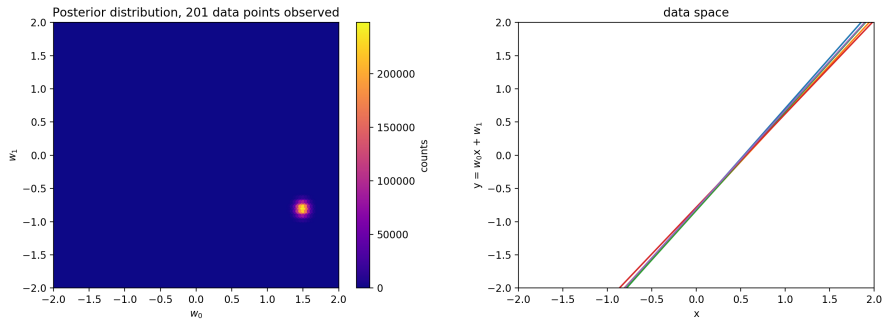


Figure 6: Posterior and function samples after observing all 201 data points.

Figure 6 is showing the posterior distribution together with five function samples after observing the whole data set. This proves that our method has recovered the true parameter values.

As we went on observing more and more data our posterior distribution got more and more sharp towards the correct values. In this case we are studying a linear regression and since two data points are sufficient to define a line,

the posterior in the top row of Figure 5, is already relatively compact. This is of course a desirable behaviour since the goal is to be able to determine the underlying parameters by only observing data.

1.2.2 Non-parametric Regression

Question 10:
1. Create a GP-prior with a squared exponential co-variance function.
2. Sample from this prior and visualize 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.

As can be seen from figure 7, the length factor controls the smoothness of the prior functions. As $l \to \infty$ the term in the exponent approaches zero and $k(x_i, x_j)$ approaches it's maximum value, disregarding σ_f. The element values in the covariance matrix, and thus how the dimensions co-vary, will be predominantly determined by the σ parameter. If we assume this parameter to be fixed, a larger l implies more correlation between the dimensions of f in $p(f|\mathbf{X}, \theta) = \mathcal{N}(\mathbf{0}, k(\mathbf{X},\mathbf{X}))$. If l takes on a relative small value, the dimensions will get more independent of each other as in the first picture in figure 7.

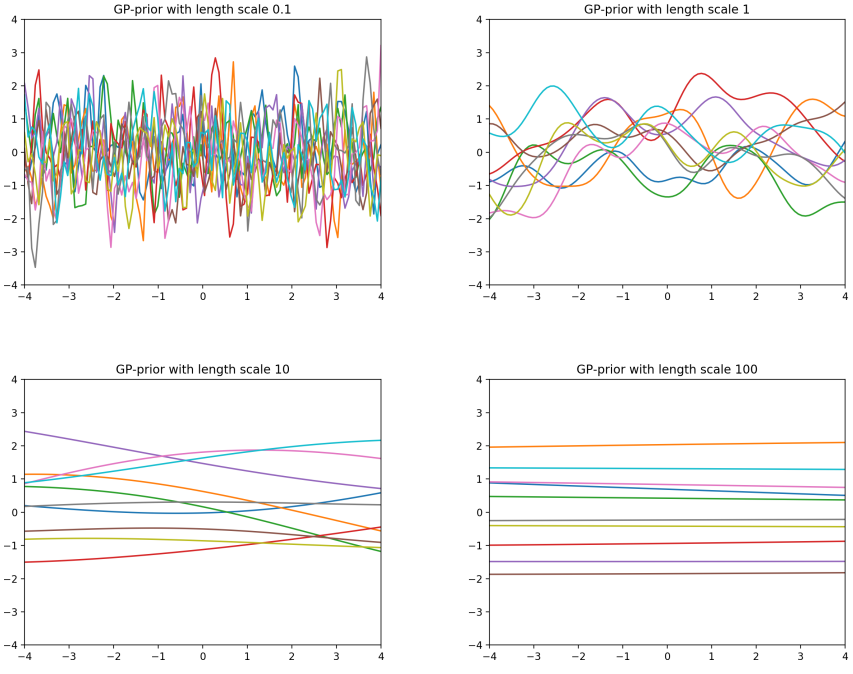


Figure 7: GP-prior with four different length-scales

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?

1.) Before we have observed any data we do not have any relevant information about the functions available, thus the posterior is interpreted as the prior,

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

2.) The predictive posterior is $p(t_{N+1}|\mathbf{t}_N)$, where $\mathbf{t}_N$ are the target values for the training set and t_{N+1} is the target value we wish to predict for a new observation x_{N+1}. We can obtain the predictive posterior from the joint distribution $p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$, where $\mathbf{t}_{N+1}$ is $(\mathbf{t}_N, t_{N+1})$, and using the steps in Bishop 2.3.1 (Bishop 309), it then takes the form,

$$p(t_{N+1}|\mathbf{t}) = \mathcal{N}(t_{N+1}|m(\mathbf{x}_{N+1}), \sigma^2(\mathbf{x}_{N+1})) \tag{15}$$

Where,

$$m(\mathbf{x}_{N+1}) = \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t}$$
$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k}$$

and

$$\mathbf{C}_{N+1} = \begin{bmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{bmatrix}$$

$\mathbf{C}_{N+1}$ is an $(N+1) \times (N+1)$ covariance matrix with elements $C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$, here β^{-1} is a hyper-parameter representing the precision of the noise, $\mathbf{k}$ is a vector with elements $k(\mathbf{x}_n, \mathbf{x}_{N+1})$ for $n = 1, ..., N$ and $c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1})$.

3.) Figure 8 is showing three samples from the posterior distribution together with the data points and the original cosine function they were generated from. In this case I used $\mathbf{C}_N$ with the elements, $C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m)$ for the case of visualization. As can be seen from the figure, the samples get much more gathered when approaching the data points, and further away, where there is no information available, the samples tends to oscillate in a much more uncontrolled way.

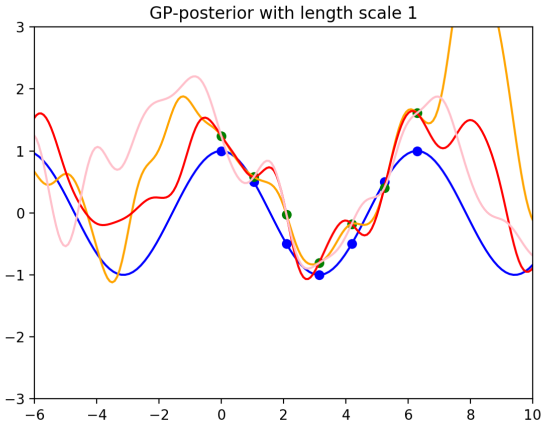


Figure 8: 3 samples (orange, pink, red) from the GP-posterior distribution. The blue curve is the $cos(\mathbf{x})$ function from which the data points (green) are obtained by addition of Gaussian noise. The blue points are the corresponding cosine points of the data points.

4.)

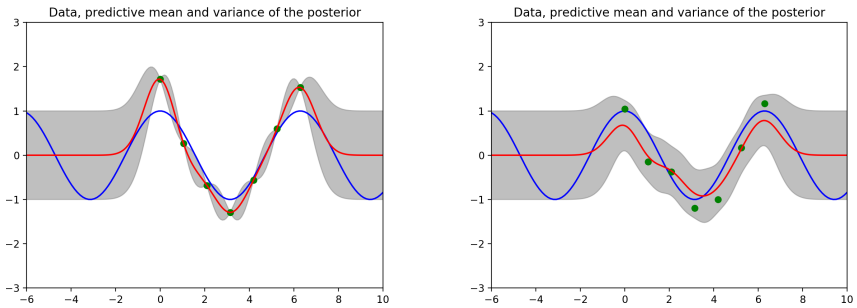


Figure 9: Data (green dots), predictive mean (red curve) and the original cosine function which generated the data (blue curve). Left image when $C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m)$ and right image with $C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$. Length scale 1 was used in this case as well.

In figure 9 I generated two different plots where one covariance matrix is without noise (left) and the other with (right). This gives an intuition of how the regression becomes more difficult when noise is added. The right image is thus the left image with a diagonal covariance matrix, with β^{-1} as diagonal elements, added to the exponential covariance matrix.
Comparing the second image in 7 with Figure 8, we can see that the difference between the behavior in the prior compared to the posterior is that the samples gets much more collected in the area where data is present. This behavior is of course desirable because it reveals information about the underlying function.

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

1.3 Theory

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

The prior over the latent variable $\mathbf{X}$ is defined as a spherical Gaussian,

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

which encodes our prior preferences about what properties the latent variable $\mathbf{X}$ should have. From this we can see that the dimension, a.k.a features, of $\mathbf{x}_i$'s don't co-vary, i.e. the off diagonal element in the covariance matrix for the observations is 0. With this preference encoded about $\mathbf{X}$ we simply achieve constraints on the $\mathbf{W}$ variable too, since there is a simple linear relationship between $\mathbf{X}$ and $\mathbf{W}$.

Question 13: *Perform the marginalisation*

$$p(\boldsymbol{Y}|\,\boldsymbol{W}) = \int p(\boldsymbol{Y}|\boldsymbol{X},\,\boldsymbol{W})p(\boldsymbol{X})d\boldsymbol{X} \tag{16}$$

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* $\boldsymbol{X}$ *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* $\boldsymbol{Y(X)}$*.*

Given that the marginal distribution is Gaussian and the relationship $\mathbf{Y(X)}$, we try to find the mean and covariance for $p(\mathbf{Y}|\mathbf{W})$. As stated in the theory part we are focusing on the same linear model as in Part I,

$$\mathbf{t}_i = \mathbf{W}\mathbf{x}_i + \epsilon \tag{17}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. From this we can construct our likelihood function as before (but with $\mathbf{t}_i \to \mathbf{y}_i$),

$$p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W}) = \mathcal{N}(\mathbf{y}_i|\mathbf{W}\mathbf{x}_i, \sigma^2\mathbf{I}) \tag{18}$$

Now we want to marginalize out the $\mathbf{x}_i$ to obtain an expression depending only on $\mathbf{y}_i$ and $\mathbf{W}$. First we note that both the likelihood (18) and the prior, $p(\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, are Gaussian distributions and thus the product will be Gaussian. Considering equation 16 for a single data point, it can be written,

$$p(\mathbf{Y}|\mathbf{W}) = \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{X})d\mathbf{X} \to p(\mathbf{y}_i|\mathbf{W}) = \int p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W})p(\mathbf{x}_i)d\mathbf{x}_i$$

Thus we aim to find the mean and covariance for $p(\mathbf{y}_i|\mathbf{W})$. Making use of equation 17, that the distributions for ϵ and $\mathbf{x}_i$ are independent and Bishop chap. 2.3 we get,

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

where I made use of the mean from the prior distribution $p(\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the mean from the noise distribution $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

$$Cov[\mathbf{y}_i|\mathbf{W}] = E[(\mathbf{y}_i - E[\mathbf{y}_i])(\mathbf{y}_i - E[\mathbf{y}_i])^T] = \{E[\mathbf{y}_i] = 0\} = E[(\mathbf{y}_i\mathbf{y}_i^T)] =$$
$$= E[(\mathbf{W}\mathbf{x}_i+\epsilon)(\mathbf{W}\mathbf{x}_i+\epsilon)^T] = E[\mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T]+E[\mathbf{W}\mathbf{x}_i\epsilon^T]+E[\epsilon\mathbf{W}^T\mathbf{x}_i^T]+E[\epsilon\epsilon^T] =$$
$$= \{E[\mathbf{x}_i] = 0, E[\epsilon] = 0\} = E[\mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T] + E[\epsilon\epsilon^T] = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Where I made use of the general relationship $E[\mathbf{x}\mathbf{x}^T] = \mu\mu^T + \Sigma$, (Bishop 83), the covariance of the prior and the noise, as well as that they both have mean $\mathbf{0}$.
Finally we can wright down the marginal distribution,

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

which for the full data set $\mathbf{Y}$, where the points are i.d.d, can be written,

$$p(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) \tag{19}$$

1.3.1 Learning

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. 25 equal?*

1.) Making use of Bishop 1.2.5 we find the following,
The negative log likelihood function:

$$-\ln p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mathbf{y}_i - \mathbf{W}^T\mathbf{x}_i)^2 + constant \tag{20}$$

From this we can see that maximizing the log likelihood function is the same as minimizing the sum-of-squares error function, i.e. the first term in equation 20 (Bishop 29).
With similar arguments we can see that maximum-a-posteriori is equal to minimizing the regularized sum-of-squares error function with the addition of a quadratic regularization term (Bishop 153),

$$-\ln p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mathbf{y}_i - \mathbf{W}^T\mathbf{x}_i)^2 + \frac{1}{2}\sum_{n=i}^{N}\mathbf{w}_i^2 + constant \tag{21}$$

As for the Type-II Maximum-Likelihood we have to minimize the first two terms in (adapted from question 15),

$$-\ln p(\mathbf{Y}|\mathbf{W}) = \frac{N}{2}ln(|\mathbf{W}\mathbf{W}^T - \sigma^2\mathbf{I}|) + \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y}_i) + constant \tag{22}$$

Comparing these three estimation procedures, we conclude that ML estimation is the least computational costly. But the other two estimation procedures have other benefits. For example, when dealing with models where two or more variables interact, as in our case, the Type-II Maximum-Likelihood estimation is a sensible approach.

2.) When more data is observed, only the first term in 21 will be affected and thus when a lot of data is observed the MAP estimate will approach the ML estimate.

3.) The posterior distribution here is given by,

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

comparing this to Bayes' theorem (Bishop 22),

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

we see that the denominators must be equal for Bayes' theorem to hold. Thus this means that the denominator in both cases are independent of $\mathbf{W}$, which we want to maximize over. Therefore the denominator can be seen as a constant when maximizing over $\mathbf{W}$,

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

Question 15:
1. Write down the objective function -log($p(\boldsymbol{Y}|\boldsymbol{W})$) = $\mathcal{L}(\boldsymbol{W})$ for the marginal distribution in Eq. 23.
2. Write down the gradients of the objective with respect to the parameters $\frac{\delta\mathcal{L}}{\delta\boldsymbol{W}}$

Here we are interested in the logarithm of the marginal distribution,

$$p(\mathbf{Y}|\mathbf{W}) = \int p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{X})d\mathbf{X} = \prod_{i=1}^{N}\mathcal{N}(\mathbf{y}_i|\mathbf{0},\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}) \qquad (23)$$

1.) Taking the negative logarithm of equation 23 we get,

$$\mathcal{L}(\mathbf{W}) = -ln(p(\mathbf{Y}|\mathbf{W})) = -ln(\prod_{i=1}^{N}\mathcal{N}(\mathbf{y}_i|\mathbf{0},\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})) = \{product-rule\} =$$

$$= -\sum_{i=1}^{N} ln(\mathcal{N}(\mathbf{y}_i|\mathbf{0},\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})) = \left\{\mathcal{N} \sim \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu))\right\} =$$

$$= -\sum_{i=1}^{N} ln(\frac{1}{(2\pi)^{D/2}|\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I}|^{1/2}}exp(-\frac{1}{2}(\mathbf{y}_i^T(\mathbf{W}\mathbf{W}^T+\sigma^2\mathbf{I})^{-1}\mathbf{y}_i)) =$$

$$= -\sum_{i=1}^{N}[(-ln((2\pi)^{D/2}|\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}|^{1/2}) + (-\frac{1}{2}(\mathbf{y}_i^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y}_i))] =$$

$$= \frac{ND}{2}ln(2\pi) + \frac{N}{2}ln(|\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}|) + \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y}_i)$$

Making use of the proof of "Theorem 1" in [1]:

$$\Lambda_n(\mu, \Omega) = -\frac{1}{2}mnlog(2\pi) - \frac{1}{2}nlog|\Omega| - \frac{1}{2}tr\Omega^{-1}Z \tag{24}$$

where,

$$Z = \sum_{i=1}^{N}(y_i - \mu)(y_i - \mu)'$$

and $\Lambda_n = logL_n$ where L_n is the likelihood function.

Comparing expression 24 with the previous one we end up with,

$$\mathcal{L}(\mathbf{W}) = \frac{ND}{2}ln(2\pi) + \frac{N}{2}ln(|\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}|) + \frac{1}{2}tr((\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{Y}^T\mathbf{Y})$$

Here tr is the trace of a matrix, $tr(A) = \sum_{i=1}^{n} a_{ii}$.

2.) Now we wish to calculate the gradients of $\mathcal{L}(\mathbf{W})$ with respect to the parameters. For this we will make use of a couple of rules from [2],

$$\begin{aligned} \partial tr(\mathbf{X}) &= tr(\partial\mathbf{X}) \\ \partial(ln(det(\mathbf{X}))) &= tr(\mathbf{X}^{-1}\partial\mathbf{X}) \\ \partial\mathbf{X}^{-1} &= -\mathbf{X}^{-1}(\partial\mathbf{X})\mathbf{X}^{-1} \end{aligned}$$

$$\frac{\partial(\mathbf{X}^T\mathbf{A})_{ij}}{\partial\mathbf{X}_{mn}} = \delta_{in}(\mathbf{A}_{mj}) = (\mathbf{J}^{nm}\mathbf{A})_{ij}$$

Where $\mathbf{J}^{nm}$ is the single-entry matrix, 1 at (n,m) and zero elsewhere.
The gradient of $\mathcal{L}(\mathbf{W})$ can now be derived using these rules,

$$\frac{\partial\mathcal{L}(\mathbf{W})}{\partial\mathbf{W}_{ij}} = \frac{N}{2}tr((\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1} \times (\mathbf{J}_{ij}\mathbf{W}^T + \mathbf{W}\mathbf{J}_{ij}^T))+$$

$$-\frac{1}{2}tr(\mathbf{Y}^T\mathbf{Y}((\mathbf{W}\mathbf{W}^T + \sigma^2 I)^{-1} \times (\mathbf{J}_{ij}\mathbf{W}^T + \mathbf{W}\mathbf{J}_{ij}^T) \times (\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}))$$

[1] Magnus & Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, Version: 07/01 January 16, 2007, Page 353

[2] Petersen & Pedersen, The Matrix Cookbook, Version: January 5, 2005, Page 7-9

1.4 Practical

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

Here we want to learn $\mathbf{W} \sim \mathbf{A}$ and then recover $\mathbf{X}'$ when only knowing the $\mathbf{Y}$-values. We know that $\mathbf{Y}$ is a linear combination of $\mathbf{X}'$ and $\mathbf{W}$. In this exercise we are finding the estimate of $\mathbf{W}$ through a type-II Maximum-likelihood estimate as

$$\hat{\mathbf{W}} = argmax_W \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{X})d\mathbf{X} = argmin_W \mathcal{L}(\mathbf{W})$$

After we have estimated the $\hat{\mathbf{W}}$-values, we can find an estimate of $\mathbf{X}'$ through the following relationships,

$$\mathbf{Y} = \mathbf{X}'\hat{\mathbf{W}}^T$$
$$\mathbf{X}' = \mathbf{Y}\hat{\mathbf{W}}(\hat{\mathbf{W}}^T\hat{\mathbf{W}})^{-1}$$

In Figure 10 we can see plots of both the true $\mathbf{X}'$-values (red curve) and the learned $\mathbf{X}'$-values (blue curve). Comparing these two we can see that have similar shapes, but they are scaled and rotated compared to each other. This is because the likelihood distribution is invariant to a rotation in the parameters $\mathbf{W}$. Thus all linear combinations $\hat{\mathbf{W}} = \mathbf{WR}$ will work. Here $\mathbf{R}$ is an orthogonal matrix. This means that there exists no unique solution for $\mathbf{W}$ under these circumstances.[3]

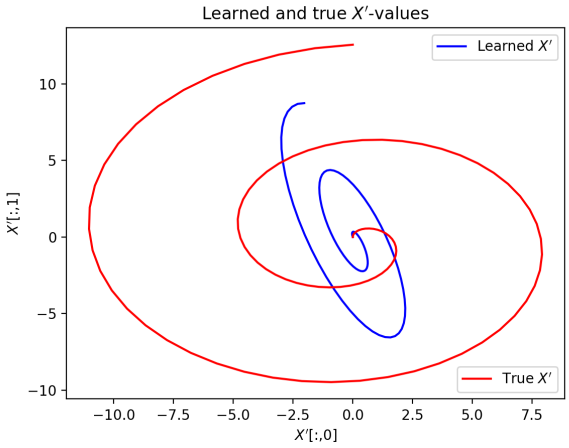


Figure 10: Learned $\mathbf{X}'$-values (blue curve) and true $\mathbf{X}'$-values (red curve)

[3]DD2434 lecture 6, November, 2017, Page 40

$$
\begin{aligned}
\hat{\mathbf{W}} &= \mathbf{WR} \\
p(\mathbf{y}_i|\hat{\mathbf{W}}) &= \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \hat{\mathbf{W}}\hat{\mathbf{W}}^T + \sigma^2\mathbf{I}) \\
&= \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathbf{WRR}^T\mathbf{W}^T + \sigma^2\mathbf{I}) \\
&= \mathcal{N}(\mathbf{y}_i|\mathbf{0}, \mathbf{WW}^T + \sigma^2\mathbf{I})
\end{aligned}
$$

III. The Evidence p(D)

1.5 Theory

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?*

$p(D|M_0, \theta_0) = \frac{1}{512}$ may be said to be the simplest model in the sense that it has no free parameters and treats all data sets equally. It defines a single distribution over all the data sets, assigning them all probability 1/152. M_0 has the largest evidence over a large range of data sets, compared to other models with more parameters such as M_1, M_2 and M_3. Though a more negative side is that M_0 is unable to assign as much probability mass to simple data sets. Simple models usually choose to concentrate their probability mass around a limited number of data sets and complex models predict that data will be drawn from a large range of probabilities. Thus, in some sense M_0 is a complex model, it assigns many different types of behaviours similar probability.[4]

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?*

M_3 is standard logistic regression, which has a bias term θ_3^3 that can account for very unequal distributions in the data set, for example a set $D_1 = \{t^i\}_{i=1}^9$ where all $t_i = +1$. This model can be seen as the most complex one in a way, since it has the most parameters to be tuned and can also realize all the other models by setting some of the parameters to zero. This implies that M_3 spreads it bulk of unit probability mass over a wider range of data sets than the other models.
M_2 is like M_3 but without a bias term, thus it does not favour very unequal distributions in the data set. But since M_2 includes both x_1 and x_2 it is capable of modelling decision boundaries for data sets that are due to rotation invariance, whereas this is not possible for M_1. Since M_2 is lacking the bias term it cannot spread its probability mass as much as M_3 or M_0 over D.
Now M_1 is M_2 but only including the first dimension of $\mathbf{x}$. Due to this M_1 can

[4]Murray & Ghahramani, A note on the evidence and Bayesian Occam's razor, August 2005, Page 3

easily model data sets that has decision boundaries which are not a function of x_2. This means that M_1 concentrates its probability mass around these types of data sets, whereas M_0 account for the probability that a data set can take on many different forms. So, data sets with decision boundaries that may look alike under M_2 is not always possible to model using M_1 due to rotation invariance.[5]

Question 19: *How have the choices we made above restricted the distribution of the model? What data sets 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.?*

Some point where already brought up in question 18. As mentioned there, M_1 concentrates its probability mass around data sets that has decision boundaries which are only a function of x_1 and will have higher probabilities for data sets where the decision boundaries crosses the origin. M_2 will also put higher probability to data sets in this area, though since it can handle decision boundaries which are functions of both x_1 and x_2 it will stretch out more than M_1. M_3 stretches its probability over even a greater amount of data sets since it is including a bias term θ_3^3 which allows decision boundaries to be offset from the origin. As mentioned in the previous question M_3 can be considered the most complex model, but at the same time M_0 can be considered complex as it assigns, many different types of behaviours, similar probability. This complexity can also be argued by looking at the distribution graph, where M_3 and M_0 grasps a greater amount of data sets. Though if we observe data that can be explained well by a simple model, more complex model such as M_3 and M_0, which have spent more of their available probability mass elsewhere, will be automatically penalized.[6] Since M_3 and M_0 allows for a greater amount of different data sets, they can be said to be more flexible when considering the amount of data sets where the decision boundaries can be modelled. But as M_3 gets more complex the risk of over fitting grows, though a complex model is not always as flexible in a general case.

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

To reach the evidence $p(D|M_i)$ of a model M_i we have to marginalize out the parameters,

$$p(D|M_i) = \int_{\forall\theta} p(D|M_i, \theta)p(\theta)d\theta$$

Here one can interpret the marginal likelihood, i.e. the evidence, as a probability of generating the data set D from a model whose parameters are sampled at random from the prior $p(\theta)$, (Bishop 162). To choose models in this way will not favour the models with the most parameters. Though this would be the case if

[5]Murray & Ghahramani, A note on the evidence and Bayesian Occam's razor, August 2005, Page 3

[6]Murray & Ghahramani, A note on the evidence and Bayesian Occam's razor, August 2005, Page 1

we used $p(D|\hat{\theta}_m)$ to select models, where $\hat{\theta}_m$ is the MLE or MAP estimate of the parameters for model m, because models with more parameters will fit the data better and hence achieve a higher likelihood. But if we instead integrate out the parameters we are automatically protected from over-fitting, models with more parameters do not necessarily have higher marginal likelihood. This is called the Bayesian Occam's razor effect.[7]

Question 21: *What does this choice of prior imply? How does the choice of the parameters of the prior μ and Σ effect the model?*

The prior over the parameters of the model is defined as,

$$\begin{aligned} p(\theta|M_i) &= \mathcal{N}(\mu, \Sigma) \\ \mu &= \mathbf{0} \\ \Sigma &= \sigma^2\mathbf{I} \\ \sigma^2 &= 10^3 \end{aligned}$$

This prior implies that, since $\Sigma = \sigma^2\mathbf{I}$, that the dimensions of the parameters are independent, since the off diagonal elements of the covariance matrix is zero. When considering a σ^2 of this magnitude, this means that we favour simpler models, since each parameter is "allowed" to vary in magnitude by a lot.[8]

1.6 Practical

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 different 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.*

At a first look, Figure 11 and 12 are very similar to the plots in [9], which is what we aimed for. Here I used 5000 samples (S) instead of 10^8.
In Figure 11 and 12 we can see that M_0 (pink) spreads over the whole data set with an equal probability for all, as expected. M_1 (blue) focus its probability mass near the origin where the decision boundaries of the data sets can be described in one dimension, x_1. M_2 (red), will spread out more than M_1, due to its ability to model decision boundaries in two dimensions. Lastly, M_3 (green) reaches over an even greater set than both M_1 and M_2. This is because M_3 has three parameters, which makes it possible to derive any other model from this one, by setting some of the parameters equal to zero. This spread over the possible sets implies that M_3 is more complex than M_1 and M_2, in the sense of numbers of parameters. Another crucial feature that only M_3 has, is that it is able to model very unequal distributions in the data set, thus the pike by the origin.

[7] Murphy, Machine Learning: A Probabilistic Perspective, 2012, Page 156
[8] Murphy, Machine Learning: A Probabilistic Perspective, 2012, Page 162
[9] Murphy, Machine Learning: A Probabilistic Perspective, 2012

The sum of the evidence for each model, over the whole data set D, all sum to one. This is because they all represent true probability distribution, the only difference is that they distribute this probability mass in different ways.

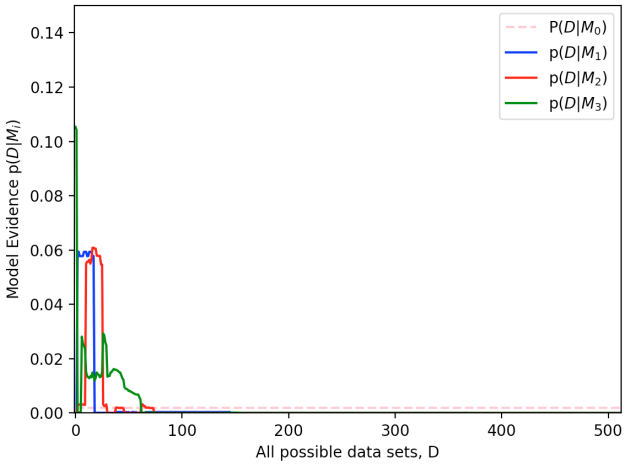


Figure 11: Evidence plot for all models over the whole data set D.

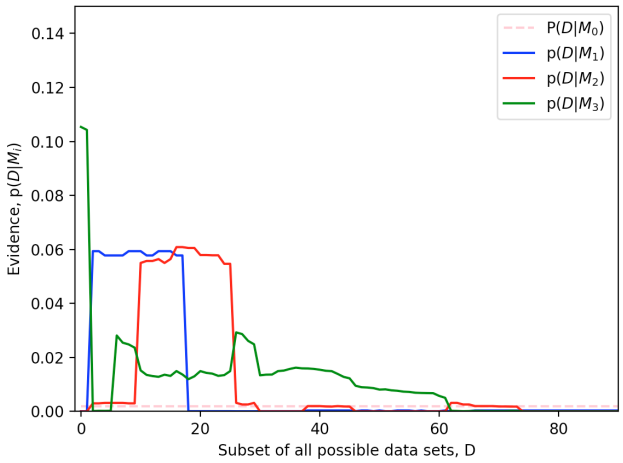


Figure 12: Evidence plot for all models over a subset of D.

Question 23: *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?*

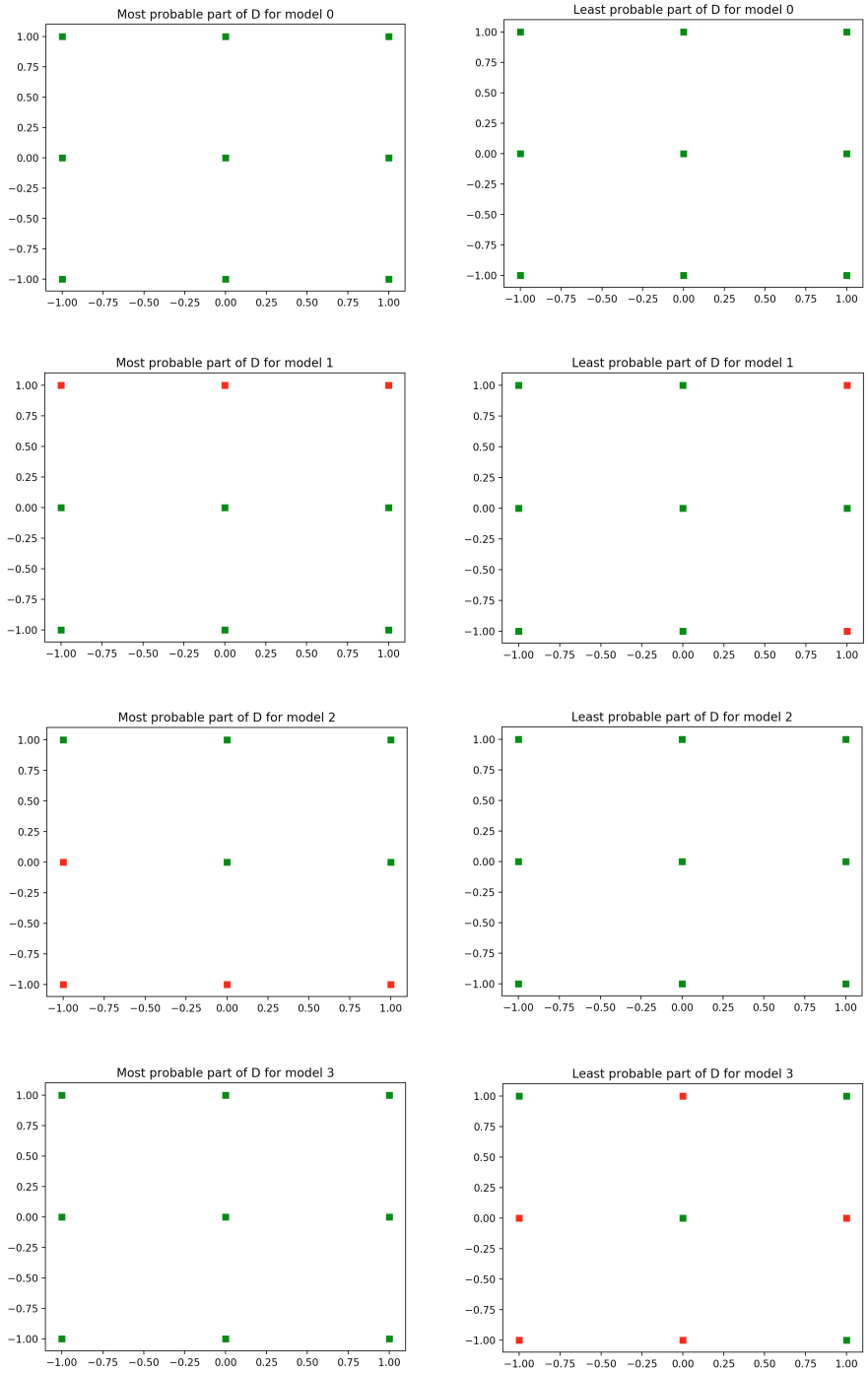


Figure 13: Most and least probable data sets for the different models, where x_2 is on the x-axis and x_1 is on the y-axis

Figure 13 supports the previous arguments about what decision boundaries each model is suited for. M_0 has the same probability for every kind of data set, therefore it is reasonable that it has same kind of set as both the most and least reasonable. M_1 can only model decision boundaries in one dimension, which is shown by the second row of 13. The least probable set is a very non linear distribution. Row three shows that the most probable data set constitutes a decision boundary described by two variables, therefore this is consistent with the capabilities of M_2. On the other hand M_2 is lacking the bias term, therefore a very uneven distribution set is the least probable one. At the last row we have M_3 which here clearly shows that it supports very uneven distributions, in line with the theory. The least probable set is a set that cannot be modeled using a linear decision boundary.

Question 24: *What is the effect of the prior $p(\theta)$.*
- *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 $\mu = [5, 5]^T$?*
- *Redo evidence plot for these and explain the changes compared to using zero-mean.*

The prior $p(\theta)$ expresses our prior guess about the parameters θ of the models. If we change the parameters of the prior, i.e. the mean and the variance, we will obtain a different prior distribution. A change in the mean will lead to that posterior distributions near that mean will contribute more to the marginal posterior. Predicted probabilities from parameters close to the mean are given higher probabilities than those further away. With a mean of zero this means that θ-values close to zero are given the largest probability. A change in the variance will affect the smoothness of the evidence function. A lower value on the variance will make the parametric models spread out more, approaching the distribution of M_0, and in a sense become more complex. But instead, large values of the variance gives rise to more spiky distributions and favour sharp decision boundaries, as shown in the previous parts.
If we use a non-diagonal covariance matrix, it implies that the dimensions of the parameters θ for each model will be dependent and co-vary with each other. This correlation will effect how probable a certain data set will be.
In Figure 14 and 15 we see the model evidences after changing the mean to $\mu = [5, 5]^T$. We can see, when comparing these two figures with Figure 11 and 12, that the evidence takes on higher values at later data sets, especially M_3. Here we expect that predicted probabilities from θ-values close to a mean of $[5, 5]^T$ should be given higher weight than those further away. Since the distributions are more peaked around other data sets than before, we can draw the conclusion that this new mean will put larger probability mass to data sets where the decision boundaries corresponds to θ-values close to this mean.

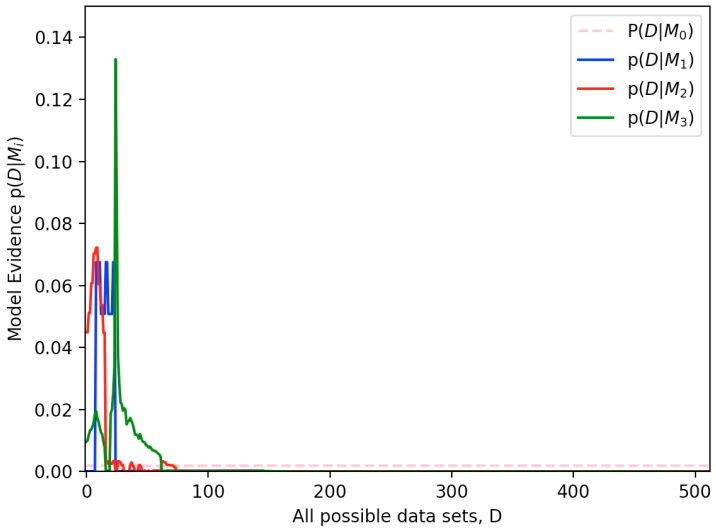


Figure 14: Evidence plot for all models over the whole data set D, with mean $\mu = [5, 5]^T$

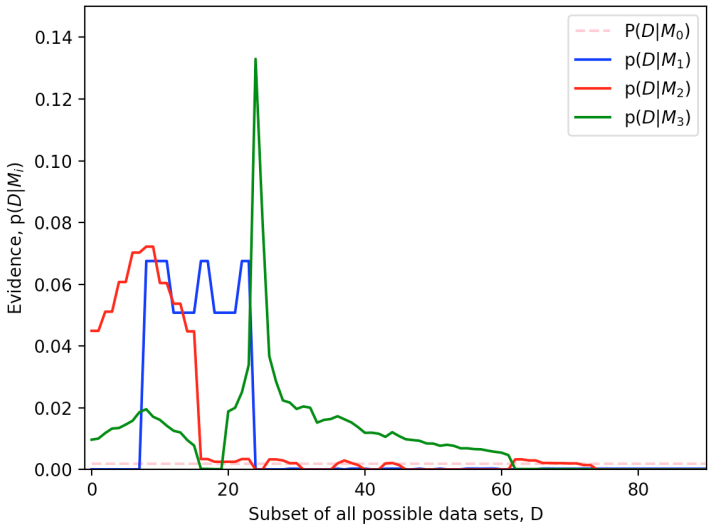


Figure 15: Evidence plot for all models over a subset of D, with mean $\mu = [5, 5]^T$