Models

Jeonghyun Kim[1] and Sizhe Sun[2]

[1] *candidate number: 97485*
[2] *candidate number: 97198*

November 2, 2018

This report focuses on basic machines learning theories and techniques, including linear regression and non-parametric methods with Gaussian process. It demonstrates their use in both supervised and unsupervised learning contexts. The third section extends the theories further to model complexities. Practical are included to consolidate our understanding of the theory and exposes some oddity of implementation details.

1 The Prior

1.1 Theory

Q1

1. The choice of Gaussian likelihood function implies the assumption that the mapping f we are estimating has some degree of uncertainty of ϵ, possibly caused by some additive Gaussian noise in measurement, i,e.

   $$\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon_i.$$

   The uncertainty is presumably modelled by a normal distribution

   $$\epsilon_i \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

   Hence the likelihood can be modelled by a normal distribution with variance of $\sigma^2\mathbf{I}$ and mean of $f(\mathbf{x}_i)$, i.e.

   $$p(\mathbf{y}_i|f, \mathbf{x}_i) \sim \mathcal{N}(f(\mathbf{x}_i), \sigma^2\mathbf{I})$$

2. By using a spherical covariance matrix for the likelihood, we assume that the random variables for both dimensions are independent, reducing the number of independent parameters in the covariance matrix from $D(D+1)/2$ of the general case, to just D. As we simplify our model, the computational complexity decreases, so as the expressiveness of our model, which means we might fail to capture some correlations in the data.

Q2 We can split the large joint distribution $p(\mathbf{y}_1|\mathbf{y}_N \cdots \mathbf{y}_2, f, \mathbf{X})p(\mathbf{y}_N \cdots \mathbf{y}_2, f, \mathbf{X})$ by applying the product rule repeatedly.

$$\begin{aligned}
&p(\mathbf{Y}|f, \mathbf{X}) \\
&= p(\mathbf{y}_N, \mathbf{y}_{N-1} \cdots \mathbf{y}_2, \mathbf{y}_1 | f, \mathbf{x}_N, \mathbf{x}_{N-1} \cdots \mathbf{x}_2, \mathbf{x}_1) \\
&= \frac{p(\mathbf{y}_N, \mathbf{y}_{N-1} \cdots \mathbf{y}_2, \mathbf{y}_1, f, \mathbf{x}_N, \mathbf{x}_{N-1} \cdots \mathbf{x}_2, \mathbf{x}_1)}{p(f, \mathbf{x}_N, \mathbf{x}_{N-1} \cdots \mathbf{x}_2, \mathbf{x}_1)} \\
&= \frac{p(\mathbf{y}_1|\mathbf{y}_N \cdots \mathbf{y}_2, f, \mathbf{X})p(\mathbf{y}_N \cdots \mathbf{y}_2, f, \mathbf{X})}{p(f, \mathbf{X})} \\
&= \frac{\prod_{i=1}^{N} p(\mathbf{y}_i | \bigcap_{j=i+1}^{N} \mathbf{y}_j, f, \mathbf{X})p(f, \mathbf{X})}{p(f, \mathbf{X})} \\
&= \prod_{i=1}^{N} p\left(\mathbf{y}_i \,\middle|\, \bigcap_{j=i+1}^{N} \mathbf{y}_j, f, \mathbf{X}\right)
\end{aligned}$$

1.1.1 Linear Regression

Q3 Note: is $\mathbf{W}$ supposed to be a column vector of parameters or a $D \times q$ matrix? The prior $p(W) \sim \mathcal{N}(\mathbf{W_0}, \tau^2\mathbf{I})$ suggests the former while the equation $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \epsilon$ suggests the later. I suspect this is result of either my misunderstanding or abusive notations.

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

Q4 Conjugate distributions are probability distributions with the same functional form
From Bayes' theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int_{\theta'} p(x|\theta')p(\theta')d\theta'}$$

For a known likelihood function $p(x|\theta)$, we can choose the prior such that the prior and the posterior are conjugate distributions, i.e. forming a conjugate pair and sharing the same functional form. In this case it is possible to avoid computing the integral for the evidence, and to treat it as a normalising constant instead. Essentially,

$$p(\theta|x) = \frac{1}{Z}p(x|\theta)p(\theta)$$

We can then derive value of the posterior since we already know the prior, the likelihood, and the functional form of the posterior.

Q5 Since we assume that the observations are corrupted by additive Gaussian noise, i.e. we have a Gaussian likelihood function, then picking a Gaussian prior gives a Gaussian posterior as the Gaussian distribution is self-conjugate. The use of a spherical covariance matrix is effectively encoding a scaled Euclidean distance.

Q6

$$\begin{aligned}
&p(\mathbf{W}|\mathbf{X},\mathbf{Y})\\
&= \frac{1}{Z}p(\mathbf{Y}|\mathbf{X},\mathbf{W})p(\mathbf{W})\\
&= \frac{1}{Z}\mathcal{N}(\mathbf{Y}|\mathbf{XW},\sigma^2\mathbf{I})\mathcal{N}(\mathbf{W}|\mathbf{W}_0,\tau^2\mathbf{I})\\
&= \frac{1}{Z}\left(\frac{1}{(2\pi)^{\frac{D}{2}}|\sigma^{\mathbf{2}}\mathbf{I}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{Y}-\mathbf{XW})^{\mathrm{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{Y}-\mathbf{XW})}\right)\\
&\quad\left(\frac{1}{(2\pi)^{\frac{D'}{2}}|\tau^{\mathbf{2}}\mathbf{I}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{W}-\mathbf{W}_0)^{\mathrm{T}}(\tau^2\mathbf{I})^{-1}(\mathbf{W}-\mathbf{W}_0)}\right)\\
&= \frac{1}{Z}\left(\frac{1}{(2\pi)^{\frac{D}{2}}|\sigma^{\mathbf{2}}\mathbf{I}|^{\frac{1}{2}}}e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{XW})^{\mathrm{T}}(\mathbf{Y}-\mathbf{XW})}\right)\\
&\quad\left(\frac{1}{(2\pi)^{\frac{D'}{2}}|\tau^{\mathbf{2}}\mathbf{I}|^{\frac{1}{2}}}e^{-\frac{1}{2\tau^2}(\mathbf{W}-\mathbf{W}_0)^{\mathrm{T}}(\mathbf{W}-\mathbf{W}_0)}\right)\\
&= \frac{1}{Z}\left(\frac{1}{(2\pi)^{\frac{D+D'}{2}}|\sigma^{\mathbf{2}}\mathbf{I}|^{\frac{1}{2}}|\tau^{\mathbf{2}}\mathbf{I}|^{\frac{1}{2}}}\right)\\
&\quad\exp(-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{XW})^{\mathrm{T}}(\mathbf{Y}-\mathbf{XW})\\
&\quad-\frac{1}{2\tau^2}(\mathbf{W}-\mathbf{W}_0)^{\mathrm{T}}(\mathbf{W}-\mathbf{W}_0))\\
&\exp(-\frac{1}{2\sigma^2}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}-2\mathbf{Y}^{\mathrm{T}}(\mathbf{XW})+(\mathbf{XW})^{\mathrm{T}}(\mathbf{XW}))\\
&\quad-\frac{1}{2\tau^2}(\mathbf{W}^{\mathrm{T}}\mathbf{W}-2\mathbf{W}_0^{\mathrm{T}}\mathbf{W}+\mathbf{W}_0^{\mathrm{T}}\mathbf{W}_0))\\
&= \exp(-\frac{1}{2\sigma^2}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}-\frac{1}{2\tau^2}\mathbf{W}_0^{\mathrm{T}}\mathbf{W}_0\\
&\quad+\mathbf{W}^{\mathrm{T}}(\frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{Y}+\frac{1}{\tau^2}\mathbf{W}_0)\\
&\quad+\mathbf{W}^{\mathrm{T}}(-\frac{1}{2\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}-\frac{1}{2\tau^2})\mathbf{W})\\
&= \mathcal{N}(\mathbf{W}|\mathbf{\Sigma}^{-1}(\frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{Y}+\frac{1}{\tau^2}\mathbf{W}_0),\mathbf{\Sigma}^{-1})
\end{aligned}$$

where $\mathbf{\Sigma} = \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}+\frac{1}{\tau^2}$.

We observe that by multiplying the likelihood (which is Gaussian) by the Gaussian prior we indeed get a Gaussian distribution, which is proportional to the posterior.

1.1.2 Non-parametric Regression

Q7 A non-parametric model, unlike a parametric one, does not require a specific predetermined form of the model with parameters in a finite dimension space. Instead, the model is built by using observed data. If we have a large amount of data, non-parametric techniques can build a very complex model that is not achievable with a parametric technique. Parametric models represent data with some predefined basis functions and parameters, whereas non-parametric models represent data with observed data as basis functions.

Non-parametric models are also less useful to be studied analytically. For example, one can reason about the parameters as linear coefficients in a simple linear regression, but a Gaussian process is much harder to reason about.

Q8 The prior $p(f|\mathbf{X},\theta)$ represents the Gaussian process with mean of $\mathbf{0}$ and the covariance function k, formulating our prior knowledge of the function space. This prior also gives our uncertainty on the function, i.e. how our belief changes as we move away from an observed output.

Q9 The prior encodes a subset of possible functions, since it assumes continuity of the modelled function.

Q10

$$\begin{aligned}
p(\mathbf{Y},\mathbf{X},f,\theta) &= p(\mathbf{Y}|\mathbf{X},f,\theta)p(f|\mathbf{X},\theta)p(\mathbf{X},\theta)\\
&= p(\mathbf{Y}|f)p(f|\mathbf{X},\theta)p(\mathbf{X})p(\theta)\\
&= \mathcal{N}(f,\epsilon^2\mathbf{I})\mathcal{N}(\mathbf{0},k(\mathbf{X},\mathbf{X}))p(\mathbf{X})p(\theta)
\end{aligned}$$

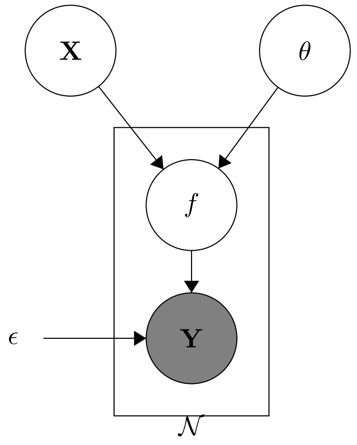

Figure 1: *Graphical representation of the model based on Gaussian Process*

Certain assumptions are made for our model:

- The function we are modelling can be represented by a Gaussian process with some certain mean function and covariance function. In our model the mean is assumed to be constant 0, and the covariance function modelling the function behaviour is a kernel function $k(\cdot, \cdot)$ controlled by a hyperparameter θ. Hence the function should be well-behaved, i.e. generally continuous and differentiable at most points.
- The observed output $\mathbf{Y}$ is corrupted by an additive Gaussian noise with zero mean and isotropic covariance matrix of $\epsilon^2\mathbf{I}$.
- All input points $\mathbf{x}_i$ are sampled independently. Also the observed output $\mathbf{Y}$ only depends on the function output $\mathbf{f}$.

Q11 Our prior, $p(f|\mathbf{X}, \theta)$ is our initial belief on the distribution of the function f, which represents the relationship between $\mathbf{Y}$ and $\mathbf{X}$. In order to integrate our prior with the likelihood $p(\mathbf{Y}|\mathbf{X}, \theta)$, which only includes θ but not f, we marginalise on f to introduce f as a new random variable, where we can write down its distribution using our prior. This way we chain our beliefs on the uncertainties of f and the observed output $\mathbf{Y}$ together.

The uncertainty in the relationship model, f, is given by the prior $p(f|\mathbf{X}, \theta) = \mathcal{N}(\mathbf{0}, k(\mathbf{X}, \mathbf{X}))$. The uncertainty in the observed output, is given by the term $p(\mathbf{Y}|f) = \mathcal{N}(f, \epsilon^2\mathbf{I})$.

The fact that θ remains un-marginalised tells us that the space of function f is controlled by the hyperparameter θ, and marginalising out f does not integrate over all possible fs under all possible θs.

1.2 Practical

Q12 (a) The following graph shows the prior $p(\mathbf{W})$.

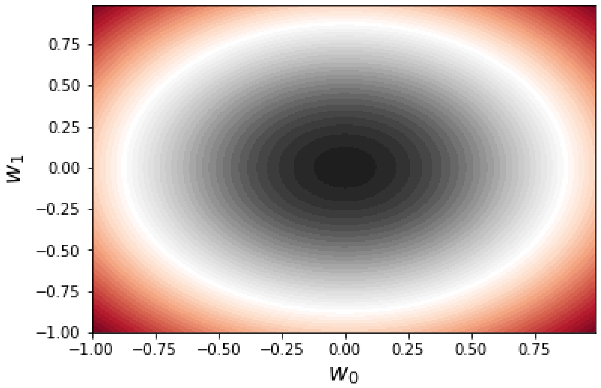

Figure 2: *Prior distribution over* ***W***

(b)

(c)

(d)

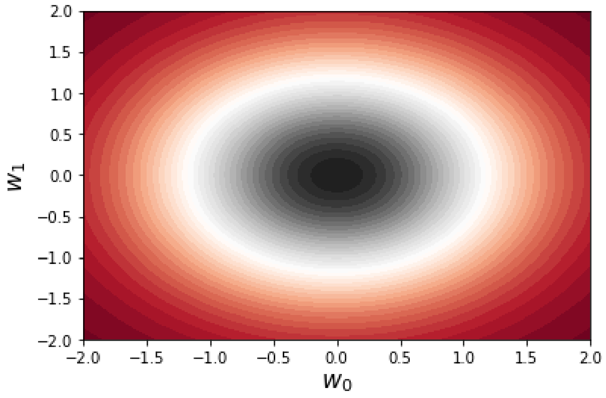

Figure 3: *Posterior distribution over* ***W*** *after seeing one data point*

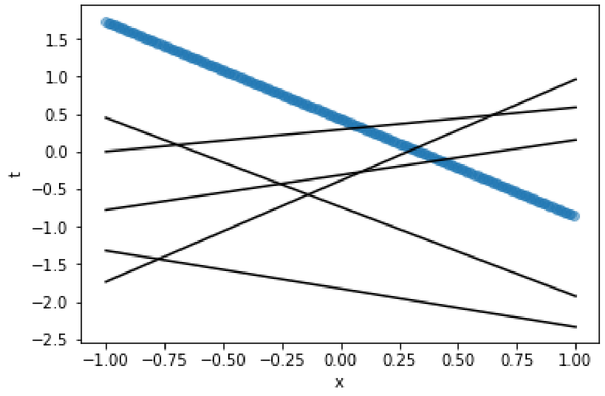

Figure 4: *Some sample functions from the posterior after seeing one data point. The blue scatter plot is sampled from the original data as reference*

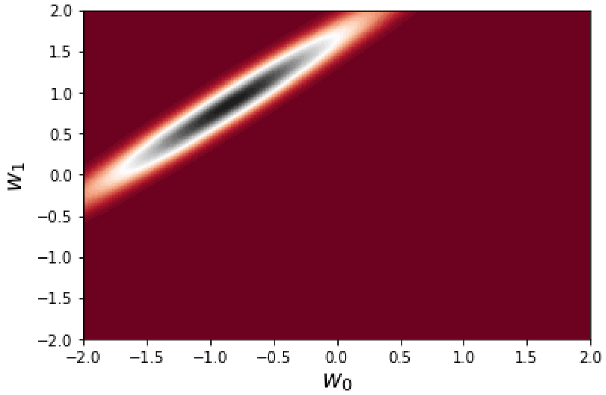

Figure 5: *Posterior distribution over* ***W*** *after seeing 5 data points*

(e) For the prior distribution, as we have seen no data the distribution is just a zero-mean Gaussian.

As we see more and more data points, the peak of the posterior Gaussian distribution becomes narrower and narrower, since we gain more certainty in distribution of $\mathbf{W}$.

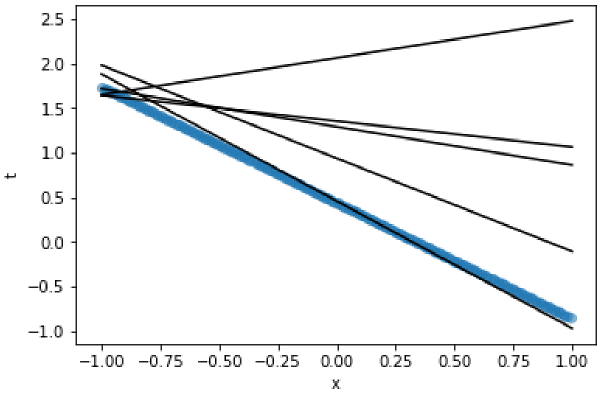

Figure 6: *Some sample functions from the posterior after seeing 5 data points*

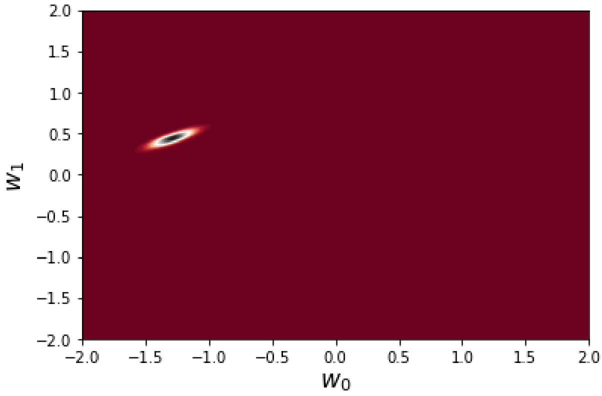

Figure 7: *Posterior distribution over* ***W*** *after seeing 100 data points*

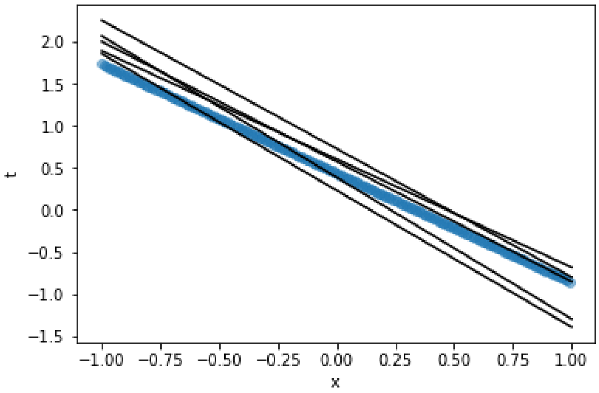

Figure 8: *Some sample functions from the posterior after seeing 100 data points*

Sampled functions taken from the posterior also suggests that, the ones taken from posterior distributions after seeing more data, are closer to the true model where the original data is sampled from. This is certainly desirable since we can gain belief about the model behind the data by seeing more data.

(f) In the posterior,

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

The Covariance matrix is $(\frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{x}+\mathbf{\Sigma^{-1}})^{-1}$ As we add more data, the covariance decreases, i.e. we are more certain with our belief of the model.

Q13 (a)

(b)

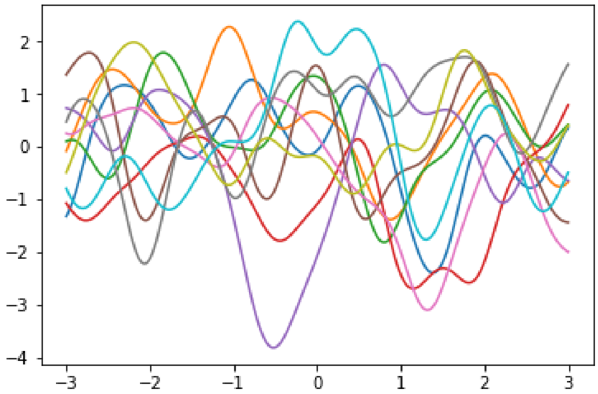

Figure 9: *Prior distribution with* $l = 0.5$

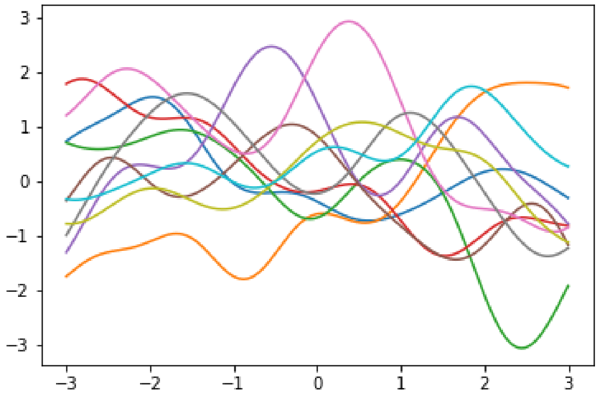

Figure 10: *Prior distribution with* $l = 1$

(c)

(d) As the length-scale of the covariance function increases, the gradient of the sampled prior distributions changes less aggressively, i.e. the second derivative is smaller in absolute value.

(e) The length-scale of the covariance function encodes our assumption of how much effect each data point has on its neighbours. The larger the length-scale, the more likely its neighbouring points take a value close to that data point.

Q14

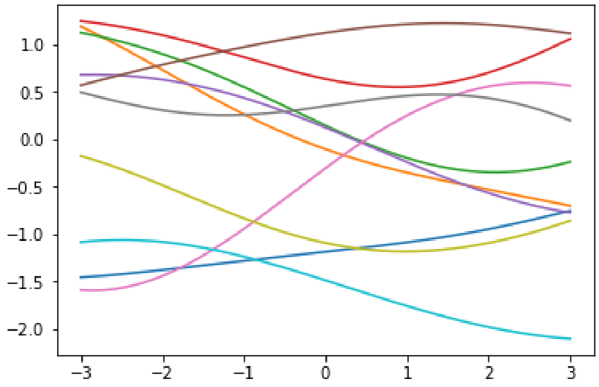

Figure 11: *Prior distribution with $l = 5$*

2 The Posterior

2.1 Theory

Q15 Assumptions are things we believe to be true with absolute certainty in our model.
Beliefs are also things we believe to be true, but with some uncertainty modelled by a probability value or probability distribution.
Preference is about our belief of, without other knowledge, what things are more probable than some other things, possibly associated with a probability distribution.

Q16 The assumption here is that the probability distribution of the latent variable $\mathbf{x}$ is modelled by a spherical Gaussian centred at $\mathbf{0}$, with the standard identity matrix as its covariance/precision matrix. We can also interpret it as a preference to probability distributions of $\mathbf{x}$ that are closer to the standard multivariate Gaussian distribution with mean $\mathbf{0}$ and isotropic covariance matrix.

Q17 To simplify the equations, without loss of generality, we assume the mean of $\mathbf{Y}$ is $\mathbf{0}$.
Assume that there is a linear relationship between $\mathbf{Y}$ and $\mathbf{X}$, corrupted by an additive Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, that is,

$$\mathbf{y}_i = \mathbf{W}^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\epsilon}$$

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

$$\begin{aligned}
p(\mathbf{Y}|\mathbf{W}) &= \int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{X})d\mathbf{X} \\
&= \int \mathcal{N}(\mathbf{Y}|\mathbf{W}^{\mathrm{T}}\mathbf{X}, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{X}|\mathbf{0}, \mathbf{I})d\mathbf{X} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\sigma^2}\mathbf{I}|^{\frac{1}{2}}}\frac{1}{(2\pi)^{\frac{D}{2}}} \\
&\quad \int e^{-\frac{1}{2}(\mathbf{Y}-\mathbf{W}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{Y}-\mathbf{W}^{\mathrm{T}}\mathbf{X})} \\
&\quad e^{-\frac{1}{2}\mathbf{X}^{\mathrm{T}}\mathbf{I}^{-1}\mathbf{X}}d\mathbf{X} \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\sigma^2}\mathbf{I}|^{\frac{1}{2}}}\frac{1}{(2\pi)^{\frac{D}{2}}} \\
&\quad \int \exp(-\frac{1}{2}(\frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} \\
&\quad - \frac{2}{\sigma^2}\mathbf{Y}\mathbf{W}^{\mathrm{T}}\mathbf{X} \\
&\quad + \mathbf{X}^{\mathrm{T}}(\frac{1}{\sigma^2}\mathbf{W}\mathbf{W}^{\mathrm{T}} + \mathbf{I})\mathbf{X}))d\mathbf{X}
\end{aligned}$$

Notice that the product within the integral is also a Gaussian, thus the integral itself is a Gaussian. Hence we can derive the covariance directly,

$$p(\mathbf{Y}|\mathbf{W}) = \mathcal{N}(\mathbf{0}, \mathbf{C})$$

where

$$\begin{aligned}
\mathbf{C} &= \mathbb{E}[(\mathbf{W}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\epsilon})(\mathbf{W}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\epsilon})^{\mathrm{T}}] \\
&= \mathbb{E}[\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}]\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}] \\
&= \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}
\end{aligned}$$

2.1.1 Learning

Q18 (a) The Maximising Likelihood(ML) approach optimises by maximising the likelihood. The Maximum-a-Posteriori(MAP) approach maximises the posterior. The Type-II Maximum-Likelihood maximises the marginal likelihood, by marginalising out the input data $\mathbf{X}$. These optimisation techniques all seek to fit a model to the data. ML only considers the data and not the prior at all. MAP considers both the data and the prior to restrict the model complexity. The Type-II ML tries to evaluate the model over all possible $\mathbf{X}$ as it marginalises it out.

(b) As we observe more data, both MAP and ML generate a model fitting better to the data, and they become closer as the implication of prior is reduced. ML tends to over-fit the model to data as there is no prior to restrict the model complexity, whereas MAP learns a model that keeps the balance between the data and the prior.

(c) The denominator, $\int p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})d\mathbf{W}$ can be regarded as a constant as we optimise $\mathbf{W}$ to maximise the whole expression, which does not affect the optimisation result.

2.1.2 Practical Optimisation

Q19 (a) Assuming that all data points $\mathbf{y}_i$ are independent,

$$\begin{aligned} & -\log(p(\mathbf{Y}|\mathbf{W})) \\ &= -\log(\prod_i p(\mathbf{y_i}|\mathbf{W})) \\ &= \sum_i (-\log(p(\mathbf{y_i}|\mathbf{W}))) \\ &= \sum_i (-\log(\mathcal{N}(\mathbf{0}, \mathbf{C}))) \\ &= \sum_i -\log\left(\frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{C}|^{\frac{1}{2}}} \exp(-\frac{1}{2}\mathbf{y}_i^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}_i)\right) \\ &= \sum_i \left[\frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{C}| + \frac{1}{2}\mathbf{y}_i^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}_i\right] \end{aligned}$$

(b) Note: since we have almost forgot all the linear algebra we did in our undergraduate studies, all what we did was basically reading through the appendix and trying to follow each step. We implemented the differentiation function by following the appendix exactly as shown in the practical question in this section.
I figure that doing matrix differentiation can be useful as gradient-based optimisation techniques, e.g. gradient descent, are powerful in the context of machine learning as many questions are transformed into optimisation problems.

$$\frac{\partial\mathcal{L}(\mathbf{W})}{\partial\mathbf{W}}$$

Q20 In the context of learning, we want to know the relationship between the data $\mathbf{Y}$ and some latent random variable $\mathbf{X}$. We model the relationship with some function value f (with an additive Gaussian noise or what), which forms the likelihood, $p(\mathbf{Y}|f)$, and prior, $p(f|\mathbf{X})$. We might include a hyperparameter θ which f depends on, as part of the prior. Additionally we have prior knowledge of what we think the latent variable is, $p(\mathbf{X})$.
We see that we have some knowledge between $\mathbf{Y}$ and f, and between f and $\mathbf{X}$, as well as some prior knowledge about $\mathbf{X}$, $\boldsymbol{\theta}$, etc. Marginalising f out naturally connects out belief about relationship between $\mathbf{Y}$ and $\mathbf{f}$, and that between f and $\mathbf{X}$ together. As featured in figure 12, f sits between $\mathbf{X}$ and $\mathbf{Y}$. Marginalising f effectively combines the two arrows, $\mathbf{X}$ to f, and f to $\mathbf{Y}$, which represent our beliefs about their relationships.
Another way to think about it is that marginalising a random variable is taking the expected value by integrating over all possible values it can take. As the domain of the latent representation $\mathbf{X}$ is usually unrestricted compared to f, which is limited by our model directly, we could argue that integrating over f is easier as the problem space is smaller and the its probability distribution is more predictable.

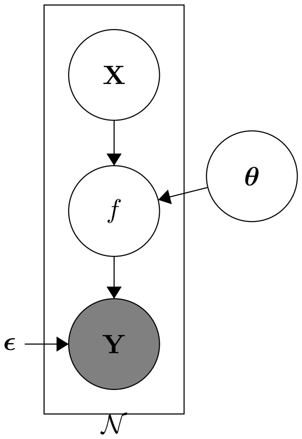

Figure 12: *Graphical representation of a model based on Gaussian Process with latent random variables* $\mathbf{X}$ *and* f

On the other hand, attempting to marginalising out $\mathbf{X}$ is effectively combining the information between $\mathbf{X}$ and f, as well as $\mathbf{X}$ and $\mathbf{Y}$, to get some belief of relationship between $\mathbf{Y}$ and f. In the general case, it is intractable to do so as the belief of relationship between $\mathbf{Y}$ and $\mathbf{X}$ is not enough to connect the other beliefs. This only becomes possible in specific simple cases, e.g. assuming a simple linear relationship, as introduced previously in the report.

2.2 Practical

Q21 The 2D representation that we learned shows a spiral pattern, which likely comes from the non-linear trigonometric functions we applied. Since we are recovering a 2D latent representation, which happen to match the dimension of x^*, i.e. $f_{\text{non-lin}}(x)$, we do expect the result to capture some features of the non-linear functions, which it does exactly.

Q22 As we see in figure 14, with some random basis functions, the projected $\mathbf{Y}$ does not retain much information about the non-linear transformation applied. Furthermore, many data points coincide on this subspace, which implies a lot of information loss.

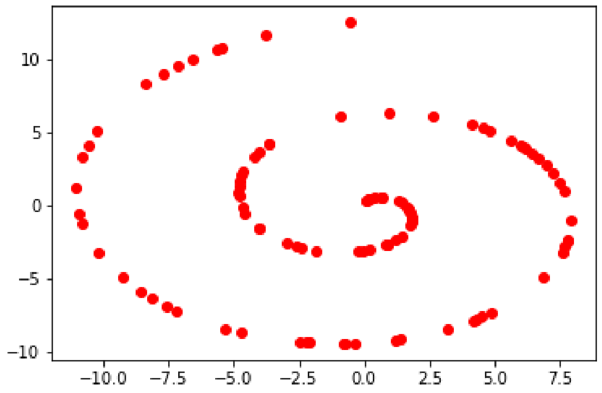

Figure 13: *Learned latent variable* $\mathbf{X}$ *with Type-II Maximising Likelihood*

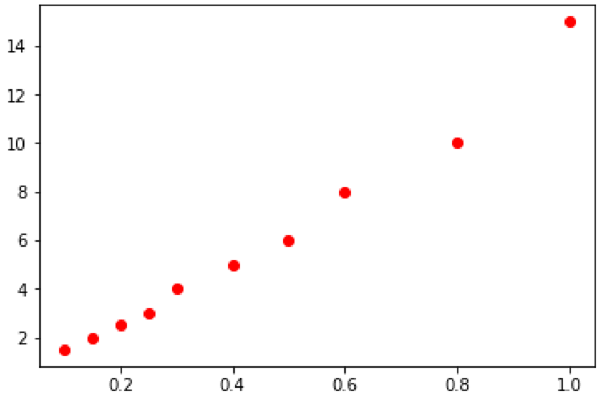

Figure 14: $\mathbf{Y}$ *projected in 2D with a randomly generated basis*

3 The Evidence

3.1 Theory

3.1.1 Data

3.1.2 Models

Q23 (a) According to Laplace's principle of insufficient reason, the equiprobable model should be assumed in the absence of any other information. This model should be considered simplest as it does not require any information, and certainly has no flexibility whatsoever.

(b) The equiprobable model should be regarded as the most complex model since this model is applicable for any case. The fact that it does not require any information also suggests that using the model as a prior does not provide you any information at all. In the context of information theory, the equiprobable model has the highest possible information entropy, thus might be thought of as the most complex model.

Q24 The assumptions made by the models above impose some constraints on the probability mass distribution. All of the models assume that each y^i are independent. Model $p(\mathcal{D}|M_1, \boldsymbol{\theta}_1)$ asserts that given that $x_1^n = 0$, conditional probabilities $p(y = 1) = p(y = -1)$. The constraints on the probability mass distribution implies that assuming such a model gives us more information (than the equiprobable model). These models contain a hyperparameter, which allows the model to more or less adjust to different cases/distributions, i.e. more flexible. In this sense the ones with more parameters, $p(\mathcal{D}|M_3, \boldsymbol{\theta}_3)$, is potentially the most flexible.

On the other hand, the model also restricts the probability mass distribution such that it deviates from the uniform distribution, which means it less average information entropy.

3.1.3 Evidence

Q25 The choice of prior implies that we know nothing but that $\boldsymbol{\theta}$ is corrupted by an additive Gaussian noise. A large covariance function $\boldsymbol{\Sigma}$ implies high uncertainty on our belief whereas the prior $\boldsymbol{\mu}$ can usually be moved out without loss of generality.

3.2 Practical

Q26

Q27

Q28

Q29

4 Final thoughts

Q30 The coursework has enhanced our understanding of basic Bayesian concepts of prior, likelihood, posterior, etc. Performing several derivations by hand greatly helps with memorisation and understanding of some important concepts and techniques, e.g. conjugate prior, marginalisation, etc. Several linear algebra techniques seem challenging but proven useful. We think the main purpose of this task is to consolidate the mathematical background theory as well as to see and practise how the theory integrates into modern machine learning applications.