MIL 780 - Assignment Two

Ryan Balshaw

05 April 2022

Foreword

All code, documents and work relevant to this assignment can be found in my Github repository.

1 Question 1

Consider the following model

$$x_n = \theta + \epsilon_n, \tag{1}$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$. Given some observed data $\mathbf{x} = [x_1, \cdots, x_N]^T$, the objective is to perform Bayesian inference through a grid-based approach and a conjugate prior approach. The observed data is shown in Figure 1.

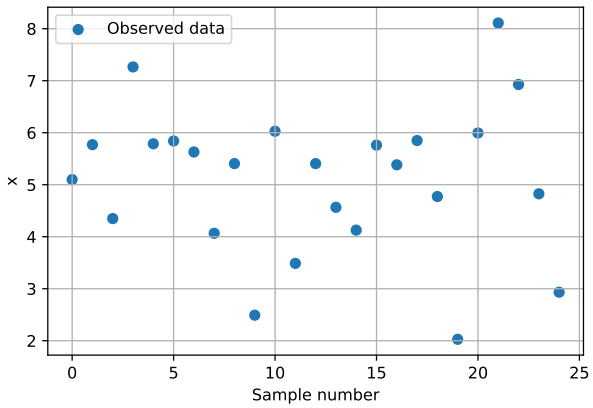


Figure 1: The observed x data for Question 1.

1.1 Part A: Grid-based Bayesian inference

Assume that σ is known and given by $\sigma = 1.5$. Assume a Laplacian prior for θ with the following form:

$$p(\theta) = L(\theta|0, 2), \tag{2}$$

where 0 is the location parameter and 2 is the scaling parameter. A grid-based approach where $\theta \in [-10, 20]$ is used in the results to follow. There are a number of aspects that are of great important to Bayesian inference:

- The generative model $p(x|\theta)$.
- The prior $p(\theta)$ over the generative model parameters.
- The likelihood function $\mathcal{L}(\mathbf{x}, \theta)$.
- The unnormalised posterior $p(\mathbf{x}|\theta)p(\theta)$.
- The model evidence $p(\mathbf{x})$.

- The posterior distribution $p(\theta|\mathbf{x})$.
- The posterior predictive distribution $p(x|\mathbf{x})$.

In this part of the assignment, the objective is to gain exposure to each of these aspects for a simple one-dimensional problem.

1.1.1 The prior distribution

The prior over θ is given analytically as

$$p(\theta) = L(\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}, \tag{3}$$

and the logarithm of this distribution is equal to

$$\log p(\theta) = -\log(2 \cdot b) - \frac{|x-\mu|}{b}. \tag{4}$$

In Figure 2 the prior and the log-prior is shown for $\mu = 0$ and $b = 2$.

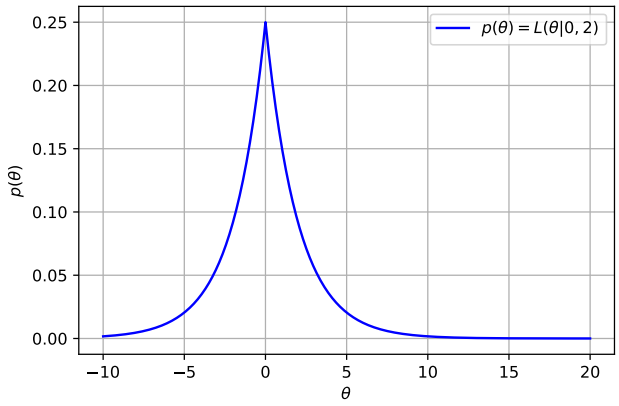


Figure 2: The prior distribution function $p(\theta)$ visualised over the domain $\theta \in [-10, 20]$.

1.1.2 The generative model and the likelihood function

Given the model in Equation (1), the model distribution over x is given as

$$\begin{aligned} p(x|\theta) &= \mathcal{N}(x|\theta, \sigma^2) \\ &= \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} e^{-\frac{1}{2\sigma^2} \cdot (x-\theta)^2}. \end{aligned} \tag{5}$$

Given some observed data $\mathbf{x}$, where $\mathbf{x}$ is a column vector $\mathbf{x} \in \mathbb{R}^{N \times 1}$ and N is the number of observed samples, the likelihood function, given θ, is

$$\mathcal{L}(\mathbf{x}, \theta) = \prod_{n=1}^{N} p(x_n|\theta). \tag{6}$$

The log-likelihood function may be given as

$$\begin{aligned} \mathcal{LL}(\mathbf{x}, \theta) &= \sum_{n=1}^{N} \log p(x_n|\theta) \\ &= \sum_{n=1}^{N} \left[-\frac{1}{2} \log(2 \cdot \pi) - \frac{1}{2} \log(\sigma^2) - \frac{(x-\theta)^2}{2\sigma^2} \right] \\ &= -\frac{N}{2} \log(2 \cdot \pi) - \frac{N}{2} \log(\sigma^2) - \sum_{n=1}^{N} \left[\frac{(x-\theta)^2}{2\sigma^2} \right]. \end{aligned} \tag{7}$$

It is important here to remind ourselves (read: myself) of the functions of the $\mathcal{LL}$ function. The observed data is typically fixed, and thus the only variable that we can alter is θ. Thus, we can plot how the likelihood/log-likelihood function varies with θ. In Figure 3 the likelihood function and log-likelihood function over θ is shown. Furthermore, the likelihood function is the joint probability of the observed data given the model parameters. In Bayesian inference, the likelihood function is key to finding the posterior distribution $p(\theta|\mathbf{x})$.

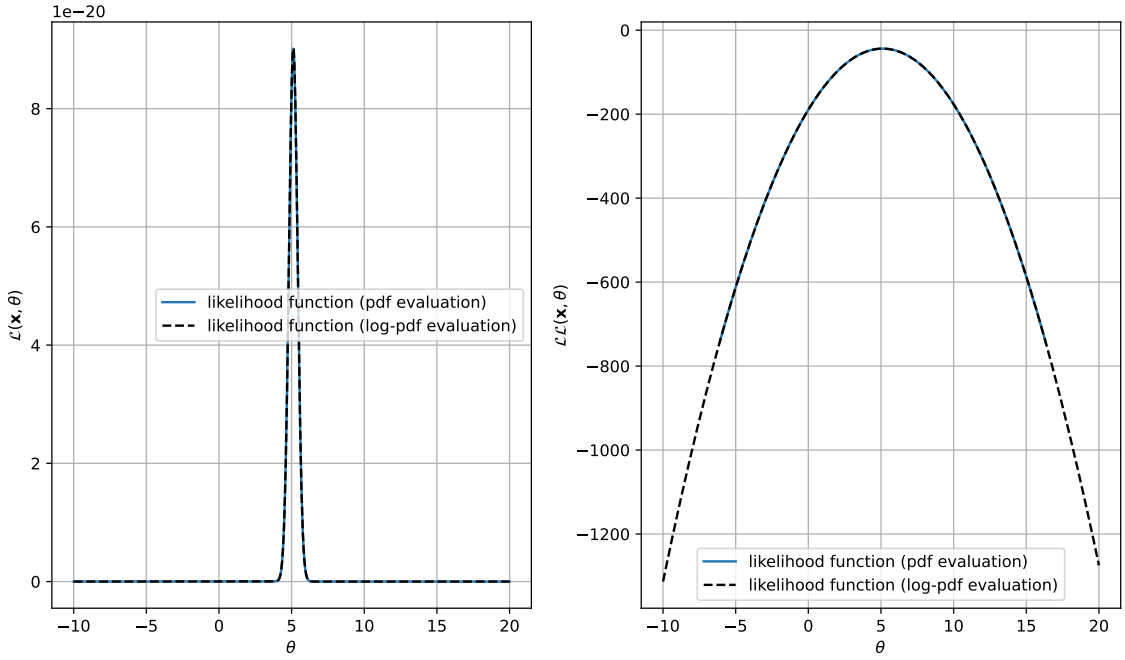


Figure 3: The likelihood function $\mathcal{L}(\mathbf{x}, \theta)$ of the generative model $p(x|\theta)$ over the domain $\theta \in [-10, 20]$. In a) the likelihood function is shown, and in b) the log-likelihood function is shown.

1.1.3 The unnormalised posterior

The unnormalised posterior is the numerator of posterior distribution over θ given $\mathbf{x}$. The posterior is given through Bayes' rule

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}. \tag{8}$$

It is now clear why the prior was the first step of the Bayesian process. We assume a prior over θ, observe some data $\mathbf{x}$ and then we can inspect the posterior $p(\theta|\mathbf{x})$. To do this, we need the unnormalised posterior $p(\mathbf{x}|\theta)p(\theta)$ and the model evidence $p(\mathbf{x})$. To determine the unnormalised posterior we can use the analytical forms of the likelihood function and the prior distribution

$$\begin{aligned} p(\mathbf{x}, \theta) &= p(\mathbf{x}|\theta)p(\theta) \\ \mathcal{F}(\theta, \mathbf{x}, , \sigma, \mu_\theta, b) &= \frac{1}{2 \cdot b} \exp\left(-\frac{|\theta - \mu_\theta|}{b}\right) \prod_{n=1}^{N} \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \cdot (x_n - \theta)^2\right), \end{aligned} \tag{9}$$

where $\mathcal{F}(\theta, \mathbf{x}, \sigma, \mu_\theta, b)$ is the notation that I have adopted to show the parameters of the unnormalised posterior. In this case, σ, μ_θ an b are fixed to $\sigma = 1.5$, $\mu_\theta = 0$ and $b = 2$. It is also possible to inspect the log-unnormalised posterior, and this is given analytically as

$$\begin{aligned} \log p(\mathbf{x}, \theta) &= \log p(\mathbf{x}|\theta) + \log p(\theta) \\ \mathcal{LF}(\theta, \mathbf{x}, , \sigma, \mu_\theta, b) &= \sum_{n=1}^{N} \left[-\frac{1}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \log(\sigma^2) - \frac{(x_n - \theta)^2}{2\sigma^2}\right] - \log(2 \cdot b) - \frac{|\theta - \mu_\theta|}{b} \\ &= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{N}{2} \cdot \log(\sigma^2) - \log(2 \cdot b) - \frac{|\theta - \mu_\theta|}{b} - \sum_{n=1}^{N} \frac{(x_n - \theta)^2}{2\sigma^2}. \end{aligned} \tag{10}$$

In Figure 4 the unnormalised posterior is shown. Notice how the prior and likelihood function influences the unnormalised posterior. There appears to be a slight shift towards the origin, and the influence of the prior is marginal in comparison to the likelihood function. The marginal influence of the prior is attributed to the prior parameters, as the prior is far from the likelihood function and the variance ensures that the prior is numerically larger over the θ domain in comparison to the likelihood function. Decreasing the variance or shifting the prior mean should increase its influence on the unnormalised posterior.

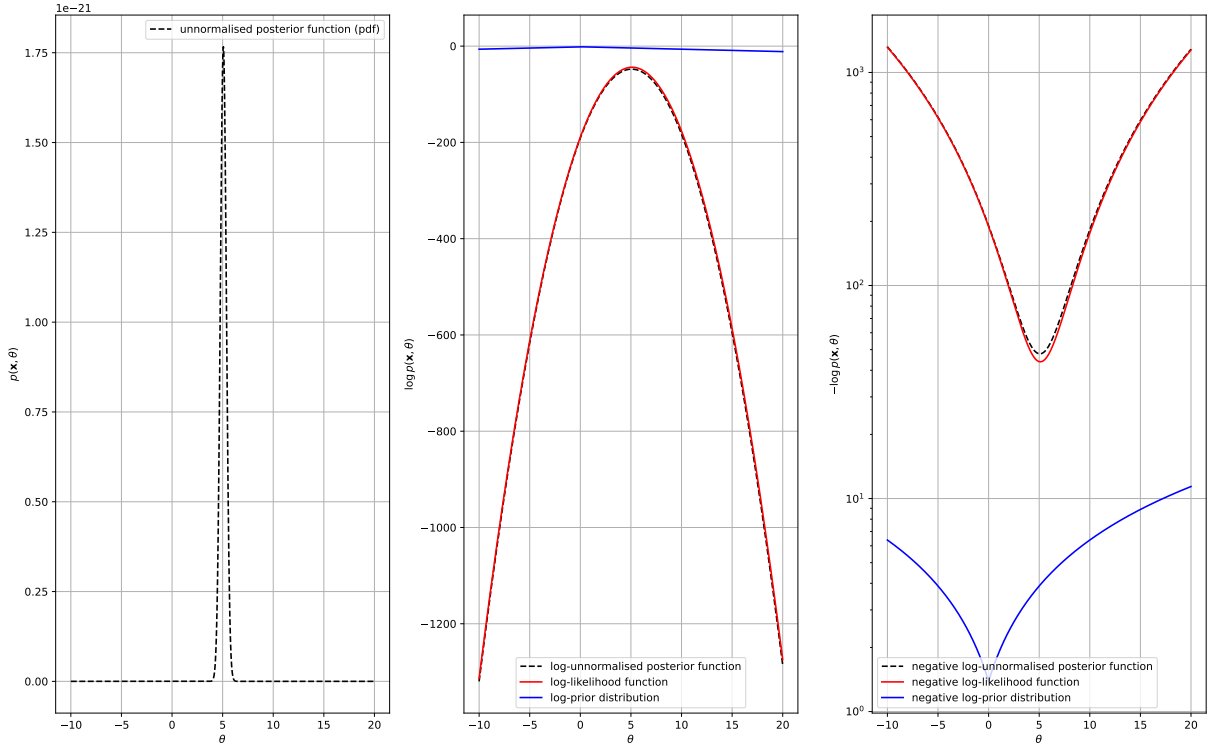


Figure 4: The unnormalised posterior distribution $\mathcal{F}(\theta, \mathbf{x}, \sigma, \mu_\theta, b)$ for $\sigma = 1.5, \mu_\theta = 0$ and $b = 2$ over the domain $\theta \in [-10, 20]$. Three variations of the unnormalised posterior is visualised here, the standard unnormalised posterior, the logarithm of the unnormalised posterior and the terms therein, and the negative of the unnormalised log-posterior.

1.1.4 The model evidence

In order to determine the posterior distribution over the parameters $p(\theta|\mathbf{x})$, we need to determine the denominator of Equation (8). The denominator is commonly referred to as the model evidence, and it is sometimes known as the marginal likelihood $p(\mathbf{x})$. Note that this marginalisation is with respect to the generative model parameters and represents the probability of the observed data marginalized over the parameters θ. This is given formally as

$$\begin{aligned} p(\mathbf{x}) &= \int_\theta p(\mathbf{x}, \theta) d\theta \\ &= \int_\theta p(\mathbf{x}|\theta) p(\theta) d\theta \\ &= 1.32856e^{-21}. \end{aligned} \tag{11}$$

1.1.5 The posterior distribution

Now that we know the model evidence, we can use the unnormalised posterior and normalise it by the model evidence to obain the posterior distribution $p(\theta|\mathbf{x})$. In Figure 5 the posterior distribution for the model parameter θ is shown. Notice how the model evidence, which is a constant, effectively rescales the unnormalised posterior. This is to ensure that the posterior distribution is a true distribution (i.e., it satisfies $\int_\theta p(\theta|\mathbf{x}) d\theta = 1$ and $p(\theta|\mathbf{x}) > 0$ $\forall\ \theta$).

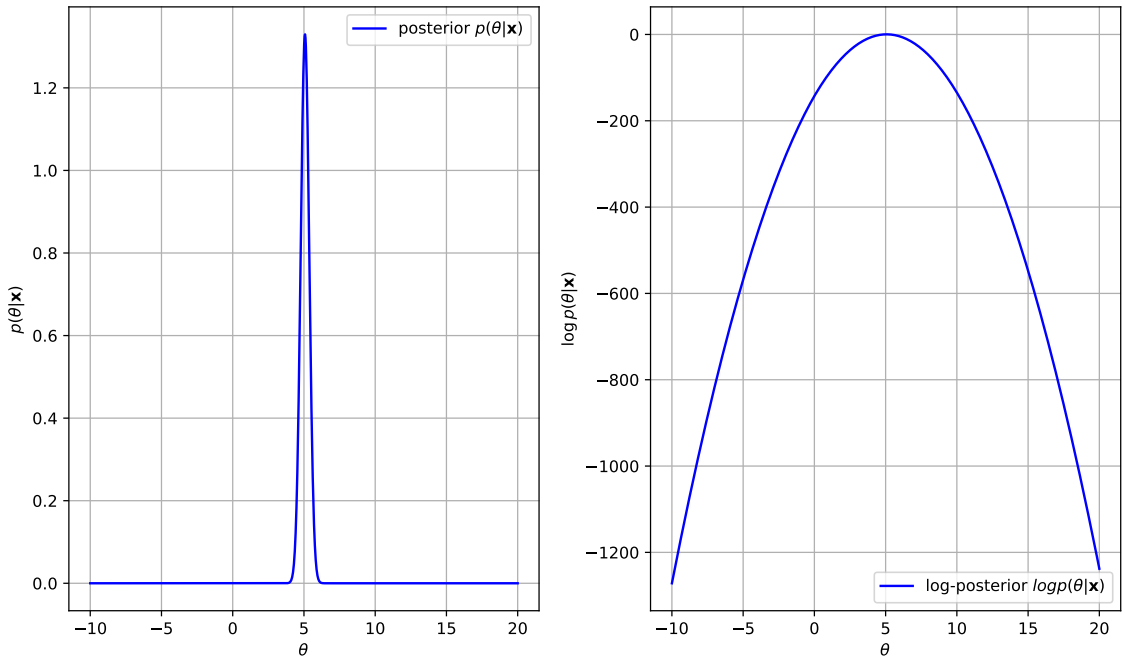


Figure 5: The posterior distribution $p(\theta|\mathbf{x})$ and the logarithm thereof over the domain $\theta \in [-10, 20]$. The MAP solution is given in Table 1.

1.1.6 The posterior predictive distribution

However, finding the posterior distribution is only only part of the Bayesian inference puzzle. We would still like to use the generative model given in Equation (1) for any newly observed data and to do so, we need to incorporate the posterior distribution in some way. A naive, but viable approach is to use the maximum-a-posteriori (MAP) estimate from the unnormalised posterior as a basis for the model parameter θ. However, this is simply a parameter re-substitution of the estimator for θ and is equivalent to maximum likelihood estimation. The power of Bayesian inference is that we can incorporate all uncertainty in the unknown parameter when observing new data. This allows us to quantify new data under all possible variations of the unknown model parameter θ rather than on a maximum estimate[1]. To use the uncertainty of the unknown parameter for newly observed data, we can use the posterior predictive distribution. This is given as

$$p(x|\mathbf{x}) = \int_{\theta} p(x|\theta)p(\theta|\mathbf{x})d\theta. \tag{12}$$

In Figure 6 the posterior predictive distribution is shown. In my implementation I used two methods to numerically estimate the integral in Equation (12). The expensive version, where I used `scipy.integrate.quad`, took 202.154 seconds to perform 1000 iterations over x. The cheap version, where I used `scipy.integrate.trapz`, took 0.039 seconds to perform 1000 iterations over x.

1.1.7 Estimating estimators

We can now compare different estimators for the unknown parameter θ using the prior distribution, the likelihood function and the unnormalised posterior function. This estimator estimation process can be performed using a grid-based approach, analytical solutions and through numerical optimisation techniques. In Table 1 the estimates for these three approaches is given. Notice how the MAP estimate is lower than the maximum likelihood (ML) estimate, which is due to the influence of the prior $p(\theta)$.

[1] I found the following StackExchange discussion very insightful.

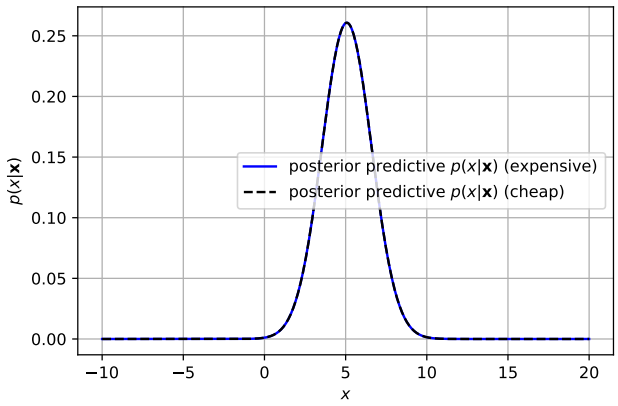


Figure 6: The posterior predictive distribution $p(x|\mathbf{x})$ determined using a cheap and expensive computational methodology over the domain $\theta \in [-10, 20]$.

Table 1: The estimates for the unknown parameter θ using the prior, maximum likelihood and maximum a posteriori. These estimates are found using a grid-based approach, analytically and through numerical optimisation.

Approach	Prior	ML	MAP
Grid-based	0.0	5.1051	5.075
Analytical	-0.0002	5.1161	N/A
Numerical	0.002737	5.1161	5.071

1.1.8 Using Bayesian inference

We can use the posterior distribution $p(\theta|\mathbf{x})$ and the posterior predictive distribution $p(x|\mathbf{x})$ to determine probabilities $p(a \leq X \leq b)$. The first probability of interest is $p(\theta \leq 4|\mathbf{x})$ and the second is $p(x \leq 4|\mathbf{x})$. As a reminder, the probability $p(a \leq x \leq b)$ may be given as

$$p(a \leq x \leq b) = F_x(b) - F_x(a), \tag{13}$$

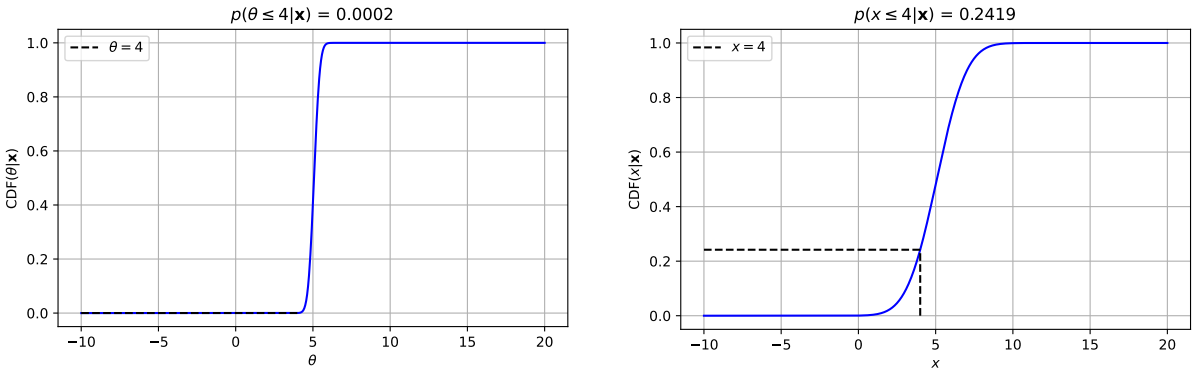
where $F_x(X) = \int_{-\infty}^{X} p(x)dx$ is the cumulative distribution function (CDF) of $p(x)$. As we do not know the analytical forms of the posterior and posterior predictive distributions, we cannot use an analytical expression for the CDFs. However, we have a discretised set of likelihoods along the θ and x domain and thus we can estimate the CDF by cumulatively adding the likelihoods. It is assumed here that the CDF varies linearly between points along each respective domains. A linear interpolation approach is then used to estimate $F_x(X)$. This process ensures that we can estimate the probabilities of interest.

Using this approach, it was found that $p(\theta \leq 4|\mathbf{x}) = 0.0002$ and $p(x \leq 4|\mathbf{x}) = 0.2419$. In Figure 7 the CDFs of interest and the resulting probabilities are shown. The posterior probability that $\theta \leq 4$ informs us that it is very unlikely that a model parameter θ in this range is responsible for generating the observed data. The posterior predictive probability that $x \leq 4$ is noticeably larger, and this indicates that it is far more probable that newly observed x samples may be drawn in the x domain $x \in [-\infty, 4]$.

1.1.9 Prior hyper-parameters

At this point in the assignment we have assumed that the prior parameters are fixed in nature. However, there are no set values that are always given to these parameters. As such, all of the distributions of interest in Bayesian inference are conditioned on these parameters. Changing the mean of the prior distribution may be a biased procedure as we are effectively biasing the model towards a certain direction. While this is not always a bad decision, we can also control the impact of the prior in the Bayesian inference process using the prior variance. The prior is parametrised as

$$p(\theta; \alpha) = L(\theta|0, \alpha) \tag{14}$$



(a) $F_{\theta|\mathbf{x}}(\theta)$.

(b) $F_{x|\mathbf{x}}(x)$.

Figure 7: The CDFs used to determine the probabilities $p(\theta \leq 4|\mathbf{x})$ and $p(x \leq 4|\mathbf{x})$ over the domain $\theta \in [-10, 20]$. In a) the CDF of the posterior distribution $p(\theta|\mathbf{x})$ is shown, while in b) the CDF of the posterior predictive distribution $p(x|\mathbf{x})$ is shown.

where α is a hyper-parameter. The prior hyper-parameter is noticeably influential in posterior distribution

$$p(\theta|\mathbf{x}, \alpha) = \frac{p(\mathbf{x}|\theta)p(\theta|\alpha)}{p(\mathbf{x}|\alpha)}, \tag{15}$$

and the model evidence

$$\begin{aligned} p(\mathbf{x}) &= \int_{\theta} p(\mathbf{x}, \theta) d\theta \\ &= \int_{\theta} p(\mathbf{x}|\theta)p(\theta|\alpha) d\theta. \end{aligned} \tag{16}$$

As such, a range of α values were considered and the effects on the posterior distribution, the MAP estimate and the model evidence were inspected. In Figure 8 the results of varying α is shown. It is clear from Figure 8b) that by increasing the variance we approach a region where the MAP estimate steadies out. Furthermore, as seen in Figure 8d) a larger variance produces a posterior distribution with smaller variation. It is clear that small model variances strongly affect the Bayesian inference process. This implies that having a 'tight' prior increases its influence in the results, whereas a prior with a larger variance decreases its influence.

1.2 Part B: Conjugate Bayesian inference

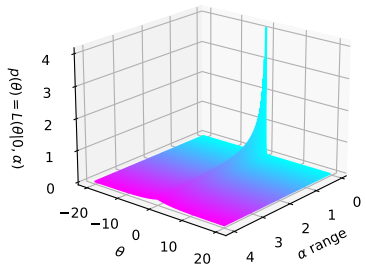
Assume that both the parameter θ and the noise variance σ^2 are unknowns in the following model

$$x_n = \theta + \epsilon_n, \tag{17}$$

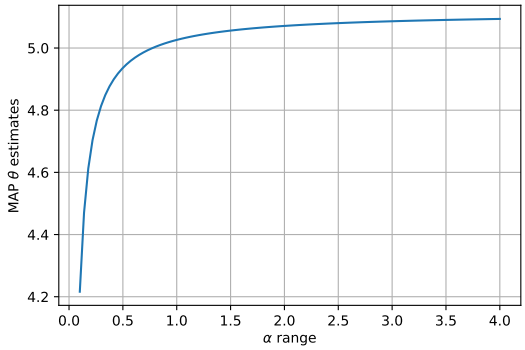
where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$. The objective here is to perform Bayesian inference on the unknown model parameters. To assist this process, conjugate priors will be used to improve the posterior condition inference process. We could use specific priors (informative or uninformative priors), or we could use conjugate priors that are specifically designed for the assumed generative model. For this assignment, the form of the generative model is assumed, and a conjugate prior for the unknown model parameters will be used. The generative model is a Gaussian model with unknown mean θ and precision $\lambda = \sigma^{-2}$. This is given as

$$p(x|\theta, \lambda) = \mathcal{N}(x|\theta, \lambda^{-1}) = \left(\frac{\lambda}{2 \cdot \pi}\right)^{\frac{1}{2}} \cdot \exp\left(-\frac{\lambda}{2} \cdot (x - \theta)^2\right). \tag{18}$$

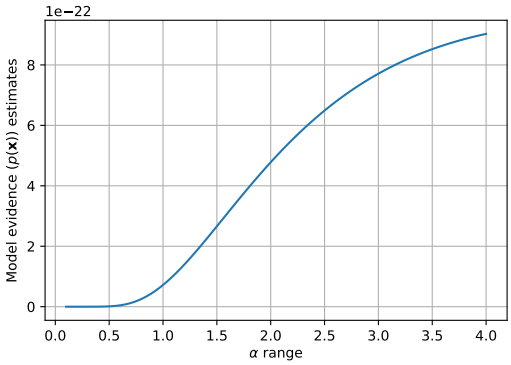
The assumed generative model, in the presence of some observed data $\mathbf{x} = [x_1, \cdots, x_N]^T$ comprising of N



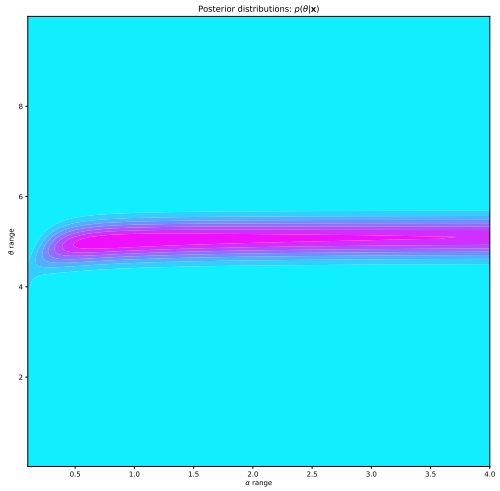
(a) $p(\theta|\alpha)$.



(b) MAP θ variation.



(c) Evidence $p(\mathbf{x})$ variation.



(d) Posterior $p(\theta|\mathbf{x})$ variation.

Figure 8: The influence of the prior variance $p(\theta|\alpha) = L(\theta|0, \alpha)$ on the Bayesian inference process. In a) the prior is visualised over α, in b) the variation of the MAP estimate for θ is shown, in c) the model evidence variation in shown, and in d) the variation of the posterior $p(\theta|\mathbf{x})$ is shown.

samples, which gives rise the the likelihood function

$$\begin{aligned} p(\mathbf{x}|\theta, \lambda) &= \prod_{n=1}^{N} p(x_n|\theta, \lambda^{-1}) \\ &= \left(\frac{\lambda}{2 \cdot \pi}\right)^{\frac{N}{2}} \cdot \exp\left(-\frac{\lambda}{2} \sum_{n=1}^{N} \cdot (x_n - \theta)^2\right). \end{aligned} \tag{19}$$

The likelihood function is proportional to

$$p(\mathbf{x}|\theta, \lambda) \propto \left[\lambda^{\frac{1}{2}} \cdot \exp\left(-\frac{\lambda \cdot \theta^2}{2}\right)\right]^N \cdot \left[\exp\left(-\frac{\lambda}{2} \cdot \sum_{n=1}^{N} x_n^2 + \lambda \cdot \theta \cdot \sum_{n=1}^{N} x_n\right)\right], \tag{20}$$

where the normalisation constant $\left(\frac{1}{2 \cdot \pi}\right)^{\frac{N}{2}}$ has been removed. To design a conjugate prior to the assumed likelihood function, we need to construct a prior $p(\theta, \lambda)$ that creates a functionally similar posterior distribution. Before the proceed, we need to think about the domains of θ and σ^2. θ can be any value on the real domain $\mathbb{R}$, while the domain of σ (and equivalently, λ) is bounded to $[0, \infty]$. Thus, the prior distribution should reflect this domain variation.

Furthermore, based on the inspection of Equation (20), the prior distribution can be expressed as $p(\theta|\lambda)p(\lambda)$. In Bishop [1], the prior of interest given as is the normal-Gamma distribution

$$p(\theta, \lambda) = \mathcal{N}\left(\theta|\theta_0, (\beta_0 \cdot \lambda)^{-1}\right) \mathrm{Gam}\left(\lambda|a_0, b_0\right), \tag{21}$$

where the Gamma distribution $\mathrm{Gam}\left(\lambda|a_0, b_0\right)$ is given as

$$\mathrm{Gam}\left(\lambda|a_0, b_0\right) = \frac{1}{\Gamma(a_0)} \cdot b_0^{a_0} \cdot \lambda^{a_0 - 1} \cdot \exp\left(-b_0 \cdot \lambda\right), \tag{22}$$

where the Gamma function $\Gamma(x)$ is given as

$$\Gamma(x) = \int_0^\infty u^{x-1} \cdot e^{-u} du. \tag{23}$$

Furthermore, the normal distribution $\mathcal{N}\left(\theta|\theta_0, (\beta_0\lambda)^{-1}\right)$ is given as

$$\mathcal{N}\left(\theta|\theta_0, (\beta_0 \cdot \lambda)^{-1}\right) = \left(\frac{\beta_0 \cdot \lambda}{2 \cdot \pi}\right)^{\frac{1}{2}} \cdot \exp\left(-\frac{\beta_0 \lambda}{2} \cdot \{\theta - \theta_0\}^2\right) \tag{24}$$

Thus expanded normal-Gamma distribution is given as

$$p(\theta, \lambda|a_0, b_0, \theta_0, \beta_0) = \frac{b_0^{a_0} \cdot \lambda^{\left(a_0 - \frac{1}{2}\right)} \cdot \beta_0^{\frac{1}{2}}}{\Gamma(a_0)\sqrt{2 \cdot \pi}} \exp\left(-\frac{\beta_0 \lambda}{2} \cdot \{\theta - \theta_0\}^2 - b_0 \cdot \lambda\right). \tag{25}$$

To produce the posterior distribution and posterior-predictive distribution, the emperical mean $\hat{x}$ is

$$\hat{x} = \frac{1}{N} \sum_{n=1}^{N} x_n, \tag{26}$$

where this mean is used to condense some of the results that follow. The posterior distribution $p(\theta, \lambda|\mathbf{x})$ is given as

$$p(\theta, \lambda|\mathbf{x}) = NG(\theta, \lambda|a_n, b_n, \theta_n, \beta_n) = N(\theta|\theta_n, (\beta_n \cdot \lambda)^{-1})\mathrm{Gam}\left(\lambda|a_n, b_n\right), \tag{27}$$

where

$$\theta_n = \frac{\beta_0 \theta_0 + N \cdot \hat{x}}{\beta_0 + N}, \tag{28}$$

$$\beta_n = \beta_0 + N, \tag{29}$$

$$a_n = a_0 + \frac{N}{2}, \tag{30}$$

$$b_n = b_0 + \frac{1}{2} \sum_{n=1}^{N} (x_i - \hat{x})^2 + \frac{\beta_0 \cdot N \cdot (\hat{x} - \theta_0)^2}{2 \cdot (\beta_0 + N)}. \tag{31}$$

Additionally, the posterior marginal distributions are given as

$$p(\theta|\mathbf{x}) = T_{2 \cdot a_n}\left(\theta|\theta_n, \frac{b_n}{a_n \cdot \beta_n}\right), \tag{32}$$

$$p(\lambda|\mathbf{x})\mathrm{Gam}(\theta|a_n, b_n), \tag{33}$$

where $t_\nu(x|\mu, \sigma^2)$ is a Student-t distribution

$$t_\nu(x|\mu, \sigma^2) = c\left[1 + \frac{1}{\nu} \cdot \left(\frac{x - \mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}, \tag{34}$$

where ν is the degrees of freedom. The parameter c is given as

$$c = \frac{\Gamma\left(\frac{\nu}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\sqrt{\nu \cdot \pi \cdot \sigma^2}}. \tag{35}$$

The posterior predictive distribution is given as

$$p(x|\mathbf{x}) = T_{2 \cdot a_n}\left(x|\theta_n, \frac{b_n \cdot (\beta_n + 1)}{a_n \cdot \beta_n}\right). \tag{36}$$

1.2.1 Maximum likelihood estimation

Given the generative model in Equation (17) and the observed data shown in Figure 1, we can obtain maximum likelihood estimates for the unknown model parameters θ and σ. These are given through

$$\begin{aligned}\hat{\theta} &= \frac{1}{N}\sum_{n=1}^{N} x_n \\ &= 5.1161,\end{aligned} \tag{37}$$

$$\begin{aligned}\hat{\sigma} &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{\theta})^2 \\ &= 1.3905 \\ \lambda &= \frac{1}{\hat{\sigma}^2} = 0.5172.\end{aligned} \tag{38}$$

1.2.2 Bayesian inference

In this problem, we will use the Normal-Gamma conjugate prior to perform Bayesian inference for the generative model parameters θ and λ. In this investigation, the prior will be explored, and from this an investigation will be performed into how the prior hyper-parameters affect the MAP estimates will be conducted. This is to ensure that reasonable parameter ranges are selected and to determine the sensitivity of the Bayesian inference process to the prior.

The prior distribution

To gain some intuition into how the prior parameters affect the Normal-Gamma prior, different initialisations of the prior is needed. In Figure 9 different initialisations of the prior are shown. It is clear that a and b affect the distribution in the λ direction, while θ_0 and β_0 affect the mean and spread along the θ dimension.

Prior hyper-parameter investigation

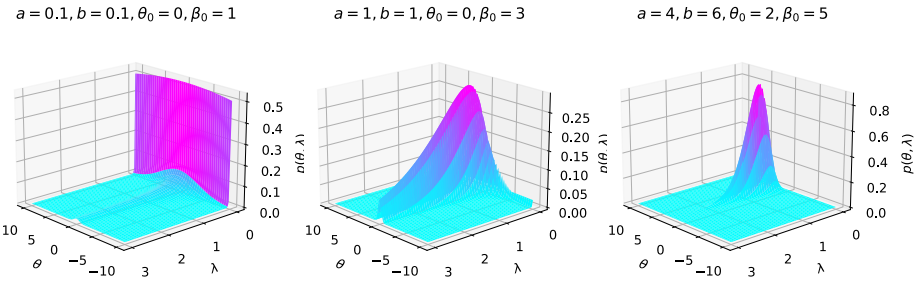
To conduct than investigation into the prior hyper-parameters, a reference prior state was used and each parameter was perturbed individually. The initial parameter state was set to $a_0 = 1, b_0 = 1, \theta_0 = 0, \lambda_0 = 3$. This was chosen at it gives a subjectively reasonable spread in each hyper-parameter dimension. This prior is visualised in Figure 10.

In Figure 11 the results from the prior hyper-parameter investigation is shown. It is clear from Figure 11 that the MAP estimates, and thus the posterior distribution, exhibits a large sensitivity to θ_0 and β_0. While the MAP estimates under a perturbation of a_0 and b_0 result in some changes, but to a lesser extent. In the perturbation of θ_0, it is interesting to note that as the θ hyper-parameter is perturbed to approach the ML estimate, the λ MAP estimate also approaches the ML λ estimate. Furthermore, as seen in Figure 11, perturbations of β_0 result in significant MAP parameter estimate variation.

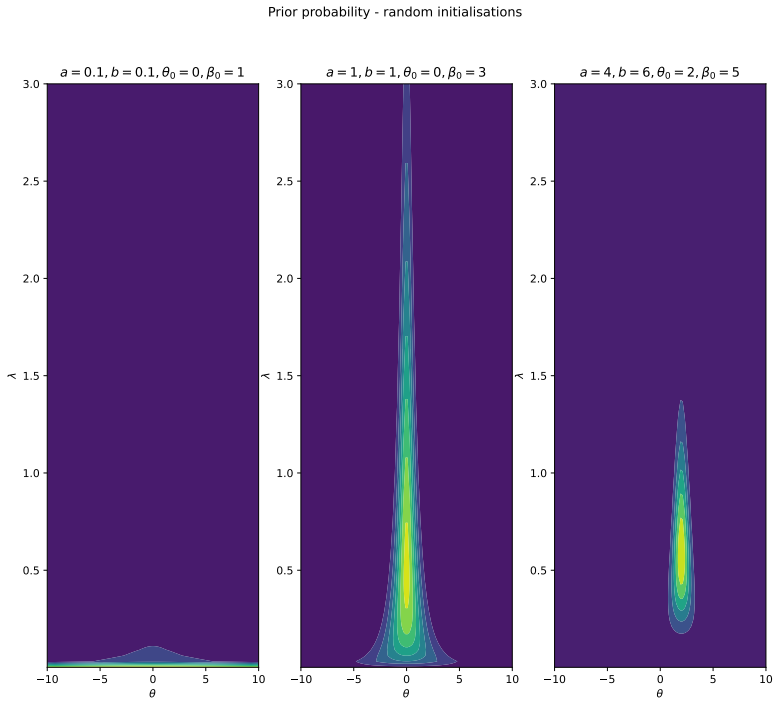
Inference results

Equations (28), (29), (30), and (31) were used to obtain the hyper-parameters of the posterior Normal-Gamma distribution. In Table 2 the prior and updated hyper-parameters are shown. Note that the λ_0 parameter was reduced so that its effect was less pronounced. Naturally, decreasing λ_0 is akin to raising the variance of a Gaussian distribution. This was done to reduce to dominance of the prior in the resulting posterior distribution. In Table 3 the MAP estimates for the unknown parameters are shown. It is clear that there are significant changes in the posterior distribution parameters. In Figure 12 the prior and posterior distributions with parameters given in Table 3 are given. It is clear from Figure 12 that the posterior distribution is highly localised around the MAP parameters.

The Bayesian inference process allows us to inspect the posterior distribution over θ and λ, and through marginalisation we can inspect the posterior marginal distribution $p(\theta|\mathbf{x})$ and $p(\lambda|\mathbf{x})$. In Figure 13 these distributions are shown. Note that Equations (32) and (33) were used to develop Figure 13. Using Equation (36), the posterior predictive distribution was determined. In Figure 14 the posterior predictive distribution is shown.



(a) Three-dimensional visualisation.



(b) Two-dimensional visualisation.

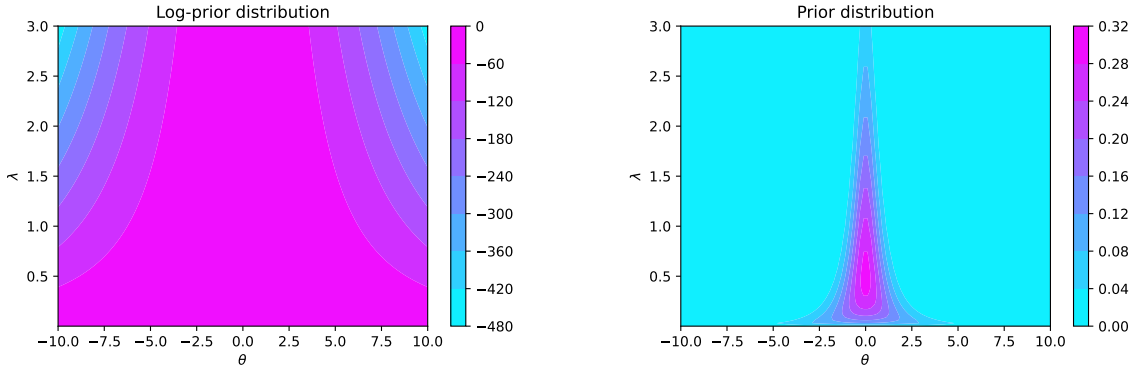
Figure 9: The Normal-Gamma prior under different hyper-parameter initialisations.

Table 2: The estimates for the hyper-parameters of the prior and posterior Normal-Gamma distributions

	a	b	θ	λ
Prior parameters	2	3	0	0.1
Posterior parameters	14.5	28.473	5.096	25.1

Using Bayesian inference

We can use the posterior distribution $p(\theta|\mathbf{x})$ and the posterior predictive distribution $p(x|\mathbf{x})$ to determine probabilities $p(a \leq X \leq b)$. The first probability of interest is $p(\theta \leq 4|\mathbf{x})$ and the second is $p(x \leq 4|\mathbf{x})$. It was found that $p(\theta \leq 4|\mathbf{x}) = 0.0003$ and $p(x \leq 4|\mathbf{x}) = 0.2247$. In Figure 15 the CDFs of interest and the resulting probabilities are shown. The posterior probability that $\theta \leq 4$ informs us that it is very unlikely that a model parameter θ in this range is responsible for generating the observed data. The posterior predictive probability that $x \leq 4$ is noticeably larger, and this indicates that it is far more probable that newly observed x samples may be drawn in the x domain $x \in [-\infty, 4]$.



(a) log-prior.

(b) prior.

Figure 10: The log Normal-Gamma prior and the Normal-Gamma prior in its initial state. The initial parameters are: $a_0 = 1, b_0 = 1, \theta_0 = 0, \lambda_0 = 3$.

Table 3: The estimates for the model parameters θ and λ using maximum likelihood and maximum a posteriori estimation. The MAP estimates were found through numerical optimisation.

Parameter	Maximum likelihood	MAP
θ	5.116	5.0957
λ	0.5172	0.4917
σ	1.3905	1.4261

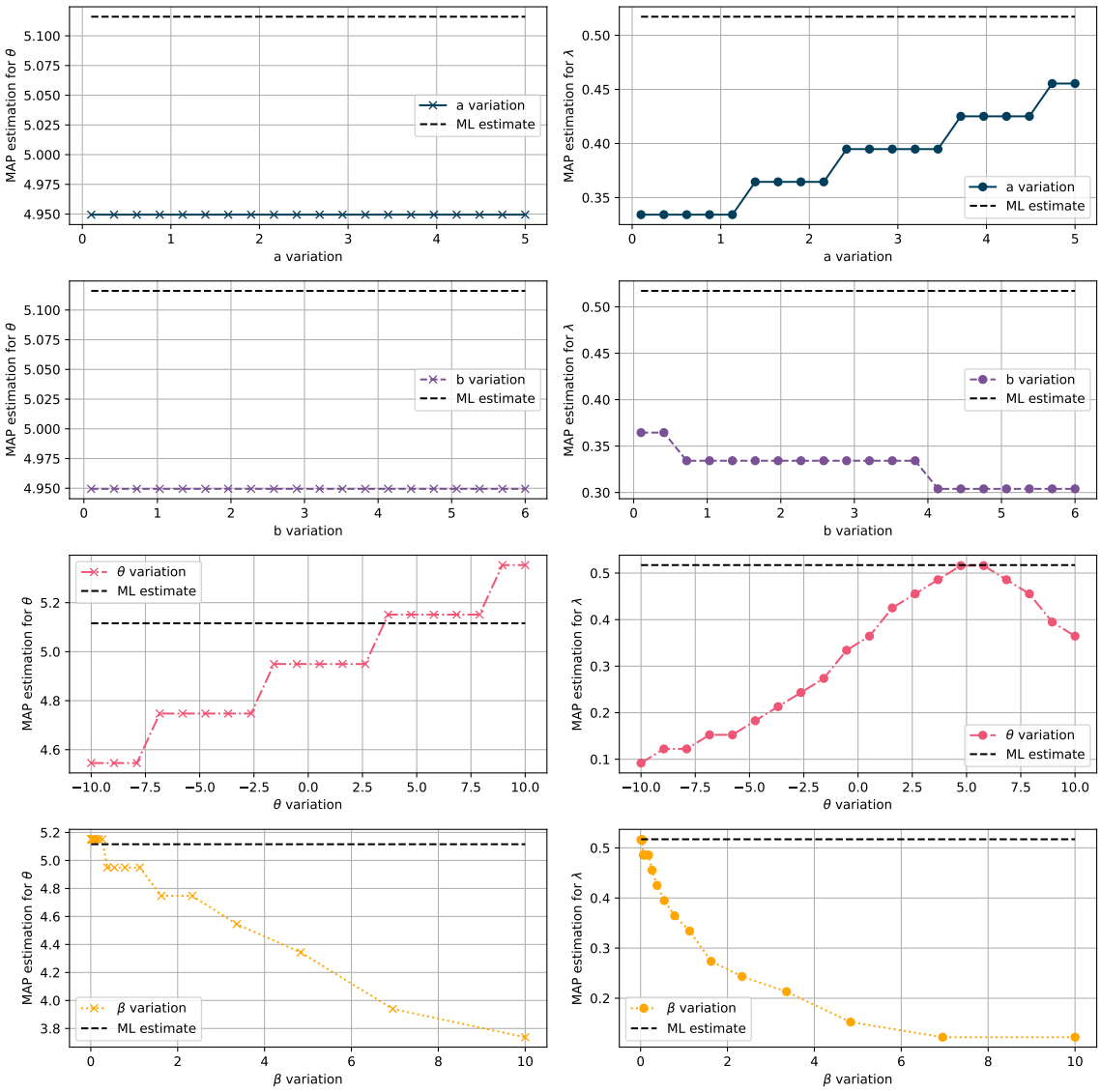
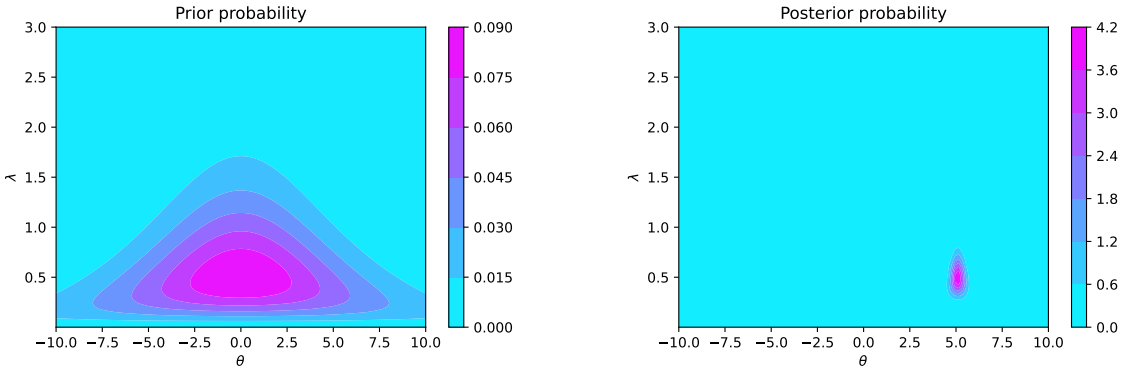


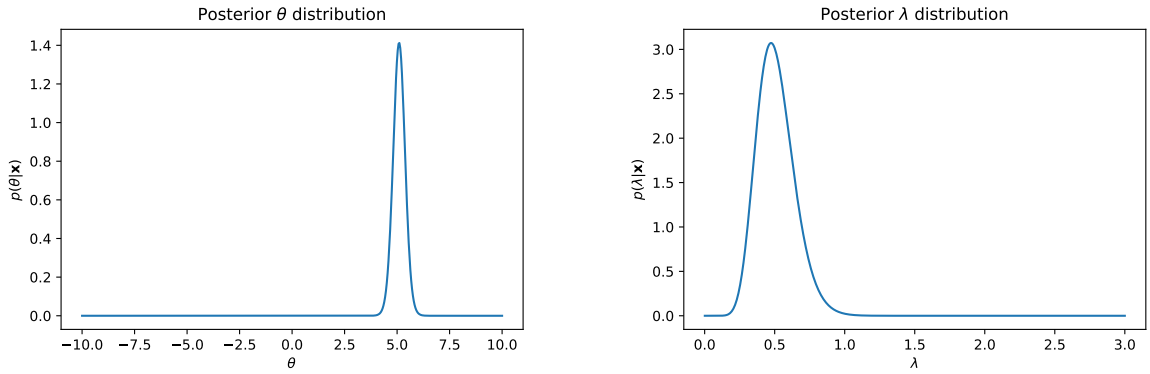
Figure 11: The resulting MAP estimates for θ and λ for each of the prior hyper-parameters.



(a) Prior distribution $p(\theta, \lambda)$.

(b) Posterior distribution $p(\theta, \lambda|\mathbf{x})$.

Figure 12: The prior and posterior Normal-Gamma distributions. The parameters for these distributions are given in Table 2.



(a) $p(\theta|\mathbf{x})$. (b) $p(\lambda|\mathbf{x})$.

Figure 13: The posterior marginal distributions for θ and λ.

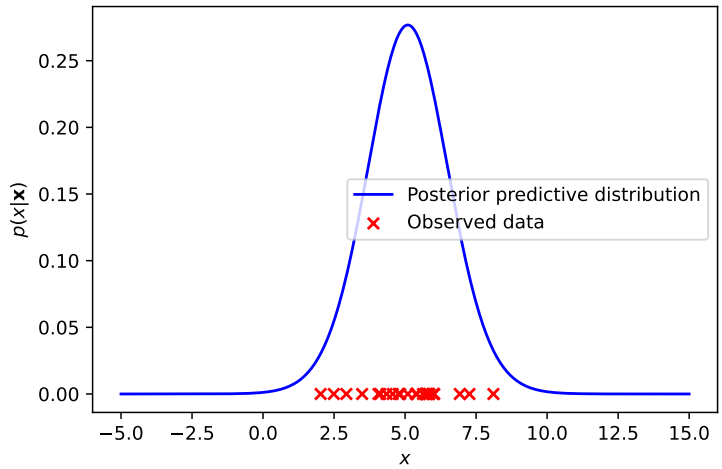
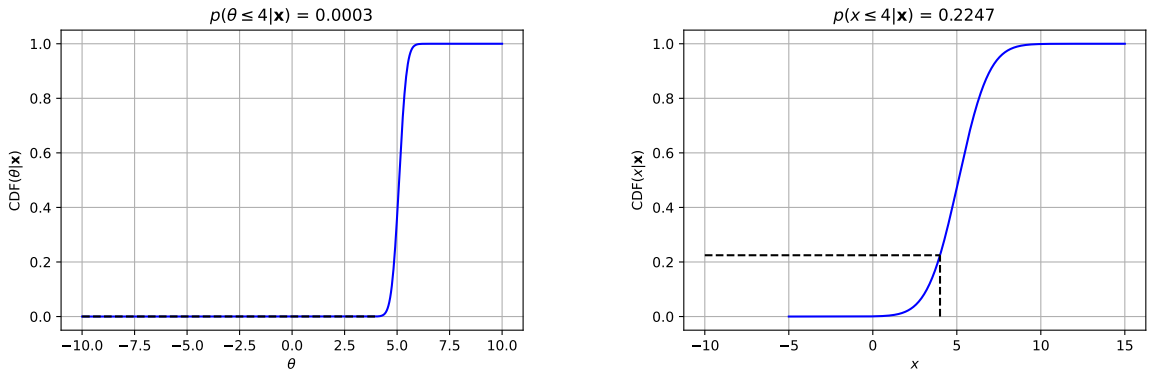


Figure 14: The posterior predictive distribution $p(x|\mathbf{x})$ for the Bayesian conjugate prior inference process.



(a) $F_{\theta|\mathbf{x}}(\theta)$. (b) $F_{x|\mathbf{x}}(x)$.

Figure 15: The CDFs used to determine the probabilities $p(\theta \leq 4|\mathbf{x})$ and $p(x \leq 4|\mathbf{x})$. In a) the CDF of the posterior distribution $p(\theta|\mathbf{x})$ is shown, while in b) the CDF of the posterior predictive distribution $p(x|\mathbf{x})$ is shown.

2 Question 2

Consider the regression model

$$\begin{aligned} y_n &= f(x_n) + \epsilon_n \\ &= \omega_0 + \omega_1 \cdot x_n + \omega_2 \cdot x_n^2 + \cdots + +\omega_K \cdot x_n^K + \epsilon_n \\ &= \boldsymbol{\omega}^T \boldsymbol{\phi}(x_n) + \epsilon_n, \end{aligned} \tag{39}$$

where y_n denotes the regression model response, x_n is the independent regression variable, $\boldsymbol{\omega}$ is the regression parameter vector, $\boldsymbol{\phi}(\cdot)$ is a basis column vector and ϵ_n is the sample noise

$$\epsilon_n \sim \mathcal{N}(0, \sigma^2), \tag{40}$$

with $\sigma = 0.5$. As the noise is an additive Gaussian distribution, the generative model is

$$p(y|x, \boldsymbol{\omega}, \sigma^2) = \mathcal{N}(y|f(x), \sigma^2). \tag{41}$$

In this problem, the goal is to perform Bayesian inference. The conjugate prior for this process is given by

$$p(\boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}), \tag{42}$$

where α is the prior precision. The observed regression data $\mathbf{y} = [y_1, y_2, \cdots, y_N]^T$ and corresponding independent variable $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T$ is detailed in Figure 16.

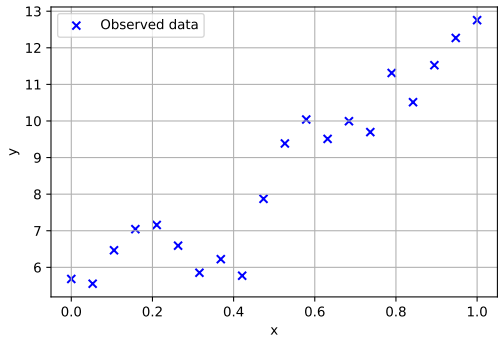


Figure 16: The observed regression data for Question 2.

2.1 Bayesian inference

In order to perform Bayesian inference, we assume that the samples in $\mathcal{D}$ satisfy the $i.i.d$ assumption. As such, we can calculate the conditional likelihood function $p(\mathbf{y}|\mathbf{x})$ as follows

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}, \sigma^2) = \prod_{n=1}^{N} p(y_n|x_n, \boldsymbol{\omega}, \sigma^2), \tag{43}$$

However, to avoid numerical underflow issues, and to simplify derivative computations, the logarithm of the likelihood function is used, which gives

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\omega}, \sigma^2) &= \sum_{n=1}^{N} p(y_n|x_n) \\ \mathcal{L}(\boldsymbol{\omega}) &= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left[y_n - f(x_n)\right]^2 \right), \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(\left[y_n - f(x_n)\right]^2 \right), \end{aligned} \tag{44}$$

where $\mathcal{L}(\mathbf{w}, \sigma^2)$ is the log-likelihood function that we wish to maximise. To perform maximum likelihood estimation we would maximise this function, we can use numerous methods (grid search, optimisation), but in this assignment we will make use of the analytical solution to the ML problem. The ML estimate for $\boldsymbol{\omega}$ is given as

$$\hat{\boldsymbol{\omega}} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \mathbf{t}, \tag{45}$$

where

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1^{(0)} & \phi_1^{(1)} & \cdots & \phi_1^{(K)} \\ \phi_2^{(0)} & \phi_2^{(1)} & \cdots & \phi_2^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_N^{(0)} & \phi_N^{(1)} & \cdots & \phi_N^{(K)} \end{bmatrix}, \tag{46}$$

such that the n^{th} row of $\boldsymbol{\Phi}$ is equal to $\boldsymbol{\phi}(x_n)^T$ and $\boldsymbol{\Phi} \in \mathbb{R}^{N \times K+1}$. Given the ML estimate for the regression parameters, the sampling distribution of the estimated parameters is given by

$$\hat{\boldsymbol{\omega}} \sim \mathcal{N}(\boldsymbol{\omega}_{ML}, \boldsymbol{\Sigma}_{ML}), \tag{47}$$

where the covariance $\boldsymbol{\Sigma}_{ML}$ is given by

$$\boldsymbol{\Sigma}_{ML} = \sigma^2 \cdot \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1}. \tag{48}$$

However, in the Bayesian inference setting we are interesting in the posterior distribution $p(\boldsymbol{\omega}|\mathbf{x}, \mathbf{y})$, which is given in Bishop [1] as

$$p(\boldsymbol{\omega}|\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\boldsymbol{\omega}|\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}\right), \tag{49}$$

where the mean of the multivariate Gaussian posterior distribution is

$$\boldsymbol{\mu}_{\boldsymbol{\omega}} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \frac{\alpha}{\beta} \cdot \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^T \mathbf{y}, \tag{50}$$

and the covariance of the multivariate Gaussian posterior distribution is

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \left(\beta \cdot \boldsymbol{\Phi}^T \boldsymbol{\Phi} + \alpha \cdot \mathbf{I}\right)^{-1}. \tag{51}$$

Note that the parameter β is the inverse of the generative model variance

$$\beta = \frac{1}{\sigma^2}. \tag{52}$$

Furthermore, the posterior predictive distribution is given as

$$p(y|x, \mathbf{x}, \mathbf{y}, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\mu}_{\boldsymbol{\omega}}^T \boldsymbol{\phi}(x), \sigma_p(x)^2), \tag{53}$$

where the posterior predictive variance is given as

$$\sigma_p(x)^2 = \frac{1}{\beta} + \boldsymbol{\phi}(x)^T \boldsymbol{\Sigma}_{\boldsymbol{\omega}} \boldsymbol{\phi}(x). \tag{54}$$

The posterior distribution

The posterior distribution over $\boldsymbol{\omega}$ is dependent on the polynomial order K. For an order of $K = 1$, the posterior mean (which is also the MAP estimate for $\boldsymbol{\omega}$) is given by

$$\boldsymbol{\mu}_{\boldsymbol{\omega}} = \begin{bmatrix} 5.6164 & 16.3821 & -119.3896 & 352.7705 & -397.5298 & 154.9652 \end{bmatrix}^T. \tag{55}$$

In comparison, the ML estimate for the optimal regression parameters is

$$\hat{\boldsymbol{\omega}} = \begin{bmatrix} 5.2873 & 30.3442 & -222.7383 & 630.0886 & -706.5113 & 276.5713 \end{bmatrix}^T. \tag{56}$$

The posterior covariance is given by

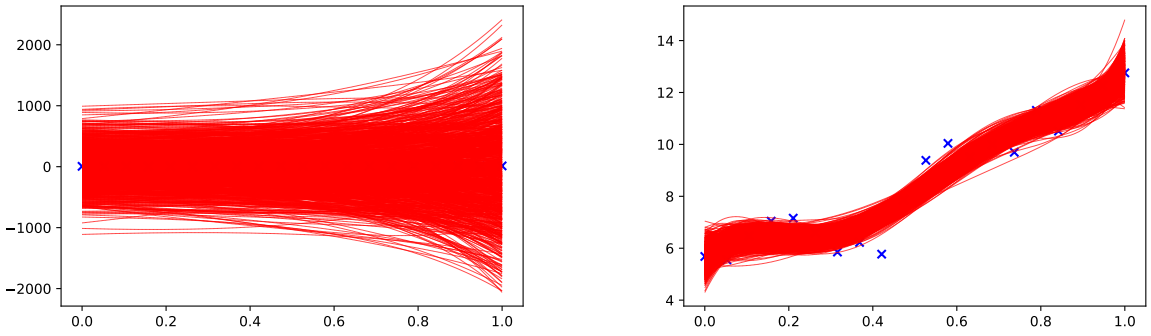
$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \begin{bmatrix} 0.192 & -2.666 & 11.817 & -22.766 & 19.903 & -6.479 \\ -2.666 & 68.785 & -399.800 & 913.643 & -905.064 & 325.313 \\ 11.817 & -399.800 & 2642.418 & -6563.822 & 6906.319 & -2599.888 \\ -22.766 & 913.643 & -6563.822 & 17239.678 & -18895.118 & 7340.081 \\ 19.903 & -905.064 & 6906.319 & -18895.118 & 21343.734 & -8487.156 \\ -6.479 & 325.313 & -2599.888 & 7340.081 & -8487.156 & 3436.762 \end{bmatrix} \tag{57}$$

In comparison, the ML estimate for the sampling distribution covariance is

$$\boldsymbol{\Sigma}_{ML} = \begin{bmatrix} 0.212 & -3.478 & 17.796 & -38.751 & 37.667 & -13.457 \\ -3.478 & 103.027 & -652.543 & 1590.580 & -1658.292 & 621.463 \\ 17.796 & -652.543 & 4510.225 & -11570.545 & 12480.551 & -4792.511 \\ -38.751 & 1590.580 & -11570.545 & 30667.479 & -33850.660 & 13224.552 \\ 37.667 & -1658.292 & 12480.551 & -33850.660 & 38005.432 & -15044.317 \\ -13.457 & 621.463 & -4792.511 & 13224.552 & -15044.317 & 6017.727 \end{bmatrix} \tag{58}$$

2.1.1 Sampling the prior and posterior

If we draw samples from the prior $p(\boldsymbol{\omega})$ and the posterior $p(\boldsymbol{\omega}|\mathbf{y})$, we can visualise how different regression models fit the observed data before and after performing inference on the observed data. In Figure 17, the models with parameters from the prior and posterior distribution is shown.



(a) $\boldsymbol{\omega} \sim p(\boldsymbol{\omega})$.

(b) $\boldsymbol{\omega} \sim p(\boldsymbol{\omega}|\mathbf{y})$.

Figure 17: A visualisation of the samples from the prior and posterior distributions over $\boldsymbol{\omega}$.

2.1.2 The posterior predictive distribution

The posterior distribution $p(\boldsymbol{\omega}|\mathbf{y})$ is only one half of the Bayesian inference process. We can use the posterior predictive distribution in Equation (53) to make predictions for new x values. In Figure 18 the posterior predictive distribution over the x domain is shown.

2.1.3 Bayesian inference interpretation

Before we continue with the assignment, it is important that we reflect and comment on the distributions that we are using and the information they contain. The distributions discussed here are:

- The prior predictive distribution.
- Functions generated from sampling the posterior distribution.
- The posterior predictive distribution.

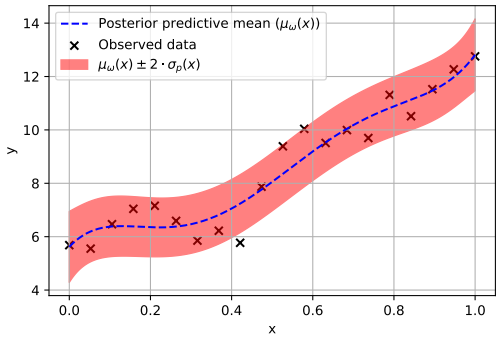


Figure 18: The posterior predictive distribution overlaid with the observed data.

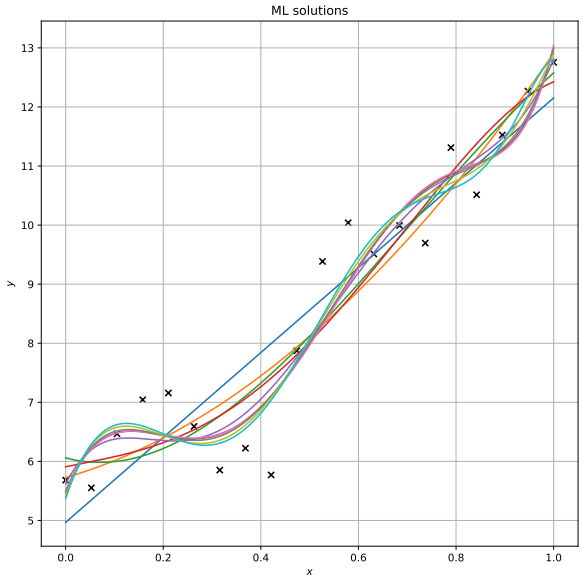
The prior predictive distribution: This distribution reflects our prior belief given to the generative model. The important aspect of this distribution is that it is a marginalisation over the model parameters for the product of the generative model and the prior distribution. As such, this distribution calculates the probability of a new data point x given our prior uncertainty about the model parameters $\boldsymbol{\omega}$. The interpretation of this distribution is subjective, as it highly depends on our prior inference built into to the prior distribution. Hypothetically, if we expect that one regression parameter is not negative (such as the model offset ω_0) then we could incorporate this prior knowledge into the prior distribution. This hypothetical process is subjective, as it depends on the user, and hence the prior predictive distribution will reflect this bias.

Functions generated from sampling the posterior distribution: Simply put, these functions represent different generative models. However, the posterior distribution represents the uncertainty in the parameter space given some observed data. As such, these functions represent are a proxy to the fit of the generative model to the observed data, where this fit has some uncertainty or variation (which is conveyed in the posterior distribution). If the functions vary greatly from sample to sample, then this is an indication that there is a large amount of uncertainty in the posterior distribution or the parameter space.

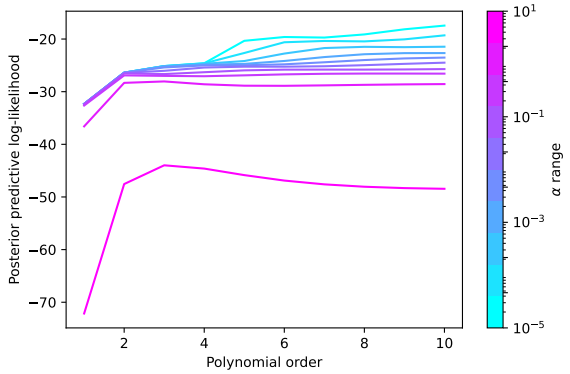
The posterior predictive distribution: The posterior predictive distribution is primarily used for predictive inference. This inference gives a distribution over an unobserved data point x, and it represents the process of integrating over an infinite number of models. Naturally, the posterior distribution conveys the uncertainty in the possible parameters for these different models, and hence gives an weighting to models that it believes are more likely given the observed data. This then allows the posterior predictive distribution to not only capture and express the uncertainty around a new data point, but it also serves as a proxy for us as to whether or not this new data is similar to the data that was previously observed.

2.1.4 The impact of the prior hyper-parameters

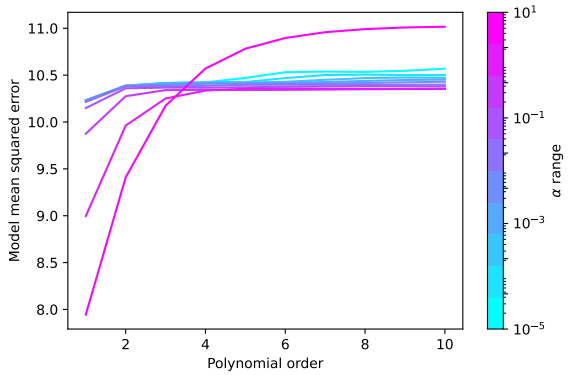
As before, the prior hyper-parameters influence the distributions obtained through the Bayesian inference process. The effects of α will be investigated. In this investigation, the model order will be linearly increased from 1 to 10. The objective is to investigate the effect of the model order on the ML estimate for $\boldsymbol{\omega}$, and the effect of the hyper-parameter α on the MAP estimate of $\boldsymbol{\omega}$ for different model orders. To quantify these effects, the mean-squared error (MSE) and the posterior-predictive log-likelihood will be inspected for different model orders and different values of α. It is clear from Figure 19a) that as the model order is increased, the generative model has more flexibility and thus appears more non-linear. In Figure 19b) one can note that increasing α causes a gradual decrease in the posterior predictive log-likelihood. This gradual decrease, based on Figure 19d), is attributed to the dominance of the prior in the Bayesian inference process. The prior enforces that the model parameters should be closer to zero, which acts as a form of regularisation on the model parameters. Hence, the expected values of the posterior predictive distribution for different model order become less non-linear as α increases. It is also clear to see that the posterior predictive log-likelihood steadies out after a model order of 2, which indicates that a second order polynomial fits that data well and additional model order do not have a significant impact. In Figure 19c) an interesting response is observed, whereby as α is increased the MSE gets better up until $\alpha = 10$. From this point higher order models fit the data poorly in comparison to the lower-order variants.



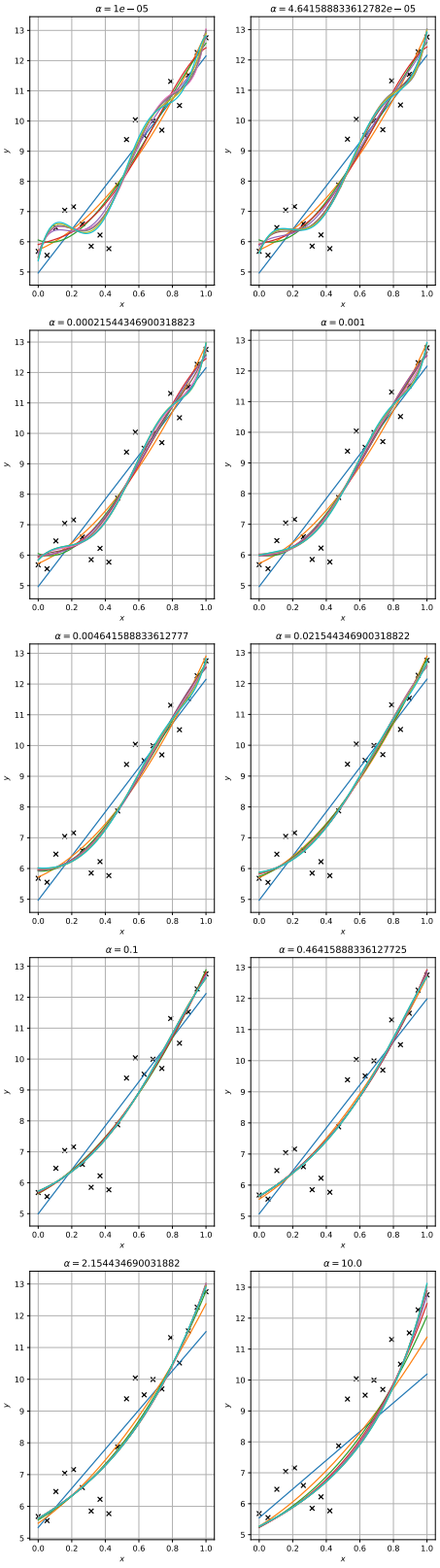
(a) ML estimates for $M = 1, \cdots, 10$.



(b) Posterior predictive log-likelihood.



(c) Mean squared error.



(d) Posterior predictive expected value visualisation.

Figure 19: The results of the α investigation. In a) the ML estimates for different model orders are shown, in b) the posterior-predictive log-likelihood for various α values and various model orders is shown, in c) the MSE for various α values and various model orders is shown, and in d) the expected values of the posterior predictive distribution are shown for the values of α considered in the investigation.

2.1.5 Interpreting the Bayesian inference results

We can now use the prior distribution, posterior distribution, and posterior-predictive distribution to as questions regarding the model parameters and the predictive inference process. As the prior and posterior distributions are Gaussian, we can utilise the known properties of the Gaussian distribution to obtain the marginal distributions for the ω_1 parameter. The details of this process are given in Bishop [1]. As a result, the marginal distribution of interest for $P(\omega_1 > 8)$ is given by

$$p(\omega_1) = \mathcal{N}\left(\omega_1|0, \frac{1}{\alpha}\right). \tag{59}$$

To calculate the probability $P(\omega_1 > 8|\mathbf{x}, \mathbf{y})$, we can use the following Gaussian distribution

$$p(\omega_1|\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\omega_1|\mu_{\omega_1}, \Sigma_{\omega_{1,1}})\right). \tag{60}$$

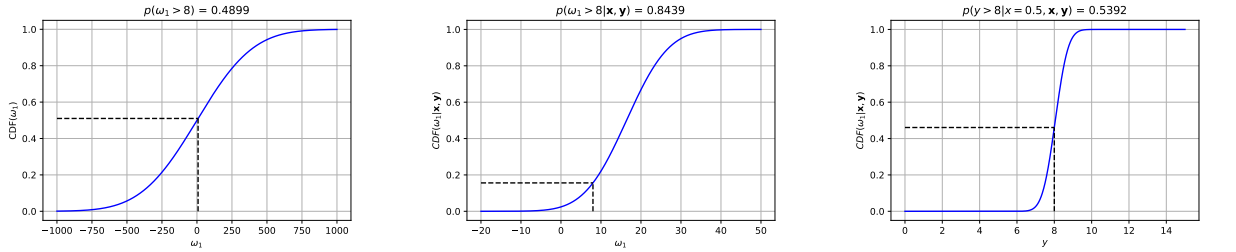
Finally, to calculate the probability $P(y > 8|x = 0.5, \mathbf{x}, \mathbf{y})$, we can use the posterior predictive distribution

$$p(y|x = 0.5, \mathbf{x}, \mathbf{y}, \alpha, \beta) = \mathcal{N}(y|\boldsymbol{\mu}_{\boldsymbol{\omega}}^T \boldsymbol{\phi}(0.5), \sigma_p(0.5)^2) \tag{61}$$

Note that our standard definition of probability is $P(x \leq a) = F_x(a)$. The probabilities of interest here, however, are $P(x > a)$ which can be expressed as

$$P(x > a) = 1 - P(x \leq a). \tag{62}$$

The first probability $P(\omega_1 > 8)$ was found to be $P(\omega_1 > 8) = 1 - 0.5101 = 0.4899$. The second probability $P(\omega_1 > 8|\mathbf{x}, \mathbf{y})$ of interest was found to be $P(\omega_1 > 8|\mathbf{x}, \mathbf{y}) = 1 - 0.1561 = 0.8439$. The final probability of interest is $P(y > 8|x = 0.5, \mathbf{x}, \mathbf{y})$ and this was determined to be $P(y > 8|x = 0.5, \mathbf{x}, \mathbf{y}) = 1 - 0.4608 = 0.5392$. In Figure 20 the CDFs of the prior, posterior and posterior-predictive distributions are shown. It is clear from Figure 20a) that the variance in the prior distribution is significant. This is expected as the ratio $\frac{1}{\alpha} = 10^5$. Thus, unless we evaluate the probability at the extremities of the prior, the probability will be close to 0.5. Hence the ω_1 parameter can be sampled from an extensive range when using the prior. From Figure 20b), it is clear that the posterior distribution has a reduced variance. As such, the probability that the $\omega_1 > 8$ indicates that the parameter value is more likely to be larger than 8. Finally, as seen in Figure 20c), the posterior predictive distribution at $x = 0.5$ gives a distribution over y that appears to be centred at 8. Hence, evaluating the probability that $y > 8$ returns a value close of 0.5, which indicates that a value of $y = 8$ is highly likely at $x = 0.5$.



(a) prior distribution. (b) posterior distribution. (c) posterior-predictive distribution.

Figure 20: The CDFs of the prior, posterior, and posterior predictive distributions.

2.1.6 The model evidence

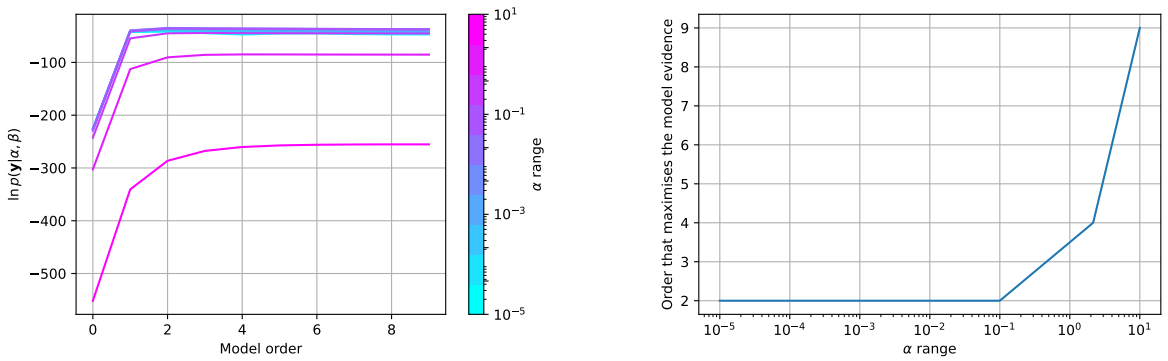
In Bishop [1], the model evidence is given by

$$\ln p(\mathbf{y}|\alpha, \beta) = \frac{M}{2} \cdot \ln \alpha + \frac{N}{2} \cdot \ln \beta - E\left(\boldsymbol{\mu}_{\boldsymbol{\omega}}\right) - \frac{1}{2} \cdot \ln |\boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{-1}| - \frac{N}{2} \cdot \ln(2 \cdot \pi), \tag{63}$$

where $E\left(\boldsymbol{\mu}_{\boldsymbol{\omega}}\right)$ is given by

$$E\left(\boldsymbol{\mu}_{\boldsymbol{\omega}}\right) = \frac{\beta}{2}\|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{\omega}}\|_2^2 + \frac{\alpha}{2} \cdot \boldsymbol{\mu}_{\boldsymbol{\omega}}^T \boldsymbol{\mu}_{\boldsymbol{\omega}}. \tag{64}$$

In Figure 21 the model evidence for different model orders and different values of the hyper-parameter α are shown. It is clear from Figure 21 that the parameter α has little to no sensitivity on the optimal model order until $\alpha = 0.1$. From this point, a higher model order is found to be desirable by the model evidence. This is attributed to the nature in which the α hyper-parameter regularises the Bayesian inference process, and hence once this regularisation is strong, models with more flexibility are needed to overcome the effects of regularisation.



(a) The model evidence.

(b) Optimal order versus α.

Figure 21: The model evidence for different values of α. In a) the model evidence is shown, and in b) the model order that maximises the model evidence for each value of α is plotted.

2.1.7 Introducing sparsity

Sparsity, in Bayesian inference, may be introduced to the model parameters through choices in the prior or through careful selection of the prior hyper-parameters. As was seen in this problem, if we increase α to a point where the prior variance is small, the resulting posterior predictive distribution and the MAP estimates of the model parameters is sparse. Alternatively, we could use a prior that is naturally sparse (i.e., it is strongly centered around zero) such as a Laplacian prior. The idea of sparsity is linked to regularisation, whereby the model parameters are constrained to ensure that the model does not overfit to the observed data. I would recommend looking into Laplacian priors or using the hyper-parameter α if one wishes to introduce sparsity into the Bayesian linear regression process.

Inducing sparsity differs from model evidence selection through the way that it interacts with the posterior distribution. A sparse prior, for example, has a direct influence on the posterior distribution, whereas the model evidence is the maximisation of the marginal likelihood function over the observed data for different model orders. While the posterior distribution does feature in this marginal likelihood, the prior has directly affects the methods used to determine the posterior distribution.

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, Jan. 2006. [Online]. Available: `https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/`.