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.

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

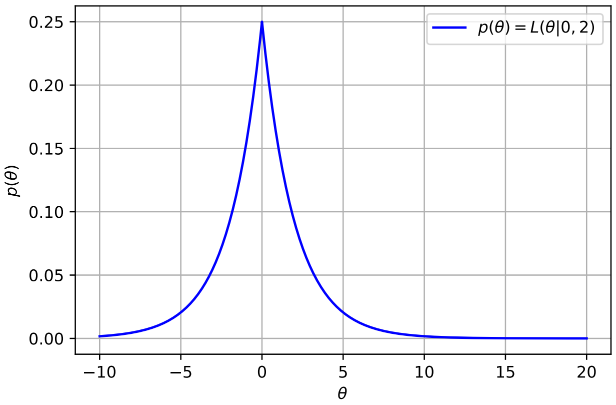


Figure 1: The prior distribution function $p(\theta)$ visualised over the domain $\theta \in [-10, 20]$.

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 2 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 $p(\theta|\mathbf{x})$.

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 $p(\mathbf{x}|\theta)p(\theta)$

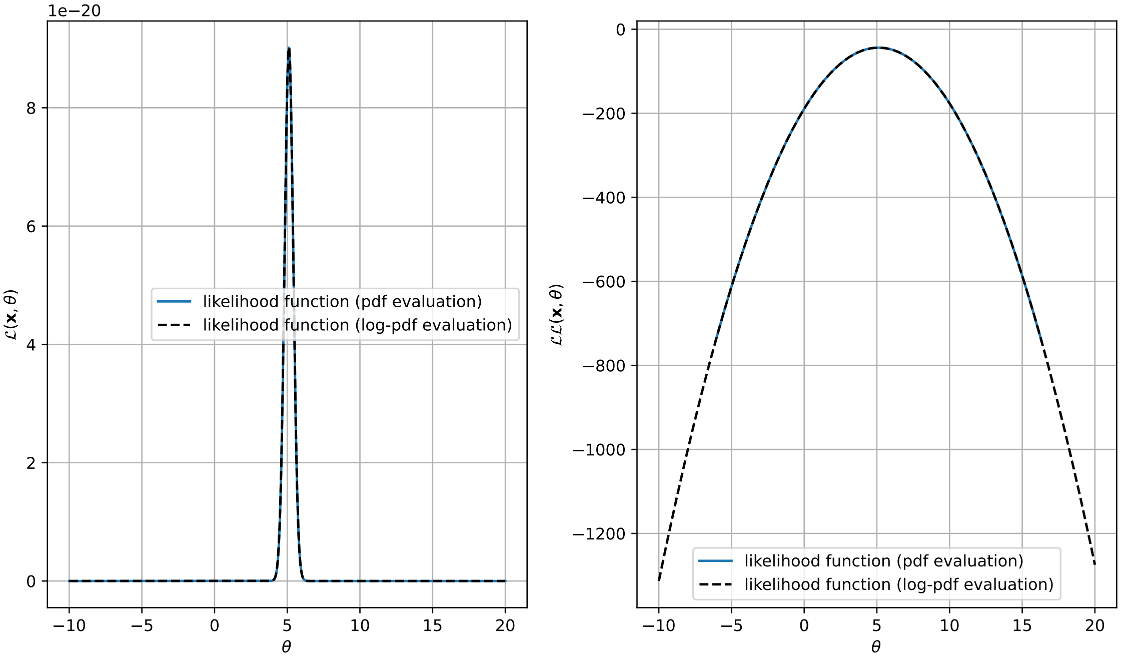


Figure 2: The likelihood function $\mathcal{L}(\mathbf{x},\theta)$ of the generative model $p(x|\theta)$ over the domain $\theta \in [-10, 20]$. In a) the likelihood function is shown, and in b) the log-likelihood function is shown.

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 3 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.

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

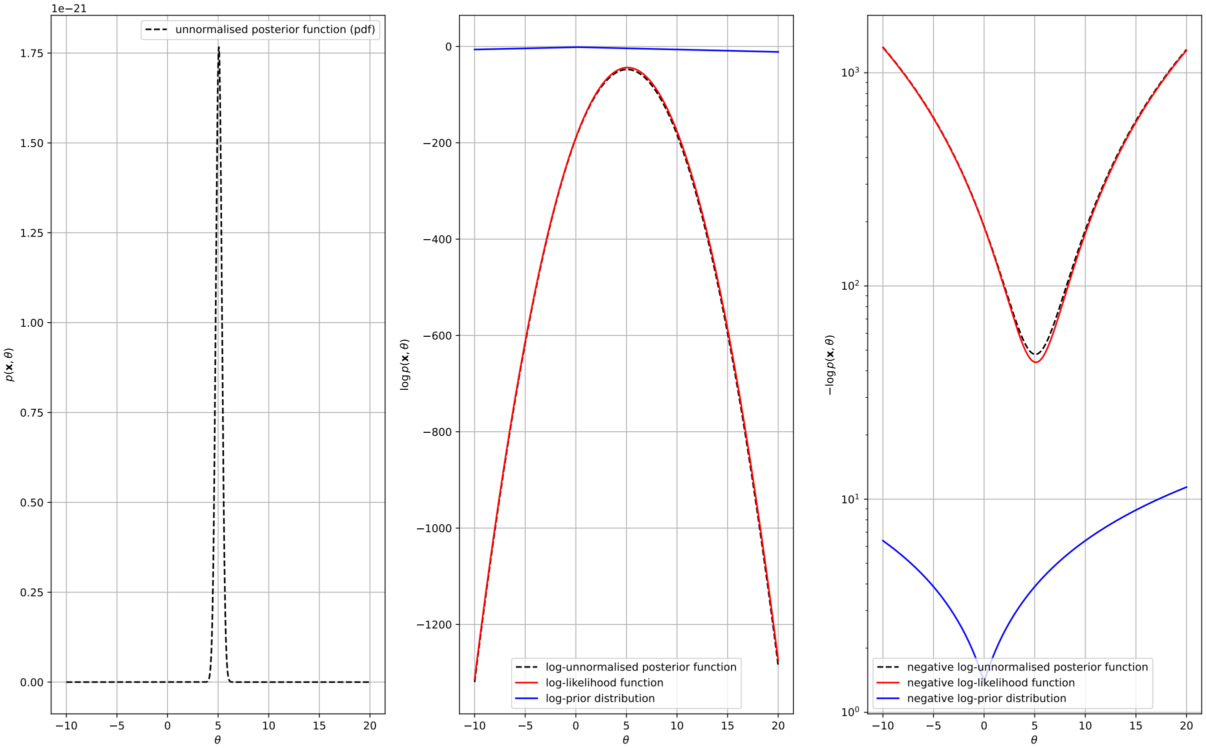


Figure 3: The unnormalised posterior distribution $\mathcal{F}(\theta, \mathbf{x}, \sigma, \mu_\theta, b)$ for $\sigma = 1.5, \mu_\theta = 0$ and $b = 2$ over the domain $\theta \in [-10, 20]$. Three variations of the unnormalised posterior is visualised here, the standard unnormalised posterior, the logarithm of the unnormalised posterior and the terms therein, and the negative of the unnormalised log-posterior.

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

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

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 4 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 $p(\theta|\mathbf{x}) > 0$ $\forall\ \theta$).

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

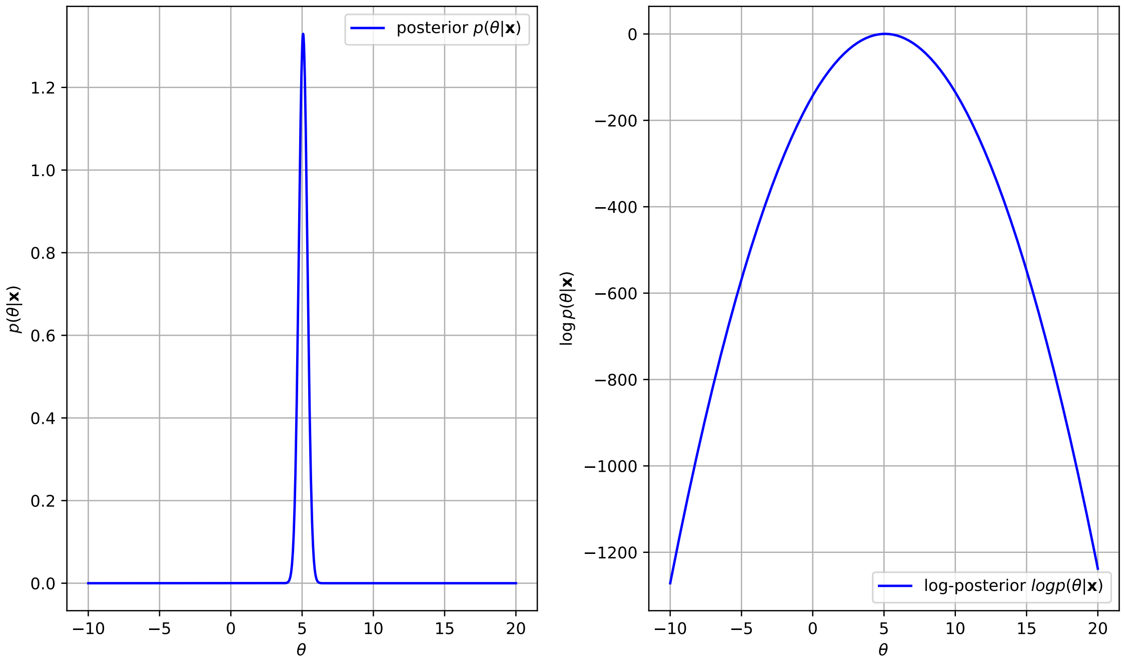


Figure 4: The posterior distribution $p(\theta|\mathbf{x})$ and the logarithm thereof over the domain $\theta \in [-10, 20]$. The MAP solution is given in Table 1.

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 5 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.

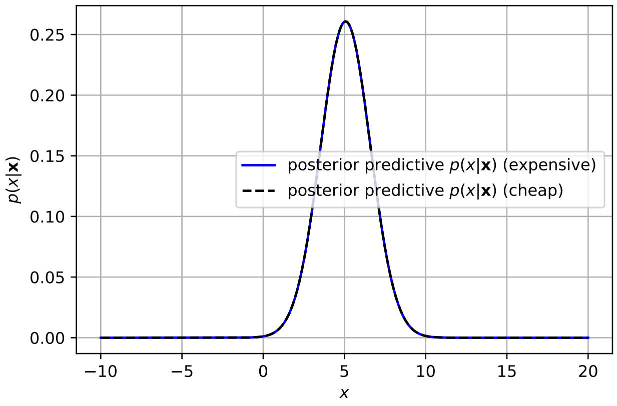
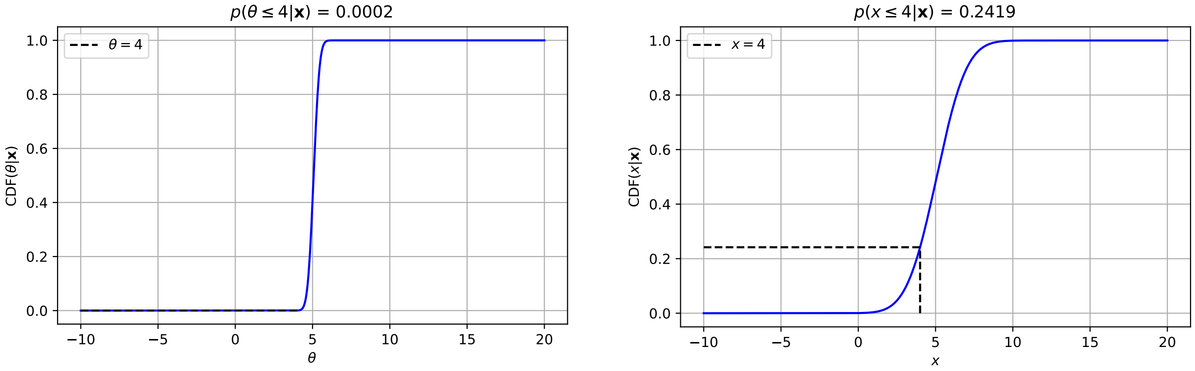


Figure 5: The posterior predictive distribution $p(x|\mathbf{x})$ determined using a cheap and expensive computational methodology over the domain $\theta \in [-10, 20]$.

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



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

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

Figure 6: 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.

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 $p(\theta)$.

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 an using 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 reminded, 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 6 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

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

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

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

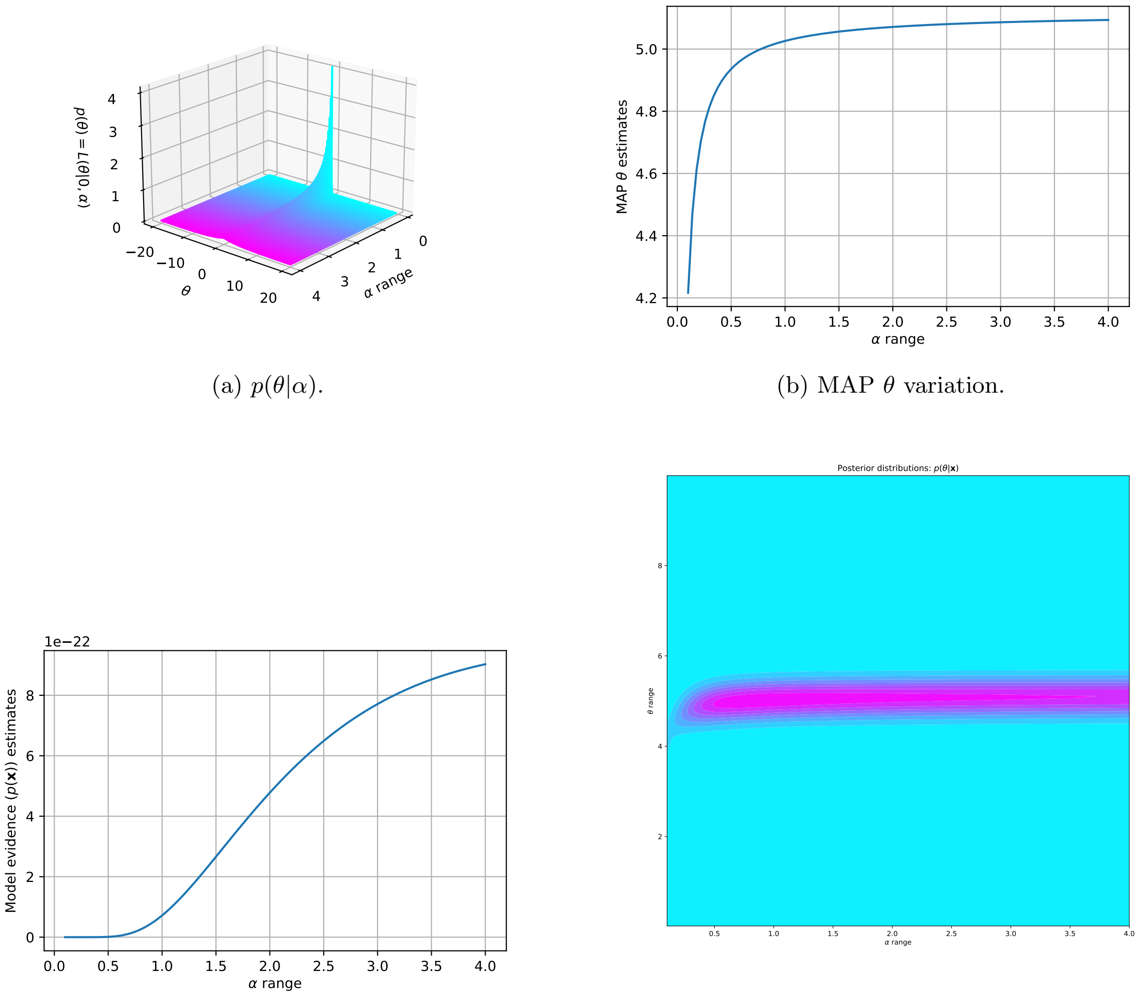
and the model evidence

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

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 7 the results of varying α is shown. It is clear from Figure 7b) that by increasing the variance we approach a region where the MAP estimate steadies out. Furthermore, as seen in Figure 7d) 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

asd



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

(b) MAP θ variation.

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

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

Figure 7: 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.