The Lighthouse Problem: S2 coursework report

Vishal Jain

March 23, 2024

Contents

1 Introduction

This coursework is based on the Lighthouse problem. Where a lighthouse is at position α along a straight coastline and a distance β out to sea. The lighthouse rotates and emits flashes at uniformly-distributed random angles θ; the light beams are narrow and (if $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$) intersect the coastline at a single point. An array of detectors spread along the coastline record the locations x_k (where $k = 1, 2, \ldots, N$) where N flashes are received; the detectors only record that a flash has occurred, not the direction from which it was received. Your task is to find the location of the lighthouse. The setup is illustrated in Fig 1.
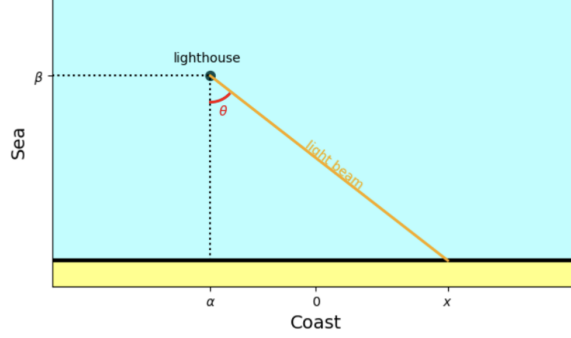

Figure 1: The lighthouse problem setup

1.1 Part i - Trigonometric relationship between variables

From basic trigonometry, the tangent of the angle θ, where θ is the angle of the light beam with respect to a line perpendicular to the coastline, is defined as the ratio of the opposite side to the adjacent side of the right angled triangle formed by the points (α, β), $(x, 0)$ and $(\alpha, 0)$. This relationship can be represented as:

$$\tan(\theta) = \frac{x - \alpha}{\beta}.$$

1.2 Part ii - Probability density function of x

the probability density function of x can be found from the following relationship:

$$Pr(\theta)d\theta = Pr(x)dx$$

Given that $\theta \sim U(-\frac{\pi}{2}, \frac{\pi}{2})$, The probability density function of θ is defined as:

$$Pr(\theta) = \begin{cases} \frac{1}{\pi} & \text{if } -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, in this interval, the probability density function of x is given by:

$$Pr(x) = Pr(\theta)\frac{d\theta}{dx}$$

$$= \frac{1}{\pi}\frac{d\theta}{dx}$$

$$= \frac{1}{\pi}\frac{d}{dx}\arctan\left(\frac{x - \alpha}{\beta}\right)$$

$$\therefore Pr(x) = \frac{1}{\pi}\left(\frac{\beta}{\beta^2 + (x - \alpha)^2}\right) \tag{1}$$

2

Where the last line follows from the standard derivative formula for the arctan function:

$$\frac{d}{dx}\arctan(x) = \frac{1}{1 + x^2}$$

Using this expression with the chain rule, we obtain the probability density function for x as shown.

2 Finding the Lighthouse

2.1 Part iii - Most likely location of a flash

Given the previously calculated probability density function of x, the distribution can be visualised for different light house locations. Figure 2 shows the probability density function of x for 3 different choices of α and β. The plots reveal several interesting properties about the PDF of x. Firstly, as the value of β decreases, the peak becomes more pronounced. This makes sense as, the closer the lighthouse to the shore, the more information its flashes would give about its location along the shore. The plots also reveals that the peak of the distribution occurs at $x = \alpha$, this can also be seen by inspection of equation 1.
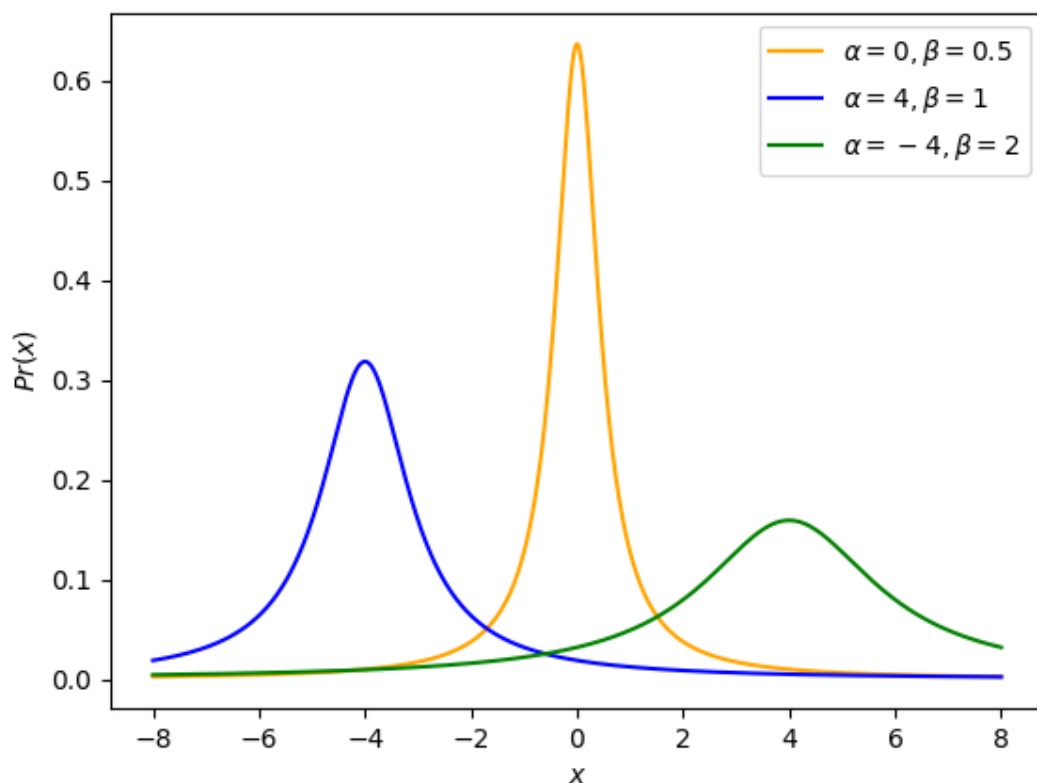

Figure 2: Probability density function of x for 3 different choices of α and β

To estimate α, one may consider using the sample mean of x, however, since x follows a cauchy distribution, the sample mean will not converge as the distribution's mean and variance are both undefined.

Another way to show that the sample mean is not a good estimator of α is to consider the maximum likelihood estimator of α, which is a good estimator.

3

The MLE estimate of α can be found by taking the derivative of the log likelihood function with respect to α and setting it to zero. The likelihood function of a set of flashes $\{x_k\}$ is given by:

$$L(\{x_k\}|\alpha, \beta) = \prod_{k=1}^{N} Pr(x_k),$$

this follows by the independence of the flashes. The log likelihood function is then given by:

$$\log L(\{x_k\}|\alpha, \beta) = \sum_{k=1}^{N} \log Pr(x_k).$$

Substituting in the expression for $Pr(x)$ from equation 1 gives:

$$\log L(\{x_k\}|\alpha, \beta) = \sum_{k=1}^{N} \log \left(\frac{\beta}{\pi(\beta^2 + (x_k - \alpha)^2)} \right).$$

Taking the derivative of this with respect to α and setting it to zero gives the expression:

$$0 = \sum_{k=1}^{N} \frac{2(x_k - \hat{\alpha})}{(\beta^2 + (x_k - \hat{\alpha})^2)^2},$$

where $\hat{\alpha}$ is the MLE estimate of α. This equation can be solved numerically to find the MLE estimate of α. Note how the MLE estimate of α is not the sample mean of x.

2.2   Part iv - Prior distribution of α and β

To build a posterior distribution of α and β, we need to define a prior distribution for these parameters. The choice of prior should capture the current state of belief regarding the values of α and β. As the lighthouse is equally likely to be at any location along the coast and at any distance from the coast, choosing a uniform prior distribution is a sensible approach. This is because a uniform prior distribution assigns equal probability to all values of α and β.

$$Pr(\alpha, \beta) = \begin{cases} \frac{1}{(a-b)(c-d)} & \text{if } a \leq \alpha \leq b \text{ and } c \leq \beta \leq d \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In theory the values for a, b, c and d should be determined in a manner which does not involve the data. However, in practice, to ensure the prior provides relevant support, the values of a, b, c and d are set by looking at the data and deciding on a conservative range. In this case, the values of a, b, c and d were set to be -100, 100, 0, 100 respectively.

2.3   Part v - Posterior distribution of α and β

The posterior $p(\alpha, \beta|\{x_k\})$ can be found using Bayes' theorem:

$$p(\alpha, \beta|\{x_k\}) = \frac{p(\{x_k\}|\alpha, \beta)p(\alpha, \beta)}{p(\{x_k\})},$$

where $p(\{x_k\}|\alpha, \beta)$ is the likelihood function, $p(\alpha, \beta)$ is the prior distribution and $p(\{x_k\})$ is the bayesian evidence. To draw samples from the posterior distribution, we can use the Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm is a Markov Chain Monte Carlo (MCMC) method that generates a sequence of samples from a target distribution. The algorithm works by first proposing a distribution Q and starting point $\mathbf{x_0}$. The proposal is selected to be a distribution which is easy to sample from. Points are drawn from this proposal

distribution and accepted or rejected based on an acceptance ratio. The acceptance ratio is a function of the proposal distribution and the target distribution. In the case below, a 2 dimensional gaussian is used as the proposal to draw points (α, β). This sampling algorithm is simple and also convenient in that it works only using ratios of the target distribution so the evidence is not required. The algorithm is outlined in Algorithm 1.

Algorithm 1 Metropolis Hastings

1: $x_0 \sim \alpha$
2: $i \leftarrow 0$
3: **while** $i \geq 0$ **do**
4:      $y \sim Q(y|x_i)$ ▷ Proposal
5:      $a \leftarrow \left(\frac{P(y)Q(x_i|y)}{P(x_i)Q(y|x_i)} \right)$ ▷ MH acceptance probability
6:      $u \sim U(0, 1)$
7:      **if** $u < a$ **then**
8:          $x_{i+1} \leftarrow y$ ▷ Accept
9:      **else**
10:          $x_{i+1} \leftarrow x_i$ ▷ Reject
11:      **end if**
12:      $i \leftarrow i + 1$
13: **end while**

Note that the expression for the acceptance ratio a simplifies to the ratio of the target distribution $\frac{P(y)}{P(x_i)}$ as the proposal distribution is symmetric. This is a key feature of the Metropolis-Hastings algorithm with symmetric proposal distributions. It is assumed that there is no correlation between the parameters α and β. As such the covariance matrix for the 2D gaussian proposal distribution is diagonal.

Since the consecutive points in the chain are generated sequentially from the previous point, they are not independent. To go from the samples returned by the Metropolis-Hastings algorithm to independent samples, there are 2 corrections that are made to the chain. Specifically, the 'burn-in' period must be discarded and the samples must be thinned. The burn-in period is the number of samples that are discarded at the start of the chain before it converges towards the target distribution, ensuring that the subsequent samples are not influenced by the choice of initial values. The burn in is normally determined manually be inspecting the trace plots. The trace plot is just the name given to the plot of the random variables values (in this case α and β) as the chain progresses. The thinning factor is the number of samples that are skipped to reduce the correlation between samples. To determine the number of samples to skip, the integrated autocorrelation time (IAT) is calculated.

2.3.1 Aside: Emcee integrated autocorrelation time

To compute IAT, the `emcee.autocorr.integrate_time` function was used. This package deviates slightly from the definition of IAT. Which is defined as:

$$\tau_f = \sum_{\tau=-\infty}^{\infty} \rho_f(\tau) \tag{3}$$

where τ_f is the IAT, τ is the lag, which is the distance between elements of the chain, $\rho_f(\tau)$ is the normalised autocorrelation as a function of τ. For a finite chain, you can estimate $\rho_f(\tau)$ as

$$\hat{\rho}_f(\tau) = \frac{\hat{c}_f(\tau)}{\hat{c}_f(0)} \tag{4}$$

where

$$\hat{c}_f(\tau) = \frac{1}{N-\tau} \sum_{n=1}^{N-\tau} (f_n - \mu_f)(f_{n+\tau} - \mu_f) \tag{5}$$

and

$$\mu_f = \frac{1}{N} \sum_{n=1}^{N} f_n. \tag{6}$$

Here N is the length of the chain and f_n is the value of the chain at the nth step and μ_f is the sample mean of the chain. The integrated autocorrelation time τ_f can now be estimated as:

$$\hat{\tau}_f = \sum_{\tau=-(N-1)}^{N-1} \hat{\rho}_f(\tau) = 1 + 2 \sum_{\tau=1}^{N-1} \hat{\rho}_f(\tau) \tag{7}$$

However, the `emcee` package uses the following estimator for τ_f.

$$\hat{\tau}_f(M) = 1 + 2 \sum_{\tau=1}^{M} \hat{\rho}_f(\tau) \tag{8}$$

For some $M \ll N$. choosing the smallest value of M where $M \geq C\hat{\tau}_f(M)$ for a constant $C \approx 5$. This approach decreases the variance of the estimator at the cost of bias. A more detailed discussion can be found ¡CITE¿.

To find the optimal standard deviation for the proposal distribution, 10 short chains were run from different starting points, sampled from the uniform prior. The acceptance rate and integrated autocorrelation time were calculated. The trade off between the acceptance rate and the integrated autocorrelation time was considered to find the optimal standard deviation. The results are shown in Figure 3. To reduce the search space, the same standard deviation was used for both α and β. The optimal standard deviation was found to be 1.
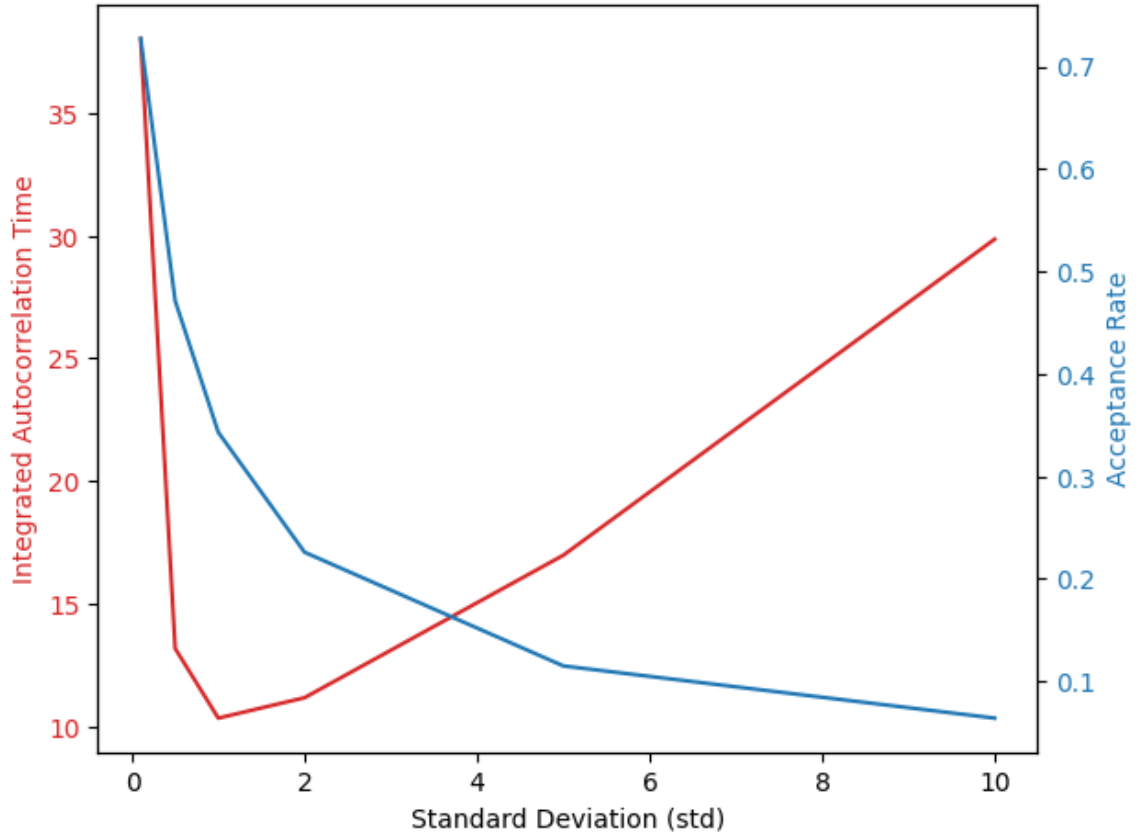
Figure 3: Acceptance rate and integrated autocorrelation time for different standard deviations

10 chains were run for 100000 samples each. The burn in period was set to 1000 from the trace plots shown in figure **??** and the thinning factor was set using the maximum auto correlation length between alpha and beta after discarding the burn in samples. The Gelman Rubin statistic [Gelman and Rubin(1992)]was calculated to be 1.0 for both α and β indicating that the chains have converged. The samples were then plotted as a joint histogram.

The samples were then plotted as a joint histogram, marginal histograms and the mean $\pm$ standard deviation of α and β were calculated. The results are shown in Figure **??**. The joint histogram shows the distribution of the samples in the α and β plane. The marginal histograms show the distribution of the samples for α and β separately. The mean $\pm$ standard deviation of α and β were calculated to be 0.0 ± 0.1 and 1.0 ± 0.1 respectively. The results show that the posterior distribution of α and β is centred around the true values of 0 and 1 respectively.

References

[Gelman and Rubin(1992)] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. doi: 10.1214/ ss/1177011136.