EXPLORING EXOPLANET ATMOSPHERES: BAYESIAN ANALYSIS FOR PARAMETER ESTIMATION

A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of

MASTER OF SCIENCE

by

ALIYYA FATHIMA RINU

to the

School of Mathematical Sciences

National Institute of Science Education and Research

Bhubaneswar

May 16, 2024

To the inhabitants of the distant world, WASP96 b

DECLARATION

I hereby declare that I am the sole author of this thesis in partial fulfillment of the requirements for a postgraduate degree from National Institute of Science Education and Research (NISER). I authorize NISER to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Student
May 16, 2024

The thesis work reported in the thesis entitled Exploring Exoplanet Atmospheres: Bayesian Analysis for parameter estimation was carried out under my supervision, in the School of Mathematical Sciences at NISER, Bhubaneswar, India.

Thesis Supervisor
Dr. Nabin Kumar Jana
School of Mathematical Sciences, NISER
May 16, 2024

Thesis co-supervisor
Dr. Jayesh M Goyal
School of Earth and Planetary Sciences, NISER
May 16, 2024

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to **Dr. Nabin Kumar Jana** and **Dr.Jayesh M Goyal** for their invaluable guidance and unwavering support throughout my project. Their expertise and encouragement have been crucial in shaping my research journey. I am also thankful to **Dr.Kaushik Majumdar** and my evaluation committee members for their valuable input and feedback.

I extend my heartfelt appreciation to all the professors who have imparted their knowledge and wisdom, contributing significantly to my academic growth. Additionally, I am deeply thankful to all my seniors for their assistance and advice, which have been invaluable during my time at NISER.

I want to acknowledge **NISER** and **DAE** for funding me through **DISHA**.

I am forever grateful to my parents, **Mrs. Junainath** and **Adv. Asraf Ali**, and my family for their love, support, and encouragement.

I am especially grateful to my brother **Naseef Mohammed** for being my pillar of strength and inspiration. His tireless belief in me has propelled me forward, and I am deeply thankful for his constant motivation and encouragement. Thank you for being there for me in every step.

A million thanks to Neha Sanal Kumar, Sumegha MT, and Anna Binoy for the laughter and good times we have shared together. Special Thanks to Shwetha Sasindran and N Devanand for their camaraderie, and to Krishnapriya KM and Meghna Manoj for making the study of Mathematics a joyous and memorable experience.

ABSTRACT

The atmospheres of Exoplanets are characterized using theoretical models with multiple parameters. Using the Bayesian framework, we retrieve the posterior probability distributions of parameters of the model based on the observed transmission spectra. This study is based on two well-known Bayesian techniques: Nested sampling and Markov Chain Monte Carlo (MCMC). We explore the parameter spaces systematically to constrain as many atmospheric characteristics as possible. Particularly in higher-dimensional scenarios, the comparative analysis between MCMC and Nested Sampling reveals their respective strengths and limitations. This statistical methodology not only contributes to our understanding of exoplanetary atmospheres but also highlights the reliability and adaptability of Bayesian analysis in retrieval problems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In real-life scenarios, dealing with data entails coping with uncertainty and randomness. Thus, accounting for the initial circumstances of random experiments becomes imperative.

Probability serves as a measure of our confidence in the truth of a proposition from a Bayesian standpoint. These concepts aid in expressing our uncertainty, wherein probabilities represent our uncertainty in parameter values after observing data. Contrasted with the classical technique, the Bayesian approach offers significant advantages. Importantly, the frequentist approach overlooks the cause of events or the underlying processes leading to them. Leveraging Bayesian statistics, we can extrapolate backward from observed data, such as transmission spectra, to make probabilistic assertions about the parameters—unknown values of the model—that generated them. This inversion process employs Bayes' rule in Bayesian statistics. The Bayesian approach treats model parameters as varying, while considering the observed data as fixed. In contrast, the frequentist approach views the unseen component—the parameters of the probability model—as fixed and the observed data as variable. By inverting the frequentist probability to obtain the "probability of the hypothesis given the actual data," the Bayes formula circumvents issues encountered by the frequentist approach.

1.0.1 Bayes Rule

Bayes' theorem constitutes a fundamental concept in Bayesian analysis. The conditional probability of A given B and I can be expressed as:

$$p(A|B,I) = \frac{p(A \cap B|I)}{p(B|I)}$$

where $p(A \cap B|I)$ denotes the joint probability of A and B given I, and $p(B|I)$ represents the probability of observing B given I. Applying the rule of conditional probability once more to $p(A \cap B|I)$ yields:

$$p(A \cap B|I) = p(B|A,I) \cdot p(A|I)$$

Substituting this expression into the first equation results in the formulation of Bayes' rule:

$$p(A|B,I) = \frac{p(B|A,I) \cdot p(A|I)}{p(B|I)}$$

1.0.2 Likelihoods

The frequentist approach to estimation, known as the method of maximum likelihood, evaluates the probability of generating a specific sample of data given the parameters in the statistical model equal to θ. Probability models incorporate parameters that, when varied, yield various system characteristics. This expression constitutes a likelihood rather than a probability since we vary θ while keeping the data constant. Maximizing the likelihood involves differentiating the function, which is simplified by first taking the logarithm of the expression. This transformation ensures that the function is maximized at the same value of θ. Although likelihood, not being a legitimate probability distribution, can be transformed into a posterior probability distribution for parameters using Bayes' rule, defining a prior distribution is necessary for this purpose.

1.0.3 Prior And Posterior Distributions

The prior probability distribution represents the pre-data uncertainty regarding the true value of a parameter. Bayes' rule facilitates the revision of initial beliefs in light of new information. Consequently, defining a starting assumption is crucial. Prior to observing a sample of a random variable X, whose distribution function depends on B, we specify our confidence level in the value of B. Following the observation of the value of the random variable A, we ascertain the posterior distribution of X. This constitutes the probability density function (or probability mass function) of B under the condition that $A = a$. Subsequently, model comparisons, hypothesis testing, and the estimation of the most probable parameter values and their uncertainties can be conducted using the posterior distribution.

1.0.4 Denominator, The Normalizing Factor

We conclude that the numerator of Bayes' rule, which comprises the likelihood multiplied by the prior, is similarly flawed due to the invalidity of the likelihood as a probability distribution. The sum or integral across all parameter values, depending on whether the parameters are discrete or continuous, never equals 1. This sum or integral's value serves to naturally normalize the numerator by division. The denominator of Bayes' rule, $p(\text{data})$, constitutes this normalizing factor:

$$p(\text{data}) = \int p(\text{data}, \theta)\, d\theta = \int p(\text{data}|\theta)\, p(\theta)\, d\theta$$

1.1 The Model

Bayesian computations, coupled with theoretical models of atmospheric processes, (Refer figure 1.1) is a powerful method for estimating parameters

A **Retrieval model** is a computational framework used to infer the atmospheric parameters of exoplanets based on observational data, such as transmission spectra or emission spectra. It utilizes statistical techniques to compare observed data with simulated spectra generated from a forward model.

On the other hand, a **Forward model** is a numerical simulation that calculates the expected spectra of an exoplanet's atmosphere based on its physical properties, such as temperature, pressure, and composition. By incorporating known atmospheric physics and chemistry, forward models predict how different atmospheric constituents interact with starlight or thermal radiation, producing observable features in the spectra.[7]

As retrieval of parameters is computationally expensive, hence we are looking for different ways in which we can optimize the probabilistic methods in retrieval to make it more computationally efficient. For this, we are using the transmission model from SANSAR [1]. It is a python-based retrieval package for unraveling the atmosphere of faraway worlds. It can retrieve both the chemical as well as physical properties of a planetary atmosphere by implying several tools to analyze planetary observed spectra. For the details on the parameters retrieved and the model refer to the documentation of the SANSAR model which is coupled with pymultinest.

1.1.1 The Data - Transmission spectra

A transmission spectrum is a plot that shows how much light from a star is absorbed by the atmosphere of an exoplanet at different wavelengths. During a planetary transit, as the exoplanet passes in front of its host star, some of the starlight travels through the exoplanet's atmosphere. The atmosphere then absorbs specific wave-

[1] *SANSAR: Transmission Spectra Model for Exoplanet atmospheres*, Verma and Goyal (manuscript under preparation)

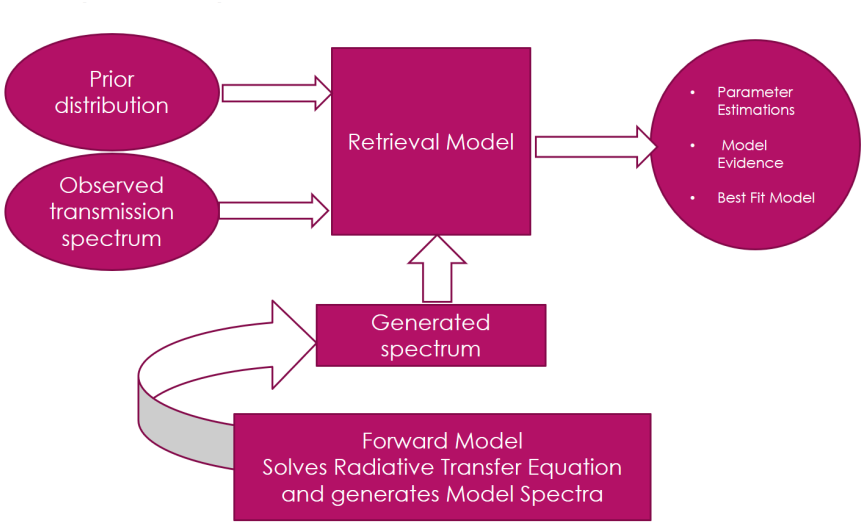

Figure 1.1: A simple representation illustrating the working of our theoretical model

lengths of light depending on the molecules present in it.

The transmission spectrum is created by comparing the intensity of light at each wavelength before and during the transit. The differences in intensity reveal which wavelengths of light were absorbed by the exoplanet's atmosphere. We have utilized the transmission spectra of WASP96 b in the wavelength range of 0.1 to 5 nm, from JWST for this project work. (Refer figure 1.2)

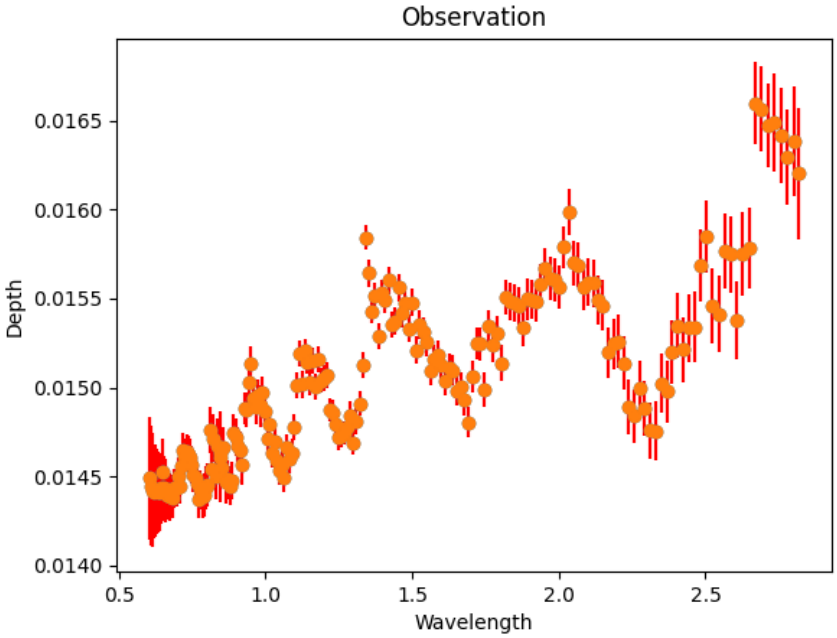

Figure 1.2: Observed Transmission spectra of WASP96 b planet, wavelength range (0.1 - 2.5 nm)

Chapter 2

Basics of Bayesian Inference

2.1 Elements of Bayesian Decision Theory

Now, that we have computed our posterior distribution, one may report posterior estimates also. like posterior mean, posterior median, and posterior mode also known as Maximum A Posteriori. No estimates are exactly the true value of the unknown parameter (Except in rare cases). Therefore each has to pay a penalty for using that certain estimate for estimating the unknown so that we can choose a function that corresponds to the minimum penalty. Decision theory deals with situations in which one has to make choices among given alternatives. First, let us go through the process of classical decision theory.

Θ - The Parameter Set: This is the set of all possible parameter values, denoted as Θ, where $\Theta \subseteq \mathbb{R}^p$.

A general estimator is an educated guess about our true unknown parameter θ, which is essentially a function of the random variables $X_1, X_2, \ldots, X_n$, represented as $T(X_1, X_2, \ldots, X_n)$.

Given a particular dataset $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, then T is known as the point estimate of θ if $T(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \in A$, where $A \subseteq \mathbb{R}^p$. Here, A is known as the Action space. Note that the action space and parameter space need not always be the same.

Quantifying the precision of these estimators is done using loss functions and risk functions. A risk function of a decision procedure (action) $\boldsymbol{a}$ is the average loss incurred by the action $\boldsymbol{a}$, where the loss function $L(\theta, \boldsymbol{a}) = W(|\theta - a|)$ for some

function W with $W(0) = 0$, and $W(x)$ is a monotone non-decreasing function of the difference between θ and $\boldsymbol{a}$.

Some examples of loss functions include:

$$\text{Squared error loss} = (\boldsymbol{a}\text{-}\theta)^2$$

$$\text{Absolute Error Loss} = (|\boldsymbol{a}\text{-}\theta|)$$

Writing down a loss function gives us an idea about the consequences of choosing a decision rule. As our samples differ, the best decision also varies, so we take the average over the data samples fixing a particular value of θ. Consider a function $R(\theta, \boldsymbol{a})$ defined as follows:

$$R(\theta, \boldsymbol{a}) = E_\theta(L(\theta, \delta(X))$$

known as the risk function. In the classical approach, in an estimation problem, if δ is an estimate of $\tau(\theta)$, then $E_\theta((\tau(\theta) - (\delta(X))^2)$ is the MSE, and in a hypothesis testing problem, where H_0 is $\theta = \theta_0$ and H_1 is otherwise, $A = \boldsymbol{a_j}$, $j = 0, 1$, where a_j means the decision to accept H_j, $L(\theta, a_j) = 0$ if θ satisfies H_j and $L(\theta, a_j) = 1$ otherwise.

If $\boldsymbol{I}(x)$ is the indicator of a rejection region for H_0, then the corresponding $\boldsymbol{I}(x)$ is equal to a_j if $\boldsymbol{I}(x) = j$, $j = 0, 1$. But in Bayesian analysis, θ has a distribution. So before we have our data, we are equipped with a prior probability, $\pi(\theta)$, with which we obtain the risk function for a particular prior distribution and a particular decision function known as the pre-posterior risk:

$$R(\pi, \delta) = \int_\Theta \pi(\theta) R(\theta, \delta(X)) d(\theta)$$

After gaining data, the probability distribution for θ is now updated to the posterior distribution. Then, using this posterior distribution in place of the prior, we obtain the posterior risk function, which is a conditional expectation over Θ.

The following **Theorem** explains these two problems in Bayesian decision theory.

(a) For any δ, $R(\pi, \delta) = E(\psi(X, \delta(X)))$.

(b) Suppose $\boldsymbol{a(x)}$ minimizes $\psi(x, a(x))$, i.e., $\psi(x, a(x)) = \inf_a \psi((x, a))$, then $a(x)$ minimizes $R(\pi, \delta)$.

Proof:

(a) $E_X\left(L(\theta, \delta(x) \mid x) = \psi(X, a(X))\right.$

$E_\Theta\left(E_X\left(L(\theta, \delta(x) \mid x)\right)\right. = \int_\Theta \pi(\theta) R(\theta, \delta(X)) d(\theta) = R(\pi, \delta)$

(b) Let $\boldsymbol{a(x)}$, as defined in the theorem, be denoted by δ_0. Then, by part (a) and definition of $\boldsymbol{a(x)}$,

$R(\pi, \delta_0) = E(\psi(X, \delta_0)) \leq E(\psi(X, \delta(X))) = R(\pi, \delta)$

2.1.1 Testing in Bayesian Analysis

Let's examine Bayesian hypothesis testing problems:

Case 1: H_0: $\theta \in \Theta_0$ and the alternative H_1: $\theta \notin \Theta_1$, where Θ_i, $i = 0, 1$, are the parameter sets.

The ratio of the posteriors of null to the alternative, $P(\Theta_0|x_i)/P(\Theta_1|x_i)$, is a crucial factor in this comparison. However, there are some issues with this approach:

(i) In the case of an improper prior (a prior function that cannot be normalized), the prior probabilities may be undefined. This is particularly true when dealing with hypotheses like $H_0 : \theta < \theta_0$ and $H_1 : \theta \geq \theta_0$.

(ii) Careful consideration is required when choosing the prior for Θ_i because, in classical approaches, we only reject the null hypothesis if we have substantial predetermined evidence, denoted as α. Consequently, assigning arbitrary prior probabilities

to Θ_i can be considered unfair.

These issues highlight the importance of making informed decisions about prior probabilities and being aware of the implications of improper priors in Bayesian hypothesis testing. Suppose Π_0 is the prior probability for Θ_0, then $1 - \Pi_0$ is the prior for Θ_1.

The probability distribution for θ, $\Pi(\theta)$, is given by:

$$\Pi(\theta) = \Pi_0 g_0(\theta) I(\theta \in \Theta_0) + (1 - \Pi_0) g_1(\theta) I(\theta \in \Theta_1)$$

where g_i is the probability density function of θ under Θ_i.

The **Bayes Factor** of H_0 relative to H_1 is defined as:

$$BF_{01} = \frac{\int_{\Theta_0} f(X|\theta) g_0(\theta)}{\int_{\Theta_1} f(X|\theta) g_1(\theta)} \tag{2.1}$$

This is equal to the ratio of **posterior odds** to **prior odds**, as follows:

Posterior Odds = $\frac{P(\Theta_0|X)}{P(\Theta_1|X)}$

Prior Odds = $\frac{\Pi_0}{1-\Pi_0}$

The smaller the value of BF_{01}, the stronger the evidence against H_0.

The posterior density function for H_i, where $i = 0, 1$, is then:

$$\Pi(H_i) = \frac{f(X|\theta)\Pi_i}{m_{\Pi_i}}$$

This implies the posterior probability of θ given data X is:

$$\frac{\Pi_i}{m_{\Pi(X)}} \int_{\Theta_i} f(x|\theta) g_i(\theta) d\theta$$

where $\Pi_1 = 1 - \Pi_0$, and m_Π is the marginal density.

Given this, we can write:

Posterior Odds = BF_{01} x Prior Odds

2.1.2 Testing a point null hypothesis

The hypothesis test of the form $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, where $\theta, \theta_0 \in$ parameter set Θ. In such tests, we encounter two types of null hypotheses:

1. **Point Null Hypothesis**: This occurs when we specify a precise value for θ_0.

2. **Interval Null Hypothesis**: In real-life situations, θ might belong to small and unspecified intervals. In such cases, it's challenging to define g_0. Therefore, we consider these situations as point null hypotheses. However, when we can specify a small yet specific interval, it's better to treat it as an interval null hypothesis.

In a point null hypothesis case, $BF_{01} = \frac{f(x|\theta_0)}{m_1(x)}$, which can be derived as follows:

The posterior density of Θ_0 is:

$$\Pi(\theta_0|X) = \frac{f(X|\theta_0)\Pi_0}{m(x)} \tag{2.2}$$

The posterior odds ratio is:

$$= \frac{\Pi_0 f(X|\theta_0)}{m_1(x)} = BF_{01} \times \left(\frac{\Pi_0}{1-\Pi_0}\right)$$

Thus, the equation for the Bayes factor above holds.

Credible Intervals: For $0 < \alpha < 1$, a $100(1-\alpha)\%$ credible set for θ is a subset $C \subset \Theta$ such that $P(C|X = x) = 1 - \alpha$. Credible intervals are particularly useful for testing sharp null hypotheses (explained in Chapter 6).

Definition: Suppose the posterior density for θ is unimodal. Then the Highest Posterior Density (HPD) interval for θ is the interval $C = \{\theta : P(\theta|X = x) \geq k\}$, where k is chosen such that $P(C|X = x) = 1 - \alpha$.

2.1.3 Prior vs. Posterior Predictive Distribution of a New Observation

For a fixed value of θ, data follows $P(X|\theta)$, which is known to us. However, since the parameter θ is unknown and has some uncertainty associated with it due to a prior probability distribution, we should consider averaging over the possible values of the parameter.

The probability for a new data point before observing a sample is thus:

$$P(x_{new}) = \int_{\Theta} P(X|\theta)P(\Theta)d\theta$$

After taking the sample, we have a better representation of the uncertainty in θ via our posterior $p(\theta|x)$. So, the posterior predictive distribution for a new data point x_{new} is:

$$P(x_{new}|\theta) = \int_{\Theta} P(x_{new}|\theta)P(\Theta|x)d\theta$$

This reflects how we would expect new data to behave or vary. If the observed data closely conforms to this pattern, it suggests that our model and prior selections have been effective.

The great thing about Bayesian statistics is that it allows us to include expert knowledge or opinions when making decisions or analyzing data. This can be super helpful, especially in situations where we don't have a lot of data to work with or when dealing with complex problems.

When we use objective Bayesian analysis, where we use objective priors, it doesn't clash with the subjective approach where we consider personal beliefs. It can even serve as a sort of benchmark to see how important our prior information is in our decision-making.

Chapter 3

The Sampling Techniques

3.1 Approximation Of Integrals By Sampling

In cases where analytical, closed-form solutions are not readily available, we often need to resort to numerical integration methods or approximations. Unfortunately, these approaches can be less practical, particularly when dealing with complex or large datasets or higher dimensions of parameter space.

This is where sampling methods come into play. **Sampling methods** are probabilistic techniques that are particularly useful when we have a sufficiently large number of data points to work with. Suppose we're interested in estimating the posterior expectation. In such a scenario, we can sample data points from the posterior distribution and then calculate the sample means from these samples.

3.2 Some Basic Methods For Generating Random Samples- Monte Carlo Integration Methods

The key insight here is that the Law of Large Numbers (LLN) comes into play. It guarantees that as we collect more and more samples from the posterior distribution, the sample means will converge to the true expectation with high probability. However, it's important to note that this convergence relies on the assumption that the posterior distribution is in a standard form which is amenable to sampling techniques. Along with Bayesian inferences, marginal likelihood calculation, elimination of nuisance parameters, and computing Bayes factor which we discussed earlier in the previous sections, involves integration problems. however, in several examples, this method

exposes its inefficiency.

3.2.1 Indepentant sampling techniques

Transformation Methods

In many situations, the posterior distribution may not be in a standard form making it difficult to sample, in that case, a smart way of sampling from a *target density* f could be to sample from another distribution function known as *Instrumental distribution* g which is a known and easy to handle distribution. such methods are collectively known as **Transformation methods**, the Accept-Reject method is one example. We thus turn to another class of methods that only require us to know the functional form of the density f of interest up to a multiplicative constant; no deep analytical study of f is necessary. However, this method demands a wise choice of the instrumental function g and the scaling factor M satisfying the condition, $f(x) \leq Mg(x)$. Another transformation method known as Inverse transform is discussed later in this section.

Accept-Reject Method

Algorithm:

1 - Sample $X \sim g(X)$ and $U \sim \mathcal{U}[0,1]$

2 - Accept the proposal if $U \leq \frac{f(x)}{Mg(x)}$, else reject it.

3- Repeat step 1

Theorem 1. *This algorithm ensures that the accepted samples follow the target distribution.*

Proof.

$$P(Y \leq y) = P(X \leq y \mid U \leq \frac{f(X)}{Mg(X)}) = \frac{P(X \leq y, U \leq \frac{f(X)}{Mg(X)})}{P(U \leq \frac{f(X)}{Mg(X)})}$$

now writing out the probability integral yields

$$P(Y \leq y) = \frac{\int_{-\infty}^{y} \int_{0}^{\frac{f(x)}{Mg(x)}} du\, g(x)\, dx}{\int \int_{0}^{\frac{f(x)}{Mg(x)}} du\, g(x)\, dx} = \frac{\frac{1}{M} \int_{-\infty}^{y} f(x)\, dx}{\frac{1}{M} \int_{-\infty}^{\infty} f(x)\, dx} = \int_{-\infty}^{y} f(x)\, dx$$

□

Inverse Transform

We'll explore this method by working through an example to gain a better understanding.

Suppose we have a Uniform distribution and we want an exponential distribution to be generated using the samples from the uniform distribution.

This method involves somehow transforming the sampled values to a value in the exponential distribution. so the goal is to find such a transformation function $T(U) = X$.

The CDF of the exponential distribution is $1 - exp(\lambda x)$ for $x \geq 0$ and 0 otherwise.

$$P(X \leq x) = P(T(u) \leq x) = P(u \leq T^{-1}(x) = T^{-1}(x)$$
$$F_x(X) = T^{-1}(x)$$
$$F_x^{-1}(X) = T(x)$$

inverse of the CDF function is the required transformation function.

$$x = \frac{ln(u)}{\lambda}$$

Note that this method is of very less practical importance as it is necessary to know the CDF function and also the inverse of the CDF function to perform sampling.

Monte Carlo Importance Sampling

To better understand this, let us look into one example.

$$E(g(\theta)|x) = \frac{\int g(\theta) f(x|\theta)\pi(\theta)d\theta}{\int f(x|\theta)\pi(\theta)d\theta}$$

The expectation of $g(\theta)$ is in this form, which means we sample from the weight of $g(\theta)$ and further work out the procedure. but the issue is that as our sample size increases the samples get accumulated near the MLE of the distribution function, which ignores the contribution from the prior as the sample size grows.in order to minimize this error in sampling if we sample from the prior distribution of θ, as the tail portion of the distribution is not as heavy as the central portion much more sampling will be required thus making the convergence slower and hence the larger error in approximation (for a fixed sample size). This suggests that we sample from the posterior distribution itself.

1 - we can do the Monte Carlo sampling without knowing the probability distribution.

2 - once we assume an importance distribution function, we can sample from this known distribution function given that we know the ratio of these two.

3 - another use is that we can use this method, even if the posterior density function is not normalized.)

4- flexibility in choosing g is also a notable advantage, as long as $supp(f) \subset supp(g)$,g could be any distribution function that is easy to handle. The choice of g that minimizes the variance of the estimator(expectation of the sample means) is, however, choosing an importance distribution is often the difficulty here. an arbitrary distribution function as importance distribution has a very high variance. the following theorem states some sufficient conditions for g however, In problems involving higher dimensions, a combination of numerical methods, Laplace approximations, and Monte Carlo sampling are used to find the importance function. Using instrumental distri-

butions g with lighter tails and unbounded ratios (f/g) in importance sampling can lead to estimators with infinite variances for many functions h. For example, consider estimating the expected value of a heavy-tailed Cauchy distribution:

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Instrumental distribution g:

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The ratio $\frac{f(x)}{g(x)}$ is unbounded as x approaches infinity because g assigns exponentially decreasing probabilities to extreme values, while f assigns substantial probability mass to them.

In importance sampling, if you sample from $g(x)$ and encounter extreme values, the ratio $\frac{f(x)}{g(x)}$ can be extremely large, leading to very large weights in the estimator:

$$\hat{E}_g[h(x)] = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} h(x_i)$$

Even a single sample with an extremely large weight can dominate the sum, causing the estimator to be highly sensitive to outliers. This results in an estimator with a very large, possibly infinite, variance, making it unreliable. To address this issue, choose an instrumental distribution g(x) with tails similar to or heavier than the target distribution when estimating heavy-tailed distributions. This ensures a stable and reliable estimator. If we are able to sample from a probability distribution independently, then we can approximate an integral (expectation). but independent sampling is not quite an easy procedure.

Rejection Sampling

This is another Monte Carlo independent sampling method, where the pdf need not be normalized. However, rejection sampling is inefficient, due to a vast majority of

sampled points being rejected, and the inefficiency due to rejections worsens as the number of dimensions increases. After doing the iterations of rejection and acceptance (sample $U(\text{range}(X))$, and y_i from $U(0,1)$, then find corresponding $\text{pdf}(x_i)$, and if $y_i \leq \text{pdf}(x_i)$, accept x_i and otherwise reject it), following this until convergence we can compute the integral. Count of the accepted points gives us the approximate area. Rejection sampling is considered a last resort for independent sampling. Unlike rejection sampling, **Inverse Transform sampling** is more efficient.

3.2.2 Dependent Sampling Techniques

So far we've explored methods where samples are drawn independently from the same distribution, ensuring speed and efficiency in covering various distribution areas than dependent sampling techniques because of the very fact that it does not account for the correlation between samples. However, envision a scenario in higher dimensions, where techniques like rejection sampling propose and potentially reject samples repeatedly, leading to a considerable number of iterations required for convergence. In such cases, the quest for an optimal sampling method becomes natural.

What if we could leverage previously sampled values to guide future sampling attempts? This notion inspires the exploration of a more efficient technique—one that enhances the probability of the algorithm searching for samples in areas highly likely in the target distribution. Enter the Markov chain, a special stochastic process whose properties significantly aid in simplifying the sampling procedure. Despite its advantages, Markov chain sampling isn't without pitfalls, prompting the crucial question: Is this algorithm truly optimal?

3.3 Markov chains

Definition 1: A stochastic process constitutes a set of state space random variables $\{X_t\}_{t\in T}$, where T can be represented as time, either discrete or continuous.

State space (S): This refers to the set of all possible values taken by $\{X_t\}$.

The initial distribution denotes the Probability Mass Function (PMF) of X_0, signifying the probability distribution of the initial state.

A Markov chain is characterized by the following conditions:

$$P(X_{n+1} = j|X_n = i, X_{n-1} = i-1, \ldots, X_0 = i) = P(X_{n+1} = j|X_n = i)$$

where P_{ij} is the **transition probability** from state i to state j, in a single step, which is the ij th component of the probability matrix $[P_{ij}]$. It's worth noting that Markov chains can also be defined using transition kernels, which is useful for Markov processes. we will focus solely on Markov chains, omitting transition kernels for the sake of clarity.

In a time-homogeneous Markov chain:

$$P(X_{n+1} = j|X_n = i) = P(X_n = j|X_{n-1} = i) = P_{ij}$$

which is true for all $n \geq 1$.

The probabilistic behavior of a Markov chain is determined by its initial distribution and the one-step transition probability matrix. Additionally, the Markov chain possesses a significant property known as the **Markov Property**.

Suppose $\{X_n\}$ is a Markov chain with an initial distribution (probability distribution for X_0 and a Transition Probability Matrix (TPM)), then, Conditioning on $\{X_n = i\}$, $\{X_{n+m}\}$ is also a Markov chain with an initial distribution δ_i and the same TPM as the former chain. The statement about determining a Markov chain follows from this Markov property and is easily derivable.

The **Chapman-Kolmogorov equation** explicitly provides the formula for this n-step transition from i to j as follows:

$$P_{ij}^{(m+n)} = \sum_k (P_{ik}^{(m)} \cdot P_{kj}^{(n)})$$

This result can be derived through the following steps:

$$P_{ij}^{(m+n)} = \sum_k P(X_{m+n} = j | X_m = k, X_0 = i) \cdot P(X_m = k | X_0 = i)$$

using Markov property,

$$\sum_k P_(X_{m+n} = j | X_m = k, X_0 = i) = \sum_k P_(X_{m+n} = j | X_m = k)$$

and substituting this back the CK equation follows.

One may also note that, *n-step Transition Probability* is defined by:
$(P_{ij})^{(n)} = P(X_n = i | X_0 = j)$, signifying the process of reaching state i precisely in n steps. Because, As $P^{(1)} = P$. By induction on n, it can be shown that $P^{(n)} = P^n$.

3.3.1 Accessibility and Communication

Definition: A state j is said to be accessible from a state i if there exists $n \geq 0$ such that $P_{ij}^{(n)} > 0$ and $P_{ij}^{(0)} = \delta_{ij}$. that means there is a non zero probability, starting from state i the chain visits state j in some n number of steps. **Definition:** If there exist $m \geq 0$ and $n \geq 0$ such that $P_{ij}^{(m)} > 0$ and $P_{ji}^{(n)} > 0$, then the states in a Markov chain are said to be **communicating**.

In other words, if a state is accessible from another state, it doesn't imply that both states communicate, but the converse is true. So, all states that are communicating with each other can be now put into one 'basket' called a **class**. more precisely, This

property communication defines an equivalence relation in the state space partitioning the state space. Symmetry and reflexivity naturally follow from the definition. It is also transitive; suppose $P_{ik}^{(n)} > 0$ and $P_{ki}^{(m)} > 0$. Additionally, let $P_{kj}^{(l)} > 0$ and $P_{jk}^{(q)} > 0$. Then,

$$P_{ij}^{(n+l)} = \sum_{i' \in S} P_{ii'}^{(n)} P_{i'j}^{(l)} \geq P_{ik}^{(n)} P_{kj}^{(l)} > 0,$$

thus proving that states i and j communicate.

A communicating class C is considered closed if for any state $i \in C$, whenever j is accessible from i, then $j \in C$.

The following definition is to be kept in mind as it is important in the following sections.

Definition: A Markov chain is **irreducible** if there exists a single communicating class, i.e., all states communicate with each other. Seems like a nice property isn't it? Now what follows are some more definitions, theorems, and results that simplify our understanding and lead us to explore further definitions and theorems that streamline sampling techniques.

Another significant quantity in Markov chains is the **hitting time**. Suppose A is a subset of S, then $T^{(A)} = \inf\{n \geq 0 \,|\, X_n \in A\}$.

The probability that starting from state i, the chain hits A is denoted by $P(T^{(A)} < \infty)$ and represented as h_i^A. If the subset A is closed, implying that once the Markov Chain (MC) enters A, it becomes absorbed, this probability is referred to as the **absorption probability**. here, let's take a moment to appreciate the usefulness of this quantity. It gives us the probability of getting stuck at a particular state i forever given that the chain started from any state j To compute this, the following theorem provides a method:**Theorem:** If $h_i^{(A)} = 1$ for $i \in A$ and $h_i^{(A)} = \sum_{j \in S}(P_{ij} \cdot h_j^{(A)})$ for $i \notin A$,

the minimum non-negative solution to this system of linear equations yields the vector

$h^{(A)} = (h_i^{(A)}, i \in S)$. where i th component of this vector is the probability of hitting A starting from state i.

Now, as T^A is a random value, the expected value of T^A, denoted as k_i^A, is given by:

$$k_i^A = \sum_{n \geq 1} (n \cdot P_i(T^A = n) + \infty \cdot P_i(T^A = \infty))$$

This represents the expected time taken by the Markov Chain to hit A starting from state i.

Similar to the previous theorem, the vector of mean hitting times K^A is the minimal non-negative solution of the system of linear equations:

$$K_i^A = 0 \text{ for } i \in A \text{ and } K_i^A = 1 + \sum_{j \notin A} P_{ij} K_j^A \text{ for } i \notin A$$

These notions lead us to another property of the Markov chain as stated below: The Strong Markov Property states that given a non-random time m, the process from that time onward behaves like a Markov chain. However, the **Strong Markov Property (SMP)** asserts that even if the time m becomes random, the conclusion remains the same as the Markov property. Here, m is referred to as a stopping time, which is essentially a random variable T, which can be finite or infinite and satisfies the condition that the event T = n depends only on $X_0, X_1, ...X_n$.

The SMP implies that when the process reaches a stopping time, the future behavior of the process beyond that point, conditioned on the stopping time, behaves like a fresh, independent Markov chain.

This property is powerful as it allows studying the process at the moment it reaches certain states or conditions, treating that moment as a fresh start for the Markov process.

3.3.2 Reccurance and Transience

Here's another intriguing set of definitions and results:

Definition 1: Passage Time: $T_i = \inf\{n \geq 1 \,|\, X_n = i\}$. One should take a while to notice the distinction between hitting time and passage time.

Definition 2: Recurrent State: A recurrent state is one where $\mathbb{P}_i(T_i < \infty) = 1$.

Definition 3: Transient State: A transient state is one where $\mathbb{P}_i(T_i < \infty) < 1$.

Now, suppose the Markov Chain (MC) is recurrent. This intuitively means that the chain visits state i in finite time with probability 1, i.e., $\mathbb{P}(X_n = i \text{ for infinitely many } n) = 1$. This conclusion is derived from using the Strong Markov Property: the chain keeps revisiting state i.

Conversely, if $\mathbb{P}(X_n = i \text{ for infinitely many } n) = 1$, then the state i is recurrent, and it is a straightforward observation.

Theorem: A state is recurrent if and only if $\sum_{n=0}^{\infty} P_{ii}^n = \infty$, and a state is transient if and only if $\sum_{n=0}^{\infty} P_{ii}^n < \infty$.

Proof: suppose $\sum_{n=0}^{\infty} P_{ii}^n = \infty$ Given the state is recurrent, let V_i be the number of visits to state i, and f_i^k be the distribution of V_i. Then, $P_i(V_i > k) = f_i^k$, and $f_i = P_i(T_i < \infty)$. From the definition of a recurrent state, $f_i = 1$, implying $P_i(V_i = \infty) = \lim_{k\to\infty} P_i(V_i > k) = 1$. We know that the expected number of visits is infinity, i.e., $\infty = E_i(V_i) = \sum_{n=0}^{\infty} P_i(X_n = i) = \sum_{n=0}^{\infty} P_{ii}^n$.

Converse and Transient Condition:

To show the converse, suppose $\sum_{n=0}^{\infty} P_{ii}^n = \infty$. Assuming state i is transient ($f_i < 1$), leads to a contradiction since $\sum_{n=0}^{\infty} P_{ii}^n = \sum_{p=0}^{\infty} \frac{1}{1-f_i} < \infty$. Thus, state i must be recurrent.

The equivalent condition for a transient state can be proved similarly.

Theorem: Assumption of theorem:

1. X_n is a Markov Chain.

2. S is a finite state space.

To show that at least one state must be recurrent, i.e., if there exists some $i_0 \in S$ such that $\sum_{j=1}^{\infty} \delta_{i_0}(X_j) = \infty$, then we are done.

Note that the total number of steps $n = \sum_{i=1}^{N} \{\delta_i(X_j)$.

Taking the limit on both sides, we obtain the desired result.

Theorem: Assumptions of the theorem:

1. C is a communicating class

2. i is a transient state

To show that all states in C are transient.

Proof: Let $i \neq j$ and $i, j \in C$, implying the existence of $n, m \geq 0$ such that $P_{ij}^n > 0$ and $P_{ji}^m > 0$ from the definition of a communicating class.

For any $r \geq 0$:

$$P_{ii}^{(n+r+m)} \geq P_{ij}^{(n)} \cdot P_{jj}^{(r)} \cdot P_{ji}^{(m)}.$$

$$\sum_{r=0}^{\infty} P_{jj}^r \leq \frac{1}{P_{ij}^n P_{ji}^m} \cdot \sum_{r=0}^{\infty} P_{ii}^{(n+r+m)} < \infty.$$

This implies that state j is transient, and consequently, all states in the communicating class C are transient.

3.3.3 Stationary Distributions

Definition: Stationary Measure in Markov Chains

An invariant (stationary) measure for a Markov Chain, denoted as π, is represented by a vector $(\pi_i : i \in S)$ with non-negative entries and the Transition Probability Matrix P. The components of this vector are given by $\pi_j = \sum_{i \in S} \pi_i \cdot p_{ij}$ for all j.

Definition: Stationary Distribution

A stationary measure is referred to as a stationary distribution if $\sum_{i \in S} \pi_i = 1$. Also, $\pi \cdot P = \pi$. in matrix form. This means that even after n time, the distribution of states remains unchanged under transition. This fundamental property of Markov chains is crucial in Metropolis algorithms, which we'll soon explore. so the task now is to design a Markov chain with a stationary distribution. the 'why' and 'how' questions to this statement will be clearly explained in the following sections.

Insight into Stationary Distribution Existence

The existence and uniqueness of a stationary distribution in a Markov Chain are pivotal for MCMC algorithms used in sampling. The question arises: how do we determine if such a stationary distribution exists? Let's delve into some relevant theorems and remarks shedding light on this:
These theorems and remarks provide valuable insights into the fundamental concepts of stationary distributions in Markov Chains, essential for understanding their role in MCMC algorithms and probabilistic sampling methods.

Theorem: Assumptions of the theorem

1. State space is finite.

2. $\lim_{n \to \infty} P_{ij}^{(n)} = \pi_j$ as n tends to infinity for some state i, for all $j \in S$.

- To show that $(\pi_j : j \in S)$ is a stationary distribution.

The proof involves establishing the following:

$$\lim_{n \to \infty} P_{ij}^{n} = \pi_j = \lim_{n \to \infty} \sum_{k \in S} P_{ik}^{(n-1)} P_{kj}.$$

Taking the limit inside the summation is possible since the state space S is finite, which underscores the significance of assuming a finite state space.

Thus, the final equation, $\sum_{k \in S} \pi_k P_{kj}$, implies that the limiting distribution, if it exists, constitutes a stationary distribution. For a Markov Chain with a finite state space, at least one stationary distribution exists. and there are two possibilities, either the stationary distribution is unique, or there exists infinitely many. this is because, if there exist two, then any convex linear combinations of these two stationary distributions is again a stationary distribution. so there can't be finitely many stationary distributions. However, in the case of a Markov Chain with an infinite state space, there might be instances where a stationary distribution does not exist. It's important to note that while a stationary measure always exists, a stationary distribution may not be in the scenario of an infinite state space. This limitation arises due to the constraint that the summation of the row vector of the invariant (stationary) measure must equal 1, which could be violated with an infinite sum.

Now, the significant question arises: how do we compute the stationary distribution given the various cases of existence or non-existence discussed now? One approach mentioned earlier involves solving $\pi \cdot P(n) = \pi$, but a practical challenge lies in the computational complexity of calculating the matrix P^n.

So, how do we address this practical challenge? The following theorems and remarks offer some useful strategies. As these strategies are pertinent for practical implementations, the theorems are stated without providing proof.
Before delving into the theorems, let's acquaint ourselves with a few terms:

For a fixed state k, the expected time spent in each state $i \in S$ is denoted as $\gamma_i^k = E_k \left(\sum_{n=0}^{T_k - 1} \delta_i(X_n) \right)$. Notably, for an irreducible recurrent Markov Chain (MC), $\gamma_k^k = 1$, with the crucial condition that $0 < \gamma_i^k < \infty$ for all $i \in S$.

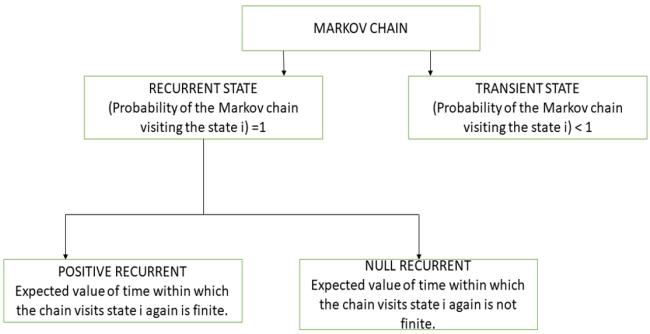

Figure 3.1: Classification of the States

Theorem: Assumptions of the theorem:

1. The MC is irreducible.

2. For a fixed state k, $\pi_k = 1$.

Then, $\pi \geq \gamma^k$

Remark: The equality holds if the chain is irreducible and recurrent.

Corollary: Assumptions:

1. The MC is irreducible and recurrent.

If π and π' are two non-zero stationary measures, then $\pi = c\pi'$, where c is a constant. Recurrent states can be further classified into two categories: positive recurrent if $E_i(T_i) < \infty$, and null recurrent otherwise. Why this distinction? Recurrence implies returning in a finite time, but note that $P_i(T_i < \infty) = 1$ doesn't necessarily mean the expected time should also be finite. Positive recurrence signifies a finite expected time, whereas null recurrence implies that it might take an extremely long time, albeit with a high probability. The following theorem emphasizes the significance of these classifications:

Theorem: Assumptions of theorem:

- The Markov chain is irreducible.

If any state is positive recurrent, then every state is positive recurrent, and there exists a stationary distribution. Moreover, $\pi_i = \frac{1}{m_i}$. This theorem essentially tells us that for an irreducible Markov Chain, if some states are positively recurrent, then a unique stationary distribution exists, precisely equal to $\frac{1}{m_i}$. where m_i is the expected stopping time.
So, in summary,

- An asymmetric simple random walk is transient, so,it cannot possess a stationary distribution.
- For a finite-state Markov chain, there always exists at least one stationary distribution.
- For a given finite-state irreducible Markov chain, the theorem guarantees a unique stationary distribution and ensures all states are positively recurrent.

This emphasizes the importance of a finite state space, as it guarantees the existence of a stationary distribution and facilitates the use of the theorem stated above.

3.4 Markov Chain Monte Carlo

Having gained a comprehensive understanding of the foundational aspects of Markov chains, let us now focus on utilizing this knowledge in designing specific Markov chains tailored for sampling techniques. Exploring practical applications and the underlying mathematical framework, we aim to understand these chains' role in sampling methods like MCMC and Gibbs sampling.

3.4.1 Metropolis-Hastings

It is a more general Algorithm.

The Metropolis-Hastings algorithm initiates with the target density f. Following this, a conditional density $q(Y|x)$ is selected.

Algorithm: Given $x(t)$:

1. Generate Y_t from $q(Y|x(t))$.
2. Take $x(t+1) = \begin{cases} Y_t & \text{with probability } \alpha(x(t), Y_t) \\ x(t) & \text{with probability } 1 - \alpha(x(t), Y_t) \end{cases}$

where

$$\alpha(x, y) = \min\left(1, \frac{f(Y)q(x|Y)}{f(x)q(Y|x)}\right).$$

The function $p(x, y) = \min\left(\frac{f(Y)q(x|Y)}{f(x)q(Y|x)}, 1\right)$ defines the transition probability from state x to state y in the Markov chain.

The distribution q is referred to as the instrumental (or proposal) distribution. Given an (unnormalized) target distribution that is difficult to sample from and a simple proposal distribution, the Metropolis-Hastings (MH) algorithm generates a sequence of iterates whose distribution approximates the target with sufficiently large iterations. i.e., after a number of iterations, if the chain reaches its stationary distribution, and if the target distribution is this stationary distribution, then as we have already seen, the chain is going to stay in this distribution, and the algorithm samples every time after this from the stationary distribution which is now the target distribution.

First, the **detailed balance criterion** of a Markov Chain with transition distribution $T\left(x_{k+1} \mid x_k\right)$ states that a stationary distribution $P(x)$ satisfies

$$T\left(x_{k+1} \mid x_k\right) P\left(x_k\right) = T\left(x_k \mid x_{k+1}\right) P\left(x_{k+1}\right).$$

This condition is necessary for a random walk to asymptotically reach a stationary distribution as we already have seen in the previous section. So at each iterations, we need to ensure that the detailed balanced condition is satisfied. additionally, irreducibility must also be satisfied for the walk to truly converge to the stationary distribution.

So, after drawing a sample, there are two possibilities: Let Case A denote the case where $x_{k+1} \neq x_k$ and let Case B indicate $x_{k+1} = x_k$. Case A transition happens with probability:

$$T\left(x_{k+1} \mid x_k\right) = \alpha\left(x_{k+1}, x_k\right) Q\left(x_{k+1} \mid x_k\right).$$

Case B transition:

$$T\left(x_k \mid x_k\right) = \alpha\left(x_k, x_k\right) Q\left(x_k \mid x_k\right) + \int_{x'} Q\left(x' \mid x_k\right)\left(1 - \alpha\left(x', x_k\right)\right) dx'$$

The first term is the probability of luckily being accepted back on the same point, and the second term is the probability of a rejection. Here, we have $x_{k+1} = x_k$, and so it is trivially satisfied.

$T\left(x_{k+1} \mid x_k\right) P\left(x_k\right) = T\left(x_k \mid x_{k+1}\right) P\left(x_{x+1}\right)$

Another perspective on the Metropolis-Hastings (MH) algorithm is to consider how we should design the acceptance probability function, $\alpha(x', x)$, to satisfy the detailed balance. for Case A. Expanding this, we need α to satisfy:

$$\begin{aligned}
&\alpha(x_{k+1}, x_k) Q(x_{k+1} \mid x_k) P(x_k) \\
&= T(x_{k+1} \mid x_k) P(x_k) \\
&= T(x_k \mid x_{k+1}) P(x_{k+1}) \\
&= \alpha(x_k, x_{k+1}) Q(x_k \mid x_{k+1}) P(x_{k+1})
\end{aligned}$$

In essence, $\frac{\alpha(x_{k+1},x_k)}{\alpha(x_k,x_{k+1})} = \frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)}$.

But what does the minimality condition mean, intuitively?

Let's consider two cases:

case 1:

$\frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)} < 1$, then

- $\alpha(x_{k+1}, x_k) = \frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)}$
- $\alpha(x_k, x_{k+1}) = 1$

so that we accept the decision of moving from x_{k+1} to x_k with probability 1.

case 2: $\frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)} \geq 1$, then,

- $\alpha(x_{k+1}, x_k) = 1$
- $\alpha(x_k, x_{k+1}) = \frac{1}{\frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)}}$

this means $\alpha(x_{k+1}, x_k) = min(1, \frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)})$

One can verify that these conditions satisfy the equations for detailed balance.

As mentioned, metropolis hastings is the general case, what I meant is that the proposal distribution Q is a symmetric distribution (for example normal distribution) so that

$Q(x_k \mid x_{k+1}) = Q(x_{k+1} \mid x_k)$, but in a general metropolis- Hastings, Q could be an asymmetric distribution. for simplicity, let's assume Q is symmetric then:

$\alpha(x_{k+1}, x_k) = min(1, \frac{P(x_{k+1})}{P(x_k})$ if

- $P(x_{k+1}) > P(x_k)$ implies probability of accepting x_{k+1} =1
- $P(x_{k+1}) < P(x_k)$ implies probability of accepting $x_{k+1} = \frac{P(x_{k+1})}{P(x_k)}$

So, one of the disadvantages is clearly visible here, suppose $P(x_{k+1}) <<< P(x_k)$, then there is a high chance that the algorithm sampling from the highly likely areas and almost ignoring tail portions of the target distribution, which is not a desirable quality in an ideal sampler. The search for an acceptance probability function, $\beta(x, y, c)$, satisfying $\beta(x, y, c) = c\beta(y, x, c)$ for all x, y, c led to the realization that while constant scaling in $(0, 1]$ satisfies this equation, other solutions might exist.

Efforts to design a function yielding fewer rejections than the Metropolis criterion resulted in challenges. The Metropolis acceptance probability maintains a maximum number of accepted steps by rejecting samples where $\alpha(x', x_k) < 1$. Attempting a higher acceptance probability $\beta(x', x_k) > \alpha(x', x_k)$ leads to $\beta(x_k, x') > 1$, posing an obstacle to fewer rejections, proving that MH algorithm is optimal.

3.4.2 Gibb's Sampling

Utilizing Gibbs Sampling for Multivariate Distributions

Gibbs Sampling serves as a potent technique for sampling from complex multivariate distributions when direct joint distribution sampling poses challenges. This method capitalizes on the relative accessibility of conditional distributions while facing difficulties in directly sampling from the joint distribution.

Algorithm - The Gibbs Sampler

Given $x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, the algorithm iterates as follows:

1. $X_1^{(t+1)} \leftarrow \text{rv } f_1\left(X_1 \mid X_2^{(t)}, \ldots, X_p^{(t)}\right)$
2. $X_2^{(t+1)} \leftarrow \text{rv } f_2\left(X_2 \mid X_1^{(t+1)}, X_3^{(t)}, \ldots, X_p^{(t)}\right)$
3. ...
4. $X_p^{(t+1)} \leftarrow \text{rv } f_p\left(X_p \mid X_1^{(t+1)}, \ldots, X_{p-1}^{(t+1)}\right)$

This algorithm, while theoretically robust, might suffer from slow convergence due to prolonged stays in low and high-probability regions, thereby presenting a computational challenge.

3.5 Nested Sampling

Nested Sampling, a sampling technique recently proposed by John Skilling, [11] transforms the high-dimensional integral over the parameter space into a one-dimensional integral over the possible values of the likelihood:

$$Z = \int_{\Theta} L(\theta; X)\, P(\theta)\, d(\theta)$$

Now, this integral represents the normalization constant, which posed a challenge in our problem. While this term was effectively addressed in the Markov Chain Monte Carlo (MCMC) technique, nested sampling tackles it cleverly using statistics. Given the difficulty of solving the high-dimensional integral, the idea of transforming it into a simpler one-dimensional integral is fundamental to nested sampling. To achieve this, we define a function known as the restricted prior mass, $X(l)$, representing the prior volume enclosed by a particular value of likelihood (iso-likelihood contour) in the parameter space:

$$X(l) = \int_{L(\theta)>l} P(\theta)\, d(\theta)$$

Thus, we can express Z as:

$$Z = \int X(l)\, d(l)$$

over the possible values of $L(\theta; X)$.

Let's illustrate these transformations with a simple example:

$$P(\theta) = \begin{cases} \frac{1}{b-a} & \text{if } \theta \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

$$L(\theta; X) = \theta^3$$

$$a = 2, \quad b = 0$$

$$Z = 8$$

$$X(l) = \frac{1}{2}(\theta)\Big|_{\theta > l^{1/3}}$$

$$Z = \int_0^4 \frac{1}{2}(l^{1/3})\, dl = 8$$

As observed, the solutions agree.

However, is it this simple? While this method utilizes the known restricted prior mass or the volume enclosed above a threshold likelihood value l to integrate the one-dimensional integral, it still faces a challenge: we do not know the restricted prior mass. We begin by generating a group of points from the parameter space, called the initial **live points**. Then, we evaluate the likelihood values of all these selected points and pick the least likely point among them. This point sets the likelihood threshold l and is discarded after storing its likelihood value, becoming a **dead point**. Next, we select another set of live points from the parameter space with the constraint that every live point should belong to the remaining volume (i.e., $1 - X(l)$), ensuring that every point is currently above the previously set iso-likelihood value. We continue this process until some stopping criterion is met. Our objective is to discretize the one-dimensional integral, which can be expressed as the sum over $n = 1$ to $n = N_{\max}$ of $(L_n(X(L_n)) - X(L_{n+1}))$. The unknown $X(L_n)$ values are computed probabilistically using order statistics. Since we have an increasing sequence of likelihood values, we uniformly sample n values from $[0, 1]$ and assign them as X_n's. As $X_n(L_n)$ is a

monotonically decreasing function of L (from the definition of $X(l)$), the sequence of X_n's becomes: $X_1(L_1) > X_2(L_2) > \ldots > X_{N_{\max}}(L_{N_{\max}})$. Now, the probability density function (pdf) of X_n is found using the order statistics.

A key result, stated without proof, is that if X_n's are uniform samples, then the pdf of the maximum follows a Beta$(N, 1)$ distribution. Thus, the X_n corresponding to L_n follows a Beta$(N, 1)$ distribution. With this, we have approximately computed the intractable multidimensional integral with the help of statistics!

Finally, for creating the histogram of samples to obtain the posterior distribution for each parameter, we collect all the dead points accumulated thus far and plot the histogram. The maximum likelihood parameters are then utilized to plot the best-fit model to the data using chi-square minimization.

The MULTINEST algorithm [3] represents a significant advancement in Bayesian inference techniques, specifically tailored for complex and multimodal probability distributions. Built upon the nested sampling framework, MULTINEST offers a robust solution for accurately estimating the Bayesian evidence, accompanied by reliable error estimates. Moreover, it excels in generating posterior samples from distributions characterized by multiple modes and intricate degeneracies, particularly in high-dimensional spaces. By efficiently exploring the parameter space, MULTINEST facilitates comprehensive analyses of complex models, making it a valuable tool for a wide range of scientific and statistical applications.

Chapter 4

Implementation of MCMC and Nested Sampling Techniques

In this chapter, we detail the implementation of Markov Chain Monte Carlo (MCMC) and nested sampling techniques in our work, building upon the theoretical foundations established in the preceding chapters. We utilized publicly available Python packages, specifically `emcee` for MCMC and `pymultinest` for nested sampling, to compute estimations for the parameter values.

4.1 MCMC

As discussed earlier, while MCMC samplers satisfy the detailed balance condition and directly sample from the posterior distribution, they are prone to get stuck at modes of the distribution, especially in the presence of degenerate parameters. Moreover, as the dimensionality of the parameter space increases, MCMC sampling becomes less efficient due to the larger exploration area and diminished acceptance probabilities.

To address these challenges, we used `emcee`[4], which employs an affine invariant proposal step known as the **Stretch move**. This approach offers improved efficiency by distributing an ensemble of Markov chain walkers across the parameter space. However, the initial distribution of walkers in the ensemble significantly impacts the convergence of the sampler to the target distribution. We initialized `emcee` with an even number of walkers, typically twice the number of dimensions, and determined the stopping criteria based on the user-defined number of steps.

4.1.1 Benchmarking Step

For the standard test scenario, we performed a benchmarking exercise to validate our implementation. This involved simulating transmission spectra using a forward model and comparing the retrieved parameter values with known true values. In our standard test, we considered four parameters: H_2O, CO_2, Radius, and Temperature. The true values for these parameters were set to $H_2O = 10^{-2.5}$, Radius = 1.2 $R_{Jupiter}$, and Temperature = 1200 K. The retrieved parameter values from our implementation aligned with these true values, confirming the accuracy of our methodology, refer to figure 4.1 and 4.6 which shows that it fits well.

4.1.2 Eight Dimensional Parmeter Space

After the benchmarking step, the parameter space is now 8 dimensional (see Figure 4.3). The parameters consist of abundances of CO, H_2O, CO_2, K, Na, Radius, Temperature, and Haze. A total of 16 chains were initiated with 20,000 samples per chain initially. However, the number of samples per chain was further increased to 35,000 until the autocorrelation time was reduced sufficiently to ensure enough independent samples were collected. After reaching 35,000 samples per chain, there were no notable changes observed in the marginal distribution of the parameters up to 65,000 samples. Therefore, as a trade-off between accuracy and computational efficiency, 35,000 samples per chain was finalized. This process took approximately 300 minutes for convergence.

4.1.3 Eleven Dimensional Parameter Space

Now, the number of parameters has further increased to 11, which include abundances of CO, H_2O, CO_2, K, Na, Radius, Temperature, Haze_alpha, Haze_gamma, P_cloud, and Phi_Cloud. The same procedure was repeated to determine the optimal number

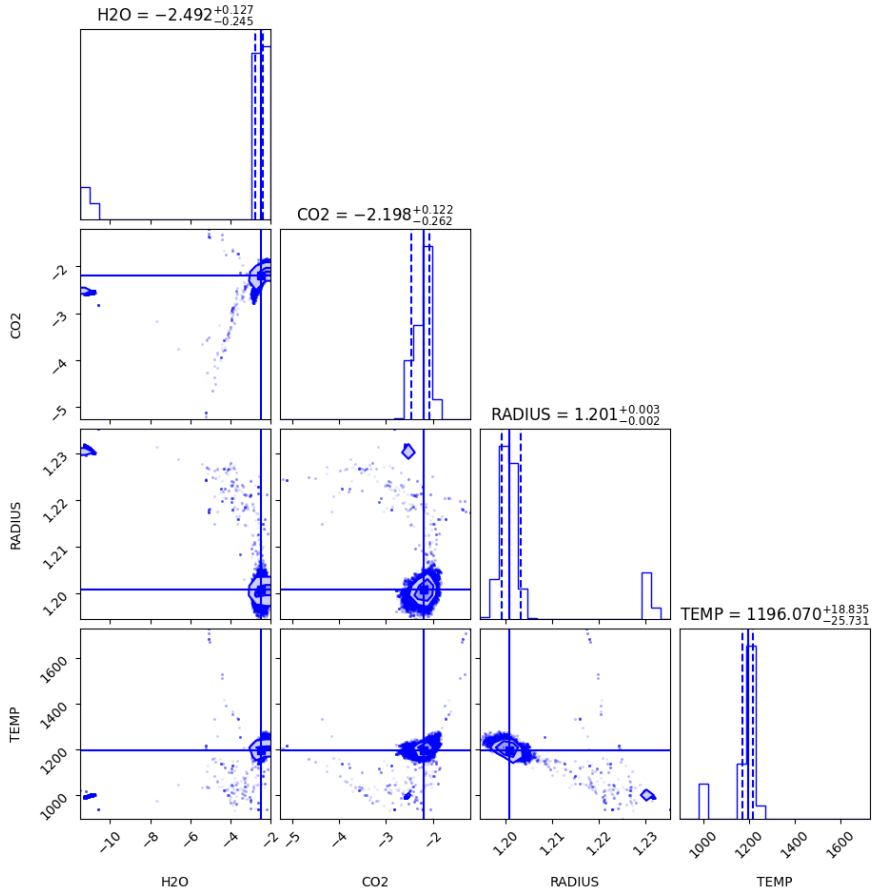

Figure 4.1: Retrieved posterior distribution using the generated transmission spectrum (standard test) with median and error (2σ) labeled

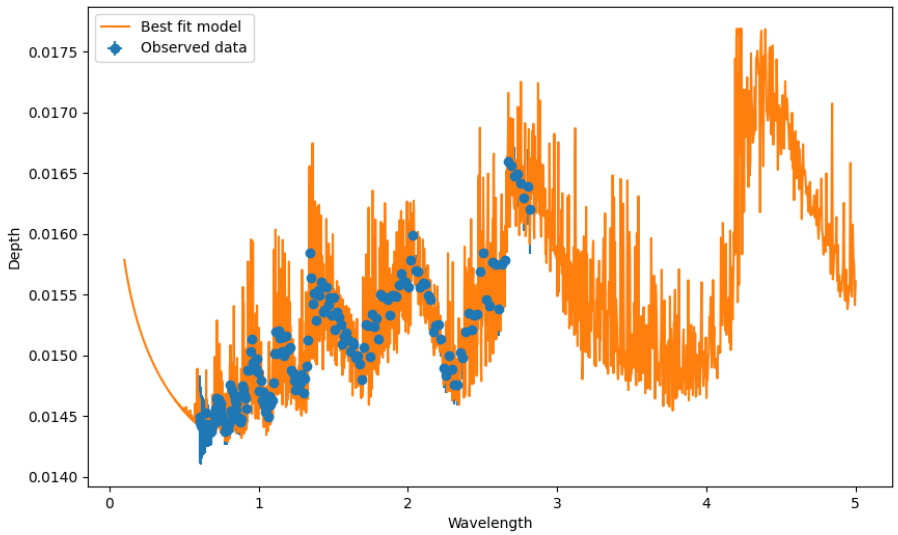

Figure 4.2: The best-fit model closely matches the observations, indicating a successful benchmarking step.

of walkers and samples per chain as in the 8-parameter case. Initially, 22 walkers and 150,000 samples per chain were used. However, even after more than 900 minutes, the autocorrelation time remained high, and convergence was not achieved. As a result, alternative techniques were adopted, such as nested sampling. A comparative study in terms of computational efficiency and accuracy was conducted.

To evaluate the reliability of the sampling techniques, an additional test was conducted. This test aimed to verify whether increasing the number of parameters would maintain the consistency of the marginal distribution for each individual parameter, and values retrieved in both cases were the same as expected.refer **??**

4.2 Nested sampling Technique

For implementing the nested sampling technique, we utilized the publicly available Python package, pymultinest [3] This versatile tool not only facilitates the execution

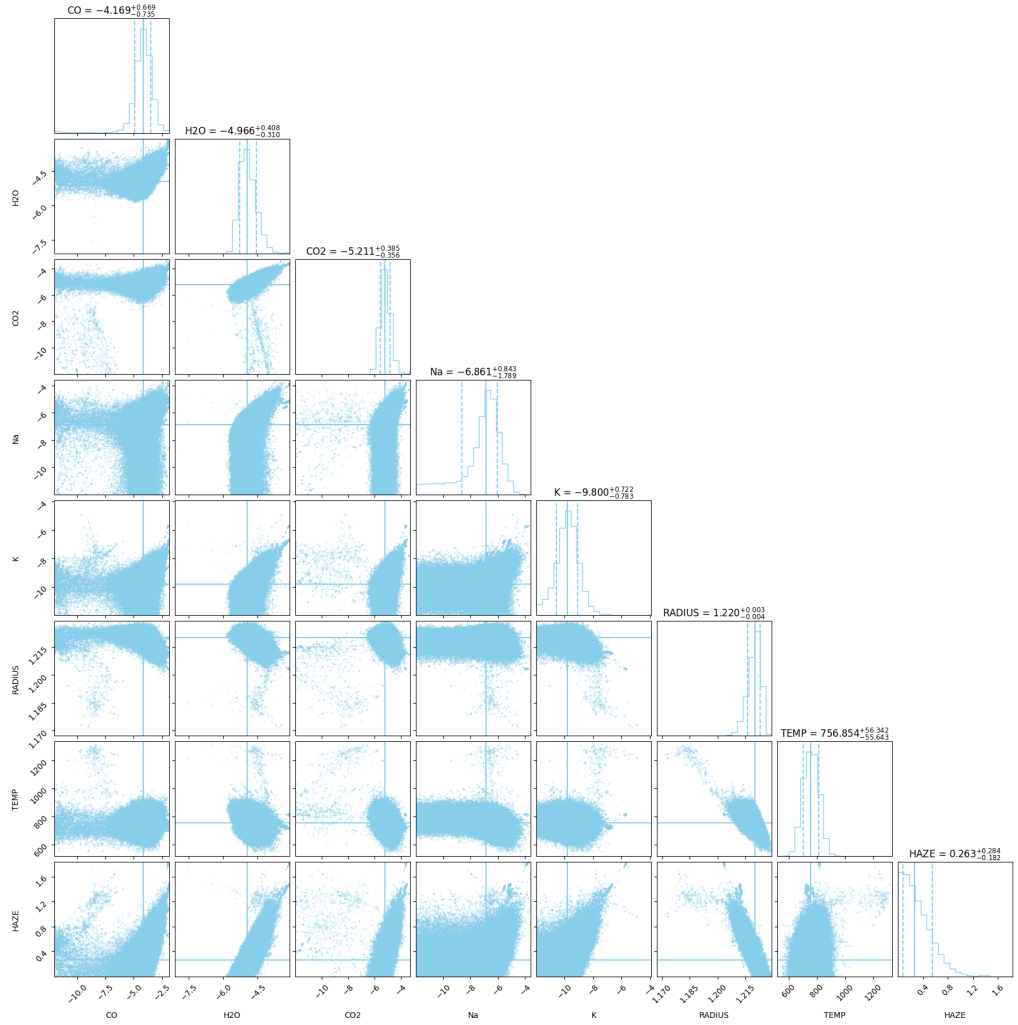

Figure 4.3: Corner plot showing the marginal distributions and correlations between the parameters in the 8-dimensional parameter space. The parameters include abundances of CO, H2O, CO2, K, Na, Radius, Temperature, and Haze.

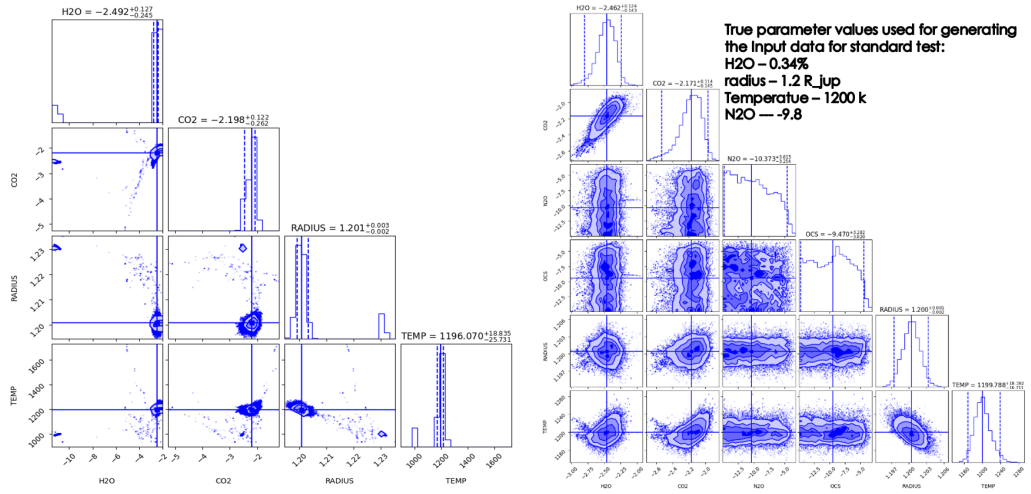

Figure 4.4: Retrieved posteriors for 4 parameters and 6 parameters respectively; for checking if the marginal values are the same in both.

of nested sampling but also possesses the capability to identify local modes within the posterior distribution.

The nested sampling process begins by assigning all active points to one group. In each iteration, these points are divided into subsets, and their associated shapes are constructed. To identify distinct regions within the data, one subset is randomly selected, and its shape is examined for intersections with others. If an intersection is found, the corresponding subset and shape are added to the analysis. This process continues until no more intersections are detected. Subsequently, if isolated regions are identified, new groups are formed, and the points are reassigned accordingly. This iterative approach ensures that all data points are adequately accounted for, leading to a comprehensive analysis of the parameter space.

As the iterations progress, the active points are continuously reassigned to new groups based on the evolving shapes and intersections. The computational efficiency is optimized by minimizing redundant intersection checks between shapes, which re-

duces processing time. Eventually, at the end of the process, each group represents a distinct mode within the data, facilitating a clear understanding of its structure and characteristics. This method allows for a systematic exploration of the parameter space, leading to valuable insights into the underlying patterns and distributions present in the data.

Similar to the MCMC technique, all analyses were conducted using the nested sampling approach, yielding results summarized below.

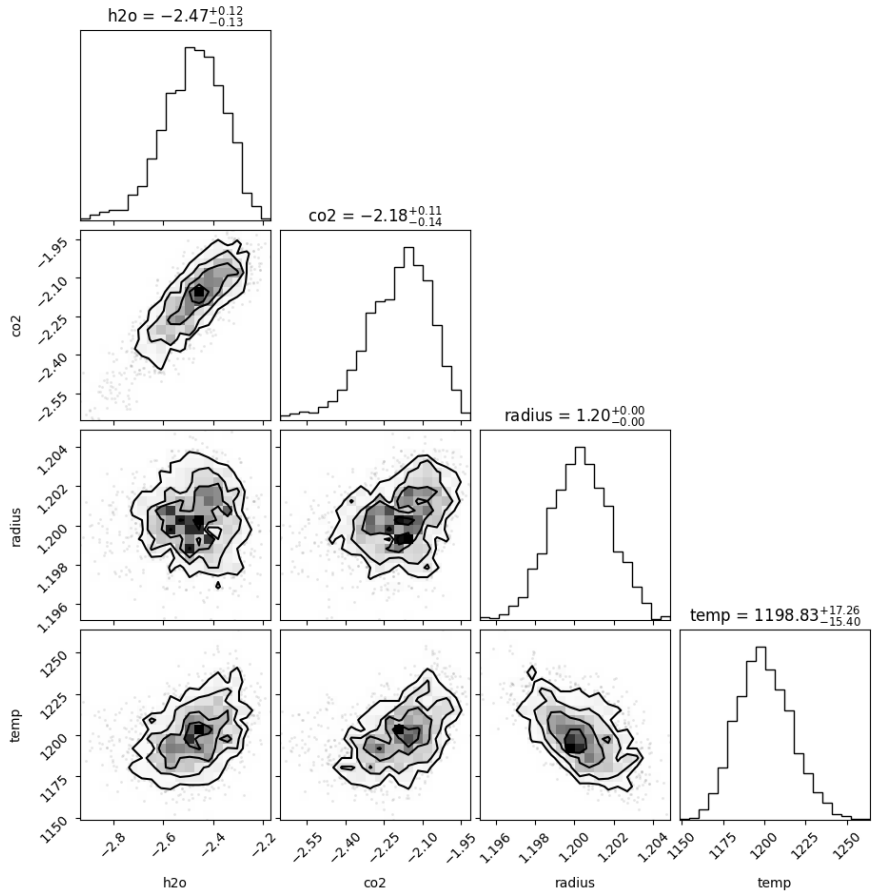

Figure 4.5: Retrieved Posterior distribution using generated transmission spectrum (standard test) with median and error (2σ) labeled

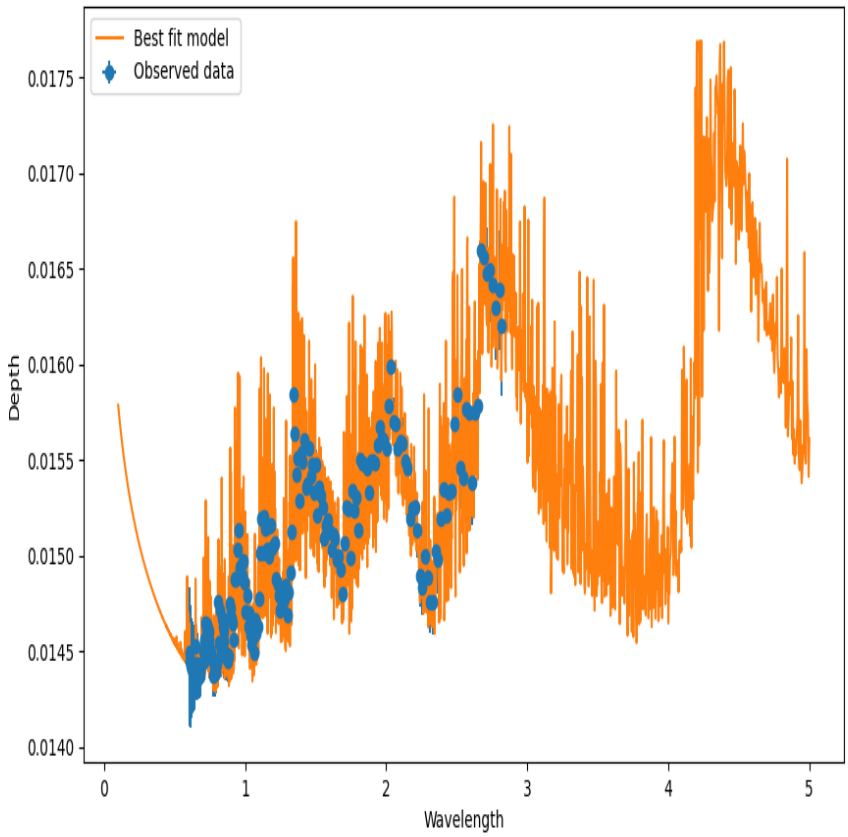

Figure 4.6: The best-fit model closely matches the observations, indicating a successful benchmarking step.

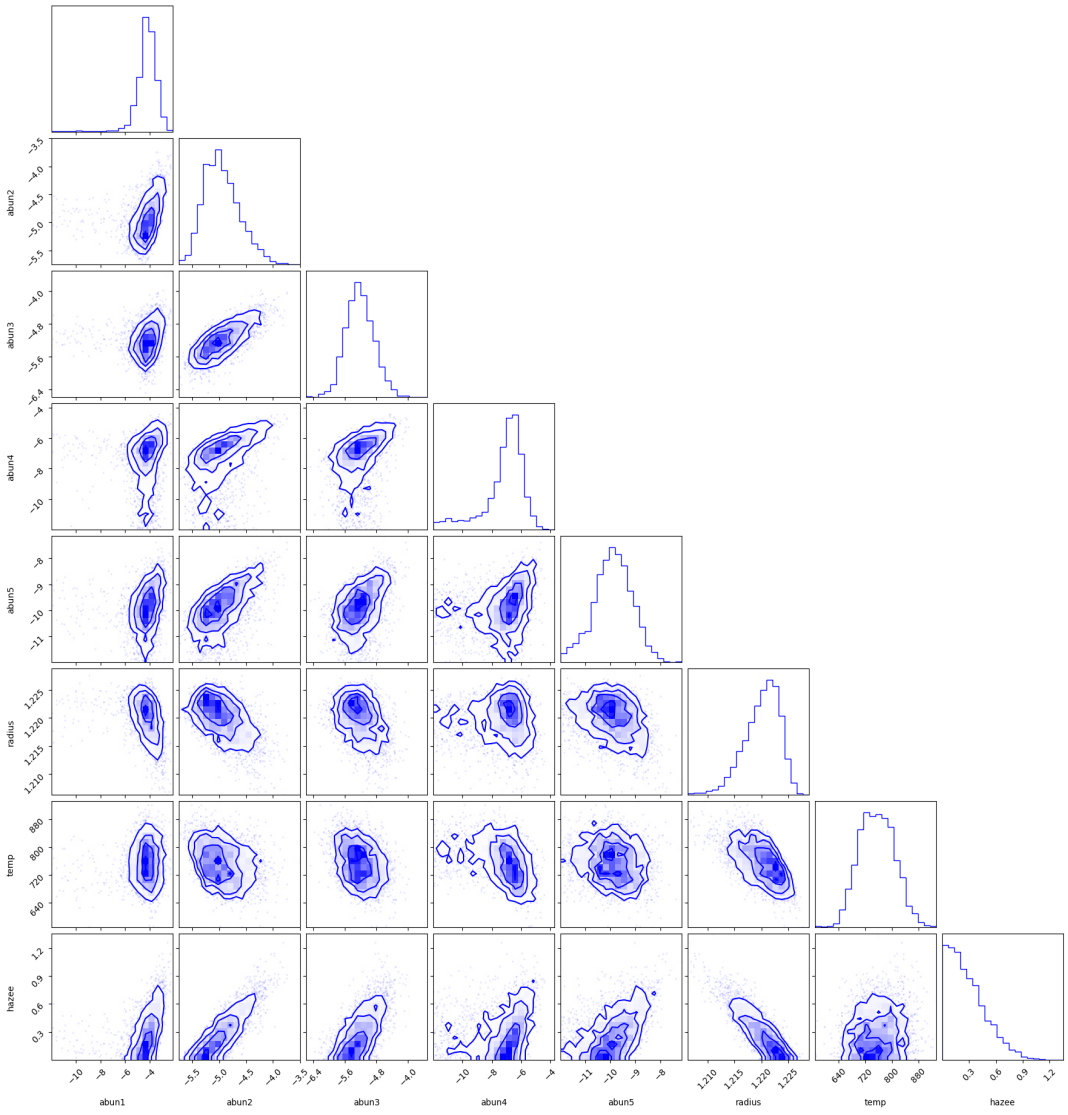

Figure 4.7: Retrieved Posterior distribution using observed transmission spectrum of WASP96 b for 8 parameters with median and error (2σ) labeled

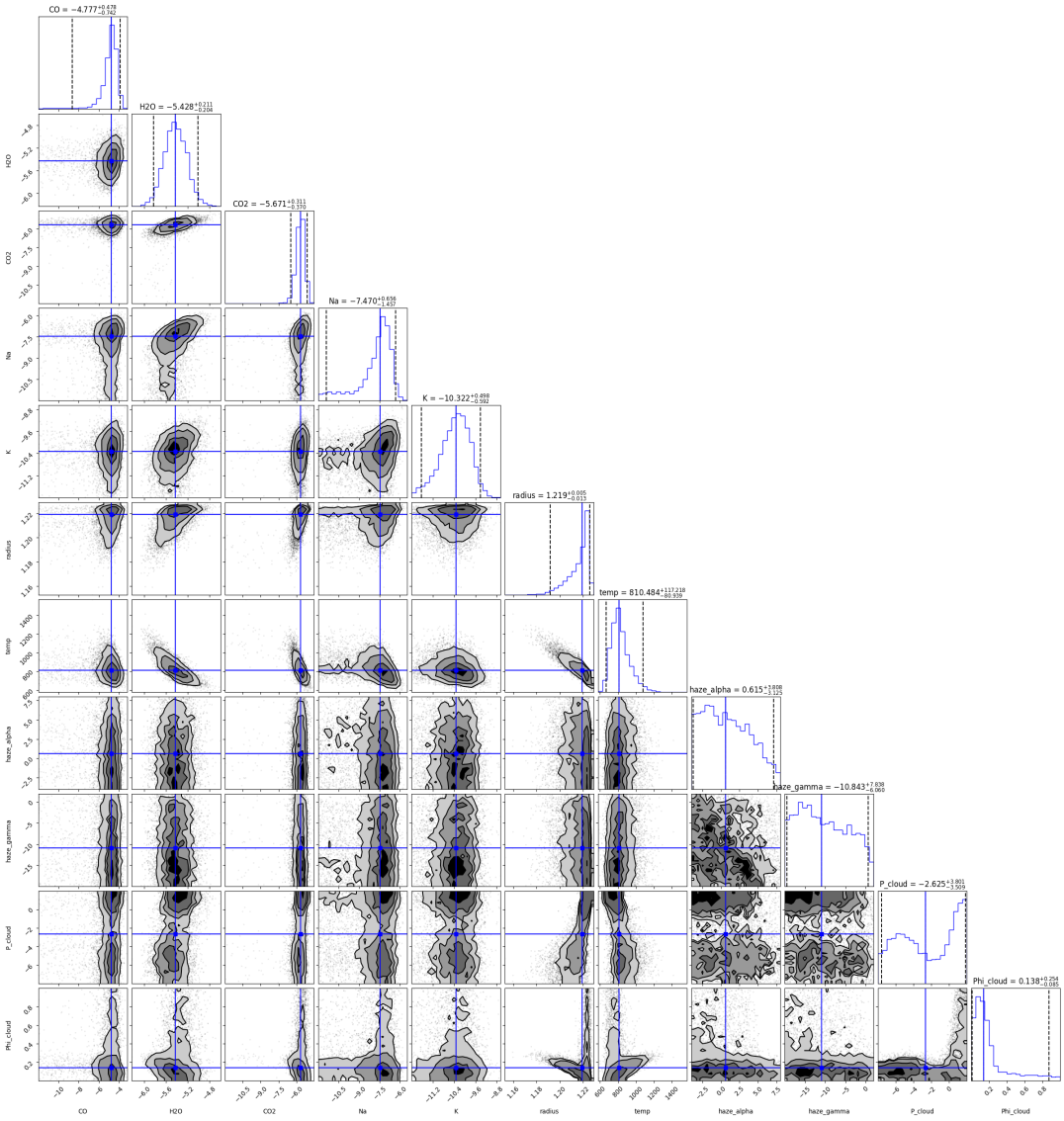

Figure 4.8: Retrieved Posterior distribution using observed transmission spectrum of WASP96 b, for 11 parameters with median and error (2σ) labeled

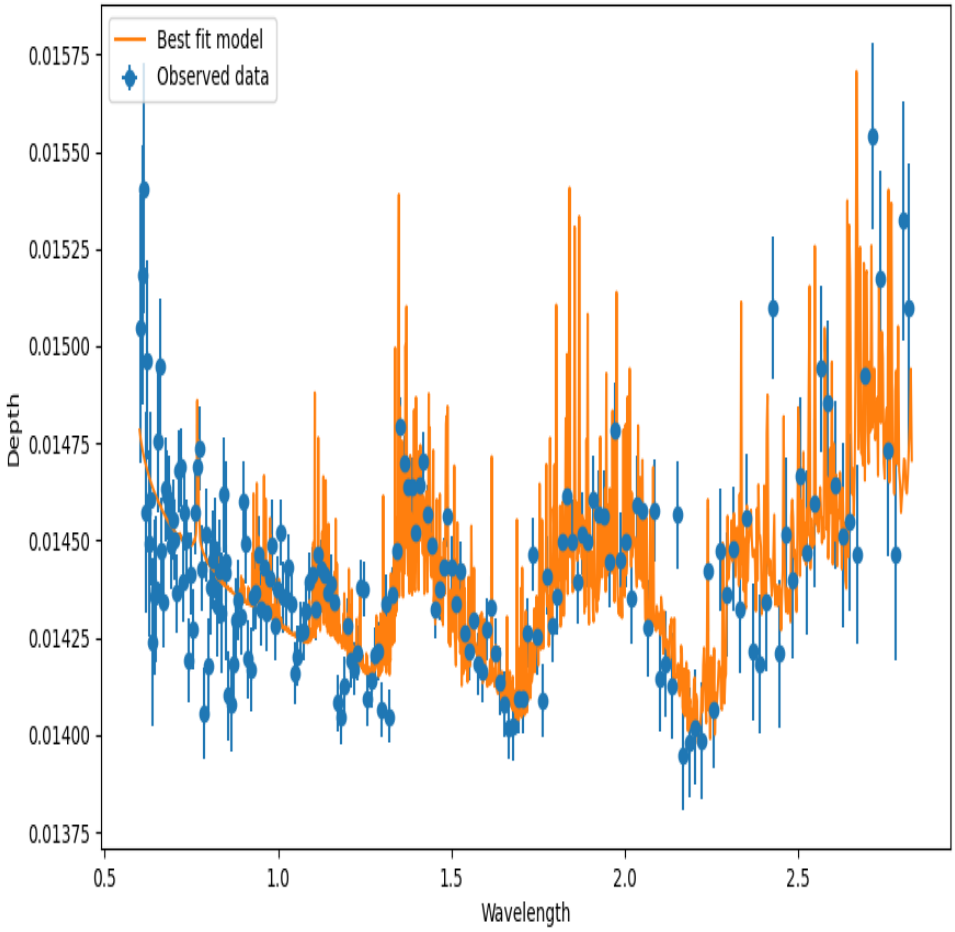

Figure 4.9: The best-fit model in the case of an 11-parameter space. The reduced chi-squared value is 1.75

Chapter 5

Results and Discussions

5.1 Benchmarking Step

The retrieved parameter values from our implementation aligned with the true values(See 4.1.1), validating our methodology, refer to figure 4.1 and 4.6 which shows that it fits well. Following is the table comparing the outputs from the standard test.

Parameter	**MCMC**		**Nested Sampling**	
	Median(log scale)	**Error**	**Median(log scale)**	**Error**
H_2O	-2.49	+0.12 -0.245	-2.47	+0.12 -0.13
CO_2	-2.2	+0.122 -0.262	-2.2	+0.11 -0.14
Radius	1.2	+0.003 -0.002	1.2	+0.00 -0.00
Temperature	1196.07	+18.835 -25.731	1194.75	+17.26 -15.40

Table 5.1: Comparison of Median Values and Errors (2 Sigma) from MCMC and Nested Sampling for the standard test. The abundance of H_2O, CO_2, Radius (Radius of Jupiter is one unit) and Temperature in K.The true values for these parameters were set to $H_2O = 10^{-2.5}$, Radius = 1.2 $R_{Jupiter}$, and Temperature = 1200 K.

5.2 Eight Dimensional Parmeter Space

The table (Table 5.2) compares the outputs for the eight-dimensional case. All parameter median values retrieved using MCMC are comparable to the corresponding median values from NS, except for CO and haze. However, this issue with the haze value was not observed when a model considering clouds was employed.

MCMC took almost 300 minutes to converge, whereas the NS algorithm converged in less than 45 minutes.

The table (Table 5.3) compares the outputs for the eight-dimensional case.

5.3 Eleven Dimensional Parameter Space

In this case, MCMC has not been successful due to the reasons discussed already in Chapter 4. Pymultinest took almost 180 minutes to converge to the posterior distribution, whereas MCMC couldn't converge to the posteriors even after 840 minutes. Furthermore, as the autocorrelation time was also high, a large value for the number of samples was set as the stopping criterion, which led to a large number of likelihood evaluations, increasing the computational cost compared to NS.

Parameter	**MCMC**		**Nested Sampling**	
	Median (log scale)	**Error**	**Median (log scale)**	**Error**
CO	-4.169	+0.669 -0.735	-2.647	+0.36 -0.53
H_2O	-4.966	+0.408 -0.310	-3.92	+0.11 -0.13
CO_2	-5.211	+0.385 -0.356	-4.356	+0.21 -0.27
Na	-6.861	+0.843 -1.789	-6.086	+0.41 -0.61
K	-9.800	+0.722 -0.783	-8.309	+0.60 -0.63
Radius	1.22	+0.003 -0.004	1.210	+0.01 -0.01
Temperature	756.854	+56.34 -55.643	791.4	+98.78 -84.25
Haze	0.263	+0.284 -0.182	1.107	+0.83 -0.12

Table 5.2: Comparison of Median Values and Errors (2 Sigma) from MCMC and Nested Sampling. The abundance of CO, H_2O, CO_2, Na, K, Radius (Radius of Jupiter is one unit), Temperature in K, and Haze

Parameter	**Nested Sampling**	
	Median (log scale)	**Error**
CO	-4.777	+0.478 -0.742
H_2O	-5.428	+0.211 -0.204
CO_2	-5.671	+0.311 -0.370
Na	-7.470	+0.656 -1.457
K	-10.322	+0.498 -0.592
Radius	1.22	+0.005 -0.013
Temperature	810.464	+117.218 -80.939
$Haze_\alpha$	0.615	+3.804 -3.125
$Haze_\gamma$	-10.843	+7.834 -6.060
Pcloud	-2.625	+3.801 -3.509
PhiCloud	0.138	+0.254 -0.085

Table 5.3: Table of Median Values and Errors (2 Sigma) from Nested Sampling. The abundance of CO, H_2O, CO_2, Na, K, Radius (Radius of Jupiter is one unit), Temperature in K, $Haze_\alpha$, $Haze_\gamma$,Pcloud, and PhiCloud

Chapter 6

Summary and Conclusions

In conclusion, this thesis has explored the principles and applications of Bayesian analysis in the context of retrieval problem. Through understanding Bayesian inference techniques, Bayes' rule, likelihoods, and prior and posterior distributions, we have demonstrated the effectiveness of Bayesian methods in parameter estimation. Through Bayesian statistics, we can make decisions based on observed data while accounting for uncertainty and incorporating prior knowledge.Unlike traditional methods, it's more robust, capable of handling situations where assumptions like data independence may not hold true. Additionally, we can make reliable predictions about future observations through Bayesian hypothesis testing and credible intervals.

In this thesis work, Bayesian analysis is done using advanced computational algorithms, such as Markov Chain Monte Carlo (MCMC) and nested sampling, and a comparitive study is done by varying the number of dimensions.The results obtained in our problem were analyzed and tabulated (Refer chapter 5).This implementation of MCMC and nested sampling methods reveals distinct advantages and limitations of these commonly used techniques in retrieving posterior distributions for the parameters in our problem.

While both MCMC and MULTINEST were successful in the benchmarking step (Refer 4.1.1), MCMC took comparatively more time to converge to the posterior distribution. As a single MCMC chain might take a long time to converge to the posterior distribution, a more efficient algorithm was employed [4], expecting faster convergence to the posteriors. However, although it is better than a single chain in

terms of computational time, pymultinest outperformed emcee especially in higher dimensions. The comparative analysis was extended to eight and further to eleven dimensions. In both cases, MCMC convergence time was far longer than pymultinest (See chapter 5). This slow convergence initially seemed to be due to the presence of local modes, as there is a chance for the MCMC chain to get stuck in the modes, especially if there are degeneracies between the parameters. Since pymultinest excels at identifying these modes efficiently (Refer chapter 3), a test was performed to count the total number of modes if present. However, from the results, we concluded that the computational time taken by MCMC is not solely due to the presence of multiple modes, and this will be one of my future works.

Although MULTINEST demonstrates faster convergence compared to MCMC in higher dimensions, MULTINEST's reliability may be compromised by errors in remaining volume estimation (as discussed in chapter 3). The performance of pymultinest in higher-dimensional parameter spaces (>11D) has to be explored as future work. In contrast, MCMC directly samples from target distributions, potentially offering greater reliability in such scenarios. Therefore, a test for ensuring the reliability of these two methods concluded that in lower dimensions, both methods are reliable, as they both gave similar expected marginal distributions for the parameters. This could also be done in higher dimensions. In conclusion, as we move to higher dimensions, MCMC might take the upper hand in terms of accurate results, albeit being very time-consuming.

Parallelizing MCMC could significantly enhance computational efficiency, provided the detailed balance condition is upheld.We compared our outputs with the values in the paper [12], in which a multinest algorithm is employed for the retrieval purpose. This could potentially introduce bias towards multinest. Therefore, the reliability of both methods has to be checked in an unbiased way. For this, we decided to perform

another test using already constrained parameter values, which is possible using the data from a very familiar planet called Earth! We will perform both MCMC and nested sampling by considering Earth as an exoplanet using the transmission spectra, which is to be done in the future.

This comprehensive analysis underscores the nuanced trade-offs between MCMC and nested sampling, informing their optimal selection based on specific modeling requirements and computational constraints. As we continue to advance our understanding of Bayesian methods and refine computational techniques, better estimations of the parameters are expected.

Bibliography

[1] M Delampady, IML Yee, and JV Zidek. "Hierarchical Bayesian analysis of a discrete time series of Poisson counts". In: *Statistics and Computing* 3 (1993), pp. 7–15.

[2] Michael Evans. "Discussion of nested sampling for Bayesian computations by John Skilling". In: *Bayesian Statistics* 8 (2007), pp. 491–524.

[3] Farhan Feroz, MP Hobson, and Michael Bridges. "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics". In: *Monthly Notices of the Royal Astronomical Society* 398.4 (2009), pp. 1601–1614.

[4] Daniel Foreman-Mackey et al. "emcee: the MCMC hammer". In: *Publications of the Astronomical Society of the Pacific* 125.925 (2013), p. 306.

[5] Jayanta K Ghosh, Mohan Delampady, and Tapas Samanta. *An introduction to Bayesian analysis: theory and methods*. Vol. 725. Springer, 2006.

[6] Jonathan Goodman and Jonathan Weare. "Ensemble samplers with affine invariance". In: *Communications in applied mathematics and computational science* 5.1 (2010), pp. 65–80.

[7] Jayesh M Goyal et al. "A library of ATMO forward model transmission spectra for hot Jupiter exoplanets". In: *Monthly Notices of the Royal Astronomical Society* 474.4 (2018), pp. 5158–5185.

[8] Edward Higson et al. "Sampling errors in nested sampling parameter estimation". In: (2018).

[9] Ben Lambert. "A student's guide to Bayesian statistics". In: *A Student's Guide to Bayesian Statistics* (2018), pp. 1–520.

[10] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*. Vol. 2. Springer, 1999.

[11] John Skilling. "Nested sampling". In: *Bayesian inference and maximum entropy methods in science and engineering* 735 (2004), pp. 395–405.

[12] Jake Taylor et al. "Awesome SOSS: atmospheric characterization of WASP-96 b using the JWST early release observations". In: *Monthly Notices of the Royal Astronomical Society* 524.1 (2023), pp. 817–834.