9 th semester Project Report

Foundations of Bayesian Inference and Markov Chains: Theoretical Insights and Computational Challenges

Aliyya Fathima Rinu (1911022)

School Of Mathematical Sciences

Batch: 2019-24
Under the supervision of
Dr. Nabin Kumar Jana and Dr. Jayesh M Goyal

Academic Year: 2023-24

Contents

Elicitation of Subjective probability through Bayesian inference, which has its foundation in a simple, powerful logical tool known as ' Baye's rule ' helps us in tackling many real-life situations needing statistical inferences. As contrasted to the classical inference method also known as the frequentist approach, where probability is interpreted as the frequency of occurrence of a certain event, Bayesian Inference interprets probability as the quantification of one's degree of belief and uses Baye's rule, to facilitate the updation of these prior beliefs as soon as the new information is available.A Bayesian takes a stand that all unknown quantities and observations come from a certain probability distribution.

Sampling from a posterior distribution using diverse techniques becomes indispensable in scenarios where analytical solutions are unattainable or when computational complexities hinder numerical methods, especially in intricate or high-dimensional models. This report presents a comprehensive theoretical exploration of Bayesian Inference, Computational Challenges, and the foundational properties of Markov Chains. It serves as an introductory discourse into the theoretical foundations of Bayesian statistics, focusing on inference methodologies. Additionally, it delves into Bayesian Computations, shedding light on methods and theoretical frameworks, alongside the theoretical aspects of Markov Chains, revealing their fundamental properties.

1 Some Basics of Bayesian Inferance

In many real-world situations, the data we observe are the results of some underlying factors or causes, which we often don't know. In statistical terms, we see these observations as the data and the unknown number that characterizes an entire population as a parameter.

Traditionally, in classical statistical methods, we calculate the probability of getting our observed data under different values of these parameters. However, this approach often relies on idealized assumptions that don't match the complexities of the real world.

What makes statistics fascinating is when we can figure out those unknown parameters that led to the data we have. This is where Bayes' rule becomes incredibly valuable in statistical inference. It allows us to reverse the traditional way of thinking and uncover hidden insights from the data.

In essence, we're shifting from the conventional approach to a more powerful and flexible way of understanding the world through statistics:

$$f(\theta \mid X) = \frac{f(X \mid \theta) f(\theta)}{f(X)}$$

where f(X|ϑ) is the likelihood function, which holds the information of how likely is the observed data to happen if the given a value for ϑ. and f(ϑ) is the probability distribution of the unknown parameter values which reflects our prior beliefs about the parameter. In some cases, when we lack any previous knowledge of the parameters, it is reasonable to use an objective prior distribution known as the uninformative priors, for example, uniform distribution. f(X) is the marginal distribution of data X. f(ϑ| X) is the conditional probability distribution of the parameters given the data (i.e., the observations) which is a realization of a probability distribution, also known as the **Posterior probability distribution**. The posterior probability depends on both data and prior, but as the size of the data increases, it tends to wash away the influence of prior on the posterior. Along with the Posterior distribution, one may report posterior mean, posterior median, or mode (maximum a posterior). in the limit of large data samples, Bayesian estimates often give results the same as frequentists. For example, Suppose we have a set of independent and identically distributed (i.i.d.) random variables, denoted as X_i, each following a Bernoulli distribution with parameter θ. Our goal is to estimate the proportion of success, θ, and derive the posterior distribution for θ using different prior distributions. Let $p(\theta)$ represent our chosen prior distribution.

For the first prior distribution, which is Beta(α, β), the posterior distribution, denoted as $\pi(\mu|X)$, is given by Beta($\alpha + r$, $\beta + (n - r)$), where r is the sum of X_i values.

The posterior mean, representing our updated estimate of θ, can be calculated as:

$$\text{Posterior Mean} = \int_{\Theta} \theta \pi(\mu|X) d\theta = \frac{\alpha + r}{\alpha + \beta + n}$$

This expression provides a way to update our estimate of θ based on the observed data and the chosen prior distribution.

where

$$MLE = \frac{r}{n}$$

Note that re-writing posterior mean as the weighted average of prior mean and maximum likelihood does not always hold true. when analyzing the effect of objective priors in this example, (uniform prior, Jeffreys prior, etc.), it is the case that the expected value of the posterior mean is exactly equal to the MLE. even for small n, the posterior mean closely resembles MLE. (which will be discussed in the chapter "Conjugate Priors").
One important notion to note down is the concept of **Exchangeability** which is a weaker condition of 'independent and identically distributed'. While computing the likelihood or the joint density the densities are split into individual density function for each data points as per the assumption that our data sample points are independent from each other. but this is not always the case, there is often various temporal or spatial correlation among the data points and in such scenarios, if it is exchangeable then the Joint density of an exchangeable sample is invariant under the permutation of data points meaning that if we have a sequence of random variables $\{x_1, x_2, x_3\}$ and if this sequence is equally as likely as the reordered sequence, $\{x_2, x_1, x_3\}$, or any other possible reordering, then the sequence of random variables is said to be *exchangeable*. then the overall likelihood is still the product of individual likelihoods. The assumption of random sampling is stronger than that of exchangeability, meaning that any random sample is automatically exchangeable. but the converse is not necessarily true.

1.1 Advantages Of Bayesian Approach

Incorporating Prior Beliefs: The Bayesian approach allows us to seamlessly incorporate prior beliefs, whether they are subjective or based on objective prior distributions. This flexibility accommodates our existing knowledge and beliefs in a structured way.

Overcoming Assumptions: Unlike classical approaches that rely on assumptions like data independence and repeatability, Bayesian analysis is more robust. It can handle situations where these assumptions do not hold, such as when data is not repeatable, and we only have a single dataset.

Subjective Bayesian Analysis: Subjective Bayesian analysis is particularly valuable in cases where classical methods might struggle due to their rigid assumptions. It provides a framework to adapt to unique or unconventional scenarios.

Predictive Power: Bayesian methods offer strong predictive capabilities. They allow us to use part of the data for analysis and another part for validation, enhancing our ability to make accurate predictions about future observations

1.2 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 $\Theta \subseteq R^p$.

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 $A \subseteq R^p$. 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(ϑ,$\boldsymbol{a}$) defined as the follows:
$R(\theta, \boldsymbol{a}) = E_\theta(L(\theta, \delta(X))$ known as the risk function.In 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 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, ϑhas 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:

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

After gaining data the probability distribution for $\vartheta$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) $E_X\left(L(\theta, \delta(x) \mid x) = \psi(X, a(X))\right.$

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

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

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

1.3 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:

$$BF_{01} = \frac{\int_{\Theta_0} f(X|\theta)g_0(\theta)}{\int_{\Theta_1} f(X|\theta)g_1(\theta)} \tag{1}$$

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

1.4 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:

$$\Pi(\theta_0|X) = \frac{f(X|\theta_0)\Pi_0}{m(x)} \tag{2}$$

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:

Definition: 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 $P(C|X = x) = 1 - \alpha$.

1.5 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.

2 APPROXIMATING INTEGRALS USING SAMPLING METHODS

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 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" that 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.

3.1 INDEPENTANT SAMPLING TECHNIQUES

3.1.1 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, $f(x) \leq Mg(x)$

3.1.2 ACCEPT-REJECT METHOD

The algorithm works as:
1 - Sample X $\sim g(X)$ and U $\sim \mathcal{U}[0,1]$

2 - Accept the proposal if $\leq \frac{f(x)}{Mg(x)}$ else reject it.
3- Repeat step 1

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

Proof. $\mathrm{P}(\mathrm{Y} \leq y) = P(X \leq y | U \leq \frac{f(X)}{Mg(X)}) = \frac{P(X \leq y, U \leq \frac{f(X)}{Mg(X)})}{P(U \leq \frac{f(X)}{Mg(X)})}$

now writing out the probability integral yields

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

□

3.1.3 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 distributions 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 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 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.

3.1.4 Rejection Sampling

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(range(X)), and y_i from U(0,1) then find corresponding pdf(x_i) and if $y_i \leq pdf(\mathrm{x}_i)$, 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 to independent sampling. unlike rejection sampling, **Inverse Transform sampling** is more efficient.

3.2 DEPENTANT SAMPLING TECHNIQUES

sampling of X_{i+1} sample depends on the sample X_i independent sampling is faster and more efficient as it covers more distribution areas than dependent sampling techniques.

4 Markov Chains

Definition 1: A stochastic process constitutes a set of random variables $\{X_t\}_{t\in T}$, where T represents 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 condition:

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

It's worth noting that while 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.

P_{ij} represents the probability of transitioning from state i to state j in a single step. The matrix $[P_{ij}]$ encompasses these transition probabilities. Similarly, a n-step transition probability is:

$$P_{ij}^n$$

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 and a Transition Probability Matrix (TPM). 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 *n-step Transition Property* is represented by $(P_{ij})^{(n)} = P(X_n = i | X_0 = j)$, signifying the process of reaching state i precisely in n steps. Additionally, it's important to note that $P^{(1)} = P$. By induction on n, it can be shown that $P^{(n)} = P^n$.

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:

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

Now by simply using the Markov property, the result follows.

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 $P_{ij}^{(0)} = \delta_{ij}$.

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. This property of communication partitions the state space into classes, meaning that communication defines an equivalence relation. 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 $j \in C$.

Definition: A Markov chain is irreducible if there exists a single communicating class, i.e., all states communicate with each other.
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.

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 $h^{(A)} = (h_i^{(A)},\, i \in S)$. 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:

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

The 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.

The SMP states 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.

4.1 Reccurance and Transience

Definition 1: Passage Time: $T_i = \inf\{n \geq 1 \,|\, X_n = i\}$. It's important to notice the distinction between hitting time and passage time.

Definition 2: Recurrent State: A recurrent state is one where $P_i(T_i < \infty) = 1$.

Definition 3: Transient State: A transient state is one where $P_i(T_i < \infty) < 1$.

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., $P(X_n = i \text{ for infinitely many } n) = 1$. This conclusion is derived from using the Strong Markov Property: the chain keeps revisiting state i.

Conversely, if $P(X_n = i \text{ for infinitely many } n) = 1$, 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 (Recurrent to $\sum_{n=0}^{\infty} P_{ii}^n = \infty$): 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: Given:

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 $n = \sum_{i=1}^{N} \sum_{\delta_i(X_j)}$.

Taking the limit on both sides, we obtain the desired result.
Theorem: Assumptions: C is a communicating class, and 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.

4.2 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 $\sum_{i \in S} \pi_i = 1$. Mathematically, $\pi \cdot P = \pi$.

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:

- State space is finite.
- $\lim_{n \to \infty} P_{ij}^{(n)} = \pi_j$ 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.

Through examples, it's evident that for a Markov Chain with a finite state space, at least one stationary distribution exists. 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 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 earlier? 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 proofs. 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 $i \in S$.

Theorem: Assumptions:

- The MC is irreducible.
- 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: 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 high probability.

The following theorem emphasizes the significance of these classifications:

Theorem: Assumptions:

- 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 $\frac{1}{m_i}$.

Considering an asymmetric simple random walk is transient, by this theorem, it cannot possess a stationary distribution. Additionally, for finite-state Markov chains, as we've established 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 this theorem.

4.3 Limitting Distribution

Law of Large Numbers for MC: Let $\{X_n\}_{n\geq 0}$ be an irreducible and recurrent MC. For a fixed state i, suppose $P(X_0 = i) = 1$, i.e., the initial distribution is δ_i. Then, for any state j, $L_j^{(n)}$ converges to $\gamma_j^{(i)}/E_i(T_i)$ if $E_i(T_i) < \infty$ and to 0 if $E_i(T_i) = \infty$, with probability 1. Here, $V_i^{(n)}$ represents the number of visits to state i during $\{0, 1, \ldots, n\}$, and $L_i^{(n)}$ is defined as $V_i^{(n)}/(n+1)$, the proportion of times the chain is in state i.

This theorem provides the limit of $L_i^{(n)}$. Notably, if a state of a recurrent chain is positive recurrent, then every state is. Considering this, the theorem also specifies that if i is positive recurrent, then $L_i^{(n)}$ converges to $1/E_i(T_i) = \pi_i$ with probability 1.
The long-run proportion of visits to state i is denoted as π_i, representing the i-th component of the stationary distribution. The stationary distribution emerges as a limit of these long-run proportions.

Another critical aspect to note is that although we may have a well-defined Markov Chain (MC), such as a simple, irreducible, positive recurrent MC with a unique stationary distribution, the existence of a limiting distribution is not guaranteed. However, if such a limiting distribution exists, it adheres to the theorem mentioned.
To comprehend the existence of a limiting distribution, we introduce the concept of a "period" for a state.

The period of a state i, denoted by d_i, is defined as the greatest common divisor (gcd) of all $n \geq 1$ for which $P_{ii}^{(n)} > 0$, if this set is non-empty. If it's an empty set, the gcd is defined as 0.

Now, a pivotal theorem states that for an irreducible, positive recurrent, and aperiodic (when all states have a period of 1) Markov chain, considering any initial distribution μ where μ represents the probability of $X_0 = i$, the limit as n approaches infinity of $P_\mu(X_n = i)$ exists. This limit is equal to π_i for all i, where π_i, for i belonging to S, represents the unique stationary distribution.

Furthermore, in an irreducible and positive recurrent chain, we know the stationary distribution exists and is unique, given by $1/E_i(T_i)$. Hence, the limiting distribution coincides with the unique stationary distribution.

5 MARKOV CHAIN MONTE CARLO

5.1 Metropolis-Hastings

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. in this section, we will try to formally prove the algorithm with mathematical reasoning.

First, the detailed balance criterion of a Markov Chain with transition distribution $T(x_{k+1} \mid x_k)$ states that a stationary distribution $P(x)$ satisfies $T(x_{k+1} \mid x_k) P(x_k) = T(x_k \mid x_{k+1}) P(x_{k+1})$. This condition is necessary for a random walk to asymptotically reach a stationary distribution as we already have seen in the previous section. additionally, considerations like irreducibility must also be satisfied for the walk to truly converge to P.

The transition distribution of the MH sampling sequence is given by distribution of x_{k+1} after each inner MH loop completes, given the value of x_k at the beginning of the loop. Examine the following equations. Let Case A denote the case where $x_{k+1} \neq x_k$ and let Case B indicate $x_{k+1} = x_k$. A Case A transition happens with probability:

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

A Case B transition:

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

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.

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

Another perspective on the Metropolis-Hastings (MH) algorithm is to consider how we should design the acceptance probability function, $\alpha(x', x)$, to satisfy 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, $\frac{\alpha(x_{k+1},x_k)}{\alpha(x_k,x_{k+1})} = \frac{Q(x_k|x_{k+1})P(x_{k+1})}{Q(x_{k+1}|x_k)P(x_k)}$ after some algebraic manipulations.

An interesting inquiry explores the creation of an acceptance probability function, $\beta(x, y, c)$, obeying $\beta(x, y, c) = c\beta(y, x, c)$ for all x, y, c. While this holds true for any constant scaling in $(0, 1]$, there might exist additional solutions.

However, the aim is to devise a function that results in fewer rejections compared to the Metropolis acceptance probability. The Metropolis criterion involves rejecting whenever $\alpha(x', x_k) = \frac{P(x')Q(x_k|x')}{P(x_k)Q(x'|x_k)}$ is less than 1. But if $\beta(x', x_k) > \alpha(x', x_k)$, $\beta(x_k, x')$ exceeds 1, creating an issue. Consequently, the Metropolis acceptance probability maximizes the number of accepted steps.

5.2 Gibb's Sampling

5.2.1 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.

5.2.2 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. $\ldots$
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.

References

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

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

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